Statistiques Asymptotiques- Notes de cours - M2

Elisabeth Gassiat

Table des matières

1 Introduction

En **probabilité**, on s'intéresse au comportement, à l'évolution, d'un processus aléatoire, dont on connait a priori la loi (ou un modèle permettant de connaitre sa loi). En **statistique**, on considère donné (ou *observé*) un processus, ou une variable aléatoire, que l'on appelle alors *observation*, et l'on cherche à en déduire quelque chose de sa loi.

On considèrera dans ce cours que l'observation est constituée de $X_1, \dots, X_n$, où $(X_n)_{n\geq 1}$ est une suite de variables aléatoires de loi $\mathbb{P}$. Il faut indiquer dans quel espace $\mathcal{X}$ les variables aléatoires X_i prennent leurs valeurs, et de quelle tribu est muni $\mathcal{X}$. L'espace $\mathcal{X}^{\mathbb{N}}$ est alors muni de la tribu cylindrique. Souvent, on se placera dans la situation où les X_i sont des variables aléatoires indépendantes et de même loi P, auquel cas $\mathbb{P} = P^{\otimes\mathbb{N}}$, et $P^{\otimes n}$ est la loi de l'observation. On fait une hypothèse de modélisation, sous la forme $P \in \mathcal{P}$, où $\mathcal{P}$ est un ensemble de lois de probabilité sur $\mathcal{X}$, et on cherche alors à estimer une quantité $\psi(P)$, par un *estimateur* T_n qui est une variable aléatoire fonction mesurable de $X_1, \dots, X_n$.

En **statistique asymptotique**, on s'intéresse aux propriétés lorsque n tend vers l'infini : consistance des estimateurs, convergence en loi (pour la construction de régions de confiance), risque et limitations intrinsèques. Cela dépendra : du modèle choisi $\mathcal{P}$ et de ce que l'on cherche à estimer $\psi(P)$.
Lorsque $\mathcal{P}$ peut être paramétré sous la forme $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ où $\Theta \subset \mathbb{R}^k$ est de dimension finie, on parle de modèle paramétrique. Lorsque ce n'est pas le cas, on parle de modèle non paramétrique. Pour un modèle paramétrique, la vitesse typique d'estimation est $\sqrt{n}$. On étudiera aussi ce que l'on appellera l'estimation semi-paramétrique, où le modèle est non paramétrique mais où ce que l'on cherche à estimer est de dimension finie.

Références bibliographiques.
Aad van der Vaart : Asymptotic Statistics (Cambridge University Press, 1998).
Alexandre Tsybakov : Introduction à l'estimation non paramétrique (Springer, collection Mathématiques et Applications, 2004).
J.K Ghosh et R.V. Ramamoorthi : Bayesian Nonparametrics (Springer, 2003).

2 Des méthodes d'estimation

2.1 Outils probabilistes

On aura besoin des notions de convergence et des outils pour prouver des convergence : la convergence en probabilité, et la convergence en loi. Pour toutes ces convergences, quand ce sera nécessaire, on notera sous quelle loi elle a lieu. Quand ce ne sera pas précisé, la convergence sera quand n tend vers l'infini.

Rappelons notions et critères pour une suite T_n de variables aléatoires à valeurs dans $\mathbb{R}^d$.
On dit que T_n **converge en probabilité vers la variable aléatoire** T si et seulement si

$$\forall \epsilon > 0, \ \lim_{n \to +\infty} \mathbb{P}\left(\|T_n - T\| \geq \epsilon \right) = 0.$$

On dit que T_n **converge en loi vers la variable aléatoire** T si et seulement si pour toute fonction f continue bornée de $\mathbb{R}^d$ dans $\mathbb{R}$,

$$\lim_{n \to +\infty} E\left[f(T_n) \right] = E\left[f(T) \right].$$

Les critères suivants sont équivalents :

- (1) T_n converge en loi vers T ;
- (2) Pour toute fonction réelle continue positive f, $\liminf_{n \to +\infty} E[f(T_n)] \geq E[f(T)]$;
- (3) La fonction caractéristique de T_n converge ponctuellement vers celle de T ;
- (4) Pour tout ensemble mesurable B tel que $P(T \in \partial B) = 0$ (∂B désigne la frontière de B i.e. sa fermeture moins son intérieur), $\lim_{n \to +\infty} P(T_n \in B) = P(T \in B)$.
- (4bis) (si il s'agit de variables aléatoires réelles) La fonction de répartition de T_n converge vers la fonction de répartition F de T en tout point de continuité de F ;
- (5) Pour tout ensemble ouvert A, $\liminf_{n \to +\infty} P(T_n \in A) \geq P(T \in A)$.
- (6) Pour tout ensemble fermé F, $\limsup_{n \to +\infty} P(T_n \in F) \leq P(T \in F)$.

Comme la convergence en loi ne concerne que les lois, si T est de loi L, on dira aussi par abus de langage " T_n converge en loi vers L".

Pour une suite $(T_n)_{n \geq 1}$ de variables aléatoires à valeurs dans $\mathbb{R}^d$, on note $T_n = o_P(1)$ si $\|T_n\|$ tend en probabilité vers 0, et on note $T_n = O_P(1)$ si $(T_n)_{n \geq 1}$ est une suite **tendue**, c'est à dire si :

$$\forall \epsilon > 0, \ \exists K \ : \forall n, \ \ \mathbb{P}\left(\|T_n\| \leq K \right) \geq 1 - \epsilon.$$

Si $(T_n)_{n\geq 1}$ est une suite tendue de variables aléatoires à valeurs dans $\mathbb{R}^d$, alors on peut en extraire une suite qui converge en loi, et si il y a une seule loi limite possible, alors T_n converge en loi.

Loi des grands nombres (LGN) : si $(Z_n)_{n\geq 1}$ est une suite de variable aléatoires indépendantes et de même loi (on dira i.i.d.) admettant un moment d'ordre 1, c'est à dire telles que $E(|Z_1|) < +\infty$, alors $\frac{1}{n}\sum_{i=1}^n Z_i$ converge en probabilité (et presque sûrement, ce que l'on notera p.s.) vers $E(Z_1)$.
On notera souvent $\overline{Z}$ la moyenne empirique $\frac{1}{n}\sum_{i=1}^n Z_i$.

Théorème de limite centrale (TLC) : si $(Z_n)_{n\geq 1}$ est une suite de variable aléatoires i.i.d. admettant un moment d'ordre 2, c'est à dire telles que $E(\|Z_1\|^2) < +\infty$, alors $\sqrt{n}(\overline{Z} - E(Z_1))$ converge en loi vers U de loi $\mathcal{N}_d(0, V)$ où V est la matrice de variance de Z_1, c'est à dire la matrice $d \times d$ donnée par $V_{i,j} = Cov((Z_1)_i, (Z_1)_j)$, $i, j = 1, \ldots, d$.
On dira par abus de langage que $\sqrt{n}(\overline{Z} - E(Z_1))$ converge en loi vers $\mathcal{N}_d(0, V)$ (bien que l'objet limite n'est pas de même nature que les éléments de la suite, et car la convergence en loi ne concerne que les lois).

Théorème de l'image continue : Soit $(T_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$ et f une fonction continue de $\mathbb{R}^d$ dans $\mathbb{R}^m$. Si T_n converge en probabilité vers la variable aléatoire T, alors $f(T_n)$ converge en probabilité vers $f(T)$. Si T_n converge en loi vers la variable aléatoire T, alors $f(T_n)$ converge en loi vers $f(T)$.

Lemme de Slutsky : Soit $(T_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$ qui converge en loi vers la variable aléatoire T, soit $(V_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^m$ qui converge en loi vers la constante $a \in \mathbb{R}^m$, alors V_n converge en probabilité vers a et (T_n, V_n) converge en loi vers (T, a).

Méthode delta : Soient $(T_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$, $(r_n)_{n\geq 1}$ une suite de réels qui tend vers l'infini, $a \in \mathbb{R}^d$ et g une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^m$ différentiable en a. On suppose que $r_n(T_n - a)$ converge en loi vers Z. Alors $r_n(g(T_n) - g(a))$ converge en loi vers $Dg(a).Z$, où $Dg(a)$ est la matrice $m \times d$ telle que $(Dg(a))_{i,j} = \frac{\partial g_i}{\partial t_j}(a)$.

2.2 Estimateurs de type moments

On considère une suite $(X_n)_{n\geq 1}$ de variables aléatoires i.i.d. de loi P à valeurs dans $\mathbb{R}^d$, et $f : \mathbb{R}^d \to \mathbb{R}^m$ une fonction mesurable. On note

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i).$$

Par la LGN, $\mathbb{P}_n f$ est un estimateur consistant de $E[f(X_1)]$, et $\sqrt{n}(\mathbb{P}_n f - E[f(X_1)])$ converge en loi vers $\mathcal{N}_m(0, V)$ par le TCL, où V est la matrice de variance de $f(X_1)$, ce

qui permet de construire des régions de confiance asymptotiques.

On choisit le modèle $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^k$, on suppose que $P \in \mathcal{P}$, donc qu'il existe $\theta_0 \in \Theta$ tel que $P = P_{\theta_0}$. Il s'agit alors d'estimer θ_0. Si l'on peut trouver $f : \mathbb{R}^d \to \mathbb{R}^k$ et $g : \mathbb{R}^k \to \mathbb{R}^k$ inversible telle que

$$\forall \theta \in \Theta,\ g(\theta) = \int f(x) dP_\theta(x) := P_\theta f := E_\theta f(X)$$

on peut choisir l'estimateur $\widehat{\theta}_n = g^{-1}(\mathbb{P}_n f)$ lorsque $\mathbb{P}_n f \in g(\Theta)$, et un point fixé T de Θ sinon.

Théorème 2.2.1. *On suppose θ_0 dans l'intérieur de Θ, que g est de classe $\mathcal{C}^1$ en θ_0, que $Dg(\theta_0)$ est inversible, et que $E_{\theta_0}[\|f(X)\|^2] < +\infty$. Alors $\widehat{\theta}_n$ est un estimateur consistant de θ_0 et $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converge en loi sous P_{θ_0} vers $\mathcal{N}_k\left(0, Dg(\theta_0)^{-1} Var_{\theta_0}[f(X)][Dg(\theta_0)^{-1}]^T\right)$.*

Remarque. On n'a pas besoin de l'inversibilité de g, seulement de son inversibilité locale. Ceci dit, comme θ_0 est inconnu....

Preuve. Comme $Dg(\theta_0)$ est une matrice inversible, il existe un voisinage V de θ_0 tel que g est inversible sur $g(V)$ voisinage de $g(\theta_0)$. Si l'on note E_n l'événement "$\mathbb{P}_n f \in g(V)$", alors

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \sqrt{n}\left(g^{-1}(\mathbb{P}_n f) - \theta_0\right) \mathbb{1}_{E_n} + \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \mathbb{1}_{E_n^C}.$$

Remarquons que par la LGN, $\mathbb{1}_{E_n^C} = o_{\mathbb{P}}(1)$ (écrire pourquoi), où $\mathbb{P} = P_{\theta_0}^{\otimes \mathbb{N}}$ est la loi de $(X_n)_{n \geq 1}$ sous P_{θ_0}, et que $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \mathbb{1}_{E_n^C} = o_{\mathbb{P}}(1)$ (écrire pourquoi). Puis par la méthode delta, $\sqrt{n}\left(g^{-1}(\mathbb{P}_n f) - \theta_0\right) \mathbb{1}_{E_n}$ converge en loi sous P_{θ_0} vers $\mathcal{N}_k\left(0, Dg(\theta_0)^{-1} Var_{\theta_0}[f(X)] Dg(\theta_0)^{-1}\right)$, et on termine par Slutzky (écrire le détail).
Précisons pour la méthode delta : comme $Dg(\theta_0)$ est inversible, g^{-1} est différentiable en $g(\theta_0)$, de matrice de dérivée $Dg(\theta_0)^{-1}$, donc

$$\begin{aligned}
\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) &= \left\{\sqrt{n}\left(Dg(\theta_0)^{-1}(\mathbb{P}_n f - P_{\theta_0} f) + o_{\mathbb{P}}(\|\mathbb{P}_n f - P_{\theta_0} f\|)\right)\right\} \mathbb{1}_{E_n} \\
&\quad + \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \mathbb{1}_{E_n^C} \\
&= \left\{Dg(\theta_0)^{-1} \sqrt{n}(\mathbb{P}_n f - P_{\theta_0} f) + o_{\mathbb{P}}(\sqrt{n}\|\mathbb{P}_n f - P_{\theta_0} f\|)\right\} \mathbb{1}_{E_n} \\
&\quad + \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \mathbb{1}_{E_n^C}.
\end{aligned}$$

2.3 M- et Z- estimateurs

D'autres idées d'estimation : par moindres carrés, par maximum de vraisemblance. Cela consiste à choisir comme estimateur un minimisant (ou maximisant) approximatif d'une fonction réelle construite à partir des données.
On peut du coup (par exemple en considérant le gradient dans la méthode par optimisation) choisir l'estimateur comme annulant approximativement une fonction à valeurs

dans $\mathbb{R}^k$ par exemple.

On considère une suite $(X_n)_{n\geq 1}$ de variables aléatoires i.i.d. de loi P à valeurs dans $\mathbb{R}^d$, et on veut estimer $\psi(P) = \theta_0 \in \Theta$. On ne précise pas pour l'instant Θ, seulement qu'il est inclus dans un ensemble métrique muni d'une distance $d(\cdot, \cdot)$.

2.3.1 Définitions

M-estimateur : soit, pour tout $\theta \in \Theta$, $m_\theta : \mathbb{R}^d \to \mathbb{R}$ une fonction réelle. Soit $M_n : \Theta \to \mathbb{R}$ telle que pour tout $\theta \in \Theta$, $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$. Soit (u_n) une suite de réels positifs (qui tend vers 0, cette suite sert pour définir un maximum "approximatif"). Le M-estimateur $\widehat{\theta}_n$ vérifie :

$$M_n\left(\widehat{\theta}_n\right) \geq \sup_{\theta\in\Theta} M_n(\theta) - u_n.$$

Z-estimateur : soit, pour tout $\theta \in \Theta$, $\phi_\theta : \mathbb{R}^d \to \mathbb{R}^k$. Soit $Z_n : \Theta \to \mathbb{R}^k$ telle que pour tout $\theta \in \Theta$, $Z_n(\theta) = \frac{1}{n}\sum_{i=1}^n \phi_\theta(X_i)$. Soit (u_n) une suite de réels positifs (qui tend vers 0, cette suite sert pour définir un zéro "approximatif"). Le Z-estimateur $\widehat{\theta}_n$ vérifie :

$$\|Z_n\left(\widehat{\theta}_n\right)\| \leq \inf_{\theta\in\Theta} \|Z_n(\theta)\| + u_n.$$

Exemples :
- Estimateurs de type moment : $\phi_\theta(x) = f(x) - g(\theta)$.
- Maximum de vraisemblance : on choisit le modèle $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ que l'on suppose dominé, c'est à dire qu'il existe une mesure μ sur $\mathbb{R}^d$ tel que pour tout $\theta \in \Theta$, il existe une fonction mesurable réelle f_θ telle que $dP_\theta(x) = f_\theta(x)d\mu(x)$. L'estimateur du maximum de vraisemblance (e.m.v.) maximise M_n où $m_\theta = \log f_\theta$. Si $\log f_\theta(x)$ est $\mathcal{C}^1$ sur Θ pour tout x, l'e.m.v. est un Z-estimateur en prenant ϕ_θ le gradient de $\log f_\theta$.
- Médiane, et plus généralement p-quantile : ici $\Theta = \mathbb{R}$, et $\phi_\theta(x) = (1-p)\mathbb{1}_{x<\theta} - p\mathbb{1}_{x>\theta}$.

2.3.2 Consistance

On suppose que pour tout $\theta \in \Theta$, $m_\theta \in L^1(P)$. Par la LGN, si on définit $M : \Theta \to \mathbb{R}$ telle que pour tout $\theta \in \Theta$, $M(\theta) = Pm_\theta$, on a que pour tout $\theta \in \Theta$, $M_n(\theta) = M(\theta) + o_P(1)$. Du coup, la méthode d'estimation est bonne si en effet θ_0 maximise M sur Θ. Pour que cela permette d'obtenir la consistance du M-estimateur, il faut un peu plus :

Théorème 2.3.1. *On suppose :*
- *(1)* $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| = o_P(1)$,
- *(2) Pour tout* $\epsilon > 0$, $\sup_{\theta\in\Theta : d(\theta,\theta_0)\geq\epsilon} M(\theta) < M(\theta_0)$.
- *(3)* $M_n\left(\widehat{\theta}_n\right) \geq \sup_{\theta\in\Theta} M_n(\theta) - u_n$ *où* u_n *tend vers* 0.

Alors $\widehat{\theta}_n$ *est consistant, c'est à dire que* $d(\widehat{\theta}_n, \theta_0) = o_P(1)$.

Preuve. Soit $\epsilon > 0$ quelconque. Notons $\delta(\epsilon) = M(\theta_0) - \sup_{\theta\in\Theta:d(\theta,\theta_0)\geq\epsilon} M(\theta)$. D'après l'hypothèse (2), $\delta(\epsilon) > 0$. On a maintenant :

$$\begin{aligned}
\mathbb{P}\left(d(\widehat{\theta}_n,\theta_0)\geq\epsilon\right) &\leq \mathbb{P}\left(\sup_{\theta\in\Theta:d(\theta,\theta_0)\geq\epsilon} M_n(\theta) \geq M_n(\theta_0) - u_n\right)\\
&= \mathbb{P}\left(\sup_{\theta\in\Theta:d(\theta,\theta_0)\geq\epsilon} M_n(\theta) - \sup_{\theta\in\Theta:d(\theta,\theta_0)\geq\epsilon} M(\theta) \geq M_n(\theta_0) - M(\theta_0) + \delta(\epsilon) - u_n\right)\\
&\leq \mathbb{P}\left(2\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \geq \delta(\epsilon) - u_n\right)\\
&\leq \mathbb{P}\left(\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \geq \frac{\delta(\epsilon)}{4}\right)
\end{aligned}$$

pour n assez grand (tel que $u_n \leq \delta(\epsilon)/2$), et qui tend donc vers 0 par (1).

On suppose que pour tout $\theta \in \Theta$, $\phi_\theta \in L^1(P)$. Par la LGN, si on définit $Z : \Theta \to \mathbb{R}$ telle que pour tout $\theta \in \Theta$, $Z(\theta) = P\phi_\theta$, on a que pour tout $\theta \in \Theta$, $Z_n(\theta) = Z(\theta) + o_P(1)$. Du coup, la méthode d'estimation est bonne si en effet θ_0 est un zéro de Z sur Θ. Pour que cela permette d'obtenir la consistance du Z-estimateur, il faut un peu plus :

Théorème 2.3.2. *On suppose :*
- *(1)* $\sup_{\theta\in\Theta}\|Z_n(\theta) - Z(\theta)\| = o_P(1)$,
- *(2) Pour tout* $\epsilon > 0$, $\inf_{\theta\in\Theta:d(\theta,\theta_0)\geq\epsilon}\|Z(\theta)\| > 0 = \|Z(\theta_0)\|$.
- *(3)* $\|Z_n\left(\widehat{\theta}_n\right)\| \leq \inf_{\theta\in\Theta}\|Z_n(\theta)\| + u_n$ *où* u_n *tend vers* 0.

Alors $\widehat{\theta}_n$ *est consistant, c'est à dire que* $d(\widehat{\theta}_n,\theta_0) = o_P(1)$.

Preuve. On applique le théorème de consistance précédent en posant $M(\theta) = -\|Z(\theta)\|$ (et en remarquant que la preuve n'utilise pas la forme particulière de moyenne empirique de M_n).

Exemple : la médiane empirique. Ici, $\Theta = \mathbb{R}$, $\phi_\theta(x) = \frac{1}{2}\mathbb{1}_{x<\theta} - \frac{1}{2}\mathbb{1}_{x>\theta}$. Pour montrer la consistance, il s'agit alors de montrer que $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i<\theta} - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i>\theta}$ converge uniformément (en θ) en probabilité vers $P(X < \theta) - P(X > \theta)$.

Question : On a besoin de convergence uniforme. Comment obtenir ce genre de résultat ? Un outil : l'entropie à crochet.

Soient ℓ et u deux fonctions mesurables de $\mathbb{R}^d$ dans $\mathbb{R}$ telles que pour tout x, $\ell(x) \leq u(x)$. On appelle **crochet** $[\ell, u]$ l'ensemble des fonctions $f : \mathbb{R}^d \to \mathbb{R}$ telles que pour tout $x \in \mathbb{R}^d$, $\ell(x) \leq f(x) \leq u(x)$.
Soit $\mathcal{F}$ un ensemble de fonctions réelles mesurables de $\mathbb{R}^d$ dans $\mathbb{R}$ inclus dans $L^p(P)$, $0 < p \leq +\infty$. Pour tout $\epsilon > 0$, on note $N_{[]}(\mathcal{F}, L^p(P), \epsilon)$ **le nombre minimal de crochets de taille** ϵ **nécessaires pour recouvrir** $\mathcal{F}$. C'est à dire : si $N \in \mathbb{N}$ et $[\ell_1, u_1], \ldots, [\ell_N, u_N]$ sont des crochets tels que $\|u_i - \ell_i\|_{L^p(P)} \leq \epsilon$ et $\mathcal{F} \subset \cup_{i=1}^N [\ell_i, u_i]$,

alors $N_{[]}(\mathcal{F}, L^p(P), \epsilon) \leq N$.
On dit que $\mathcal{F}$ est P-**Glivenko-Cantelli** si $\mathcal{F} \subset L^1(P)$, et $\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n f(X_i) - E_P f(X)| = o_P(1)$.

Proposition 2.3.1. *Soit $\mathcal{F} \subset L^1(P)$. On suppose que pour tout $\epsilon > 0$, $N_{[]}(\mathcal{F}, L^1(P), \epsilon) < +\infty$. Alors $\mathcal{F}$ est P-Glivenko-Cantelli.*

Preuve. Soit $\epsilon > 0$, et $[\ell_1, u_1], \ldots, [\ell_N, u_N]$ des crochets tels que $\|u_i - \ell_i\|_{L^1(P)} \leq \epsilon$ et $\mathcal{F} \subset \cup_{i=1}^N [\ell_i, u_i]$. Pour tout $f \in \mathcal{F}$, il existe j tel que pour tout $x \in \mathbb{R}^d$, $\ell_j(x) \leq f(x) \leq u_j(x)$. On a donc

$$\frac{1}{n}\sum_{i=1}^n \ell_j(X_i) \leq \frac{1}{n}\sum_{i=1}^n f(X_i) \leq \frac{1}{n}\sum_{i=1}^n u_j(X_i),$$

et $E_P\ell_j(X) \leq E_P f(X) \leq E_P u_j(X)$. Comme $0 \leq E_P u_j(X) - E_P\ell_j(X) \leq \epsilon$, on obtient

$$\frac{1}{n}\sum_{i=1}^n \ell_j(X_i) - E_P\ell_j(X) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n f(X_i) - E_P f(X) \leq \frac{1}{n}\sum_{i=1}^n u_j(X_i) - E_P u_j(X) + \epsilon,$$

et donc

$$\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - E_P f(X)\right| \leq \max\left\{\left|\frac{1}{n}\sum_{i=1}^n \ell_j(X_i) - E_P\ell_j(X)\right|; \left|\frac{1}{n}\sum_{i=1}^n u_j(X_i) - E_P u_j(X)\right|\right\} + \epsilon.$$

Du coup,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - E_P f(X)\right| \leq \max_{j=1,\ldots,N} \max\left\{\left|\frac{1}{n}\sum_{i=1}^n \ell_j(X_i) - E_P\ell_j(X)\right|; \left|\frac{1}{n}\sum_{i=1}^n u_j(X_i) - E_P u_j(X)\right|\right\} + \epsilon.$$

Mais par la LGN, pour tout $j = 1, \ldots, N$, $\left|\frac{1}{n}\sum_{i=1}^n \ell_j(X_i) - E_P\ell_j(X)\right| = o_P(1)$ et $\left|\frac{1}{n}\sum_{i=1}^n u_j(X_i) - E_P u_j(X)\right| = o_P(1)$, donc (exercice : le démontrer)

$$\max_{j=1,\ldots,N} \max\left\{\left|\frac{1}{n}\sum_{i=1}^n \ell_j(X_i) - E_P\ell_j(X)\right|; \left|\frac{1}{n}\sum_{i=1}^n u_j(X_i) - E_P u_j(X)\right|\right\} = o_P(1)$$

Donc

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - E_P f(X)\right| \geq 2\epsilon\right) \leq \mathbb{P}\left(\max_{j=1,\ldots,N} \max\left\{\left|\frac{1}{n}\sum_{i=1}^n \ell_j(X_i) - E_P\ell_j(X)\right|; \left|\frac{1}{n}\sum_{i=1}^n u_j(X_i) - E_P u_j(X)\right|\right\} \geq \epsilon\right),$$

et donc pour tout $\epsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E_P f(X)\right| \geq 2\epsilon\right) = 0.$$

Voici un exemple simple d'application de ce résultat.

Proposition 2.3.2. *Soit $\mathcal{F} = \{f_\theta,\ \theta \in \Theta\}$, où les f_θ sont des fonctions de $\mathbb{R}^d$ dans R. On suppose que*
- *Θ est une partie compacte d'un espace métrique,*
- *Pour tout x, $\theta \mapsto f_\theta(x)$ est continue,*
- *$\sup_{\theta\in\Theta} |f_\theta| \in L^1(P)$.*

Alors $\mathcal{F}$ est P-Glivenko-Cantelli.

Preuve. On montre que pour tout $\epsilon > 0$, $N_{[]}(\mathcal{F}, L^1(P), \epsilon) < +\infty$. et on utilise la proposition précédente. Soit donc $\epsilon > 0$. Soit $\theta \in \Theta$, et soit $(B_n)_{n\geq 1}$ une suite décroissante de boules ouvertes d'intersection $\{\theta\}$ (par exemple, centrées en θ et de rayon $1/n$). Pour tout $n \geq 1$ et tout x, on note $\tilde{\ell}_n(x) = \inf_{s\in B_n} f_s(x)$ et $\tilde{u}_n(x) = \sup_{s\in B_n} f_s(x)$. Ce sont des fonctions mesurables telles que si n tend vers l'infini, $\tilde{\ell}_n(x)$ et $\tilde{u}_n(x)$ tendent vers $f_\theta(x)$ (par continuité), donc telles que $\tilde{u}_n(x) - \tilde{\ell}_n(x)$ tend vers 0 pour tout x. De plus,

$$|\tilde{u}_n - \tilde{\ell}_n| \leq 2\sup_{\theta\in\Theta} |f_\theta| \in L^1(P),$$

donc par convergence dominée, $\int |\tilde{u}_n - \tilde{\ell}_n| dP$ tend vers 0 quand n tend vers l'infini. Donc il existe n tel que $\int |\tilde{u}_n - \tilde{\ell}_n| dP \leq \epsilon$. Autrement dit, pour tout $\theta \in \Theta$, il existe une boule ouverte B_θ contenant θ telle que $\int |\sup_{s\in B_\theta} f_s - \inf_{s\in B_\theta} f_s| dP \leq \epsilon$. Maintenant, $\cup_{\theta\in\Theta} B_\theta$ est un recouvrement de Θ par des ouverts dont, par compacité, on peut extraire un recouvrement fini $B_{\theta_1} \cup \ldots \cup B_{\theta_N}$. On note $\ell_i = \inf_{s\in B_{\theta_i}} f_s$ et $u_i = \sup_{s\in B_{\theta_i}} f_s$, $i = 1, \ldots, N$, et pour tout $\theta \in \Theta$, il existe i tel que $\theta \in B_{\theta_i}$, et $f_\theta \in [\ell_i, u_i]$.

Lorsque l'on a une régularité plus grande que la continuité, on sait évaluer $N_{[]}(\mathcal{F}, L^p(P), \epsilon)$.

Proposition 2.3.3. *Soit $\mathcal{F} = \{f_\theta,\ \theta \in \Theta\} \subset L^p(P)$, où les f_θ sont des fonctions de $\mathbb{R}^d$ dans R. On suppose que Θ est une partie bornée de $\mathbb{R}^k$, et qu'il existe $\alpha > 0$ et $h \in L^p(P)$ tels que pour tous θ_1 et θ_2 dans Θ, tout $x \in \mathbb{R}^d$,*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \|\theta_1 - \theta_2\|^\alpha h(x).$$

Alors, il existe C_k (qui ne dépend que de k) telle que

$$N_{[]}(\mathcal{F}, L^p(P), \epsilon) \leq \left(C_k diam(\Theta)\left(\frac{2\|h\|_{L^p(P)}}{\epsilon}\right)^{1/\alpha} \vee 1\right)^k,$$

où $diam(\Theta)$ est le diamètre de Θ.

Preuve. Si θ est dans la boule centrée en θ_0 et de rayon δ, alors m_θ est dans le crochet $[\ell, u]$ de taille ϵ avec $\ell = f_{\theta_0} - \delta^\alpha h$, $u = f_{\theta_0} + \delta^\alpha h$, et $\epsilon = 2\|h\|_{L^p(P)}\delta^\alpha$. On obtient donc un recouvrement de $\mathcal{F}$ par des crochets de taille ϵ à partir d'un recouvrement de Θ par des boules de taille $\delta = (\epsilon/2\|h\|_{L^p(P)})^{1/\alpha}$. Mais le nombre minimal N nécessaire pour recouvrir Θ par des boules de rayon δ vérifie

$$N \leq \left(C_k \frac{diam(\Theta)}{\delta} \vee 1 \right)^k,$$

et le résultat s'en suit.

Tout ceci ne nous permet pas de traiter la question de la médiane, car les fonctions $\theta \mapsto \mathbb{1}_{x<\theta}$ ne sont pas continues. On va évaluer directement le nombre de crochets.

Proposition 2.3.4. *Soit $\mathcal{F} = \{\mathbb{1}_{x<\theta},\ \theta \in \mathbb{R}\}$. Soit P une probabilité sur $\mathbb{R}$. Alors pour tout $p > 0$ et $\epsilon \leq 1$,*

$$N_{[]}(\mathcal{F}, L^p(P), \epsilon) \leq \frac{2}{\epsilon^p}$$

Preuve. Soient $t_1 < \ldots < t_k$ des réels. Alors, si $\theta \in]t_i, t_{i+1}]$ pour un $i = 1, \ldots, k-1$, alors $\mathbb{1}_{x<\theta} \in [\mathbb{1}_{x\leq t_i}, \mathbb{1}_{x<t_{i+1}}]$, si $\theta \leq t_1$, $\mathbb{1}_{x<\theta} \in [0, \mathbb{1}_{x<t_1}]$ et si $\theta > t_k$, alors $\mathbb{1}_{x<\theta} \in [\mathbb{1}_{x\leq t_k}, 1]$, ce qui nous fait $k+1$ crochets. Ils sont de taille

$$\|\mathbb{1}_{x<t_{i+1}} - \mathbb{1}_{x\leq t_i}\|^p_{L^p(P)} = P\left(t_i < X < t_{i+1}\right),$$

$$\|\mathbb{1}_{x<t_1}\|^p_{L^p(P)} = P\left(X < t_1\right),\ \|1 - \mathbb{1}_{x\leq t_k}\|^p_{L^p(P)} = P\left(t_k < X\right).$$

On choisit les t_i de façon que ces quantités soient inférieures ou égales à ϵ^p. Ce qui est possible avec k entier tel que $k+1 \geq 1/\epsilon^p$.

Il est clair que le résultat est analogue pour $\mathcal{F} = \{\mathbb{1}_{x>\theta},\ \theta \in \mathbb{R}\}$, et donc que pour $\mathcal{F} = \{\mathbb{1}_{x<\theta} - \mathbb{1}_{x>\theta},\ \theta \in \mathbb{R}\}$, on a $N_{[]}(\mathcal{F}, L^p(P), \epsilon) \leq 2^p/\epsilon^p + 1$ pour $\epsilon \leq 1$.

Maximum de vraisemblance : On suppose que $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ avec Θ compact dans un espace métrique, que le modèle est dominé et identifiable. On suppose que si p_θ est la densité de P_θ par rapport à la mesure dominante, pour tout $x > 0$, pour tout θ, $p_\theta(x) > 0$, $\theta \mapsto p_\theta(x)$ est continue, et $\sup_{\theta\in\Theta} |\log p_\theta| \in L^1(P_{\theta_0})$. Alors l'e.m.v. est consistant en θ_0 (le démontrer !).

2.3.3 Normalité asymptotique

On considère $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi P, $\Theta \subset \mathbb{R}^k$, et $Z_n(\theta) = \mathbb{P}_n\phi_\theta$. On considère le Z-estimateur $\widehat{\theta}_n$, que l'on suppose consistant (convergeant en probabilité vers θ_0), et l'on veut comprendre si et comment obtenir une loi asymptotique du genre $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converge en loi vers une gaussienne, comme on a obtenu pour les estimateurs de type moment qui sont des cas particuliers de Z-estimateurs. On peut écrire Taylor :

$$Z_n\left(\widehat{\theta}_n\right) = Z_n\left(\theta_0\right) + \int_0^1 D_1 Z_n\left[\theta_0 + t\left(\widehat{\theta}_n - \theta_0\right)\right]\left(\widehat{\theta}_n - \theta_0\right) dt.$$

Ici D_1 désigne l'opérateur différentiel, et $D_1 Z_n$ est la matrice $k \times k$ dont chaque colonne constitue les dérivées de Z_n par rapport à une coordonnée de θ. Avec la même notation on a pour tout θ

$$D_1 Z_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} D_1 \phi_\theta(X_i).$$

Comme $P\phi_{\theta_0} = 0$, si de plus $P\|\phi_{\theta_0}\|^2 < +\infty$, alors par le TLC, $\sqrt{n} Z_n(\theta_0)$ converge en loi vers $\mathcal{N}_k\left(0, P\phi_{\theta_0}\phi_{\theta_0}^T\right)$, et on peut écrire

$$\left[\int_0^1 D_1 Z_n\left[\theta_0 + t\left(\widehat{\theta}_n - \theta_0\right)\right] dt\right] \sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -\sqrt{n} Z_n(\theta_0) + \sqrt{n} Z_n(\widehat{\theta}_n).$$

Si $\widehat{\theta}_n$ converge en probabilité vers θ_0, on se dit que pour tout t, $\theta_0 + t\left(\widehat{\theta}_n - \theta_0\right)$ est proche de θ_0, et comme par la LGN, $D_1 Z_n(\theta_0)$ converge en probabilité vers $P D_1 \phi_{\theta_0}$, on se dit que $\left[\int_0^1 D_1 Z_n\left[\theta_0 + t\left(\widehat{\theta}_n - \theta_0\right)\right] dt\right]$ doit converger en probabilité vers $P D_1 \phi_{\theta_0}$. Si c'est le cas, et si cette limite est une matrice inversible, si en plus $\sqrt{n} Z_n(\widehat{\theta}_n) = o_P(1)$, on pourra par Slutzky obtenir la converge en loi de $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$.

Pour obtenir que $\left[\int_0^1 D_1 Z_n\left[\theta_0 + t\left(\widehat{\theta}_n - \theta_0\right)\right] dt\right]$ converge en probabilité vers $P D_1 \phi_{\theta_0}$, on peut supposer (le démontrer en exercice) qu'il existe un voisinage A de θ_0 tel que

— Pour tout x, $\theta \mapsto \phi_\theta(x)$ est $\mathcal{C}^1$ sur A,

— Il existe $h \in L^1(P)$ telle que pour tout x, $\sup_{\theta \in A} \|D_1 \phi_\theta(x)\| \leq h(x)$.

Mais avec ce résultat, on ne peut obtenir la convergence en loi de la médiane empirique : $\theta \mapsto \mathbb{1}_{x<\theta} - \mathbb{1}_{x>\theta}$ n'est pas dérivable pour tout x. On peut faire mieux !

Outils de processus empirique.
Pour $f \in L^1(P)$, on note

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf).$$

Du coup, $\mathbb{P}_n f = Pf + \frac{1}{\sqrt{n}} \mathbb{G}_n f$, et c'est une façon de décomposer une somme (ou moyenne) en la partie biais et la partie variance (qui est centrée). Si $f \in L^2(P)$, par le TLC, $\mathbb{G}_n f$ converge en loi vers une gaussienne.

Si on suppose que $\widehat{\theta}_n$ vérifie le Théorème 2.3.2 avec $u_n << 1/\sqrt{n}$, on a $\sqrt{n} Z_n(\widehat{\theta}_n) = o_P(1)$, et cela se réécrit avec ces notations en

$$\begin{aligned} o_P(1) &= \sqrt{n} \mathbb{P}_n \phi_{\widehat{\theta}_n} \\ &= \mathbb{G}_n\left(\phi_{\widehat{\theta}_n}\right) + \sqrt{n} P \phi_{\widehat{\theta}_n}. \end{aligned}$$

Noter que $P\phi_{\widehat{\theta}_n} = \int \phi_{\widehat{\theta}_n}(x) dP(x)$ est une variable aléatoire, et c'est à elle que l'on veut appliquer Taylor, et comme on a vu des lois des grands nombres uniformes, on aimerait

avoir des TLC uniformes pour traiter $\mathbb{G}_n\left(\phi_{\widehat{\theta}_n}\right)$.

L'outil que l'on utilisera dans ce cours est **l'inégalité maximale** suivante :

Théorème 2.3.3. *Soit $\mathcal{F} \subset L^2(P)$, on suppose que $\mathcal{F}$ admet une fonction enveloppe de carré intégrable, c'est à dire qu'il existe $F \in L^2(P)$ telle que $\forall f \in \mathcal{F}, |f(x)| \leq F(x)$ pour P-presque tout x. Alors*

$$E^*\left[\sup_{f\in\mathcal{F}} |\mathbb{G}_n f|\right] \leq C_{IM} \int_0^{\|F\|_{L^2(P)}} \sqrt{\log N_{[]}(\mathcal{F}, L^2(P), u)}du.$$

pour une constante universelle C_{IM}.

La petite étoile signifie qu'il peut y avoir des problèmes de mesurabilité, mais on prend alors la mesure extérieure. On ne s'en inquiètera pas.
Noter que l'espérance du sup N'EST PAS le sup des espérances.

Revenons à notre problème de normalité asymptotique des Z-estimateurs.
On a :

Théorème 2.3.4. *On suppose que $\widehat{\theta}_n$ est un estimateur consistant de θ_0, que $P\phi_{\theta_0} = 0$, $\sqrt{n}\mathbb{P}_n\phi_{\widehat{\theta}_n} = o_P(1)$ et $P\|\phi_{\theta_0}\|^2 < +\infty$.*
On suppose en outre qu'il existe un voisinage A de θ_0 tel que :

- *$\theta \mapsto P\phi_\theta$ est D^1 sur A, et $D_1 P\phi_{\theta_0}$, notée V, est inversible,*
- *en notant pour tout $j = 1, \ldots, k$, $\mathcal{F}_j = \{\phi_{j,\theta}, \theta \in A\}$, $\sqrt{\log N_{[]}(\mathcal{F}_j, L^2(P), u)}$ est intégrable en 0,*
- *$\int \sup_{\|\theta-\theta_0\|\leq\delta} \|\phi_\theta - \phi_{\theta_0}\|^2 dP$ tend vers 0 quand δ tend vers 0.*

Alors

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -V^{-1}\mathbb{G}_n\left(\phi_{\theta_0}\right) + o_P(1),$$

et $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$ converge en loi vers $\mathcal{N}_k\left(0; V^{-1}P[\phi_{\theta_0}\phi_{\theta_0}^T](V^{-1})^T\right)$.

Preuve. On écrit Taylor pour la fonction $\theta \mapsto P\phi_\theta$ au voisinage de θ_0 :

$$P\phi_\theta = P\phi_{\theta_0} + V\left(\theta - \theta_0\right) + o(\|\theta - \theta_0\|).$$

Comme $\|\widehat{\theta}_n - \theta_0\| = o_P(1)$, on en déduit

$$P\phi_{\widehat{\theta}_n} = P\phi_{\theta_0} + V\left(\widehat{\theta}_n - \theta_0\right) + o_P(\|\widehat{\theta}_n - \theta_0\|),$$

et on a donc

$$\mathbb{G}_n\left(\phi_{\widehat{\theta}_n}\right) + (V + o_P(1))\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = o_P(1).$$

Mais

$$\mathbb{G}_n\left(\phi_{\widehat{\theta}_n}\right) = \mathbb{G}_n\left(\phi_{\theta_0}\right) + \mathbb{G}_n\left(\phi_{\widehat{\theta}_n}\right) - \mathbb{G}_n\left(\phi_{\theta_0}\right),$$

donc si l'on montre que

$$\mathbb{G}_n\left(\phi_{\widehat{\theta}_n}\right) - \mathbb{G}_n\left(\phi_{\theta_0}\right) = o_P(1), \tag{2.1}$$

on aura

$$V^{-1}\mathbb{G}_n\left(\phi_{\theta_0}\right) + \left(I_k + o_P(1)\right)\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = o_P(1)$$

ce qui permet de conclure par Slutzky. Montrons donc (2.1).
Pour tous $\alpha > 0$ et $\delta > 0$, pour toute coordonnée j,

$$\mathbb{P}\left(|\mathbb{G}_n\left(\phi_{j,\widehat{\theta}_n}\right) - \mathbb{G}_n\left(\phi_{j,\theta_0}\right)| \geq \alpha\right) \leq \mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \geq \delta\right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}_\delta} |\mathbb{G}_n f| \geq \alpha\right),$$

en notant $\mathcal{F}_\delta = \{\phi_{j,\theta} - \phi_{j,\theta_0}, \|\theta - \theta_0\| \leq \delta\}$. Pour δ assez petit, $\mathcal{F}_\delta \subset \mathcal{F}_j - \phi_{j,\theta_0}$, donc $N_{[]}(\mathcal{F}_\delta, L^2(P), u) \leq N_{[]}(\mathcal{F}_j, L^2(P), u)$. Aussi, $F_\delta = \sup_{\|\theta-\theta_0|\leq\delta} |\phi_{j,\theta} - \phi_{j,\theta_0}|$ est une fonction enveloppe de $\mathcal{F}_\delta$. Par Markov et en utilisant l'inégalité maximale,

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}} \mathbb{G}_n f| \geq \alpha\right) &\leq \frac{C_{IM}}{\alpha}\int_0^{\|F_\delta\|_{L^2(P)}} \sqrt{\log N_{[]}(\mathcal{F}_\delta, L^2(P), u)}du \\ &\leq \frac{C}{\alpha}\int_0^{\|F_\delta\|_{L^2(P)}} \sqrt{\log N_{[]}(\mathcal{F}_j, L^2(P), u)}du,\end{aligned}$$

donc pour tout $\delta > 0$ assez petit,

$$\limsup_{n\to+\infty} \mathbb{P}\left(|\mathbb{G}_n\left(\phi_{j,\widehat{\theta}_n}\right) - \mathbb{G}_n\left(\phi_{j,\theta_0}\right)| \geq \alpha\right) \leq \frac{C_{IM}}{\alpha}\int_0^{\|F_\delta\|_{L^2(P)}} \sqrt{\log N_{[]}(\mathcal{F}_j, L^2(P), u)}du$$

et l'on conclut par le fait que par hypothèse $\|F_\delta\|_{L^2(P)}$ tend vers 0 quand δ tend vers 0 et l'intégrabilité de $\sqrt{\log N_{[]}(\mathcal{F}_j, L^2(P), u)}$ en 0.

Application à la médiane.
Si la loi P a une densité f positive, $P\phi_\theta = P(X < \theta) - P(X > \theta)$ est dérivable de dérivée $2f(\theta)$. On suppose que cette densité est strictement positive. On a $P\phi_\theta^2 = 1$, et on a déjà vu que $N_{[]}(\mathcal{F}_j, L^2(P), u) \leq 4/u^2$, donc $\sqrt{\log N_{[]}(\mathcal{F}_j, L^2(P), u)}$ est intégrable en 0. De plus

$$\int \sup_{\|\theta-\theta_0\|\leq\delta} \|\phi_\theta - \phi_{\theta_0}\|^2 dP \leq 2P\left(\theta_0 - \delta \leq X \leq \theta_0 + \delta\right)$$

qui tend vers 0 quand δ tend vers 0 (car densité), donc

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -\frac{1}{2f(\theta_0)}\mathbb{G}_n\left(\phi_{\theta_0}\right) + o_P(1),$$

et $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$ converge en loi vers $\mathcal{N}\left(0, 1/4f^2(\theta_0)\right)$.

On s'intéresse maintenant aux M-estimateurs. Si on veut leur appliquer le résultat des Z-estimateurs, on doit supposer que le maximisant est un zéro du gradient, et supposer

donc que le gradient existe, soit que $\theta \mapsto m_\theta(x)$ est dérivable pour tout x. En fait, on va obtenir mieux, en analysant (encore) selon biais/variance. On décompose

$$\mathbb{P}_n m_\theta = Pm_\theta + \frac{1}{\sqrt{n}}\mathbb{G}_n m_\theta.$$

En fait, en supposant le biais $(Pm_\theta - Pm_{\theta_0})$ d'ordre polynomial α et la partie fluctuations $(\mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0})$ d'ordre polynomial β, on peut obtenir une vitesse de convergence $n^{1/2(\alpha-\beta)}$. Ceci vaut "en général", pas forcément en situation paramétrique : l'ensemble des θ n'est pas supposé de dimension finie, il est supposé métrique, muni d'une distance d.
La preuve du théorème qui suit utilise la technique classique de découpage en rondelles (peeling).

Théorème 2.3.5. *On suppose qu'il existe $C > 0$, $\alpha > \beta > 0$, $\delta_0 > 0$ tels que, pour tout n et tout $\delta \leq \delta_0$:*

$$\sup_{\delta_0 \geq d(\theta,\theta_0) \geq \delta} Pm_\theta - Pm_{\theta_0} \leq -C\delta^\alpha$$

et

$$E \sup_{d(\theta,\theta_0)\leq\delta} |\mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0}| \leq C\delta^\beta.$$

On suppose que $\widehat{\theta}_n$ converge en probabilité vers θ_0, et vérifie

$$\mathbb{P}_n m_{\widehat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - O_P\left(n^{-\alpha/2(\alpha-\beta)}\right).$$

Alors

$$n^{1/2(\alpha-\beta)} d\left(\widehat{\theta}_n, \theta_0\right) = O_P(1).$$

Preuve. Posons $v_n = n^{1/2(\alpha-\beta)}$, et notons $R_n = \mathbb{P}_n m_{\widehat{\theta}_n} - \mathbb{P}_n m_{\theta_0}$, on a $R_n \geq -O_P(v_n^{-\alpha})$. Soit M quelconque fixé, et notons j_n tel que $2^{j_n} \leq \delta_0 v_n$ et $2^{j_n+1} > \delta_0 v_n$. Notons

$$S_{j,n} = \{\theta \ : \ 2^j \leq v_n d\left(\theta, \theta_0\right) < 2^{j+1}\}.$$

Soit $K > 0$ quelconque. On a

$$\mathbb{P}\left(v_n d(\widehat{\theta}_n, \theta_0) \geq 2^M\right) \leq \mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \geq \delta_0\right) + \mathbb{P}\left(v_n^\alpha R_n \leq -K\right) + \sum_{j=M}^{j_n} \mathbb{P}\left(\widehat{\theta}_n \in S_{j,n} \text{ et } v_n^\alpha R_n \geq -K\right).$$

Mais si $\widehat{\theta}_n \in S_{j,n}$, $R_n \leq \sup_{\theta\in S_{j,n}} \mathbb{P}_n m_\theta - \mathbb{P}_n m_{\theta_0}$, donc

$$\mathbb{P}\left(\widehat{\theta}_n \in S_{j,n} \text{ et } v_n^\alpha R_n \geq -K\right) \leq \mathbb{P}\left(\sup_{\theta\in S_{j,n}} \mathbb{P}_n m_\theta - \mathbb{P}_n m_{\theta_0} \geq -\frac{K}{v_n^\alpha}\right).$$

Maintenant on écrit

$$\sup_{\theta\in S_{j,n}} \mathbb{P}_n m_\theta - \mathbb{P}_n m_{\theta_0} \leq \sup_{\theta\in S_{j,n}} Pm_\theta - Pm_{\theta_0} + \frac{1}{\sqrt{n}}\left(\sup_{\theta\in S_{j,n}} \mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0}\right)$$
$$\leq -C\frac{2^{j\alpha}}{v_n^\alpha} + \frac{1}{\sqrt{n}}\left(\sup_{\theta\in S_{j,n}} \mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0}\right)$$

de sorte que

$$\mathbb{P}\left(\widehat{\theta}_n \in S_{j,n} \text{ et } v_n^\alpha R_n \geq -K\right) \leq \mathbb{P}\left(\frac{1}{\sqrt{n}}\sup_{\theta\in S_{j,n}} \mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0} \geq C\frac{2^{j\alpha}-K}{v_n^\alpha}\right)$$

et pour M tel que $2^{M\alpha} - K \geq 2^{M\alpha-1}$, on a

$$\mathbb{P}\left(\widehat{\theta}_n \in S_{j,n} \text{ et } v_n^\alpha R_n \geq -K\right) \leq \mathbb{P}\left(\sup_{\theta\in S_{j,n}} \mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0} \geq \sqrt{n}C\frac{2^{j\alpha}}{2v_n^\alpha}\right).$$

On récapitule et on utilise l'inégalité de Markov pour obtenir :

$$\mathbb{P}\left(v_n d(\widehat{\theta}_n, \theta_0) \geq 2^M\right) \leq \mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \geq \delta_0\right) + \mathbb{P}\left(v_n^\alpha R_n \leq -K\right)$$
$$+ \sum_{j=M}^{j_n} \frac{2v_n^\alpha}{\sqrt{n}C2^{j\alpha}}\left(\frac{2^j}{v_n}\right)^\beta.$$

Or

$$\frac{2v_n^\alpha}{\sqrt{n}C2^{j\alpha}}\left(\frac{2^j}{v_n}\right)^\beta = 2^{j(\beta-\alpha)}\frac{v_n^{\alpha-\beta}}{\sqrt{n}}\frac{2}{C} = 2^{j(\beta-\alpha)}\frac{2}{C}$$

sommable car $\beta - \alpha < 0$. On peut donc rendre $\mathbb{P}\left(v_n d(\widehat{\theta}_n, \theta_0) \geq 2^M\right)$ petit en choisissant M assez grand et n assez grand (et K assez grand). Précisément : pour tout $\epsilon > 0$, il existe K tel que $\mathbb{P}\left(v_n^\alpha R_n \leq -K\right) \leq \epsilon/3$, et M tel que

$$\sum_{j\geq M} 2^{j(\beta-\alpha)}\frac{2}{C} \leq \frac{\epsilon}{3},$$

et n_0 tel que si $n \geq n_0$, $\mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \geq \delta_0\right) \leq \frac{\epsilon}{3}$ (car $\widehat{\theta}_n$ est consistant). On a alors pour tout $n \geq n_0$, $\mathbb{P}\left(v_n d(\widehat{\theta}_n, \theta_0) \geq 2^M\right) \leq \epsilon$. Puis pour $j = 1, \ldots, n_0$, il existe M_j tel que $\mathbb{P}\left(v_j d(\widehat{\theta}_j, \theta_0) \geq 2^{M_j}\right) \leq \epsilon$, et l'on choisit le max de M et des M_j pour obtenir l'inégalité pour tout n.

En situation paramétrique, on peut vérifier les hypothèses par Taylor pour la partie biais et l'inégalité maximale pour la partie fluctuations. On obtient :

Proposition 2.3.5. *On suppose $\Theta \subset \mathbb{R}^k$, et qu'il existe un voisinage A de θ_0 et une fonction $h \in L^2(P)$ tels que*

$$\forall(\theta_1,\theta_2) \in A^2,\ |m_{\theta_1}(\cdot) - m_{\theta_2}(\cdot)| \leq h(\cdot)\|\theta_1 - \theta_2\|.$$

On suppose que $\theta \mapsto Pm_\theta$ est maximum en θ_0, est D^2 en θ_0, de matrice hessienne $D_2Pm_{\theta_0}$ inversible. On suppose que $\widehat{\theta}_n$ converge en probabilité vers θ_0, et vérifie

$$\mathbb{P}_n m_{\widehat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - O_P\left(\frac{1}{n}\right).$$

Alors $\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_P(1)$.

Preuve. On va montrer que le Théorème 2.3.5 s'applique avec $\alpha = 2$ et $\beta = 1$.
Par Taylor, $Pm_\theta = Pm_{\theta_0} + (\theta-\theta_0)^T D_2Pm_{\theta_0}(\theta-\theta_0) + o(\|\theta-\theta_0\|^2)$ car max en θ_0 (donc gradient nul) ; aussi comme max en θ_0, $D_2Pm_{\theta_0}(\theta-\theta_0)$ est définie négative, donc il existe $\lambda > 0$ tel que $(\theta-\theta_0)^T D_2Pm_{\theta_0}(\theta-\theta_0) \leq -\lambda\|\theta-\theta_0\|^2 \leq -\lambda\delta^2$ si $\|\theta-\theta_0\| > \delta$. On choisit δ_0 tel que si $\|\theta-\theta_0\| \leq \delta_0$, $o(\|\theta-\theta_0\|^2) \leq \lambda\|\theta-\theta_0\|^2/2$, et il suffit ensuite de prendre $C < \lambda/2$.
Soit ensuite $\mathcal{F}_\delta = \{m_\theta - m_\theta,\ \|\theta-\theta_0\| < \delta\}$. δh est une fonction enveloppe de $\mathcal{F}_\delta$, et

$$N_{[]}\left(\mathcal{F}_\delta, L^2(P), u\right) \leq \left(C_k\delta\frac{2\|h\|_{L^2(P)}}{u}\right)^k$$

donc par l'inégalité maximale, pour une constante D_k,

$$E\sup_{d(\theta,\theta_0)<\delta}|\mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0}| \leq C_{IM}\int_0^{\delta\|h\|_{L^2(P)}}\sqrt{k\log\left(\frac{D_k\delta}{u}\right)}du$$
$$\leq C_{IM}\delta\int_0^{\|h\|_{L^2(P)}}\sqrt{k\log\left(\frac{D_k}{s}\right)}ds$$

par changement de variable $u = \delta s$, et il suffit de choisir $C < C_{IM}\int_0^{\|h\|_{L^2(P)}}\sqrt{k\log\left(\frac{D_k}{s}\right)}ds$.

On obtient finalement le théorème de normalité asymptotique des M-estimateurs :

Théorème 2.3.6. *On suppose $\Theta \subset \mathbb{R}^k$, et qu'il existe un voisinage A de θ_0 et une fonction $h \in L^2(P)$ tels que*

$$\forall(\theta_1,\theta_2) \in A^2,\ |m_{\theta_1}(\cdot) - m_{\theta_2}(\cdot)| \leq h(\cdot)\|\theta_1 - \theta_2\|.$$

On suppose que $\theta \mapsto Pm_\theta$ est maximum en θ_0, est D^2 en θ_0, de matrice hessienne $V = D_2Pm_{\theta_0}$ inversible. On suppose que pour P-presque tout x, $\theta \mapsto m_\theta(x)$ est D^1 en θ_0, de gradient $\dot{m}_{\theta_0}(x)$. On suppose que $\widehat{\theta}_n$ converge en probabilité vers θ_0, et vérifie

$$\mathbb{P}_n m_{\widehat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - o_P\left(\frac{1}{n}\right).$$

Alors

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = -V^{-1}\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1)$$

et $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$ *converge en loi vers* $\mathcal{N}_k(0, V^{-1}(P\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T)V^{-1})$.

Preuve. Voir exercices 2.4.12 et 2.4.13.

On peut appliquer ce théorème à la médiane avec $m_\theta(x) = -|x-\theta|$ (voir exercice 2.4.9).

2.4 Exercices

Exercice 2.4.1. Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires, on note $o_P(X_n)$ pour $\|X_n\| o_P(1)$, et $O_P(X_n)$ pour $\|X_n\| O_P(1)$. Montrer que

$$o_P(1)+o_P(1) = o_P(1),\ o_P(1)+O_P(1) = O_P(1),\ O_P(1)\,o_P(1) = o_P(1),\ o_P(O_P(1)) = o_P(1).$$

Soit R une fonction réelle telle que $R(u) = o(\|u\|^p)$ quand $u \to 0$ pour un $p > 0$, et $X_n = o_P(1)$. Montrer qu'alors $R(X_n) = o_P(\|X_n\|^p)$. Si maintenant $R(u) = O(u^p)$ quand $u \to 0$ pour un $p > 0$, montrer qu'alors $R(X_n) = O_P(\|X_n\|^p)$.

Exercice 2.4.2. Méthode de stabilisation de la variance
Soit $(P_\theta^n)_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}$, un modèle statistique, et T_n un estimateur de θ tel que $\sqrt{n}(T_n-\theta)$ converge en loi sous P_θ^n vers $\mathcal{N}(0,\sigma^2(\theta))$. Montrer que si ϕ est une primitive de $\frac{1}{\sigma(\theta)}$, $\sqrt{n}(\phi(T_n) - \phi(\theta))$ converge en loi sous P_θ^n vers $\mathcal{N}(0,1)$ et en déduire un intervalle de confiance pour θ de niveau asymptotique α.
Application : intervalle de confiance asymptotique pour le paramètre d'une loi binomiale ; d'une loi de Poisson.

Exercice 2.4.3. Région de confiance pour la variance d'une loi
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. de loi ayant des moments jusqu'à l'ordre 4. Soit σ^2 sa variance, et soit $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$, $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

1. Si l'on suppose que les X_i sont gaussiens, proposer un intervalle de confiance I_n pour σ^2 de niveau de confiance égal à $1-\alpha$.
2. Montrer que si $(Z_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires réelles qui converge en loi vers une variable Z de fonction de répartition continue, si u_n tend vers u quand n tend vers l'infini, alors $P(Z_n \le u_n)$ tend vers $P(Z \le u)$.
3. Montrer que $\sqrt{n}(\frac{S^2}{\sigma^2} - 1)$ converge en loi vers $\mathcal{N}(0, \kappa+2)$, où $\kappa = \frac{E[(X_1 - E(X_1))^4]}{\sigma^4} - 3$.
4. Si la loi des X_i n'est pas gaussienne, quel est le niveau asymptotique de I_n ?

Exercice 2.4.4. Modèles exponentiels Soit t de $\mathcal{X}$ dans $\mathbb{R}^k$, μ une mesure positive sur $\mathcal{X}$, h une fonction réelle positive ou nulle sur $\mathcal{X}$ et

$$\Theta = \left\{ \theta \in \mathbb{R}^k \ : \ c\left(\theta\right)^{-1} = \int h\left(x\right) \exp\left[\left\langle \theta, t(x) \right\rangle\right] d\mu\left(x\right) < +\infty \right\}.$$

$(P_\theta)_{\theta \in \Theta}$ telle que $P_\theta(dx) = p_\theta(x) d\mu(x)$ avec

$$p_\theta\left(x\right) = c\left(\theta\right) h\left(x\right) \exp\left[\left\langle \theta, t(x) \right\rangle\right]$$

est un modèle exponentiel k-dimensionnel de statistique exhaustive $t(X)$.
La fonction $c(\cdot)$ est de classe C^∞ sur l'intérieur de Θ, et ses dérivées se calculent en dérivant sous le signe $\int$.
Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires i.i.d. de loi P_{θ_0}, θ_0 dans l'intérieur de Θ. On suppose que $Var(t(X_1))$ est inversible. Montrer que l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ est un estimateur de type moments tel que $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converge en loi vers une gaussienne centrée de variance $[Var(t(X_1))]^{-1}$.

Exercice 2.4.5. Divergence (Information) de Kullback
Soient P, Q deux mesures de probabilité définies sur un même espace, et p, q leur densité par rapport à une mesure dominante μ.

1. Montrer que $\int_{pq>0} p \left(\log \frac{p}{q}\right)_- d\mu$ est toujours finie.
2. En déduire que l'on peut définir

$$K(P,Q) = \left\{ \begin{array}{l} \displaystyle\int \log \frac{dP}{dQ} dP \ si \ P \ll Q \\ +\infty \ sinon \end{array} \right. ,$$

   et que si $P \ll Q$, alors

$$K(P,Q) = \int_{pq>0} p \left(\log \frac{p}{q}\right)_+ d\mu - \int_{pq>0} p \left(\log \frac{p}{q}\right)_- d\mu.$$

   On appelle $K(P,Q)$ la divergence (ou l'information) de Kullback entre P et Q
3. Vérifier que si $P \ll Q$ alors

$$K(P,Q) = Q\phi\left(\frac{dP}{dQ}\right),$$

   où $\phi(x) = x \log x + 1 - x$. En déduire que $K(P,Q) \geq 0$ quelles que soient P et Q, puis que $K(P,Q) = 0$ si et seulement si $P = Q$.

Exercice 2.4.6. Médiane
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles i.i.d. de loi P ayant une unique médiane θ_0, donc telle que pour tout $\epsilon > 0$, $P(X_1 < \theta_0 - \epsilon) < \frac{1}{2} < P(X_1 < \theta_0 + \epsilon)$.
Soit $\widehat{\theta}_n$ vérifiant

$$\frac{1}{n}\sum_{i=1}^{n}\left(1_{X_i<\widehat{\theta}_n} - 1_{X_i>\widehat{\theta}_n}\right) = o_P(1).$$

En utilisant la monotonie de la fonction $\sum_{i=1}^{n}\left(1_{X_i<\theta} - 1_{X_i>\theta}\right)$, montrer que $\widehat{\theta}_n$ est un estimateur consistant de θ_0 .
On suppose que X_1 admet un moment d'ordre 1, et $\tilde{\theta}_n$ maximise

$$M_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}|X_i - \theta|$$

sur Θ compact contenant θ_0. Montrer que $\tilde{\theta}_n$ est un estimateur consistant de θ_0 (utiliser le fait que la fonction $\theta \mapsto |x-\theta|$ est lipschitzienne).

Exercice 2.4.7. Un théorème de consistance
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. de loi P. Soit $(m_\theta(\cdot))_{\theta\in\Theta}$ des fonctions réelles mesurables telles que $\theta \mapsto m_\theta(x)$ soit semi-continue supérieurement pour P-presque tout x. On définit

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i),\ M(\theta) = \int m_\theta(x)\, dP(x).$$

On suppose que Θ est compact dans un espace métrique, et on définit

$$\Theta_0 = \left\{\theta\in\Theta :\ M(\theta) = \sup_{u\in\Theta} M(u)\right\}.$$

Soit θ_0 dans Θ_0, et $\widehat{\theta}_n$ un élément de Θ tel que $M_n\left(\widehat{\theta}_n\right) \geq M_n(\theta_0) + o_P(1)$ (par exemple : $\widehat{\theta}_n$ maximise $M_n(\theta)$ sur Θ).
On suppose qu'il existe une fonction h telle que

$$\forall\theta\in\Theta,\ m_\theta \leq h \text{ et } \int h(x)\, dP(x) < +\infty.$$

Montrer que $\widehat{\theta}_n$ converge en probabilité vers Θ_0, c'est à dire que pour tout $\epsilon > 0$,

$$\lim_{n\to+\infty} P\left(d\left(\widehat{\theta}_n, \Theta_0\right) \geq \epsilon\right) = 0.$$

Application : donner des conditions suffisantes de consistance de l'estimateur du maximum de vraisemblance.
Remarque : on ne suppose pas ici le modèle paramétrique.

Exercice 2.4.8. Modèle de censure
Soient T et C deux variables aléatoires réelles indépendantes de fonction de répartition F_0 et G respectivement. Soit $X = (C, 1_{T \leq C})$. Soit μ la mesure produit tensoriel de la mesure de Lebesgue et de la mesure de comptage sur $\{0, 1\}$. On suppose que G a une densité connue g par rapport à Lebesgue. Le paramètre d'intérêt est donc F_0. Soit $(C_n, \Delta_n)_{n \in \mathbb{N}}$ une suite de variables i.i.d. de même loi que X.

1. Quelle est la densité p_F de X par rapport à μ ? En déduire que pour n observations un estimateur du maximum de vraisemblance maximise, sur les fonctions de répartitions F :

$$\ell_n(F) = \sum_{i=1}^{n} \log\left[\Delta_i F(C_i) + (1 - \Delta_i)(1 - F(C_i))\right]$$

2. Montrer qu'il existe un et un seul estimateur du maximum de vraisemblance $\hat{F}$ qui soit la fonction de répartition d'une mesure de probabilité de support les points C_i, $i = 1, \ldots, n$.
3. Soit $m_n(F) = \ell_n(F) - \ell_n\left(\frac{F+F_0}{2}\right)$. Montrer que $m_n\left(\hat{F}\right) \geq m_n(F_0)$.
4. Si l'on considère le modèle restreint aux fonctions de répartition sur un intervalle compact K, montrer que $\hat{F}$ converge en probabilité, pour la topologie de la convergence faible, vers l'ensemble $\mathcal{F}_0$ des fonctions de répartition sur K qui maximisent

$$M(F) = \int p_{F_0} \log\left(\frac{2p_F}{p_F + p_{F_0}}\right) d\mu.$$

5. On veut montrer que $\mathcal{F}_0$ est l'ensemble des F égales à F_0 G-p.p.
   a) Montrer que $p_F = p_{F_0}$ μ presque partout si et seulement si $F = F_0$ G-p.p.
   b) Montrer que si p et p_0 sont deux densités de probabilité par rapport à une même mesure dominante λ,

$$\int p_0 \log\left(\frac{2p}{p + p_0}\right) d\lambda \leq 0,$$

   avec égalité si et seulement si $p = p_0$, λ p.p.
6. Conclure

Exercice 2.4.9. Médiane
Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires réelles i.i.d. de densité de probabilité f sur un intervalle I de $\mathbb{R}$, telle que f est continue et strictement positive sur l'intérieur de I.

Le paramètre d'intérêt est l'unique médiane θ_0 de f. Soit $\widehat{\theta}_n$ tel que $\mathbb{P}_n\psi_{\widehat{\theta}_n} = o_P\left(n^{-1/2}\right)$, avec $\psi_\theta(x) = \mathbb{1}_{x<\theta} - \mathbb{1}_{x>\theta}$. Montrer que $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converge en loi vers $\mathcal{N}\left(0, \frac{1}{4f^2(\theta_0)}\right)$. On suppose maintenant que X_1 admet un moment d'ordre 1, et $\tilde{\theta}_n$ maximise

$$M_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} |X_i - \theta|.$$

Montrer que $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ converge en loi vers $\mathcal{N}\left(0, \frac{1}{4f^2(\theta_0)}\right)$.

Exercice 2.4.10. (Partiel 2010).
Soit Z une variable aléatoire de loi G sur $\mathbb{R}_+^*$. Soit (X, Y) une variable aléatoire sur $\mathbb{R}^2$ telle que, conditionnellement à Z, X et Y sont indépendantes de loi exponentielle de paramètre Z et θZ respectivement, $\theta > 0$. On note alors $P_{\theta,G}$ la loi de (X, Y).

1. Montrer que le modèle $(P_{\theta,G})_{\theta\in\mathbb{R}_+^*, G\in\mathcal{G}}$, où $\mathcal{G}$ est un ensemble de lois sur $\mathbb{R}_+^*$, est dominé par la mesure de Lebesgue sur $\mathbb{R}_+^2$, et que la densité peut s'écrire

$$p_{\theta,G}(x, y) = \int_0^{+\infty} \theta z^2 \left[\exp - (x + \theta y) z\right] dG(z).$$

2. Soit pour tout réel $a \geq 0$ la fonction

$$\psi_a(x, y) = \frac{x - ay}{x + ay}.$$

   Montrer que pour tous $x > 0, y > 0$, $-1 \leq \psi_a(x, y) \leq 1$. On fixe $\theta > 0$ et $G \in \mathcal{G}$. Montrer que $F(a) = E_{\theta,G}(\psi_a(X, Y))$ est bien définie, continue, dérivable et strictement décroissante sur $\mathbb{R}_+^*$.
3. Montrer que si l'on pose $U = X - \theta Y$ et $V = X + \theta Y$, sous $P_{\theta,G}$, $\frac{U}{V}$ est, conditionnellement à (V, Z), de loi uniforme sur $[-1, 1]$. En déduire que F admet un unique zéro en $a = \theta$.
4. Soit $(X_n, Y_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. de loi $P_{\theta,G}$. Montrer que $F_n(a) = \frac{1}{n}\sum_{i=1}^{n} \psi_a(X_i, Y_i)$ est une fonction qui admet un unique zéro, que l'on note $\widehat{\theta}$. Montrer que $\widehat{\theta}$ est un estimateur consistant de θ.
5. Montrer que

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) = \frac{3\theta}{\sqrt{n}}\sum_{i=1}^{n} \frac{X_i - \theta Y_i}{X_i + \theta Y_i} + o_{P_{\theta,G}}(1).$$

6. En déduire que $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ converge en loi, sous $P_{\theta,G}$, vers $\mathcal{N}(0, 3\theta^2)$.

Exercice 2.4.11. (Partiel 2009). Estimateur de Huber.
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles i.i.d. de loi P_{θ_0} de densité $f(\cdot-\theta_0)$, où f est une fonction strictement positive et paire sur $\mathbb{R}$ telle que $\int_{-\infty}^{+\infty} f(u)fu = 1$. Soit k un réel fixé, et soit ϕ la fonction

$$\phi(x) = x\mathbb{1}_{|x|\leq k} + k\mathbb{1}_{x>k} - k\mathbb{1}_{x<-k}.$$

On pose $\psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \phi(X_i-\theta)$, et on choisit $\hat{\theta}_n$ tel que $\psi_n(\hat{\theta}_n) = o_P(n^{-1/2})$.

1. Donner $\psi(\theta)$ telle que pour tout $\theta\in\mathbb{R}$, $\psi_n(\theta)$ converge en probabilité sous P_{θ_0} vers $\psi(\theta)$.
2. Montrer que $\psi(\theta_0)=0$, que si $\theta>\theta_0$, $\psi(\theta)<\psi(\theta_0)$, et que si $\theta<\theta_0$, $\psi(\theta)>\psi(\theta_0)$.
3. Montrer que ψ_n est décroissante. En déduire que $\hat{\theta}_n$ est consistant.
4. Montrer que pour tous θ_1, θ_2, x, réels, $|\phi(x-\theta_1)-\phi(x-\theta_2)|\leq|\theta_1-\theta_2|$.
5. Montrer que ψ est dérivable et calculer $\psi(\theta_0)$.
6. Montrer que $\sqrt{n}(\hat{\theta}_n-\theta_0)$ converge en loi sous P_{θ_0} vers une gaussienne centrée dont on précisera la variance.

Exercice 2.4.12. Le but de cet exercice est de démontrer le théorème de normalité asymptotique des M-estimateurs. Soit donc $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. de loi P dans $\mathcal{X}$ espace polonais. Soit A un voisinage de θ_0 dans $\mathbb{R}^d$. Pour tout $\theta\in A$, soit m_θ une fonction mesurable de $\mathcal{X}$ dans $\mathbb{R}$. On suppose que pour P-presque tout x, $\theta\mapsto m_\theta(x)$ est différentiable en θ_0 de dérivée $\dot{m}_{\theta_0}(x)$, et qu'il existe une fonction réelle $h\in L_2(P)$ telle que pour tous θ_1, θ_2 dans A,

$$|m_{\theta_1}(x)-m_{\theta_2}(x)|\leq h(x)\|\theta_1-\theta_2\|.$$

On admettra qu'alors, si $(U_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$ telle que $U_n=O_P(1)$, et si $(r_n)_{n\in\mathbb{N}}$ est une suite de réels qui tend vers $+\infty$ quand n tend vers $+\infty$, alors

$$\mathbb{G}_n\left(r_n(m_{\theta_0+U_n/r_n}-m_{\theta_0})-U_n^T\dot{m}_{\theta_0}\right)=o_P(1). \tag{2.2}$$

On suppose de plus que $\theta\mapsto Pm_\theta$ est deux fois différentiable en θ_0 où elle admet un maximum, et que la hessienne V_0 est symétrique inversible. On suppose enfin que $\mathbb{P}_n m_{\widehat{\theta}_n}\geq\sup_\theta \mathbb{P}_n m_\theta - o_P(1/n)$, et que $\widehat{\theta}_n$ tend en probabilité vers θ_0.

1. Montrer que si $(U_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$ telle que $U_n=O_P(1)$, alors

$$n\mathbb{P}_n\left(m_{\theta_0+U_n/\sqrt{n}}-m_{\theta_0}\right)=\frac{1}{2}U_n^TV_0U_n+U_n^T\mathbb{G}_n(\dot{m}_{\theta_0})+o_P(1).$$

2. En déduire que

$$n\mathbb{P}_n\left(m_{\theta_0 - V_0^{-1}G_n(\dot{m}_{\theta_0})/\sqrt{n}} - m_{\theta_0}\right) = -\frac{1}{2}\mathbb{G}_n(\dot{m}_{\theta_0})^T V_0^{-1}\mathbb{G}_n(\dot{m}_{\theta_0}) + o_P(1).$$

et

$$n\mathbb{P}_n\left(m_{\widehat{\theta}_n} - m_{\theta_0}\right) = \frac{n}{2}(\widehat{\theta}_n - \theta_0)^T V_0(\widehat{\theta}_n - \theta_0) + \sqrt{n}(\widehat{\theta}_n - \theta_0)^T\mathbb{G}_n(\dot{m}_{\theta_0}) + o_P(1).$$

3. Montrer que cela implique que

$$\frac{1}{2}\left(\sqrt{n}(\widehat{\theta}_n - \theta_0) + V_0^{-1}\mathbb{G}_n(\dot{m}_{\theta_0})\right)^T V_0\left(\sqrt{n}(\widehat{\theta}_n - \theta_0) + V_0^{-1}\mathbb{G}_n(\dot{m}_{\theta_0})\right) + o_P(1) \geq 0.$$

4. En déduire que $\sqrt{n}(\widehat{\theta}_n - \theta_0) + V_0^{-1}\mathbb{G}_n(\dot{m}_{\theta_0}) = o_P(1)$ et conclure.

3 Théorie de la vraisemblance

On s'intéresse maintenant au cas où le modèle est paramétrique et dominé : $\mathcal{P} = \{P_\theta = p_\theta \mu, \theta \in \Theta\}$ avec μ une mesure sur $\mathbb{R}^d$, et $\Theta \subset \mathbb{R}^k$. L'objectif est d'étudier l'estimation par maximum de vraisemblance, et d'étudier l'optimalité asymptotique des estimateurs : au sens du risque quadratique, et au sens de la loi limite. On sait que dans le cadre de l'estimation sans biais, l'inverse de l'information de Fisher est la variance minimale (inégalité de Cramer-Rao). A-t-on une généralisation asymptotique, et qui porte sur tous les estimateurs, sans contrainte de biais ?
On va voir que d'une part, $\sqrt{n}$ est la vitesse typique d'estimation, et que d'autre part, on peut généraliser asymptotiquement l'inégalité de Cramer-Rao en un sens minimax local, puis que l'estimateur du maximum de vraisemblance est optimal (sous des hypothèses de régularité).

3.1 Modèles différentiables en moyenne quadratique et inégalité de Cramer-Rao

On dit que le modèle est **différentiable en moyenne quadratique** (ce que l'on écrira **d.m.q.**) en θ_0 si il existe un vecteur de k fonctions $\dot{\ell}_{\theta_0}$ (appelé **score en** θ_0) tel que

$$\int \left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 d\mu = o(\|h\|^2). \tag{3.1}$$

Proposition 3.1.1. *Si le modèle est d.m.q. en θ_0, alors $\dot{\ell}_{\theta_0} \in (L^2(P_{\theta_0}))^k$, et $P_{\theta_0} \dot{\ell}_{\theta_0} = 0$.*

Le score est centré et admet une variance : on la note I_{θ_0}, et on l'appelle **information de Fisher en** θ_0. On a

$$I_{\theta_0} = Var_{\theta_0} \dot{\ell}_{\theta_0} = P_{\theta_0} \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T.$$

En particulier, on peut appliquer le TLC de sorte que $\mathbb{G}_n \dot{\ell}_{\theta_0}$ converge en loi sous P_{θ_0} vers $\mathcal{N}_k(0, I_{\theta_0})$.

Preuve. Fixons $h \in \mathbb{R}^k$. $h/\sqrt{n}$ tend vers 0, donc en appliquant (3.1)

$$\int \left(\sqrt{p_{\theta_0+h/\sqrt{n}}} - \sqrt{p_{\theta_0}} - \frac{1}{2\sqrt{n}} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 d\mu = o(\frac{1}{n}),$$

soit

$$\int \left(\sqrt{n}(\sqrt{p_{\theta_0+h/\sqrt{n}}} - \sqrt{p_{\theta_0}}) - \frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 d\mu = o(1).$$

La suite $[\sqrt{n}(\sqrt{p_{\theta_0+h/\sqrt{n}}}-\sqrt{p_{\theta_0}})]_{n\geq 1}$ est une suite de $L^2(\mu)$ qui converge vers $\frac{1}{2}h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}$ dans $L^2(\mu)$, qui est complet, donc pour tout $h\in\mathbb{R}^k$, $h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\in L^2(\mu)$, soit $h^T\dot{\ell}_{\theta_0}\in L^2(P_{\theta_0})$, donc $\dot{\ell}_{\theta_0}\in(L^2(P_{\theta_0}))^k$.
De même, $\sqrt{p_{\theta_0+h/\sqrt{n}}}$ converge vers $\sqrt{p_{\theta_0}}$ dans $L^2(\mu)$, et par continuité du produit scalaire,

$$\lim_{n\to+\infty}\int\left(\sqrt{n}(\sqrt{p_{\theta_0+h/\sqrt{n}}}-\sqrt{p_{\theta_0}})\right)\left(\sqrt{p_{\theta_0+/\sqrt{n}}}+\sqrt{p_{\theta_0}}\right)d\mu=\int h^T\dot{\ell}_{\theta_0}p_{\theta_0}d\mu.$$

Or pour tout n,

$$\int\left(\sqrt{n}(\sqrt{p_{\theta_0+h/\sqrt{n}}}-\sqrt{p_{\theta_0}})\right)\left(\sqrt{p_{\theta_0+/\sqrt{n}}}+\sqrt{p_{\theta_0}}\right)d\mu=\sqrt{n}\int\left(p_{\theta_0+h/\sqrt{n}}-p_{\theta_0}\right)d\mu=0,$$

donc pour tout $h\in\mathbb{R}^k$, $\int h^T\dot{\ell}_{\theta_0}p_{\theta_0}d\mu=0$, et $P_{\theta_0}\dot{\ell}_{\theta_0}=0$.

En gros, le score fois racine de la densité est deux fois la dérivée (par rapport à θ) de la racine carrée de la dérivée. Quand on dérive $2\sqrt{p_\theta}$, on obtient $\dot{p}_\theta/\sqrt{p_\theta}$, donc le score est le quotient de la dérivée de p_θ par p_θ (mais la dérivation est au sens L^2). On retrouve les connaissances antérieures : Fisher est la variance de la dérivée de la log-densité. Mais peut-on relier plus précisément tout ça ?

Proposition 3.1.2. *On suppose que θ_0 est dans l'intérieur de Θ, et que, si l'on note $s_\theta(x)=\sqrt{p_\theta(x)}$: il existe A, voisinage de θ_0 tel que*
— *Pour μ-presque tout x, $\theta\mapsto s_\theta(x)$ est $\mathcal{D}^1$ sur A, de gradient $\dot{s}_\theta(x)$,*
— *Pour tout $\theta\in A$, $\dot{s}_\theta\in L^2(\mu)$, et $\theta\mapsto\int[\dot{s}_\theta\dot{s}_\theta^T]d\mu$ est continue en θ_0.*
Alors le modèle est d.m.q. en θ_0 de score $\dot{\ell}_{\theta_0}=2\dot{s}_{\theta_0}/s_{\theta_0}$.

Remarques. Comme $s_\theta(x)\geq 0$, si $s_\theta(x)=0$ c'est un minimum de $\theta\mapsto s_\theta(x)$, et comme θ_0 est dans l'intérieur de Θ, si il y a un gradient, alors il est nul. Donc μ presque partout, $\dot{s}_\theta(x)=0$ quand $s_\theta(x)=0$.
Du coup, $p_\theta(x)=s_\theta^2(x)$ est $\mathcal{D}^1$ sur A pour μ-presque tout x, de gradient $\dot{p}_\theta=2\dot{s}_\theta(x)s_\theta(x)$. Compte-tenu de la remarque précédente, $\frac{\dot{p}_\theta}{p_\theta}$ est bien défini P_θ-p.s. et vaut $\dot{\ell}_\theta$ P_θ-p.s.

Preuve. On veut montrer (3.1) qui s'écrit

$$\int\left(s_{\theta_0+h}-s_{\theta_0}-h^T\dot{s}_{\theta_0}\right)^2d\mu=o(\|h\|^2),$$

soit, en notant $t=\|h\|$ et $u_t=h/\|h\|$,

$$\int\left(s_{\theta_0+tu_t}-s_{\theta_0}-tu_t^T\dot{s}_{\theta_0}\right)^2d\mu=o(t^2). \tag{3.2}$$

Il suffit de le montrer pour toute suite u_t qui converge (vers un vecteur u) quand t tend vers 0, en effet, si alors ce n'était pas vrai pour toute suite de vecteurs de norme 1, comme la boule unité est compacte, on pourrait extraire une sous-suite qui converge

(vers un vecteur u) et obtenir une contradiction.
On pose alors

$$r_t = \frac{s_{\theta_0+tu_t} - s_{\theta_0}}{t} - u_t^T \dot{s}_{\theta_0},$$

et on veut montrer que $\int r_t^2 d\mu = o(1)$. Posons aussi

$$g_t = 2\left(\frac{s_{\theta_0+tu_t} - s_{\theta_0}}{t}\right)^2 + 2\left(u_t^T \dot{s}_{\theta_0}\right)^2 - r_t^2.$$

On a $g_t \geq 0$, et μ presque partout, par l'hypothèse de différentiabilité, quand t tend vers 0,

$$\frac{s_{\theta_0+tu_t} - s_{\theta_0}}{t} = u_t^T \dot{s}_{\theta_0} + o(1) = u^T \dot{s}_{\theta_0} + o(1),$$

et g_t tend vers

$$2\left(u^T \dot{s}_{\theta_0}\right)^2 + 2\left(u^T \dot{s}_{\theta_0}\right)^2 - \left(u^T \dot{s}_{\theta_0} - u^T \dot{s}_{\theta_0}\right)^2 = 4\left(u^T \dot{s}_{\theta_0}\right)^2.$$

Donc par le lemme de Fatou,

$$\int \liminf_{t\to 0} g_t d\mu \leq \liminf_{t\to 0} \int g_t d\mu,$$

soit :

$$\limsup_{t\to 0} \int r_t^2 d\mu \leq 2 \liminf_{t\to 0} \int \left(\frac{s_{\theta_0+tu_t} - s_{\theta_0}}{t}\right)^2 d\mu - 2u^T \left(\int \dot{s}_{\theta_0} \dot{s}_{\theta_0}^T d\mu\right) u.$$

Maintenant :

$$s_{\theta_0+tu_t} - s_{\theta_0} = \int_0^1 t u_t^T \dot{s}_{\theta_0+vtu_t} dv,$$

donc

$$\left(\frac{s_{\theta_0+tu_t} - s_{\theta_0}}{t}\right)^2 = \left(\int_0^1 u_t^T \dot{s}_{\theta_0+vtu_t} dv\right)^2 \leq \int_0^1 \left(u_t^T \dot{s}_{\theta_0+vtu_t}\right)^2 dv$$

et par Fubini,

$$\int \left(\frac{s_{\theta_0+tu_t} - s_{\theta_0}}{t}\right)^2 d\mu \leq \int_0^1 \left(\int (u_t^T \dot{s}_{\theta_0+vtu_t})^2 d\mu\right) dv = \int_0^1 \left(u_t^T \left(\int \dot{s}_{\theta_0+vtu_t} \dot{s}_{\theta_0+vtu_t}^T d\mu\right) u_t\right) dv.$$

par l'hypothèse de continuité, $\int \dot{s}_{\theta_0+vtu_t} \dot{s}_{\theta_0+vtu_t}^T d\mu$ converge vers $\int \dot{s}_{\theta_0} \dot{s}_{\theta_0}^T d\mu$ quand t tend vers 0 et est majorée dans un voisinage de θ_0, donc par convergence dominée, $\int_0^1 \left(u_t^T \left(\int \dot{s}_{\theta_0+vtu_t} \dot{s}_{\theta_0+vtu_t}^T d\mu\right) u_t\right) dv$ tend vers $u^T(\int \dot{s}_{\theta_0} \dot{s}_{\theta_0}^T d\mu)u$ quand t tend vers 0, et l'on a donc obtenu

$$\limsup_{t\to 0} \int r_t^2 d\mu \leq 0.$$

On va pouvoir maintenant énoncer **l'inégalité de Cramer-Rao** dans le cadre des modèles d.m.q. comme une conséquence d'un résultat de dérivabilité :

Théorème 3.1.1. *On suppose que θ_0 est dans l'intérieur de Θ, et que le modèle est d.m.q. en θ_0. Soit $T : \mathbb{R}^d \to \mathbb{R}$ mesurable. Si il existe un voisinage A de θ_0 tel que*

$$\sup_{\theta \in A} P_\theta \left(T^2(X)\right) < +\infty,$$

alors la fonction g de Θ dans $\mathbb{R}$ donnée par $\theta \mapsto g(\theta) = P_\theta (T(X))$ est différentiable en θ_0 de gradient

$$\dot{g}_{\theta_0} = P_{\theta_0} \left(T \dot{\ell}_{\theta_0}\right).$$

Si de plus l'information de Fisher I_{θ_0} est inversible, alors

$$Var_{\theta_0} (T) \geq \dot{g}_{\theta_0}^T I_{\theta_0}^{-1} \dot{g}_{\theta_0}.$$

Remarque. $P_\theta T = \int T(x) p_\theta(x) dx$, et si on dérive sous le signe somme et qu'on interprète le score comme $\frac{\dot{p}_\theta}{p_\theta}$ on obtient le résultat, mais il n'est pas obtenu sous les hypothèses habituelles de dérivation sous le signe somme.

Preuve. On veut montrer

$$D(h) := \int \left(T p_{\theta_0+h} - T p_{\theta_0} - h^T T \dot{\ell}_{\theta_0} p_{\theta_0}\right) d\mu = o(\|h\|).$$

Posons

$$r_h = \sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}},$$

on sait que $\int r_h^2 d\mu = o(\|h\|^2)$. On a

$$\begin{aligned} D(h) &= \int T \left[\left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}}\right) \left(\sqrt{p_{\theta_0+h}} + \sqrt{p_{\theta_0}}\right) - h^T \dot{\ell}_{\theta_0} p_{\theta_0} \right] d\mu \\ &= \int T \left[(r_h + \frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}}) \left(\sqrt{p_{\theta_0+h}} + \sqrt{p_{\theta_0}}\right) - h^T \dot{\ell}_{\theta_0} p_{\theta_0} \right] d\mu \\ &= \int T r_h \left(\sqrt{p_{\theta_0+h}} + \sqrt{p_{\theta_0}}\right) d\mu + \int T \frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}}\right) d\mu. \end{aligned}$$

Par Cauchy-Schwarz, et si h est tel que $\theta_0 + h \in A$,

$$\begin{aligned} \int T r_h \left(\sqrt{p_{\theta_0+h}} + \sqrt{p_{\theta_0}}\right) d\mu &\leq \left(\int r_h^2 d\mu\right)^{1/2} \left(\int T^2 \left(\sqrt{p_{\theta_0+h}} + \sqrt{p_{\theta_0}}\right)^2 d\mu\right)^{1/2} \\ &\leq \left(\int r_h^2 d\mu\right)^{1/2} \left(4 \sup_{\theta \in A} P_\theta \left(T^2(X)\right)\right)^{1/2} = o(\|h\|). \end{aligned}$$

Pour le deuxième terme, on décompose selon que $|T| \leq K$ ou $|T| > K$ et l'on obtient (encore par Cauchy-Schwarz)

$$\begin{aligned}
&\int T\frac{1}{2}h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\left(\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}}\right)d\mu \\
&\qquad\leq \frac{K}{2}\left(\int (h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}})^2 d\mu\right)^{1/2}\left(\int (\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}})^2 d\mu\right)^{1/2} \\
&\quad+\left(\int T^2(\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}})^2 d\mu\right)^{1/2}\left(\int (\mathbb{1}_{|T|>K}\frac{1}{2}h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}})^2 d\mu\right)^{1/2} \\
&\qquad\leq O\left(K\|h\|^2\right)+\|h\|O\left(\left(\int (\mathbb{1}_{|T|>K}\|\dot{\ell}_{\theta_0}\|^2 p_{\theta_0}d\mu\right)^{1/2}\right)=o(\|h\|)
\end{aligned}$$

en prenant par exemple $K_h = 1/\sqrt{\|h\|}$.

Ce résultat se généralise pour des fonctions g à valeur dans $\mathbb{R}^m$. Rappelons qu'alors la différentielle de g est une matrice, dont les lignes sont les gradients de ses fonctions coordonnées.

Théorème 3.1.2. *On suppose que θ_0 est dans l'intérieur de Θ, et que le modèle est d.m.q. en θ_0. Soit $T:\mathbb{R}^d\to\mathbb{R}^m$ mesurable. Si il existe un voisinage A de θ_0 tel que*

$$\sup_{\theta\in A} P_\theta\left\|T(X)\right\|^2 < +\infty,$$

alors la fonction g de Θ dans $\mathbb{R}^m$ donnée par $\theta\mapsto g(\theta)=P_\theta\left(T(X)\right)$ est différentiable en θ_0 de matrice différentielle

$$Dg(\theta_0)=P_{\theta_0}\left(T\dot{\ell}_{\theta_0}^T\right).$$

La matrice suivante est semi-définie positive :

$$\begin{pmatrix} Var_{\theta_0}(T) & Dg(\theta_0) \\ Dg(\theta_0)^T & I_{\theta_0}\end{pmatrix}$$

Si de plus l'information de Fisher I_{θ_0} est inversible, alors

$$Var_{\theta_0}(T)-Dg(\theta_0)I_{\theta_0}^{-1}Dg(\theta_0)^T$$

est semi-définie positive.

Preuve. La première partie du théorème (différentiabilité) est une application du Théorème 3.1.1 appliqué à g coordonnée par coordonnée.
Puis, compte-tenu de ce résultat, la première matrice est une matrice de variance

$$\begin{pmatrix} Var_{\theta_0}(T) & Dg(\theta_0) \\ Dg(\theta_0)^T & I_{\theta_0}\end{pmatrix}=Var_{\theta_0}\left[\begin{pmatrix} T \\ \dot{\ell}_{\theta_0}\end{pmatrix}\right]$$

Pour la deuxième, le résultat vient de :

$$Var_{\theta_0}(T)-Dg(\theta_0)I_{\theta_0}^{-1}Dg(\theta_0)^T=\begin{pmatrix} I_m & -Dg(\theta_0)I_{\theta_0}^{-1}\end{pmatrix}\begin{pmatrix} Var_{\theta_0}(T) & Dg(\theta_0) \\ Dg(\theta_0)^T & I_{\theta_0}\end{pmatrix}\begin{pmatrix} I_m \\ -I_{\theta_0}^{-1}Dg(\theta_0)^T\end{pmatrix}.$$

Que se passe-t-il quand on dispose d'un n-échantillon ?

Théorème 3.1.3. *On suppose le modèle $\{P_\theta = p_\theta \mu, \theta \in \Theta\}$ d.m.q. en θ_0 de score $\dot{\ell}_{\theta_0}$ et d'information de Fisher I_{θ_0}. Alors pour tout $n \in \mathbb{N}$, le modèle $\{P_\theta^{\otimes n}, \theta \in \Theta\}$ est d.m.q. en θ_0 de score*

$$\dot{\ell}_{\theta_0,n}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(x_i)$$

et d'information de Fisher nI_{θ_0}.

Corollaire 3.1.1. *Si de plus $T : (\mathbb{R}^d)^n \to \mathbb{R}$ est mesurable et tel que il existe un voisinage A de θ_0 tel que*

$$\sup_{\theta \in A} P_\theta^{\otimes n}\left(T^2(X_1, \ldots, X_n)\right) < +\infty,$$

si on pose $g(\theta) = P_\theta^{\otimes n} T$, alors

$$E_{\theta_0}\left[\sqrt{n}\,(T - g(\theta_0))\right]^2 \geq \dot{g}_{\theta_0}^T I_{\theta_0}^{-1} \dot{g}_{\theta_0}.$$

Preuve. A faire en exercice !!! Et à compléter par le résultat multidimensionnel !!!

3.2 L'estimateur du maximum de vraisemblance

On s'intéresse à l'estimateur du maximum de vraisemblance et à son asymptotique. On va commencer par montrer un développement de la log-vraisemblance sous la seule hypothèse de différentiabilité en moyenne quadratique. On note $\ell_n(\theta)$ la log-vraisemblance :

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(X_i).$$

Théorème 3.2.1. *Si le modèle est d.m.q. en θ_0, alors pour toute suite $(h_n)_n$ de $\mathbb{R}^k$ convergeant vers un $h \in \mathbb{R}^k$:*

$$\ell_n\left(\theta_0 + \frac{h_n}{\sqrt{n}}\right) - \ell_n(\theta_0) = \mathbb{G}_n\left(h^T \dot{\ell}_{\theta_0}\right) - \frac{1}{2} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1).$$

Preuve. Posons pour tout n et tout $i = 1, \ldots, n$:

$$W_{n,i} = 2\left(\sqrt{\frac{p_{\theta_0 + \frac{h_n}{\sqrt{n}}}(X_i)}{p_{\theta_0}(X_i)}} - 1\right)$$

de sorte que

$$\ell_n\left(\theta_0 + \frac{h_n}{\sqrt{n}}\right) - \ell_n(\theta_0) = 2\sum_{i=1}^{n} \log\left(1 + \frac{W_{n,i}}{2}\right).$$

Taylor donne : $\log(1+u) = u - \frac{u^2}{2} + u^2 R(u)$ où $R(u)$ tend vers 0 quand u tend vers 0. Du coup

$$\ell_n\left(\theta_0 + \frac{h_n}{\sqrt{n}}\right) - \ell_n(\theta_0) = \sum_{i=1}^n W_{n,i} - \frac{1}{4}\sum_{i=1}^n W_{n,i}^2 + \frac{1}{2}\sum_{i=1}^n W_{n,i}^2 R(W_{n,i}).$$

On va montrer

$$\sum_{i=1}^n W_{n,i} = \mathbb{G}_n\left(h^T \dot{\ell}_{\theta_0}\right) - \frac{1}{4} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1), \tag{3.3}$$

$$\frac{1}{4}\sum_{i=1}^n W_{n,i}^2 = \frac{1}{4} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1) \tag{3.4}$$

et

$$\sum_{i=1}^n W_{n,i}^2 R(W_{n,i}) = o_{P_{\theta_0}}(1) \tag{3.5}$$

ce qui suffit à prouver le théorème.
Montrons (3.3).

$$E_{\theta_0}(\sum_{i=1}^n W_{n,i}) = 2n \int (\sqrt{p_{\theta_0 + \frac{h_n}{\sqrt{n}}}}\sqrt{p_{\theta_0}} - 1) d\mu = -n \int (\sqrt{p_{\theta_0 + \frac{h_n}{\sqrt{n}}}} - \sqrt{p_{\theta_0}})^2 d\mu.$$

Mais dans $L^2(\mu)$ (par d.m.q.)

$$\sqrt{p_{\theta_0 + \frac{h_n}{\sqrt{n}}}} = \sqrt{p_{\theta_0}} + \frac{1}{2\sqrt{n}} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} + o(\frac{1}{\sqrt{n}}).$$

Donc $n \int (\sqrt{p_{\theta_0 + \frac{h_n}{\sqrt{n}}}} - \sqrt{p_{\theta_0}})^2 d\mu$ tend vers $\int (\frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}})^2 d\mu$, soit $\frac{1}{4} h^T I_{\theta_0} h$. Par ailleurs,

$$\begin{aligned} Var_{\theta_0}\left(\sum_{i=1}^n W_{n,i} - \mathbb{G}_n\left(h^T \dot{\ell}_{\theta_0}\right)\right) &\leq E_{\theta_0}\left[\left(\sqrt{n} W_{n,i} - h^T \dot{\ell}_{\theta_0}(X_i)\right)^2\right] \\ &= 4 \int \left(\sqrt{n}(\sqrt{p_{\theta_0 + \frac{h_n}{\sqrt{n}}}} - \sqrt{p_{\theta_0}}) - \frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}}\right)^2 d\mu = o(1), \end{aligned}$$

et on déduit facilement (3.3). Montrons maintenant (3.4).
Puisque $\sqrt{n} W_{n,i} - h^T \dot{\ell}_{\theta_0}(X_i)$ tend vers 0 dans $L^2(P_{\theta_0})$, on peut écrire

$$n W_{n,i}^2 = \left(h^T \dot{\ell}_{\theta_0}(X_i)\right)^2 + A_{n,i}$$

où pour $i = 1, \ldots, n$ sont i.i.d. tels que $E_{\theta_0}|A_{n,i}|$ tend vers 0 quand n tend vers l'infini, et donc

$$\begin{aligned} \frac{1}{4}\sum_{i=1}^n W_{n,i}^2 &= \frac{1}{4n}\sum_{i=1}^n \left(h^T \dot{\ell}_{\theta_0}(X_i)\right)^2 + \frac{1}{4n}\sum_{i=1}^n A_{n,i} \\ &= \frac{1}{4} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1) + \frac{1}{4n}\sum_{i=1}^n A_{n,i} = \frac{1}{4} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1) \end{aligned}$$

par la LGN, puis la convergence vers 0 dans $L^1(P_{\theta_0})$ de $\frac{1}{n}\sum_{i=1}^n A_{n,i}$.
Montrons enfin (3.5). On a

$$\left|\sum_{i=1}^n W_{n,i}^2 R(W_{n,i})\right| \leq \max_{1\leq i\leq n}|R(W_{n,i})|\sum_{i=1}^n W_{n,i}^2 = O_{P_{\theta_0}}\left(\max_{1\leq i\leq n}|R(W_{n,i})|\right).$$

Montrons donc que $\max_{1\leq i\leq n}|R(W_{n,i})| = o_{P_{\theta_0}}(1)$. Comme $R(u)$ tend vers 0 quand u tend vers 0, pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que si $|R(u)| \geq \epsilon$, alors $|u| \geq \delta$. Puis

$$\begin{aligned}
P_{\theta_0}\left(\max_{1\leq i\leq n}|R(W_{n,i})| \geq \epsilon\right) &\leq \sum_{i=1}^n P_{\theta_0}\left(|W_{n,i}| \geq \delta\right) \\
&= nP_{\theta_0}\left(\left(h^T\dot{\ell}_{\theta_0}(X_1)\right)^2 + A_{n,1} \geq n\delta^2\right) \\
&\leq nP_{\theta_0}\left(\left(h^T\dot{\ell}_{\theta_0}(X_1)\right)^2 \geq n\frac{\delta^2}{2}\right) + nP_{\theta_0}\left(A_{n,1} \geq n\frac{\delta^2}{2}\right) \\
&\leq \frac{2}{\delta^2}E_{\theta_0}\left[\left(h^T\dot{\ell}_{\theta_0}(X_1)\right)^2 \mathbb{1}_{\left(h^T\dot{\ell}_{\theta_0}(X_1)\right)^2 \geq n\frac{\delta^2}{2}}\right] + \frac{2}{\delta^2}E_{\theta_0}\left|A_{n,1}\right|
\end{aligned}$$

tend vers 0 quand n tend vers l'infini.

On peut maintenant obtenir le théorème asymptotique de l'**e.m.v. (estimateur du maximum de vraisemblance)**.

Théorème 3.2.2. *On suppose que $\widehat{\theta}_n$ vérifie $\ell_n(\widehat{\theta}_n) \geq \sup_\theta \ell_n(\theta) - o_{P_{\theta_0}}(1)$. On suppose qu'il existe un voisinage A de θ_0 tel que :*

1. *$\theta \mapsto \sqrt{p_\theta(x)}$ est $\mathcal{D}^1$ sur A*
2. *$\theta \mapsto I_\theta$ est bien définie et continue sur A,*
3. *Il existe $H \in L^2(P_{\theta_0})$ tel que pour tous $(\theta_1, \theta_2) \in A^2$,*

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \|\theta_1 - \theta_2\|H(x),$$

4. *I_{θ_0} est inversible*
5. *$\widehat{\theta}_n$ converge en P_{θ_0}-probabilité vers θ_0.*

Alors

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = I_{\theta_0}^{-1}\mathbb{G}_n\dot{\ell}_{\theta_0} + o_{P_{\theta_0}}(1).$$

En particulier, $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$ converge en loi vers $\mathcal{N}_k(0, I_{\theta_0}^{-1})$.

Remarque. Si Θ est compact et si l'hypothèse 3. vaut sur Θ (et pas seulement sur A), alors $\widehat{\theta}_n$ converge en P_{θ_0}-probabilité vers θ_0 (exercice : le démontrer !).

Preuve. On va vérifier les hypothèses du Théorème 2.3.6. On a bien, comme remarqué lors de la Proposition 3.1.2, que $\theta \mapsto \log p_\theta(x)$ est $\mathcal{D}^1$ sur A pour μ-presque tout x, et son gradient en θ_0 est $\dot{\ell}_{\theta_0}$. Il reste à montrer que $\theta \mapsto P_{\theta_0}\log p_\theta$ est $\mathcal{D}^2$ en θ_0 de hessienne

$-I_{\theta_0}$.
Par la Proposition 3.1.2, le modèle est d.m.q. en θ_0, et donc, par le Théorème 3.2.1, pour toute suite $(h_n)_n$ de $\mathbb{R}^k$ convergeant vers un $h \in \mathbb{R}^k$:

$$\mathbb{G}_n\left(\sqrt{n}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) - h^T\dot{\ell}_{\theta_0}\right) + nP_{\theta_0}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) = -\frac{1}{2}h^T I_{\theta_0} h + o_{P_{\theta_0}}(1).$$

On va montrer que

$$\mathbb{G}_n\left(\sqrt{n}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) - h^T\dot{\ell}_{\theta_0}\right) = o_{P_{\theta_0}}(1),$$

ce qui donnera que : pour toute suite $(h_n)_n$ de $\mathbb{R}^k$ convergeant vers un $h \in \mathbb{R}^k$,

$$nP_{\theta_0}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) = -\frac{1}{2}h^T I_{\theta_0} h + o(1),$$

ce qui est suffisant pour obtenir le résultat souhaité. On a

$$Var_{\theta_0}\left[\mathbb{G}_n\left(\sqrt{n}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) - h^T\dot{\ell}_{\theta_0}\right)\right] = \int\left(\sqrt{n}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) - h^T\dot{\ell}_{\theta_0}\right)^2 d\mu.$$

$\sqrt{n}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) - h^T\dot{\ell}_{\theta_0}$ tend vers 0 μ-presque partout, et est dominée pour n assez grand par $\|h\|(H + \|\dot{\ell}_{\theta_0}\|)$ donc par convergence dominée

$$Var_{\theta_0}\left[\mathbb{G}_n\left(\sqrt{n}\left(\log p_{\theta_0+\frac{h_n}{\sqrt{n}}} - \log p_{\theta_0}\right) - h^T\dot{\ell}_{\theta_0}\right)\right] = o(1).$$

Questions :
L'e.m.v. est il optimal, et en quel sens ? Si T_n est un estimateur de θ basé sur $X_1, \ldots, X_n$, la variance asymptotique de $\sqrt{n}(T_n - \theta)$ est-elle toujours minorée par I_θ^{-1} ? La loi gaussienne centrée de variance I_θ^{-1} est-elle optimale comme limite en loi de $\sqrt{n}(T_n - \theta)$ et en quel sens ?

Contre-exemple de Hodge :
Considérons la situation où $P_\theta = \mathcal{N}(\theta, 1)$. L'information de Fisher en tout θ est $I_\theta = 1$, l'e.m.v. est $T_n = \frac{1}{n}\sum_{i=1}^n X_i$. Soit maintenant

$$S_n = \begin{cases} T_n & \text{si} \quad |T_n| \geq n^{-1/4} \\ 0 & \text{si} \quad |T_n| < n^{-1/4}. \end{cases}$$

Si $\theta \neq 0$, alors $\sqrt{n}(S_n - \theta)$ converge en loi vers $\mathcal{N}(0, 1)$, et si $\theta = 0$, $\sqrt{n}(S_n - \theta)$ converge en probabilité vers 0, et même, pour toute suite r_n tendant vers l'infini, $r_n(S_n - \theta)$ converge en probabilité vers 0 (exercice : le démontrer !). Au sens de la convergence en loi regardée ponctuellement (θ par θ), S_n est "meilleur" que T_n, en ayant privilégié (arbitrairement) une valeur (la valeur 0). Par contre, si on regarde le risque quadratique au voisinage de 0,

$$E_{\frac{h}{\sqrt{n}}}\left[\sqrt{n}\left(S_n - \frac{h}{\sqrt{n}}\right)^2\right] \sim_{n\to+\infty} h^2$$

(exercice : le démontrer !) peut être arbitrairement grand !

3.3 Estimateurs efficaces au sens du risque asymptotique quadratique local

Pour minorer un risque maximum, on utilise classiquement que le risque maximum est plus grand que le risque bayésien. On introduit une probabilité sur $A \subset \Theta$ de densité q par rapport à Lebesgue, et si T est un estimateur de $g(\theta)$, on a toujours

$$\sup_{\theta \in A} E_\theta \left[(T - g(\theta))^2 \right] \geq \int_A E_\theta \left[(T - g(\theta))^2 \right] q(\theta) d\theta.$$

3.3.1 Inégalité de van Trees

On va commencer par le cas où Θ est un intervalle (a, b) de $\mathbb{R}$ et $g : \Theta \to \mathbb{R}$.

Théorème 3.3.1 (Inégalité de van Trees)**.** *On suppose que le modèle $\{p_\theta \mu, \theta \in \Theta\}$ est d.m.q. en tout θ, et que $\theta \mapsto p_\theta(x)$ est $\mathcal{C}^1$ pour tout x. On suppose que q est dérivable sur $[a, b]$, nulle au bord de Θ, et on note*

$$J(q) = \int_\Theta \frac{(q'(\theta))^2}{q(\theta)} d\theta.$$

On suppose en outre que g est $\mathcal{C}^1$ sur $[a, b]$ et telle que $\int_\Theta |g'(\theta)| \, q(\theta) d\theta < +\infty$. Alors si T est un estimateur :

$$\int_\Theta E_\theta \left[(T - g(\theta))^2 \right] q(\theta) d\theta \geq \frac{\left(\int_\Theta g'(\theta) q(\theta) d\theta \right)^2}{\int_\Theta I_\theta q(\theta) d\theta + J(q)}.$$

Preuve. La preuve est simple : Fubini, intégration par parties, et Cauchy-Schwarz. Tout d'abord, il n'y a rien à démontrer si $\int_\Theta I_\theta q(\theta) d\theta = +\infty$ ou $J(q) = +\infty$ ou $\int_\Theta E_\theta \left[(T - g(\theta))^2 \right] q(\theta) d\theta = +\infty$. Puis

$$\begin{aligned}
\int g'(\theta) q(\theta) d\theta &= \int \int g'(\theta) q(\theta) p_\theta(x) d\mu(x) d\theta \\
&= \int \left\{ \left[(g(\theta) - T(x)) \, q(\theta) p_\theta(x) \right]_{\theta=a}^{\theta=b} - \int (g(\theta) - T(x)) \, \frac{d}{d\theta} \left(q(\theta) p_\theta(x) \right) d\theta \right\} d\mu(x) \\
&= - \int \int (g(\theta) - T(x)) \left(q'(\theta) p_\theta(x) + q(\theta) \dot{p_\theta}(x) \right) d\theta d\mu(x) \\
&= - \int \int (g(\theta) - T(x)) \left(\frac{q'(\theta)}{q(\theta)} + \frac{\dot{p_\theta}(x)}{p_\theta(x)} \right) p_\theta d\mu(x) q(\theta) d\theta.
\end{aligned}$$

Puis par Cauchy-Schwarz dans $L^2(p_\theta d\mu(x) q(\theta) d\theta)$:

$$\begin{aligned}
\left(\int g'(\theta) q(\theta) d\theta \right)^2 &\leq \int \int (g(\theta) - T(x))^2 \, p_\theta d\mu(x) q(\theta) d\theta \times \int \int \left(\frac{q'(\theta)}{q(\theta)} + \frac{\dot{p_\theta}(x)}{p_\theta(x)} \right)^2 p_\theta d\mu(x) q(\theta) d\theta \\
&= \int_\Theta E_\theta \left[(T - g(\theta))^2 \right] q(\theta) d\theta \times \left[\int_\Theta I_\theta q(\theta) d\theta + J(q) \right].
\end{aligned}$$

Remarque. $J(q)$ est l'information de Fisher pour le modèle $\{q(\theta-\alpha)d\theta, \alpha \in \mathbb{R}\}$ (en étendant q sur $\mathbb{R}$, nulle en-dehors de (a,b)).
On peut étendre ce résultat à $(a,b) = \mathbb{R}$.

L'inégalité de van Trees a une généralisation multi-dimensionnelle, et peut se démontrer sous des hypothèses plus faibles. Voici un énoncé général (que l'on admettra). Ici $\Theta \subset \mathbb{R}^k$, et q est une densité de probabilité sur Θ.
On utilise la notion de fonction absolument continue, voir plus loin.

Théorème 3.3.2. *On suppose que le modèle $\{p_\theta \mu, \theta \in \Theta\}$ est d.m.q. en tout θ, que $q : \Theta \to \mathbb{R}$ et $g : \Theta \to \mathbb{R}^m$ sont absolument continues, on note ∇q le gradient de q et $Dg(\theta)$ la matrice différentielle de g (qui existent presque partout), et on note $J(q)$ la matrice*

$$J(q) = \int_\Theta \frac{\nabla q(\theta) \nabla q(\theta)^T}{q(\theta)} d\theta.$$

On suppose en outre que :
- *La trace de $J(q)$ est finie*
- *Les fonctions $\theta \mapsto q(\theta)$ et $\theta \mapsto q(\theta)g(\theta)$ tendent vers 0 quand θ tend vers le bord de Θ*
- *Les intégrales $\int_\theta \|g(\theta)\|^2 q(\theta)d\theta$ et $\int_\theta |Dg(\theta)_{i,j}| q(\theta)d\theta$, $i = 1, \dots, m$, $j = 1, \dots, k$, sont finies.*

Alors si T est un estimateur, la matrice suivante est semi-définie positive :

$$\begin{pmatrix} \int_\Theta E_\theta \left[(T - g(\theta)) (T - g(\theta))^T \right] q(\theta)d\theta & \left(\int_\Theta Dg(\theta) q(\theta) d\theta \right) \\ \left(\int_\Theta Dg(\theta) q(\theta) d\theta \right)^T & \int_\Theta I_\theta q(\theta) d\theta + J(q) \end{pmatrix}.$$

Si de plus $\int_\Theta I_\theta q(\theta)d\theta + J(q)$ est inversible, alors

$$\int_\Theta E_\theta \left[(T - g(\theta)) (T - g(\theta))^T \right] q(\theta)d\theta - \left(\int_\Theta Dg(\theta) q(\theta) d\theta \right) \left(\int_\Theta I_\theta q(\theta) d\theta + J(q) \right)^{-1} \left(\int_\Theta Dg(\theta) q(\theta) d\theta \right)^T$$

est semi-définie positive.

La preuve de ce théorème repose sur les mêmes étapes que précédemment. On montre tout d'abord que l'on peut appliquer le principe "intégration par parties couplé à Fubini" pour obtenir

$$\int \int (T(x) - g(\theta)) \left(\frac{\nabla q(\theta)}{q(\theta)} + \dot{\ell}_\theta(x) \right)^T p_\theta(x) q(\theta) d\mu(x) d\theta = \int Dg(\theta) q(\theta) d\theta.$$

Ensuite on a alors

$$\int \int \begin{pmatrix} T(x) - g(\theta) \\ \frac{\nabla q(\theta)}{q(\theta)} + \dot{\ell}_\theta(x) \end{pmatrix} \begin{pmatrix} (T(x) - g(\theta))^T & (\frac{\nabla q(\theta)}{q(\theta)} + \dot{\ell}_\theta(x))^T \end{pmatrix} p_\theta(x) q(\theta) d\mu(x) d\theta$$
$$= \begin{pmatrix} \int_\Theta E_\theta \left[(T - g(\theta)) (T - g(\theta))^T \right] q(\theta)d\theta & \left(\int_\Theta Dg(\theta) q(\theta) d\theta \right) \\ \left(\int_\Theta Dg(\theta) q(\theta) d\theta \right)^T & \int_\Theta I_\theta q(\theta) d\theta + J(q) \end{pmatrix}$$

qui montre que cette matrice est semi-définie positive et on achève comme pour l'inégalité de Cramer-Rao multidimensionnelle.

Quelques mots sur les fonctions absolument continues.
Une fonction réelle F est absolument continue sur un intervalle I de $\mathbb{R}$ si pour tout $\epsilon > 0$ il existe $\delta > 0$ tel que, pour toute suite $([a_n, b_n])_{n\geq 1}$ d'intervalles de I disjoints,

$$\sum_{n\geq 1} |b_n - a_n| < \delta \Longrightarrow \sum_{n\geq 1} |F(b_n) - F(a_n)| < \epsilon.$$

F est absolument continue sur $[a, b]$ si et seulement si il existe une fonction f intégrable telle que pour tout $x \in [a, b]$, $F(x) - F(a) = \int_a^x f(t)dt$. Alors F est presque partout dérivable de dérivée f.
Si $F : \mathbb{R}^k \to \mathbb{R}^m$ on dit que F est absolument continue si pour tout $j = 1, \ldots, m$, et tout $i = 1, \ldots, k$, $x \mapsto F_j(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_k)$ est absolument continue.

3.3.2 Estimateurs localement asymptotiquement minimax

On considère un modèle d.m.q. au voisinage de θ_0, et l'on cherche à minorer le risque quadratique minimax sur un voisinage de θ_0 de taille d'ordre $1/\sqrt{n}$ (voir contre-exemple de Hodge).

Considérons tout d'abord le cas simple unidimensionnel, avec $\Theta \subset \mathbb{R}$ et g fonction réelle. SI T_n est une suite d'estimateurs (basés sur $X_1, \ldots, X_n$) on veut donc minorer

$$\sup_{|\theta-\theta_0|\leq \frac{c}{\sqrt{n}}} E_\theta \left[n \left(T_n - g(\theta) \right)^2 \right].$$

Il s'agit donc de mettre une loi a priori sur $[\theta_0 - \frac{c}{\sqrt{n}}; \theta_0 + \frac{c}{\sqrt{n}}]$. Soit donc q une probabilité sur $[-1, 1]$, qui vérifie les hypothèses du Théorème 3.3.1, avec $J(q) < +\infty$. Soit q_n la densité de la loi de la variable aléatoire $\theta_0 + \frac{c}{\sqrt{n}}U$ où U a pour densité q, q_n est une densité de probabilité sur $[\theta_0 - \frac{c}{\sqrt{n}}; \theta_0 + \frac{c}{\sqrt{n}}]$,

$$q_n(\theta) = \frac{\sqrt{n}}{c} q\left(\frac{\sqrt{n}}{c}(\theta - \theta_0) \right).$$

Si l'on applique l'inégalité de van-Trees, on obtient

$$\begin{aligned} \sup_{|\theta-\theta_0|\leq \frac{c}{\sqrt{n}}} E_\theta \left[n \left(T_n - g(\theta) \right)^2 \right] &\geq \int_{\theta_0 - \frac{c}{\sqrt{n}}}^{\theta_0 + \frac{c}{\sqrt{n}}} E_\theta \left[n \left(T_n - g(\theta) \right)^2 \right] q_n(\theta)\, d\theta \\ &\geq \frac{n \left(\int_{\theta_0 - \frac{c}{\sqrt{n}}}^{\theta_0 + \frac{c}{\sqrt{n}}} g'(\theta) q_n(\theta)\, d\theta \right)^2}{\int_{\theta_0 - \frac{c}{\sqrt{n}}}^{\theta_0 + \frac{c}{\sqrt{n}}} n I_\theta q_n(\theta)\, d\theta + J(q_n)}. \end{aligned}$$

Mais $J(q_n) = \frac{n}{c^2} J(q)$. Donc on obtient facilement :

Théorème 3.3.3. *On suppose $\Theta \subset \mathbb{R}$, et que le modèle est d.m.q. au voisinage de θ_0, d'information de Fisher I_θ continue au voisinage de θ_0, avec $I_{\theta_0} \neq 0$. On suppose que la fonction réelle g est de classe $\mathcal{C}^1$ au voisinage de θ_0. Alors pour toute suite T_n d'estimateurs*

$$\liminf_{c\to+\infty} \liminf_{n\to+\infty} \sup_{|\theta-\theta_0|\leq \frac{c}{\sqrt{n}}} E_\theta \left[n \left(T_n - g(\theta)\right)^2 \right] \geq \frac{(g'(\theta_0))^2}{I_{\theta_0}}.$$

On dit alors que T_n **est localement asymptotiquement minimax** si

$$\liminf_{c\to+\infty} \liminf_{n\to+\infty} \sup_{|\theta-\theta_0|\leq \frac{c}{\sqrt{n}}} E_\theta \left[n \left(T_n - g(\theta)\right)^2 \right] = \frac{(g'(\theta_0))^2}{I_{\theta_0}}.$$

Remarque. Si l'on revient au contre-exemple de Hodge, on a

$$\liminf_{c\to+\infty} \liminf_{n\to+\infty} \sup_{\theta\leq \frac{c}{\sqrt{n}}} E_\theta \left[n \left(S_n - \theta\right)^2 \right] = +\infty,$$

et S_n n'est pas localement asymptotiquement minimax.

On peut étendre ce résultat de minoration asymptotique en situation multidimensionnelle.
On se place dans la situation du Théorème 3.3.2, $\Theta \subset \mathbb{R}^k$ et $g : \Theta \to \mathbb{R}^m$. Soit q une densité bornée et différentiable sur la boule unité $B_k(0,1)$ dans $\mathbb{R}^k$, telle que $q(v) = 0$ si v est sur la frontière de $B_k(0,1)$, et telle que la trace de $J(q)$ soit finie. Soit $c > 0$. Alors pour n assez grand, la boule ouverte centrée en θ_0 et de rayon $\frac{c}{\sqrt{n}}$ est incluse dans Θ. Soit q_n la densité de la variable aléatoire $U_n = \theta_0 + \frac{cV}{\sqrt{n}}$ où V a pour densité q. Alors $Jq_n = \frac{n}{c^2} J(q)$. En appliquant le Théorème 3.3.2, si $(T_n)_{n\geq 1}$est une suite d'estimateurs, pour n assez grand, si $\frac{1}{c^2} J(q) + \int_\Theta I_\theta q(\theta) d\theta$ est inversible, alors

$$\int_{\|v\|\leq 1} E_{\theta_0 + \frac{cv}{\sqrt{n}}} \left[\left(T_n - g(\theta_0 + \frac{cv}{\sqrt{n}})\right) \left(T_n - g(\theta_0 + \frac{cv}{\sqrt{n}})\right)^T \right] q(v)\, dv$$
$$- \left(\int_{\|v\|\leq 1} Dg(\theta_0 + \frac{cv}{\sqrt{n}})\, q(v)\, dv \right) \left(\frac{n}{c^2} J(q) + n \int_{\|v\|\leq 1} I_{\theta_0 + \frac{cv}{\sqrt{n}}}\, q(v)\, dv \right)^{-1} \left(\int_{\|v\|\leq 1} Dg(\theta_0 + \frac{cv}{\sqrt{n}})\, q(v)\, dv \right)^T$$

est semi-définie positive. On obtient alors facilement :

Théorème 3.3.4. *On suppose $\Theta \subset \mathbb{R}^k$, et que le modèle est d.m.q. au voisinage de θ_0, d'information de Fisher I_θ continue au voisinage de θ_0 et inversible. On suppose que la fonction g à valeurs dans $\mathbb{R}^m$ est de classe $\mathcal{C}^1$ au voisinage de θ_0. Alors pour toute suite T_n d'estimateurs, pour tout $U \in \mathbb{R}^m$,*

$$\liminf_{c\to+\infty} \liminf_{n\to+\infty} \sup_{\|v\|\leq 1} U^T E_{\theta_0 + \frac{cv}{\sqrt{n}}} \left[n \left(T_n - g(\theta_0 + \frac{cv}{\sqrt{n}})\right) \left(T_n - \psi(\theta_0 + \frac{cv}{\sqrt{n}})\right)^T \right] U$$
$$\geq U^T Dg(\theta_0) I_{\theta_0}^{-1} Dg(\theta_0)^T U.$$

Aussi,

$$\liminf_{c\to+\infty}\liminf_{n\to+\infty}\sup_{\|v\|\leq 1} E_{\theta_0+\frac{cv}{\sqrt{n}}}\left[n\|T_n - g(\theta_0+\frac{cv}{\sqrt{n}})\|^2\right] \geq Tr\ \left[Dg(\theta_0)I_{\theta_0}^{-1}Dg(\theta_0)^T\right].$$

3.4 Estimateurs réguliers et efficaces au sens du théorème de convolution

On va maintenant chercher à donner un sens à une optimalité asymptotique "en loi". Le cadre est celui de suites $(X_n)_{n\geq 1}$ i.i.d., si on dit que T_n est un estimateur, cela sous-entend une fonction mesurable de $X_1,\ldots,X_n$.

3.4.1 Estimateurs réguliers et théorème de convolution

Soit $g:\Theta\to\mathbb{R}^m$ une fonction. On dit que T_n est un **estimateur régulier** en θ_0 de $g(\theta)$ si, pour tout $c\in\mathbb{R}^k$,

$$\sqrt{n}\left(T_n - g\left(\theta_0+\frac{c}{\sqrt{n}}\right)\right)$$

converge en loi sous $P_{\theta_0+\frac{c}{\sqrt{n}}}$ vers L_{θ_0} (qui ne dépend pas de c, autrement dit la même loi pour tout c).

Remarque. La convergence a lieu sous une loi qui varie avec n. Donc on ne peut pas l'obtenir en appliquant un TLC comme on l'a fait souvent. Donc pour obtenir la régularité d'un estimateur, on pourra avoir besoin de nouveaux outils : ce sera la contiguïté, que l'on définira et étudiera plus loin.

Le théorème essentiel (que l'on admettra) est le suivant.

Théorème 3.4.1. *On suppose le modèle d.m.q. en θ_0 intérieur à Θ, d'information de Fisher I_{θ_0} inversible. On suppose aussi g différentiable en θ_0. Alors, si T_n est un estimateur régulier en θ_0 de $g(\theta)$, il existe une probabilité M sur $\mathbb{R}^m$ telle que*

$$L_{\theta_0} = M\star\mathcal{N}_m\left(0, Dg\left(\theta_0\right)I_{\theta_0}^{-1}Dg\left(\theta_0\right)^T\right).$$

Ici, L_{θ_0} est la loi limite de $\sqrt{n}\left(T_n - g\left(\theta_0+\frac{c}{\sqrt{n}}\right)\right)$ pour tout $c\in\mathbb{R}^k$.

Remarques. Comme la convolution "étale" la loi, $\mathcal{N}_m\left(0, Dg\left(\theta_0\right)I_{\theta_0}^{-1}Dg\left(\theta_0\right)^T\right)$ est la loi optimale pour les estimateurs réguliers.
L'estimateur de Hodge S_n est régulier en $\theta\neq 0$ mais n'est pas régulier en 0 (exercice : le démontrer !).

Comment voir si T_n est régulier ? Comment obtenir la loi asymptotique sous $P_{\theta_0+\frac{c}{\sqrt{n}}}$? Il s'agit de trouver une probabilité L sur $\mathbb{R}^m$ telle que pour toute fonction réelle φ continue bornée,

$$\lim_{n\to+\infty} E_{\theta_0+\frac{c}{\sqrt{n}}}\varphi\left(T_n\right) = \int\varphi(t)dL(t)$$

Comme

$$E_{\theta_0+\frac{c}{\sqrt{n}}}\varphi\left(T_n\right) = \int \varphi\left(T_n(x_1,\ldots,x_n)\right) dP^{\otimes n}_{\theta_0+\frac{c}{\sqrt{n}}}(x_1,\ldots,x_n)$$
$$= \int \varphi\left(T_n(x_1,\ldots,x_n)\right) \left(\frac{\prod_{i=1}^n p_{\theta_0+\frac{c}{\sqrt{n}}}(x_i)}{\prod_{i=1}^n p_{\theta_0}(x_i)}\right) dP^{\otimes n}_{\theta_0}(x_1,\ldots,x_n),$$

il s'agit de manière générale de trouver la loi asymptotique d'une v.a. Z_n sous Q_n quand on sait dire des choses asymptotiques sous P_n en utilisant, si Q_n est absolument continue par rapport à P_n,

$$E_{Q_n}\left[\varphi\left(Z_n\right)\right] = E_{P_n}\left[\varphi\left(Z_n\right)\frac{dQ_n}{dP_n}\right].$$

Il est logique de se dire pour cela qu'il faut connaitre des choses sur la loi limite jointe de $\left(Z_n, \frac{dQ_n}{dP_n}\right)$ sous P_n. C'est l'objet de la contiguïté introduite et étudiée par Hajek et Le Cam.

3.4.2 Contiguïté

Soient $(Q_n)_{n\geq 1}$ et $(P_n)_{n\geq 1}$ deux suite de probabilités sur les mêmes espaces (pouvant changer avec n). On dit que Q_n est **contiguë** par rapport à P_n si pour toute suite d'événements $(A_n)_{n\geq 1}$, si $P_n(A_n)$ tend vers 0, alors $Q_n(A_n)$ tend vers 0.

Remarque. La contiguïté est une notion d'absolue continuité asymptotique.
Si P et Q sont deux probabilités sur un même espace, Q est absolument continue par rapport à P si, pour tout événement A, si $P(A) = 0$, alors $Q(A) = 0$.
Si Q est absolument continue par rapport à P, alors il existe une fonction mesurable q telle que $Q = qP$, q est la densité de Q par rapport à P.
On peut toujours décomposer Q en la somme $Q^a + Q^o$, où Q^a est absolument continue par rapport à P et Q^o est étrangère à P, c'est à dire qu'il existe A tel que $Q^o(A) = 0$ et $P(\overline{A}) = 0$ ($\overline{A}$ est le complémentaire de A). La notation $\frac{dQ}{dP}$ désigne la densité de Q^a par rapport à P. C'est une fonction mesurable $x \mapsto \frac{dQ}{dP}(x)$. Donc $\frac{dQ}{dP}(X)$ est une variable aléatoire, que l'on notera encore $\frac{dQ}{dP}$. Par exemple : $E_P\left(\frac{dQ}{dP}\right)$ signifie l'espérance de $\frac{dQ}{dP}(X)$ lorsque X est de loi P.
On a

$$Q \text{ absolument continue par rapport à } P \Longleftrightarrow E_P\left(\frac{dQ}{dP}\right) = 1 \Longleftrightarrow Q\left(\frac{dP}{dQ} > 0\right) = 1.$$

Remarquons que considérées sous P_n, $\frac{dQ_n}{dP_n}$ est toujours une suite tendue : pour tout $M > 0$, par l'inégalité de Markov,

$$P_n\left(\frac{dQ_n}{dP_n} \geq M\right) \leq \frac{1}{M} E_{P_n}\left(\frac{dQ_n}{dP_n}\right) \leq \frac{1}{M},$$

qui a donc au moins une valeur d'adhérence pour la topologie de la convergence en loi (valeur d'adhérence = limite d'une suite extraite).

Théorème 3.4.2 ("Premier Lemme de Le-Cam"). *Les propriété suivantes sont équivalentes :*

1. *Q_n est contiguë par rapport à P_n*
2. *Si V est une valeur d'adhérence (pour la convergence en loi) de $\frac{dQ_n}{dP_n}$ sous P_n, alors $E(V) = 1$.*
3. *Si U est une valeur d'adhérence (pour la convergence en loi) de $\frac{dP_n}{dQ_n}$ sous Q_n, alors $P(U > 0) = 1$.*
4. *Si T_n converge en probabilité vers 0 sous P_n, alors T_n converge en probabilité vers 0 sous Q_n.*

Preuve. Voir livre de van der Vaart.

Théorème 3.4.3 ("Troisième Lemme de Le-Cam"). *Soit Z_n une variable aléatoire telle que sous P_n, $(Z_n, \frac{dQ_n}{dP_n})$ converge en loi vers la variable aléatoire (Z, V). Si Q_n est contiguë par rapport à P_n, alors la mesure L définie par $L(B) = E(\mathbb{1}_{Z\in B}V)$ pour tout événement B est une probabilité, et Z_n converge en loi vers L sous Q_n.*

Preuve. Tout d'abord, L est bien une mesure positive, et de masse égale à 1 par le point 2. du premier Lemme de Le-Cam. On a alors pour toute fonction φ mesurable bornée (ou mesurable positive), $\int \varphi(u)dL(u) = E(\varphi(Z)V$. Pour montrer que Z_n converge en loi vers L sous Q_n il suffit donc de montrer que pour toute fonction φ mesurable positive,

$$\liminf_{n\to+\infty} E_{Q_n}\left[\varphi\left(Z_n\right)\right] \geq E\left[\varphi\left(Z\right)V\right].$$

Remarquons que l'on n'a pas supposé que Q_n est absolument continue par rapport à P_n, donc on a seulement l'inégalité (qui n'est pas forcément une égalité) :

$$E_{Q_n}\left[\varphi\left(Z_n\right)\right] \geq E_{P_n}\left[\varphi\left(Z_n\right)\frac{dQ_n}{dP_n}\right].$$

Mais la fonction $(z, v) \mapsto \varphi(z)v$ est continue positive, donc comme $(Z_n, \frac{dQ_n}{dP_n})$ converge en loi vers la variable aléatoire (Z, V),

$$\liminf_{n\to+\infty} E_{P_n}\left[\varphi\left(Z_n\right)\frac{dQ_n}{dP_n}\right] \geq E\left[\varphi\left(Z\right)V\right].$$

et la preuve est finie.

3.4.3 Application aux modèles d.m.q.

On va voir que dans les modèles d.m.q., sous les hypothèses de normalité asymptotique, l'e.m.v. est efficace au sens du théorème de convolution, ainsi que tous les estimateurs de fonction régulières de θ obtenus par "plug-in" de l'e.m.v.

Proposition 3.4.1. *Si le modèle est d.m.q. en θ, alors pour tout $c \in \mathbb{R}^k$, $P_{\theta+\frac{c}{\sqrt{n}}}^{\otimes n}$ est contiguë par rapport à $P_\theta^{\otimes n}$.*

Preuve. Notons $Q_n = P_{\theta+\frac{c}{\sqrt{n}}}^{\otimes n}$ et $P_n = P_\theta^{\otimes n}$. On a, si le modèle est d.m.q.

$$\log \frac{dQ_n}{dP_n} = \mathbb{G}_n\left(h^T \dot{\ell}_\theta\right) - \frac{1}{2} c^T I_{\theta_0} c + o_{P_\theta}(1).$$

Donc par le TCL, $\log \frac{dQ_n}{dP_n}$ converge en loi sous P_n vers W de loi $\mathcal{N}(-\frac{1}{2}\sigma^2, \sigma^2)$, en notant $\sigma^2 = c^T I_{\theta_0} c$; donc par image continue $\frac{dQ_n}{dP_n}$ converge en loi sous P_n vers $V = e^W$ (unique valeur d'adhérence). On écrit $W = -\frac{1}{2}\sigma^2 + \sigma U$ avec U de loi $\mathcal{N}(0,1)$, et

$$E(V) = E\left(e^{-\frac{1}{2}\sigma^2 + \sigma U}\right) = e^{-\frac{1}{2}\sigma^2 + \frac{1}{2}\sigma^2} = 1$$

donc Q_n est contiguë par rapport à P_n par le premier Lemme de Le-Cam.

Proposition 3.4.2. *On suppose que Z_n est une variable aléatoire à valeurs dans $\mathbb{R}^m$. Si*

$$\left(Z_n, \log \frac{dP_{\theta+\frac{c}{\sqrt{n}}}^{\otimes n}}{dP_\theta^{\otimes n}}\right)$$

converge en loi sous $P_\theta^{\otimes n}$ vers

$$\mathcal{N}_{m+1}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}; \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right)$$

alors Z_n converge en loi sous $P_{\theta+\frac{c}{\sqrt{n}}}$ vers $\mathcal{N}_m(\mu+\tau; \Sigma)$.

Preuve. Avec les notations de la proposition précédente, Q_n est contiguë par rapport à P_n, donc par le troisième Lemme de Le-Cam, Z_n converge en loi vers L donnée par $\int \varphi(x) dL(x) = E[\varphi(Z)e^W]$, avec (Z,W) de loi $\mathcal{N}_{m+1}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}; \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right)$. Calculons la fonction caractéristique de L. Pour tout $t \in \mathbb{R}^m$,

$$\begin{aligned}
\int e^{i\langle t,x\rangle_{\mathbb{R}^m}} dL(x) &= E\left(e^{i\langle t,Z\rangle_{\mathbb{R}^m}} e^W\right) \\
&= E\left(e^{i\langle (t,-i),(Z,W)\rangle_{\mathbb{R}^{m+1}}}\right) \\
&= \exp\left[i\langle (t,-i),(\mu,-\frac{1}{2}\sigma^2)\rangle_{\mathbb{R}^{m+1}} - \frac{1}{2}(t,-i)^T \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix} \begin{pmatrix} t \\ -i \end{pmatrix}\right] \\
&= \exp\left[i\langle t, \mu+\tau\rangle_{\mathbb{R}^m} - \frac{1}{2} t^T \Sigma t\right]
\end{aligned}$$

qui est la fonction caractéristique de $\mathcal{N}_m(\mu+\tau;\Sigma)$.

Corollaire 3.4.1. *Si le modèle est d.m.q. en θ_0 et que $\widehat{\theta}_n$ vérifie*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = I_{\theta_0}^{-1} \mathbb{G}_n\left(\dot{\ell}_{\theta_0}\right) + o_{P_{\theta_0}}(1).$$

alors $\widehat{\theta}_n$ est efficace au sens du théorème de convolution.
Si de plus $g : \Theta \to \mathbb{R}^m$ est différentiable en θ_0, alors $g(\widehat{\theta}_n)$ est efficace au sens du théorème de convolution.

Preuve. A faire en exercice !

3.5 Exercices

Exercice 3.5.1. Vitesse d'estimation paramétrique

Soit $\{P_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$, un modèle dominé sur $\mathbb{R}^d$ qui soit en tout θ : identifiable, DMQ (différentiable en moyenne quadratique) et d'information de Fisher I_θ inversible. Soient f_θ la densité de P_θ par rapport à μ, $\dot{l}_\theta$ la fonction score en θ, θ_0 un point intérieur à Θ. Soit $X_1, \ldots, X_n$ un n-échantillon de P_{θ_0}. Le but de l'exercice est de montrer que sous ces hypothèses et si en plus Θ est compact, alors il existe un estimateur $\sqrt{n}$-consistant de θ_0.

1. a) On note H la distance de Hellinger, c'est à dire $H(P_{\theta_1}, P_{\theta_2}) = (\int(\sqrt{f_{\theta_1}} - \sqrt{f_{\theta_2}})^2 d\mu)^{1/2}$. Montrer que

$$0 < \liminf_{\|h\|\to 0} \frac{H(P_{\theta_0+h}, P_{\theta_0})}{\|h\|} \leq \limsup_{\|h\|\to 0} \frac{H(P_{\theta_0+h}, P_{\theta_0})}{\|h\|} < \infty.$$

   b) En déduire que les deux assertions suivantes sont équivalentes pour $(h_n)_{n\geq 1}$ tel que $\|h_n\| \to 0$.
      i. $0 < \liminf_{n\to\infty} \sqrt{n} H(P_{\theta_0+h_n}, P_{\theta_0}) \leq \limsup_{n\to\infty} \sqrt{n} H(P_{\theta_0+h_n}, P_{\theta_0}) < \infty$,
      ii. $0 < \liminf_{n\to\infty} \sqrt{n}\|h_n\| \leq \limsup_{n\to\infty} \sqrt{n}\|h_n\| < \infty$.

Dans la suite, on suppose Θ compact. On définit la distance entre deux lois sur $\mathbb{R}^d$ par

$$d(P, Q) = \sup_{x\in\mathbb{R}^d} |F_P(x) - F_Q(x)|,$$

où F_P (resp. F_Q) est la fonction de répartition associée à P (resp. Q). Soient $X_1, \ldots, X_n$ un n-échantillon de P_{θ_0}, $\mathbb{P}_n$ la mesure empirique associée, et T_n un élément de Θ tel que

$$d(P_{T_n}, \mathbb{P}_n) \leq d(P_{\theta_0}, \mathbb{P}_n) + O\left(\frac{1}{\sqrt{n}}\right).$$

On va montrer que T_n est $\sqrt{n}$-consistant de θ_0, i.e. $\sqrt{n}(T_n - \theta_0) = O_{P_{\theta_0}}(1)$. On rappelle (on admettra) que $\sqrt{n} d(P_{\theta_0}, \mathbb{P}_n) = O_{P_{\theta_0}}(1)$.

2. Montrer que pour tout θ,

$$\int \mid f_{\theta+h} - f_\theta - h^T \dot{l}_\theta f_\theta \mid d\mu = o(\|h\|).$$

3. Montrer que $\theta \mapsto d(P_\theta, P_{\theta_0})$ est continue en θ, puis que T_n est consistant.
4. Montrer que

$$\liminf_{h\to 0} \frac{1}{\|h\|} d(P_{\theta_0+h}, P_{\theta_0}) > 0.$$

   En déduire qu'il existe $\epsilon > 0$, $c > 0$, tels que

$$\|\theta - \theta_0\| \leq \epsilon \Longrightarrow d(P_\theta, P_{\theta_0}) \geq c\|\theta - \theta_0\|.$$

5. Montrer que T_n est $\sqrt{n}$-consistant.

Exercice 3.5.2. Une application de l'inégalité de Van Trees : estimation de θ^α avec un échantillon de $\mathcal{N}(\theta, 1)$.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes de loi $\mathcal{N}(\theta, 1)$, avec $\theta \geq 0$. On s'intéresse à l'estimation de θ^α pour un $0 < \alpha < 1$.

1. Soit $\theta_0 > 0$. Montrer que sous $\theta = \theta_0$, $\sqrt{n}(|\overline{X}|^\alpha - \theta_0^\alpha)$ converge en loi vers $\mathcal{N}(0, \alpha^2\theta_0^{2\alpha-2})$.
   Montrer que si ψ_n est un estimateur,
   $$\liminf_{c\to+\infty} \liminf_{n\to+\infty} \sup_{|\theta-\theta_0|\leq \frac{c}{\sqrt{n}}} nE_\theta \left(\psi_n - \theta^\alpha\right)^2 \geq \alpha^2\theta_0^{2\alpha-2}.$$

2. Soit $\delta > 0$, et soit λ une densité de probabilité sur $[0, M]$ continûment dérivable, telle que $\lambda(0) = \lambda(M) = 0$. On note $J(\lambda) = \int_0^M \frac{\lambda'^2(u)}{\lambda(u)} du$. En utilisant la densité de probabilité $\lambda_{\delta,c}(\theta) = \frac{1}{c}\lambda(\frac{\theta-\delta}{c})$, montrer que pour tout entier n, si ψ_n est un estimateur,
   $$\liminf_{n\to+\infty} \sup_{\theta\geq\delta} nE_\theta \left(\psi_n - \theta^\alpha\right)^2 \geq \frac{\alpha^2}{\delta^{2-2\alpha}},$$
   en déduire que
   $$\lim_{n\to+\infty} \sup_{\theta\geq 0} nE_\theta \left(\psi_n - \theta^\alpha\right)^2 = +\infty.$$

3. Montrer que $u^{\alpha-1}\lambda(u)$ est intégrable en 0, puis que
   $$\lim_{\delta\to 0} \int_0^M (\delta + cu)^{\alpha-1} \lambda(u)\, du = c^{\alpha-1} \int_0^M u^{\alpha-1}\lambda(u)\, du.$$
   En déduire que si ψ_n est un estimateur,
   $$\sup_{\theta\geq 0} E_\theta \left(\psi_n - \theta^\alpha\right)^2 \geq \frac{Ac^{2(\alpha-1)}\alpha^2}{n + J(\lambda)c^{-2}}$$
   où $A = (\int_0^M u^{\alpha-1}\lambda(u)du)^2$, puis que
   $$\sup_{\theta\geq 0} E_\theta \left(\psi_n - \theta^\alpha\right)^2 \geq \frac{(1-\alpha)^{1-\alpha}\alpha^{2+\alpha}A}{n^\alpha J(\lambda)^{1-\alpha}}.$$
   Remarquer la vitesse d'estimation et comparer avec la question 1. Commentaire ?

4. En étudiant la fonction $f(h) = (x+h)^\alpha - x^\alpha - h^\alpha$ pour $h \geq 0$ et $x > 0$, montrer que pour tous réels x et y,
   $$||y|^\alpha - |x|^\alpha| \leq |y - x|^\alpha.$$

5. Montrer que $\psi_n = |\overline{X}|^\alpha$ réalise la vitesse, i.e. il existe une constante $C(\alpha)$ telle que

$$n^\alpha \sup_{\theta \geq 0} E_\theta \left(|\overline{X}|^\alpha - \theta^\alpha \right)^2 \leq C(\alpha).$$

Exercice 3.5.3. Soit $P_n = P = \mathcal{N}(0,1)$ et $Q_n = \mathcal{N}(m_n, 1)$. Montrer que P_n est contigüe par rapport à Q_n si et seulement si Q_n est contigüe par rapport à P_n si et seulement si la suite $(m_n)_{n\in\mathbb{N}}$ est bornée.

Exercice 3.5.4. Soit P_n la loi de la moyenne empirique d'un n-échantillon de $\mathcal{N}(0,1)$ et Q_n la loi de la moyenne empirique d'un n-échantillon de $\mathcal{N}(m_n, 1)$. Montrer que P_n et Q_n sont mutuellement contigües si et seulement si la suite $(\sqrt{n}m_n)_{n\in\mathbb{N}}$ est bornée.

Exercice 3.5.5. Soit P_n la loi d'un n-échantillon de la loi uniforme sur $[0,1]$ et Q_n la loi d'un n-échantillon de la loi uniforme sur $[0, 1 + \frac{1}{n}]$. Montrer que P_n est contigüe par rapport à Q_n. Q_n est-elle contigüe par rapport à P_n ?

Exercice 3.5.6. On note $\|\cdot\|$ la distance en variation, i.e. si P et Q sont deux probabilités sur un même espace probabilisable,

$$\|P - Q\| = \sup_A |P(A) - Q(A)|.$$

1. Soit μ une mesure dominante (par exemple $\frac{P+Q}{2}$) et p et q les densités respectives de P et Q par rapport à μ. Montrer que

$$\|P - Q\| = \frac{1}{2} \int |p - q| \, d\mu = \int (p - q)_+ \, d\mu.$$

2. Montrer que si $\|P_n - Q_n\|$ tend vers 0 quand n tend vers l'infini, alors P_n et Q_n sont mutuellement contigües.
3. Montrer que si P_n et Q_n sont mutuellement contigües, alors $\limsup_{n\to+\infty} \|P_n - Q_n\| < 1$.
4. Soit $\epsilon > 0$. Trouver une suite de probabilités P_n et Q_n qui sont mutuellement contigües mais telles que $\|P_n - Q_n\|$ tend vers $1 - \epsilon$ quand n tend vers l'infini.

Exercice 3.5.7. Soit $P_{\theta,f}$ la loi de $\theta + \epsilon$, ϵ de densité f paire, continûment dérivable et telle que $\int \frac{(f')^2}{f} dx < \infty$ (modèle de translation vu au TD 5). On veut estimer $g(\theta_0) = P_{\theta_0,f_0}(X \leq z)$ pour un réel z donné. On suppose que f_0 est connue, et que $\int x^2 f_0(x) dx < +\infty$. On considère les deux estimateurs

— $U_n = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq z}$
— $V_n = F_0(z - \overline{X})$, F_0 fonction de répartition associée à la densité f_0.

1. Montrer que U_n et V_n sont consistants.
2. Montrer que $\sqrt{n}(U_n - g(\theta_0))$ converge en loi vers une gaussienne centrée dont on précisera la variance.
   Montrer que $\sqrt{n}(V_n - g(\theta_0))$ converge en loi vers une gaussienne centrée dont on précisera la variance.
3. Soit c un réel non nul. Sous $P_{\theta_0+\frac{c}{\sqrt{n}},f_0}$, quelle est la loi asymptotique de $\sqrt{n}(U_n - g(\theta_0+\frac{c}{\sqrt{n}}))$? Sous $P_{\theta_0+\frac{c}{\sqrt{n}},f_0}$, quelle est la loi asymptotique de $\sqrt{n}(V_n - g(\theta_0+\frac{c}{\sqrt{n}}))$? Ces deux estimateurs sont-ils réguliers ? Efficaces ?

4 Estimation semi-paramétrique

Le cadre est toujours celui d'une suite de variables aléatoires i.i.d. $(X_n)_{n\geq 1}$ de loi $P \in \mathcal{P}$. On s'intéresse maintenant au cas où $\mathcal{P}$ n'est pas nécessairement paramétrique, et où l'on veut estimer $\psi(P)$ avec ψ une fonction (connue) de $\mathcal{P}$ dans $\mathbb{R}$ ou $\mathbb{R}^k$. On va explorer la situation où on peut estimer $\psi(P)$ avec vitesse $\sqrt{n}$.

Quelques exemples :
- Moment : $\psi(P) = \int f dP$, f étant une fonction connue. On sait estimer avec l'estimateur empirique, mais existe-t-il un estimateur efficace et lequel si $\mathcal{P}$ est l'ensemble de toutes les lois de probabilité ?
- Centre de symétrie : $\mathcal{P}$ est l'ensemble des lois sur $\mathbb{R}$ de densité par rapport à Lebesgue $f(\cdot - \theta)$, $f \in \mathcal{F}$, $\mathcal{F}$ ensemble des densités de probabilité strictement positives sur $\mathbb{R}$ et centrées en 0. On peut estimer θ par la moyenne empirique, mais peut-on trouver un estimateur "meilleur" ? Et en restreignant $\mathcal{F}$?
- Régression : si on observe $X = (Y, Z)$, avec $Y = g_\theta(Z) + \epsilon$, (Z, ϵ) de loi η inconnue mais telle que ϵ soit centré et de variance 1.

Si $\mathcal{M}$ est un sous-modèle de $\mathcal{P}$ paramétré par un réel, i.e. $\mathcal{M} = \{P_\theta, \theta \in \Theta\} \subset \mathcal{P}$ avec $\Theta \subset \mathbb{R}$ et que la fonction g est définie par $g(\theta) = \psi(P_\theta)$, si $\mathcal{M}$ est d.m.q. en θ et g dérivable en θ, on sait que pour estimer $\psi(P_\theta)$, la borne inférieure de variance (que ce soit pour l'efficacité au sens du risque minimax local asymptotique, ou au sens du théorème de convolution) est $g'(\theta)^2/I_\theta$. L'idée est de maximiser cette borne inférieure en les sous-modèles $\mathcal{M}$ possibles de dimension 1 (pour lesquels $\Theta \subset \mathbb{R}$) contenant le point $P = P_\theta$.

4.1 Ensembles tangents et fonctions d'influence

On note P_0 la probabilité de $\mathcal{P}$ pour laquelle on s'intéresse à l'efficacité des estimateurs pour estimer $\psi(P_0)$. Si $\mathcal{M}$ est un sous-modèle de dimension 1 passant par P_0 et d.m.q. pour le paramètre en ce point et de score g, alors on notera $P_{\theta,g}$ la paramétrisation de $\mathcal{M}$, et on conviendra (quitte à opérer une translation du paramètre) que c'est en $\theta = 0$ que l'on passe par P_0, i.e. $P_{0,g} = P_0$.

On dit que $\mathbb{T}$ est un **ensemble tangent à $\mathcal{P}$ en** P_0 si c'est un sous-ensemble de $L^2(P_0)$ tel que, pour tout $g \in \mathbb{T}$, il existe un sous-modèle de dimension 1 $\{P_{\theta,g}, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$, tel que $P_{0,g} = P_0$, qui soit d.m.q. en 0 et de score g.

Remarques. Bien que la notation ne l'indique pas, un ensemble tangent à $\mathcal{P}$ en P_0 dépend de P_0.
L'information de Fisher en 0 dans le modèle $\{P_{\theta,g}, \theta \in \Theta\}$ est $P_0 g^2$.
Si $a \neq 0$ est un réel fixé, le sous-modèle de dimension 1 $\{P_{a\theta,g}, \theta \in \Theta\}$ est d.m.q. en 0 et de score ag. Donc on peut toujours considérer qu'un ensemble tangent est un cône.
La borne inférieure de variance pour estimer $\psi(P_0)$ est, si $\theta \mapsto \psi(P_{\theta,g})$ est dérivable, $\left(\frac{d}{d\theta}\psi(P_{\theta,g})|_{\theta=0}\right)^2 / P_0 g^2$ dans le modèle $\{P_{\theta,g}, \theta \in \Theta\}$, c'est la même dans le modèle $\{P_{a\theta,g}, \theta \in \Theta\}$. Donc chercher à maximiser cette borne, c'est chercher toutes les "directions" d'approche de P_0 par des sous-modèles de dimension 1.

On dit que la fonction $\psi : \mathcal{P} \to \mathbb{R}$ est **différentiable à P_0 relativement à l'ensemble tangent** $\mathbb{T}$ si il existe $\dot{\psi}_{P_0}$ linéaire et continue de $L^2(P_0)$ dans $\mathbb{R}$ telle que : pour tout $g \in \mathbb{T}$, score en $\theta = 0$ du sous-modèle de dimension 1 $\{P_{\theta,g}, \theta \in \Theta\}$,

$$\lim_{\theta \to 0} \frac{\psi(P_{\theta,g}) - \psi(P_0)}{\theta} = \dot{\psi}_{P_0} \cdot g.$$

Remarque. Cela signifie que la fonction $\theta \mapsto \psi(P_{\theta,g})$ est dérivable en 0 dans tous les sous-modèles, et que la dérivée peut s'interpréter comme une composition de différentielles.

Dans les espaces de Hilbert, on sait (Théorème de Riesz) que les applications linéaires continues à valeurs dans $\mathbb{R}$ sont les produits sacalaires avec une fonction fixe de l'espace de Hilbert. Donc si $\psi : \mathcal{P} \to \mathbb{R}$ est différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$, il existe $\tilde{\psi}_{P_0} \in L^2(P_0)$ telle que pour tout $g \in \mathbb{T}$, $\dot{\psi}_{P_0} \cdot g = \int \tilde{\psi}_{P_0} g dP_0$. Y a-t-il unicité de cette fonction $\tilde{\psi}_{P_0}$? Si $\tilde{\psi}_1 \in L^2(P_0)$ et $\tilde{\psi}_2 \in L^2(P_0)$ sont telles que pour tout $g \in \mathbb{T}$, $\dot{\psi}_{P_0} \cdot g = \int \tilde{\psi}_1 g dP_0 = \int \tilde{\psi}_2 g dP_0$, alors pour tout $g \in \mathbb{T}$, $\int (\tilde{\psi}_1 - \int \tilde{\psi}_2) g dP_0 = 0$. Si l'on note $H\mathbb{T}$ l'espace de Hilbert engendré par $\mathbb{T}$ (c'est à dire la fermeture, dans $L^2(P_0)$, de l'ensemble des combinaisons linéaires de fonctions de $\mathbb{T}$), et $H\mathbb{T}^o$ son orthogonal dans $L^2(P_0)$, les fonctions $\tilde{\psi}_i$, $i = 1, 2$ se décomposent comme la somme d'une fonction de $H\mathbb{T}$ et d'une fonction de $H\mathbb{T}^o$, et il y a unicité de la fonction dans $H\mathbb{T}$ dans cette décomposition. Comme par ailleurs la partie de la décomposition dans $H\mathbb{T}^o$ n'intervient pas dans $\int \tilde{\psi}_i g dP_0$, on peut choisir pour $\tilde{\psi}_{P_0}$ la fonction de $H\mathbb{T}$ qui représente $\dot{\psi}_{P_0}$ sur $\mathbb{T}$. C'est ce que dit la définition suivante.

Si ψ est différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$, la **fonction d'influence efficace** $\tilde{\psi}_{P_0}$ est l'unique fonction qui vérifie :
— $\tilde{\psi}_{P_0} \in H\mathbb{T}$, espace de Hilbert engendré par $\mathbb{T}$,
— $\forall g \in \mathbb{T}$, $\dot{\psi}_{P_0} \cdot g = \int \tilde{\psi}_{P_0} g dP_0$.

Remarques. Les fonctions scores sont toujours centrées, donc les fonctions de $H\mathbb{T}$ aussi, et on a donc toujours

$$\int \tilde{\psi}_{P_0} dP_0 = 0.$$

La borne inférieure de variance pour estimer $\psi(P_0)$ dans le sous-modèle $\{P_{\theta,g}, \theta \in \Theta\}$

est

$$CR(g) = \frac{\left(\int \tilde{\psi}_{P_0} g dP_0\right)^2}{\int g^2 dP_0} \leq \int \tilde{\psi}_{P_0}^2 dP_0$$

par Cauchy-Schwarz. Et si on peut approcher $\tilde{\psi}_{P_0}$ dans $L^2(P_0)$ par des fonctions de $\mathbb{T}$, alors la maximisation de $CR(g)$ dans les sous-modèles conduit à $\int \tilde{\psi}_{P_0}^2 dP_0$.

On peut étendre tout cela à l'estimation de quantités multidimensionnelles.

On dit que la fonction $\psi : \mathcal{P} \to \mathbb{R}^m$ est **différentiable à P_0 relativement à l'ensemble tangent** $\mathbb{T}$ si chacune des coordonnées ψ_j, $j = 1, \ldots, m$ de ψ est différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$. La **fonction d'influence efficace** est l'unique m-uplet $\tilde{\psi} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_m)$ de fonctions qui vérifie :
— $\tilde{\psi} \in H\mathbb{T}^m$,
— $\forall g \in \mathbb{T},\ \forall j = 1, \ldots, m,\ \lim_{\theta \to 0} \frac{\psi(P_{\theta,g})_j - \psi(P_0)_j}{\theta} = \int \tilde{\psi}_j g dP_0$.

4.2 Efficacité

On s'intéresse tout d'abord à l'efficacité au sens du risque minimax asymptotique local.

Théorème 4.2.1. *Soit $\psi : \mathcal{P} \to \mathbb{R}$ différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$ et de fonction d'influence efficace $\tilde{\psi}$. On suppose que $\mathbb{T}$ est un espace linéaire. Alors, si $(T_n)_{n\geq 1}$ est une suite d'estimateurs,*

$$\liminf_{c\to+\infty} \liminf_{n\to+\infty} \sup_{g\in\mathbb{T},\ \|g\|\leq c} E_{P_{\frac{1}{\sqrt{n}},g}} \left[n \left(T_n - \psi(P_{\frac{1}{\sqrt{n}},g}) \right)^2 \right] \geq \int \tilde{\psi}^2 dP_0.$$

Preuve. Soit $(g_p)_{p\geq 1}$ une suite de fonctions de $\mathbb{T}$ qui converge dans $L^2(P_0)$ vers $\tilde{\psi}$, et pour tout $p \geq 1$, soit $h_p = \frac{g_p}{c_p}$ avec $c_p^2 = \int g_p^2 dP_0$. On fixe p. Dans le sous-modèle de dimension 1 et de score h_p, on a $P_{\frac{c}{\sqrt{n}},h_p} = P_{\frac{1}{\sqrt{n}},ch_p}$ et donc

$$\sup_{g\in\mathbb{T},\ \|g\|\leq c} E_{P_{\frac{1}{\sqrt{n}},g}} \left[n \left(T_n - \psi(P_{\frac{1}{\sqrt{n}},g}) \right)^2 \right] \geq \sup_{|\theta|\leq \frac{c}{\sqrt{n}}} E_{P_{\theta,h_p}} \left[n \left(T_n - \psi(P_{\theta,h_p}) \right)^2 \right]$$

et donc

$$\liminf_{c\to+\infty} \liminf_{n\to+\infty} \sup_{g\in\mathbb{T},\ \|g\|\leq c} E_{P_{\frac{1}{\sqrt{n}},g}} \left[n \left(T_n - \psi(P_{\frac{1}{\sqrt{n}},g}) \right)^2 \right] \geq \frac{\left(\frac{d}{d\theta}\psi(P_{\theta,h_p})|_{\theta=0}\right)^2}{\int h_p^2 dP_0}$$
$$= \frac{\left(\int \tilde{\psi} h_p dP_0\right)^2}{\int h_p^2 dP_0}$$

et l'on obtient le théorème en faisant tendre p vers l'infini.

On dit que T_n **est localement asymptotiquement minimax pour estimer** $\psi(P_0)$, ou **efficace au sens du risque local minimax asymptotique pour estimer** $\psi(P_0)$ si

$$\liminf_{c\to+\infty}\liminf_{n\to+\infty}\sup_{g\in\mathbb{T},\ \|g\|\leq c} E_{P_{\frac{1}{\sqrt{n}},g}}\left[n\left(T_n-\psi(P_{\frac{1}{\sqrt{n}},g})\right)^2\right]=\int\tilde{\psi}^2dP_0.$$

Remarque. Comme $P_{c\theta,g}$ a pour score cg on pourrait aussi écrire de manière équivalente

$$\liminf_{c\to+\infty}\liminf_{n\to+\infty}\sup_{g\in\mathbb{T},\ \|g\|\leq 1} E_{P_{\frac{c}{\sqrt{n}},g}}\left[n\left(T_n-\psi(P_{\frac{c}{\sqrt{n}},g})\right)^2\right]\geq\int\tilde{\psi}^2dP_0.$$

En multidimensionnel, on obtient

Théorème 4.2.2. *Soit $\psi:\mathcal{P}\to\mathbb{R}^m$ différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$ et de fonction d'influence efficace $\tilde{\psi}$. On suppose que $\mathbb{T}$ est un espace linéaire. Alors, si $(T_n)_{n\geq1}$ est une suite d'estimateurs, pour tout $U\in\mathbb{R}^m$,*

$$\liminf_{c\to+\infty}\liminf_{n\to+\infty}\sup_{g\in\mathbb{T},\ \|g\|\leq c} E_{P_{\frac{1}{\sqrt{n}},g}}\left[n\left(U^T(T_n-\psi(P_{\frac{1}{\sqrt{n}},g}))\right)^2\right]\geq U^TVar_{P_0}\left(\tilde{\psi}\right)U$$

et

$$\liminf_{c\to+\infty}\liminf_{n\to+\infty}\sup_{g\in\mathbb{T},\ \|g\|\leq c} E_{P_{\frac{1}{\sqrt{n}},g}}\left[n\|T_n-\psi(P_{\frac{1}{\sqrt{n}},g}))\|^2\right]\geq Tr\left[Var_{P_0}\left(\tilde{\psi}\right)\right].$$

On s'intéresse maintenant à l'efficacité au sens du théorème de convolution.

On dit que T_n **est régulier pour estimer** $\psi(P_0)$ si il existe une loi de probabilité L telle que, pour tout $g\in\mathbb{T}$, $\sqrt{n}\left(T_n-\psi(P_{\frac{1}{\sqrt{n}},g})\right)$ converge en loi sous $P_{\frac{1}{\sqrt{n}},g}$ vers L.

Théorème 4.2.3. *Soit $\psi:\mathcal{P}\to\mathbb{R}^m$ différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$ et de fonction d'influence efficace $\tilde{\psi}$. On suppose que $\mathbb{T}$ est un espace linéaire. Alors, si $(T_n)_{n\geq1}$ est régulier pour estimer $\psi(P_0)$, il existe une loi de probabilité M sur $\mathbb{R}^m$ telle que*

$$L=M\star\mathcal{N}_m\left(0;Var_{P_0}\left(\tilde{\psi}\right)\right).$$

Si M est la masse de Dirac en 0, on dit que T_n **est efficace pour estimer** $\psi(P_0)$ **au sens du théorème de convolution**.

Preuve. On commence par faire la preuve pour $m=1$.
Soit $(g_p)_{p\geq1}$ une suite de fonctions de $\mathbb{T}$ qui converge dans $L^2(P_0)$ vers $\tilde{\psi}$. Pour tout p fixé, par le théorème de convolution paramétrique (Théorème 3.4.1), il existe une loi de probabilité M_p sur $\mathbb{R}$ telle que

$$L=M_p\star\mathcal{N}\left(0,\sigma_p^2\right)$$

avec

$$\sigma_p^2 = \frac{\left(\int \tilde{\psi} g_p dP_0\right)^2}{\int g_p^2 dP_0},$$

qui tend vers $\sigma^2 = \int (\tilde{\psi})^2 dP_0$. Donc si ϕ_L est la fonction caractéristique de L et ϕ_p celle de M_p, alors pour tout $t \in \mathbb{R}$,

$$\phi_L(t) = \phi_p(t) \exp\left(-\frac{\sigma_p^2}{2} t^2\right),$$

et quand p tend vers l'infini, $\phi_p(t)$ converge vers $\phi(t) = \phi_L(t) \exp\left(\frac{\sigma^2}{2} t^2\right)$ qui est continue en 0 et vaut 1 en 0, donc par le Théorème de Lévy, c'est la fonction caractéristique d'une probabilité M sur $\mathbb{R}$. On a alors, pour tout $t \in \mathbb{R}$, $\phi_L(t) = \phi(t) \exp\left(-\frac{\sigma^2}{2} t^2\right)$, ce qui signifie que $L = M \star \mathcal{N}\left(0; \sigma^2\right)$.
Si maintenant $m > 1$: soit $U \in \mathbb{R}^m$ quelconque. On applique le résultat précédent à $U^T T_n$ qui est un estimateur régulier de $U^T \psi(P_0)$, de fonction d'influence efficace $U^T \tilde{\psi}_0$, et l'on obtient que pour tout réel t,

$$\phi_L(tU) = \phi_U(t) \exp\left(-\frac{U^T Var_{P_0}\left(\tilde{\psi}\right) U}{2} t^2\right),$$

où $\phi_U(\cdot)$ est la fonction caractéristique d'une probabilité M_U sur $\mathbb{R}$. Mais alors, pour tout $U \in \mathbb{R}^m$, la fonction $U \mapsto \phi(U) = \phi_U(1)$ vérifie $\phi(U) = \phi_L(U) \exp\left(\frac{U^T Var_{P_0}(\tilde{\psi}) U}{2}\right)$, elle est continue en 0 et vaut 1 en 0, donc par le Théorème de Lévy, c'est la fonction caractéristique d'une probabilité M sur $\mathbb{R}^m$. On a alors pour tout $U \in \mathbb{R}^m$

$$\phi_L(U) = \phi(U) \exp\left(-\frac{U^T Var_{P_0}\left(\tilde{\psi}\right) U}{2}\right),$$

ce qui signifie que $L = M \star \mathcal{N}_m\left(0; Var_{P_0}\left(\tilde{\psi}\right)\right)$.

On peut maintenant se demander quand un estimateur régulier est efficace au sens du théorème de convolution. On a :

Théorème 4.2.4. *Soit $\psi : \mathcal{P} \to \mathbb{R}^m$ différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$ et de fonction d'influence efficace $\tilde{\psi}$. On suppose que pour une fonction $f \in (L^2(P_0))^m$ telle que $\int f dP_0 = 0$,*

$$\sqrt{n}\left(T_n - \psi(P_0)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f\left(X_i\right) + o_{P_0}(1).$$

Alors T_n est régulier et efficace au sens du théorème de convolution si et seulement si $f = \tilde{\psi}$.

Preuve. Soit g quelconque dans $\mathbb{T}$. Alors dans le sous-modèle $(P_{\theta,g})_\theta$,

$$\log\left[\frac{dP^{\otimes n}_{\frac{1}{\sqrt{n}},g}}{dP_0^{\otimes n}}\right](X_1,\ldots,X_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) - \frac{1}{2}\int g^2 dP_0 + o_{P_0}(1),$$

et donc

$$\left(\sqrt{n}\left(T_n - \psi(P_0)\right), \log\left[\frac{dP_{\frac{1}{\sqrt{n}},g}^{\otimes n}}{dP_0^{\otimes n}}\right](X_1,\ldots,X_n)\right)$$

converge en loi sous P_0 vers

$$\mathcal{N}_{m+1}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\int g^2 dP_0 \end{pmatrix}; \begin{pmatrix} Var_{P_0}(f) & (\int f g dP_0) \\ (\int f g dP_0)^T & \int g^2 dP_0 \end{pmatrix}\right).$$

Donc par le troisième Lemme de Le-Cam, $\sqrt{n}\left(T_n - \psi(P_0)\right)$ converge en loi sous $P_{\frac{1}{\sqrt{n}},g}$ vers $\mathcal{N}_m\left(\int f g dP_0; Var_{P_0}(f)\right)$. Par ailleurs, $\psi\left(P_{\frac{1}{\sqrt{n}},g}\right) = \psi(P_0) + \frac{1}{\sqrt{n}}\int \tilde{\psi} g dP_0$, donc $\sqrt{n}\left(T_n - \psi(P_{\frac{1}{\sqrt{n}},g})\right)$ converge en loi sous $P_{\frac{1}{\sqrt{n}},g}$ vers $\mathcal{N}_m\left(\int (f-\tilde{\psi}) g dP_0; Var_{P_0}(f)\right)$. Cette loi ne dépend pas de g si $\int (f-\tilde{\psi}) g dP_0$ ne dépend pas de g, donc vaut 0, c'est à dire si $f-\tilde{\psi}$ est orthogonale (coordonnée par coordonnée) à $\mathbb{T}$. T_n est régulier et efficace si de plus $Var_{P_0}(f) = Var_{P_0}(\tilde{\psi})$. Ces deux conditions sont vérifiées si et seulement $f = \tilde{\psi}$.

Exemple : modèle paramétrique.
Soit $\mathcal{P} = \{P_\theta = p_\theta \mu, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ un modèle paramétrique d.m.q. en θ de score $\dot{\ell}_\theta$. Alors

$$\mathbb{T} = \left\{h^T \dot{\ell}_\theta,\ h \in \mathbb{R}^k\right\}$$

est un ensemble tangent à $\mathcal{P}$ en P_θ. C'est l'ensemble tangent maximal.
Si on suppose l'information de Fisher I_θ inversible, soit $g : \Theta \to \mathbb{R}^m$ différentiable et notons $\psi(P_\theta) = g(\theta)$. Alors ψ est différentiable à P_θ relativement à l'ensemble tangent $\mathbb{T}$, la fonction d'influence efficace est

$$\tilde{\psi} = D_1 g(\theta) I_\theta^{-1} \dot{\ell}_\theta,$$

et la borne inférieure de variance pour estimer $g(\theta)$ est

$$Var_{P_\theta}\left[\tilde{\psi}\right] = D_1 g(\theta) I_\theta^{-1} D_1 g(\theta)^T.$$

On retrouve les résultats déjà vus en paramétrique.
(Exercice : démontrer tout ça !).

Exemple : modèle non paramétrique "maximum".
Soit $\mathcal{P}$ l'ensemble des mesures de probabilité sur $\mathbb{R}^d$ et soit P_0 une probabilité sur $\mathbb{R}^d$. Alors

$$\mathbb{T} = \left\{ g \in L^2(P_0) \ : \ \int g dP_0 = 0 \right\}$$

est un ensemble tangent à $\mathcal{P}$ en P_θ. C'est l'ensemble tangent maximal. Pour le démontrer, considérer le sous-modèle $\{p_\theta P_0, \ \theta \in \mathbb{R}\}$ où $p_\theta(x) = c(\theta)H(\theta g(x))$, la fonction H étant donnée par $H(u) = 2/(1+e^{-2u})$, et voir qu'il est d.m.q. en 0 de score g. (Exercice : le faire ! Commencer par montrer que H et H' sont comprises entre 0 et 2 et valent 1 en $u = 0$).
Soit ensuite $f \in L^2(P_0)$ fixée, et posons $\psi(P) = \int f dP$. Alors ψ est différentiable à P_0 relativement à l'ensemble tangent $\mathbb{T}$, de fonction d'influence efficace $\tilde{\psi} = f - \int f dP_0$, et la borne inférieure de variance pour estimer $\int f dP_0$ est $Var_{P_0}(f(X))$. La moyenne empirique $\frac{1}{n}\sum_{i=1}^n f(X_i)$ est efficace (dans les deux sens d'efficacité) !
(Exercice : démontrer tout ça !).

4.3 Modèles semi-paramétriques

Il s'agit du cas où

$$\mathcal{P} = \{P_{\theta,\eta}, \ \theta \in \Theta, \ \eta \in H\}$$

où $\Theta \subset \mathbb{R}^k$ et H est un ensemble qui peut être de dimension infinie. On s'intéresse alors à $\psi : \mathcal{P} \to \mathbb{R}^k$ donnée par

$$P_{\theta,\eta} \mapsto \psi\left(P_{\theta,\eta}\right) = \theta.$$

Exemples :
- Centre de symétrie : $\mathcal{P}$ est l'ensemble des lois sur $\mathbb{R}$ de densité par rapport à Lebesgue $\eta(\cdot - \theta)$, $\eta \in H = \mathcal{F}$, $\mathcal{F}$ ensemble des densités de probabilité strictement positives sur $\mathbb{R}$ et centrées en 0.
- Régression : si on observe $X = (Y, Z)$, avec $Y = g_\theta(Z) + \epsilon$, (Z, ϵ) de loi η inconnue mais telle que ϵ soit centré et de variance 1.

On cherche un ensemble tangent, la fonction d'influence efficace, et la variance minimale pour estimer θ.

Regardons pour commencer le cas paramétrique où η est un paramètre "de nuisance" de dimension finie, c'est à dire quand $H \subset \mathbb{R}^m$.
On note $\alpha = (\theta, \eta) \in \mathbb{R}^{k+m}$, et $\alpha_0 = (\theta_0, \eta_0)$. On notera aussi $P_0 = P_{\theta_0,\eta_0}$. Si le modèle est d.m.q. en α_0, de score $\begin{pmatrix} \dot{\ell}_0 \\ \dot{h}_0 \end{pmatrix}$, avec $\dot{\ell}_0$ score en θ_0 du modèle $(P_{\theta,\eta_0})_\theta$ et $\dot{h}_0$ score en η_0 du modèle $(P_{\theta_0,\eta})_\eta$ alors pour tout $u \in \mathbb{R}^k$, pour tout $v \in \mathbb{R}^m$, le modèle $(P_{\theta_0+tu,\eta_0+tv})_t$ est d.m.q. en $t = 0$ de score $u^T\dot{\ell}_0 + v^T\dot{h}_0$. Donc ici, l'ensemble tangent à $\mathcal{P}$ en P_{α_0} (ensemble tangent maximal) est

$$\mathbb{T} = \left\{ u^T\dot{\ell}_0 + v^T\dot{h}_0, \ u \in \mathbb{R}^k, \ v \in \mathbb{R}^m \right\}.$$

Par ailleurs, $\frac{d}{dt}(\theta_0 + tu) = u$, donc on cherche $\tilde{\psi} \in \mathbb{T}^k$ telle que :

$$\forall u \in \mathbb{R}^k,\ \forall v \in \mathbb{R}^m,\ \int \tilde{\psi} \left(u^T \dot{\ell}_0 + v^T \dot{h}_0 \right) dP_0 = u.$$

Ceci implique :

$$\int \tilde{\psi} \dot{\ell}_0^T dP_0 = I_k,$$

la matrice identité de dimension k, et

$$\forall v \in \mathbb{R}^m,\ \int \tilde{\psi} v^T \dot{h}_0 dP_0 = 0.$$

Notons

$$F = \left\{ v^T \dot{h}_0,\ v \in \mathbb{R}^m \right\}$$

(sous-ensemble fermé de $L^2(P_0)$) et Π la projection orthogonale (au sens de $L^2(P_0)$) sur F, puis $\tilde{\ell} = (\tilde{\ell}_i)_{1 \leq i \leq k}$ avec

$$\tilde{\ell}_i = \left(\dot{\ell}_0 \right)_i - \Pi \left(\dot{\ell}_0 \right)_i .$$

Notons aussi $\tilde{I} = Var_{P_0} \tilde{\ell}$. Alors, si $\tilde{I}$ est inversible, on a

$$\tilde{\psi} = \tilde{I}^{-1} \tilde{\ell},$$

et $Var_{P_0} \tilde{\psi} = \tilde{I}^{-1}$.
Exercice : démontrer ces deux formules

Maintenant, si I_{α_0} est l'information de Fisher (paramétrique), avec

$$I_{\alpha_0} = \begin{pmatrix} J_{\theta_0} & C \\ C^T & K_{\eta_0} \end{pmatrix}$$

on a

$$\tilde{\ell} = \dot{\ell}_0 - C K_{\eta_0}^{-1} \dot{h}_0,\ \tilde{I} = J_{\theta_0} - C K_{\eta_0}^{-1} C^T.$$

Exercice : le démontrer.

$\tilde{I}^{-1}$ est la variance minimale pour estimer θ_0 d'après la théorie semi-paramétrique. Ceci coincide avec ce que nous dit la théorie paramétrique, qui dit que la variance minimale pour estimer θ_0 est la matrice $k \times k$ en haut à gauche de $I_{\alpha_0}^{-1}$.
Exercice : démontrer cette affirmation, et que cela coincide avec $[J_{\theta_0} - C K_{\eta_0}^{-1} C^T]^{-1}$.

Si l'on connait η_0, par contre, c'est à dire dans le modèle $(P_{\theta,\eta_0})_\theta$, la variance minimale pour estimer θ_0 est $J_{\theta_0}^{-1}$. Donc il n'y a pas de perte (pour l'estimation) venant du fait de ne pas connaitre η_0 si et seulement si $C K_{\eta_0}^{-1} C^T = 0$, soit si et seulement si $C = 0$, c'est à dire si et seulement si les scores relatifs à θ et les scores relatifs à η sont orthogonaux,

soit si $\dot{\ell}_0$ et $\dot{h}_0$ ont des coordonnées deux à deux orthogonales dans $L^2(P_0)$.

Revenons maintenant au cas général. Supposons que pour tout $u \in \mathbb{R}^k$ le modèle $(P_{\theta_0+tu,\eta_0})_t$ est d.m.q. en $t = 0$ de score $u^T\dot{\ell}_0$, et que $\mathbb{G}$ est un ensemble tangent à $(P_{\theta_0,\eta})_\eta$ en $P_{\theta_0,\eta_0} = P_0$. Alors souvent (mais il faut le vérifier dans chaque situation), un ensemble tangent à $\mathcal{P}$ en P_0 est

$$\mathbb{T} = \left\{ u^T\dot{\ell}_0 + g,\ u \in \mathbb{R}^k,\ g \in \mathbb{G} \right\},$$

$u^T\dot{\ell}_0 + g$ étant le score en $t = 0$ dans le sous-modèle $(P_{\theta_0+tu,\eta_t})_t$ quand g est le score en $t = 0$ dans le sous-modèle $(P_{\theta_0,\eta_t})_t$.

On note Π la projection orthogonale de $L^2(P_0)$ sur la fermeture de l'espace linéaire engendré par $\mathbb{G}$. Soit $\tilde{\ell} = (\tilde{\ell}_i)_{1\le i\le k}$ avec

$$\tilde{\ell}_i = \left(\dot{\ell}_0\right)_i - \Pi\left(\dot{\ell}_0\right)_i,$$

et soit $\tilde{I} = Var_{P_0}\tilde{\ell}$. Alors, si $\tilde{I}$ est inversible, on a

$$\tilde{\psi} = \tilde{I}^{-1}\tilde{\ell}.$$

Exercice : le démontrer !

On appelle $\tilde{\ell}$ le **score efficace** et $\tilde{I}$ l'**information de Fisher efficace**. Si maintenant T_n est un estimateur qui vérifie

$$\sqrt{n}\,(T_n - \theta) = \frac{1}{\sqrt{n}}\tilde{I}^{-1}\sum_{i=1}^{n}\tilde{\ell}\,(X_i) + o_{P_0}(1),$$

alors T_n est régulier et efficace.

4.4 Exercices

Exercice 4.4.1. Modèle de translation
Soit $\mathcal{F}$ l'ensemble des densités de probabilités f par rapport à Lebesgue sur $\mathbb{R}$ qui sont paires, strictement positives sur $\mathbb{R}$, de racine carrée continûment dérivable sur $\mathbb{R}$, et telles que $I_f = \int \frac{f'(x)^2}{f(x)}dx$ est finie. Soit $(P_{\theta,f})_{\theta\in\mathbb{R},f\in\mathcal{F}}$ le modèle donné par $dP_{\theta,f} = f(x-\theta)dx$.

1. Montrer que le modèle est identifiable.
2. Montrer que pour tout $f \in \mathcal{F}$, le modèle $(P_{\theta,f})_{\theta\in\mathbb{R}}$ est différentiable en moyenne quadratique en tout θ, d'information de Fisher I_f.
3. Soit $f_0 \in \mathcal{F}$, et soit $\mathcal{G}$ l'ensemble des fonctions paires, continûment dérivables, telles que
   $\int g\,(x)\,f_0\,(x)\,dx = 0$, $\int g\,(x)^2\,f_0\,(x)\,dx < +\infty$ et $\int g'\,(x)^2\,f_0\,(x)\,dx < +\infty$.

Soit k la fonction donnée par $k(x) = \frac{2}{1+\exp(-2x)}$. On définit pour toute $g \in \mathcal{G}$ et tout réel h, $f_{h,g}(x) = c(h)f_0(x)k[hg(x)]$, où $c(h)^{-1} = \int f_0(x)k[hg(x)]dx$. Montrer qu'il existe un voisinage V de 0 tel que le modèle $(P_{\theta+h,f_{h,g}})_{h\in V}$ soit différentiable en moyenne quadratique, de score

$$-\frac{f_0'}{f_0}(x-\theta) + g(x-\theta).$$

4. Montrer que l'information de Fisher efficace est I_f. (Conséquence : si on trouve un estimateur efficace, il n'y a pas de perte asymptotique due au fait que l'on ne connait pas f).

Exercice 4.4.2. Modèle de Neyman-Scott
Soit (X, Y) un couple de variables aléatoires réelles et Z une variable aléatoire réelle telles que, conditionnellement à Z, X et Y sont indépendantes de loi $\mathcal{N}(Z, \theta)$, $\theta > 0$. On s'intéresse à l'estimation de θ, lorsque la loi η de Z est inconnue, sur la base d'un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de (X, Y), dont la loi est notée $P_{\theta,\eta}$.

1. Montrer que $E[(X-Y)^2] = 2\theta$, et calculer $Var[(X-Y)^2]$.
2. En déduire que
$$T_n = \frac{1}{2n}\sum_{i=1}(X_i - Y_i)^2$$
est un estimateur consistant de θ, et construire un intervalle de confiance pour θ asymptotiquement de niveau de confiance $1-\alpha$.
3. Quelle est la loi de $X - Y$ sous $P_{\theta,\eta}$?
4. Montrer que sous $P_{\theta+\frac{c}{\sqrt{n}},\eta}$, $\sqrt{n}(T_n - (\theta + \frac{c}{\sqrt{n}}))$ converge en loi vers une gaussienne centrée et de variance $2\theta^2$.
5. On pose $U = \frac{X-Y}{\sqrt{2}}$ et $V = \frac{X+Y}{\sqrt{2}}$. Soit $p_{\theta,\eta}$ la densité de $P_{\theta,\eta}$ par rapport à Lebesgue. Montrer que
$$p_{\theta,\eta}(X, Y) = f_\theta(U)\, q_{\theta,\eta}(V)$$
où f_θ est la densité de U et $q_{\theta,\eta}$ celle de V, que l'on déterminera. En déduire en particulier que sous $P_{\theta,\eta}$, U et V sont des variables aléatoires indépendantes.
6. Soit $\theta_0 > 0$ et η_0 une loi de probabilité sur $\mathbb{R}$. Soit $\mathcal{T}$ l'ensemble des fonctions réelles t telles que $\int t(z)d\eta_0(z) = 0$ et $\int t(z)^2 d\eta_0(z) < +\infty$. Soit ensuite
$$\mathcal{G} = \left\{ g \;:\; g(x,y) = \frac{\int \frac{1}{\sqrt{2\pi\theta}}\left(\exp - \frac{(\frac{x+y}{\sqrt{2}} - z\sqrt{2})^2}{2\theta}\right) t(z)\, d\eta_0(z)}{\int \frac{1}{\sqrt{2\pi\theta}}\left(\exp - \frac{(\frac{x+y}{\sqrt{2}} - z\sqrt{2})^2}{2\theta}\right) d\eta_0(z)},\ t \in \mathcal{T} \right\}.$$
Soit $g \in \mathcal{G}$, soit t la fonction de $\mathcal{T}$ associée à g, et soit le modèle $(P_{\theta_0+h,\eta_{h,g}})_{h\in V}$, $V =]-\theta_0, \theta_0[$ avec $d\eta_{h,g}(z) = c(h)k(ht(z))d\eta_0(z)$, où k est la fonction donnée par

$k(x) = \frac{2}{1+\exp(-2x)}$ et $c(h)^{-1} = \int k(ht(z))d\eta_0(z)$. Montrer que ce modèle vérifie les hypothèses du premier exercice avec

$$\dot{\ell}_0(X,Y) = \frac{U^2}{2\theta_0^2} - \frac{1}{2\theta_0} + \frac{\int [\frac{(V-z\sqrt{2})^2}{2\theta_0^2} - \frac{1}{2\theta_0}] \frac{1}{\sqrt{2\pi\theta_0}} \left(\exp - \frac{(V-z\sqrt{2})^2}{2\theta_0}\right) d\eta_0(z)}{q_{\theta_0,\eta_0}(V)}.$$

7. On suppose que le support de η_0 contient un intervalle ouvert. On admettra qu'alors la fermeture de $\mathcal{G}$ est l'ensemble des fonctions $m\left(\frac{x+y}{\sqrt{2}}\right)$ telles que, sous P_{θ_0,η_0}, $m(V)$ est centrée et a une variance finie.
   Calculer le score efficace et l'information de Fisher efficace.
8. Montrer que T_n est un estimateur asymptotiquement efficace de θ_0 au sens du théorème asymptotique minimax local ainsi qu'au sens du théorème de convolution.

5 Estimation Bayésienne

5.1 Généralités

On dispose d'une observation $X^n = (X_1, \ldots, X_n)$, et d'un modèle $\{P_{n,\theta},\ \theta \in \Theta\}$, ensemble de lois "possibles" pour l'observation.
On met sur Θ une loi Π, dite **loi a priori**. Cela suppose que Θ est probabilisable (muni d'une tribu). Cela revient à considérer que θ est une variable aléatoire T de loi Π, et que le modèle décrit les lois de X^n conditionnellement à $T = \theta$. On se placera dans la situation où le modèle est dominé, et l'on notera, pour tout $\theta \in \Theta$, $p_{n,\theta}$ la densité de $P_{n,\theta}$ par rapport à la mesure dominante.

La méthodologie bayésienne consiste à baser l'inférence statistique sur la **loi a posteriori** : loi de T conditionnellement à X^n, que l'on notera $\Pi(\cdot|X^n)$. On a pour tout ensemble mesurable A

$$\Pi(T \in A|X^n) = \frac{\int_A \Pi(d\theta)\, p_{n,\theta}(X^n)}{\int_\Theta \Pi(d\theta)\, p_{n,\theta}(X^n)}. \tag{5.1}$$

Remarque importante : **La loi a posteriori est aléatoire.**

Exemple : On considère le cas où $\Theta = \mathbb{R}$, $P_{n,\theta} = \mathcal{N}(\theta, 1)^{\otimes n}$, la loi a priori est $\Pi = \mathcal{N}(0, \sigma^2)$. Alors la loi a posteriori est

$$\Pi(\cdot|X_1, \ldots, X_n) = \mathcal{N}\left(\frac{\sigma^2 \sum_{i=1}^n X_i}{n\sigma^2 + 1}, \frac{\sigma^2}{n\sigma^2 + 1}\right).$$

Exercice : le démontrer.

A partir de la loi a posteriori, on peut bâtir des estimateurs : par exemple moyenne a posteriori, médiane a posteriori, en minimisant un risque a posteriori (estimateurs bayésiens : voir cours de M1).
On peut aussi construire des **ensembles de crédibilité** : I est un **ensemble de crédibilité pour θ de niveau de crédibilité $1 - \alpha$ si**

$$\Pi(T \in I|X_1, \ldots, X_n) \geq 1 - \alpha.$$

Reprise de l'exemple : dans l'exemple précédent, si $u_{1-\alpha/2}$ est un quantile d'ordre $1 - \alpha/2$ de la loi gaussienne centrée réduite,

$$I = \left[\frac{\sigma^2 \sum_{i=1}^n X_i}{n\sigma^2 + 1} - \sqrt{\frac{\sigma^2}{n\sigma^2 + 1}} u_{1-\alpha/2}; \frac{\sigma^2 \sum_{i=1}^n X_i}{n\sigma^2 + 1} + \sqrt{\frac{\sigma^2}{n\sigma^2 + 1}} u_{1-\alpha/2}\right]$$

est un ensemble de crédibilité pour θ de niveau de crédibilité $1-\alpha$. On peut le comparer avec l'intervalle de confiance obtenu avec la moyenne empirique. (Le faire !).

Mise en oeuvre des méthodes bayésiennes : calcul de la loi a posteriori par algorithmes performants (Metropolis-Hasting, Gibbs).
Cela dépend quand même de la complexité de la loi a priori : si Θ est non paramétrique, choisir une loi a priori est un sujet en soi.
Le calcul d'ensembles de crédibilité ne demande rien d'autre que la calcul de la loi a posteriori ; pas besoin de calculer d'information de Fisher par exemple, avec le problème de son estimation (plug-in, etc...).

Etude fréquentiste : Si $X_1, \ldots, X_n$ est de loi $P_n^\star$, que peut-on dire de la loi a posteriori ? En particulier, si $P_n^\star = P_{\theta^\star,n}$ pour un $\theta^\star \in \Theta$, la loi a posteriori se concentre-t-elle en $\theta^\star$ quand n tend vers l'infini ?

On va adopter le point de vue fréquentiste, et étudier la question de la concentration de la loi a posteriori (consistance), ainsi que la vitesse de concentration ; quelles sont les conditions que cela impose sur la loi a priori.

En paramétrique : comment cela se compare à l'estimation par maximum de vraisemblance ?
En non paramétrique : obtient-on les vitesses minimax ? Peut-on obtenir des résultats adaptatifs ?

5.2 Estimation bayésienne paramétrique

On se place dans la situation où $\Theta \subset \mathbb{R}^k$, et le modèle est celui de variables i.i.d. , soit $P_{n,\theta} = P_\theta^{\otimes n}$.

Souvent on considérera que le modèle $(P_\theta)_{\theta\in\Theta}$ est dominé par une mesure dominante λ, donc que $P_{n,\theta} = (p_\theta \lambda)^{\otimes n}$, et que la loi a priori Π admet une densité π par rapport à Lebesgue. La loi a posteriori dans ce cas a donc une densité $\pi(\cdot|X_1, \ldots, X_n)$ par rapport à Lebesgue qui est donnée par :

$$\pi(\theta|X^n) = \frac{\pi(\theta)\prod_{i=1}^n p_\theta(X_i)}{\int_\Theta \pi(s)\prod_{i=1}^n p_s(X_i)\,ds}. \tag{5.2}$$

On voit que la compréhension du comportement de la loi a posteriori passe par la compréhension du comportement de la vraisemblance sur Θ.

5.2.1 Consistance

Si $\theta^\star \in \Theta$, on note $\mathbb{P}_{\theta^\star}$ la loi de $(X_n)_{n\geq 1}$, suite de variables aléatoires i.i.d. de loi $P_{\theta^\star}$.
On dit que **la suite de lois a posteriori** $(\Pi(\cdot|X_1, \ldots, X_n))_{n\geq 1}$ **est** $\mathbb{P}_{\theta^\star}$**-consistante** si elle converge en loi vers $\delta_{\theta^\star}$, en $\mathbb{P}_{\theta^\star}$-probabilité.

On dit que **la suite de lois a posteriori** $(\Pi(\cdot|X_1,\dots,X_n))_{n\geq 1}$ **est fortement** $\mathbb{P}_{\theta^\star}$**-consistante** si elle converge en loi vers $\delta_{\theta^\star}$, $\mathbb{P}_{\theta^\star}$-ps.

Remarque. : attention à ce que cela signifie. Si $\rho(\cdot,\cdot)$ métrise la convergence en loi (exemples de telles métriques ?), cela signifie que $\rho(\Pi(\cdot|X_1,\dots,X_n),\delta_{\theta^\star})$ converge en $\mathbb{P}_{\theta^\star}$-probabilité (ou $\mathbb{P}_{\theta^\star}$-ps) vers 0.

On a un résultat de consistance très général, qui ne vaut pas seulement pour $\Theta\subset\mathbb{R}^k$, mais si Θ est métrique, séparable et complet.

Théorème 5.2.1 (Théorème de consistance de Doob)**.** *On suppose que le modèle $(P_\theta)_{\theta\in\Theta}$ est identifiable. Alors il existe $\Theta_0\subset\Theta$ mesurable tel que $\Pi(\Theta_0)=1$, et pour tout $\theta\in\Theta_0$, la suite de loi a posteriori $(\Pi(\cdot|X_1,\dots,X_n))_{n\geq 1}$ est $\mathbb{P}_\theta$-consistante.*

Remarque. La preuve de ce théorème est omise mais n'est pas constructive : elle ne construit pas l'ensemble Θ_0. Donc ce résultat ne dit rien pour un θ particulier, sauf si Π est une mesure discrète.

Question : comparer les situations de consistance de l'estimateur du maximum de vraisemblance et de la loi a posteriori ? Il est possible d'avoir l'un et pas l'autre : voir exercices 6.4.3 et 6.4.4. Voici des conditions suffisantes sous lesquelles les deux consistances ont lieu. On se place dans le cadre d'un modèle dominé.

Théorème 5.2.2. *On suppose Θ compact, que le modèle $(P_\theta)_{\theta\in\Theta}$ est identifiable, et que pour tout x, $\theta\mapsto\log p_\theta(x)$ est continue. Soit $\theta^\star\in\Theta$. On suppose que $\sup_{\theta\in\Theta}|\log p_\theta|\in L_1(P_{\theta^\star})$. Alors*

1. *L'estimateur du maximum de vraisemblance est fortement consistant.*
2. *Si $\theta^\star$ est dans le support de Π, alors la suite de lois a posteriori $(\Pi(\cdot|X_1,\dots,X_n))_{n\geq 1}$ est fortement $\mathbb{P}_{\theta^\star}$-consistante.*

Preuve. Soit Ω l'événement sur lequel

$$\lim_{n\to+\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)+K(P_{\theta^\star},P_\theta)\right|=0.$$

On a $\mathbb{P}_{\theta^\star}(\Omega)=1$ (exercice : dire pourquoi). Sur cet événement, la seule valeur d'adhérence possible de la suite $(\widehat{\theta}_n)_{n\geq 1}$ ($\widehat{\theta}_n$ est l'estimateur du maximum de vraisemblance avec n observations) est $\theta^\star$, donc $\widehat{\theta}_n$ converge $\mathbb{P}_{\theta^\star}$-ps vers $\theta^\star$.

Soit maintenant U un voisinage de $\theta^\star$. On a

$$\begin{aligned}
\Pi\left(U|X_1,\ldots,X_n\right) &= \frac{\int_U \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right)}{\int_\Theta \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right)} \\
&= \frac{\int_U \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right)}{\int_U \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right) + \int_{U^C} \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right)} \\
&= \frac{1}{1+\frac{\int_{U^C} \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi(d\theta)}{\int_U \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi(d\theta)}}.
\end{aligned}$$

Comme pour tout x, $\theta \mapsto \log p_\theta(x)$ est continue, et que $\sup_{\theta\in\Theta} |\log p_\theta| \in L_1(P_{\theta^\star})$, la fonction $\theta \mapsto K\left(P_{\theta^\star}, P_\theta\right)$ est continue, donc en notant $A = \sup_{\theta\in U^C} -K\left(P_{\theta^\star}, P_\theta\right)$, $A < 0$. Soit maintenant $\epsilon > 0$ tel que $A + 2\epsilon < 0$. Pour tout $\omega \in \Omega$, il existe $n_1(\omega)$ tel que si $n \geq n_1(\omega)$ alors, pour tout $\theta \in U^C$,

$$\frac{1}{n}\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right) \leq A + \epsilon$$

et donc si $n \geq n_1(\omega)$,

$$\int_{U^C} \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right) \leq \Pi\left(U^C\right) e^{n(A+\epsilon)}.$$

Aussi, on peut choisir $V \subset U$ ouvert et contenant $\theta^\star$ tel que

$$\inf_{\theta\in V} -K\left(P_{\theta^\star}, P_\theta\right) > A + 2\epsilon.$$

(Exercice : dire pourquoi). Pour tout $\omega \in \Omega$, il existe $n_2(\omega)$ tel que si $n \geq n_2(\omega)$ alors, en posant $B = A + 2\epsilon$,

$$\forall \theta \in V,\ \frac{1}{n}\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right) \geq B - \frac{\epsilon}{2}.$$

Par ailleurs, comme $\theta^\star$ est dans le support de Π, $\Pi\left(V\right) > 0$. On a alors, si $n \geq n_2(\omega)$,

$$\int_U \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right) \geq \int_V \exp\left(\sum_{i=1}^n \log\left(\frac{p_\theta(X_i)}{p_{\theta^\star}(X_i)}\right)\right)\Pi\left(d\theta\right) \geq \Pi\left(V\right) e^{n(B-\frac{\epsilon}{2})}.$$

Donc si $n \geq n_1(\omega) \vee n_2(\omega)$,

$$\Pi\left(U|X_1,\ldots,X_n\right) \geq \frac{1}{1+\frac{\Pi(U^C)e^{n(A+\epsilon)}}{\Pi(V)e^{n(B-\frac{\epsilon}{2})}}} \geq \frac{1}{1+\frac{1}{\Pi(V)}e^{n(A+3\frac{\epsilon}{2}-B)}} = \frac{1}{1+\frac{1}{\Pi(V)}e^{-n\frac{\epsilon}{2}}}$$

et donc, pour tout ouvert U contenant $\theta^\star$, $\Pi\left(U|X_1,\ldots,X_n\right)$ tend vers 1 quand n tend vers l'infini sur l'événement Ω. (Exercice : dire pourquoi cela suffit pour conclure).

5.2.2 Théorème de Bernstein-von Mises

On veut savoir maintenant comment la loi a posteriori se concentre en $\theta^\star$. On suppose que Π a une densité par rapport à Lebesgue, et donc la loi a posteriori a une densité donnée par (5.2). Si l'on note $\ell_n(\theta)$ la log-vraisemblance, cela se réécrit

$$\frac{\pi(\theta)\exp\ell_n(\theta)}{\int_\Theta \pi(s)\exp\ell_n(s)ds}.$$

Si l'on note $H=\sqrt{n}(T-\theta^\star)$, alors la loi de H conditionnellement à $X_1,\dots,X_n$ a pour densité

$$\frac{\pi(\theta^\star+\frac{h}{\sqrt{n}})\exp\ell_n(\theta^\star+\frac{h}{\sqrt{n}})}{\int\pi(\theta^\star+\frac{s}{\sqrt{n}})\exp\ell_n(\theta^\star+\frac{s}{\sqrt{n}}))ds}=\frac{\pi(\theta^\star+\frac{h}{\sqrt{n}})\exp\left\{\ell_n(\theta^\star+\frac{h}{\sqrt{n}})-\ell_n(\theta^\star)\right\}}{\int\pi(\theta^\star+\frac{s}{\sqrt{n}})\exp\left\{\ell_n(\theta^\star+\frac{s}{\sqrt{n}})-\ell_n(\theta^\star)\right\}ds}.$$

Maintenant, si le modèle $(P_\theta)_{\theta\in\Theta}$ est d.m.q. en $\theta^\star$, le Théorème 3.2.1 donne pour tout s :

$$\ell_n\left(\theta^\star+\frac{s}{\sqrt{n}}\right)-\ell_n\left(\theta^\star\right)=\mathbb{G}_n\left(s^T\dot\ell_{\theta^\star}\right)-\frac{1}{2}s^TI_{\theta^\star}s+o_{\mathbb{P}_{\theta^\star}}(1)$$

avec $\dot\ell_{\theta^\star}$ le score et $I_{\theta^\star}$ l'information de Fisher du modèle en $\theta^\star$. Si l'on suppose que $I_{\theta^\star}$ est inversible et que l'on note

$$\Delta_n\left(\theta^\star\right)=I_{\theta^\star}^{-1}\mathbb{G}_n\left(\dot\ell_{\theta^\star}\right)=I_{\theta^\star}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\dot\ell_{\theta^\star}\left(X_i\right),$$

la densité de H a posteriori, pour n assez grand, devrait être approximativement

$$\frac{\pi(\theta^\star+\frac{h}{\sqrt{n}})\exp\left\{\langle h,I_{\theta^\star}\Delta_n\left(\theta^\star\right)\rangle-\frac{1}{2}h^TI_{\theta^\star}h\right\}}{\int\pi(\theta^\star+\frac{s}{\sqrt{n}})\exp\left\{\langle s,I_{\theta^\star}\Delta_n\left(\theta^\star\right)\rangle-\frac{1}{2}s^TI_{\theta^\star}s\right\}ds},$$

soit, si π est continue et strictement positive en $\theta^\star$, approximativement

$$\frac{\exp\left\{\langle h,I_{\theta^\star}\Delta_n\left(\theta^\star\right)\rangle-\frac{1}{2}h^TI_{\theta^\star}h\right\}}{\int\exp\left\{\langle s,I_{\theta^\star}\Delta_n\left(\theta^\star\right)\rangle-\frac{1}{2}s^TI_{\theta^\star}s\right\}ds}.$$

Cette densité est la densité de $\mathcal{N}(\Delta_n\left(\theta^\star\right),I_{\theta^\star}^{-1})$. Donc si cette heuristique est valide, cela signifie que la loi a posteriori de $\sqrt{n}(T-\theta^\star)$ est approximativement $\mathcal{N}(\Delta_n\left(\theta^\star\right),I_{\theta^\star}^{-1})$. Par ailleurs, on a vu que sous de bonne hypothèses, si $\widehat\theta_n$ est l'estimateur du maximum de vraisemblance, on a

$$\sqrt{n}\left(\widehat\theta_n-\theta^\star\right)=\Delta_n\left(\theta^\star\right)+o_{\mathbb{P}_{\theta^\star}}(1),$$

donc si tout cela est valide, cela signifie que la loi a posteriori de T est approximativement $\mathcal{N}(\widehat\theta_n,\frac{I_{\theta^\star}^{-1}}{n})$.
Le Théorème de Bernstein-von Mises dit quand et en quel sens ce résultat est valide. **Le Théorème de Bernstein-von Mises implique qu'une région de crédibilité pour $\theta^\star$ de niveau de crédibilité $1-\alpha$ est une région de confiance asymptotique pour $\theta^\star$ de niveau de confiance $1-\alpha$.** C'est ce que nous allons voir maintenant.

Théorème 5.2.3 (Théorème de Bernstein-von Mises)**.** *On suppose que la densité a priori π est continue et strictement positive en $\theta^\star$.*
On suppose que le modèle $(P_\theta)_{\theta\in\Theta}$ est d.m.q. en $\theta^\star$, que l'information de Fisher $I_{\theta^\star}$ est inversible, et que pour tout $\epsilon > 0$, il existe une suite de tests ϕ_n tels que

$$\lim_{n\to+\infty} \mathbb{P}_{\theta^\star}\phi_n = 0$$

et

$$\lim_{n\to+\infty} \sup_{\|\theta-\theta^\star\|\geq\epsilon} \mathbb{P}_\theta\left(1-\phi_n\right) = 0.$$

Alors, si l'on note $\mathcal{L}\left(\sqrt{n}(T-\theta^\star)|X_1,\ldots,X_n\right)$ la loi a posteriori de $\sqrt{n}(T-\theta^\star)$,

$$\left\|\mathcal{L}\left(\sqrt{n}(T-\theta^\star)|X_1,\ldots,X_n\right) - \mathcal{N}\left(\Delta_n\left(\theta^\star\right), I_{\theta^\star}^{-1}\right)\right\|_{VT} = o_{\mathbb{P}_{\theta^\star}}(1).$$

Remarques.
— La norme en variation totale est invariante par translation et changement d'échelle.
— L'hypothèse sur l'existence de tels tests est faible, voir par exemple le livre de Van der Vaart pages 145-146.

Preuve. On admettra (voir livre de Van der Vaart pages 143-144) que sous les hypothèses du Théorème 5.2.3, pour toute suite M_n tendant vers $+\infty$, il existe une suite de tests $(\psi_n)_{n\geq1}$, une constante $c > 0$, et un entier n_0 tels que

$$\lim_{n\to+\infty} \mathbb{P}_{\theta^\star}\psi_n = 0$$

et pour tout $\theta \in \Theta$ tel que $\|\theta-\theta^\star\| \geq \frac{M_n}{\sqrt{n}}$, si $n \geq n_0$,

$$\mathbb{P}_\theta\left(1-\psi_n\right) \leq e^{-cn(\|\theta-\theta^\star\|^2\wedge1)}. \tag{5.3}$$

On va faire la preuve du Théorème 5.2.3 sous l'hypothèse supplémentaire qu'il existe un voisinage A de $\theta^\star$ et une fonction $H \in L^2(P_{\theta^\star})$ tels que

$$\forall(\theta_1,\theta_2) \in A^2,\ |\log p_{\theta_1} - \log p_{\theta_2}| \leq \|\theta_1-\theta_2\|H.$$

On admettra qu'alors, si l'on note

$$R_n\left(h\right) = \ell_n\left(\theta^\star + \frac{h}{\sqrt{n}}\right) - \ell_n\left(\theta^\star\right) - h^T\Delta_n\left(\theta^\star\right) + \frac{1}{2}h^T I_{\theta^\star}h,$$

on a pour tout $M > 0$,

$$\sup_{\|h\|\leq M} |R_n\left(h\right)| = o_{\mathbb{P}_{\theta^\star}}(1).$$

Soit $M > 0$ quelconque. Si U est de loi $\mathcal{N}\left(\Delta_n\left(\theta^\star\right), I_{\theta^\star}^{-1}\right)$, on note $Q_{n,M}$ la loi de U conditionnellement à $\|U\| \leq M$. On note aussi $L_{n,M}$ la loi a posteriori de H conditionnellement à $\|H\| \leq M$ (H est la variable aléatoire $H = \sqrt{n}(T-\theta^\star)$). Ces deux lois sont aléatoires (fonctions de $X_1,\ldots,X_n$). On va commencer par montrer que si on note

$$Z_n\left(M\right) = \|L_{n,M} - Q_{n,M}\|_{VT},$$

alors

$$Z_n(M) = o_{\mathbb{P}_{\theta^\star}}(1).$$

$Q_{n,M}$ a pour densité par rapport à Lebesgue

$$\frac{\mathbb{1}_{\|h\|\leq M}\exp[h^T\Delta_n(\theta^\star) - \frac{1}{2}h^T I_{\theta^\star}h]}{\int_{\|g\|\leq M}\exp[g^T\Delta_n(\theta^\star) - \frac{1}{2}g^T I_{\theta^\star}g]dg}$$

et $L_{n,M}$ a pour densité par rapport à Lebesgue

$$\frac{\mathbb{1}_{\|h\|\leq M}\pi(\theta^\star + \frac{h}{\sqrt{n}})\exp[\ell_n(\theta^\star + \frac{h}{\sqrt{n}}) - \ell_n(\theta^\star)]}{\int_{\|g\|\leq M}\pi(\theta^\star + \frac{g}{\sqrt{n}})\exp[\ell_n(\theta^\star + \frac{g}{\sqrt{n}}) - \ell_n(\theta^\star)]dg}$$

On sait que $\pi(\theta^\star) > 0$. Notons

$$\delta_n = \sup_{\|h\|\leq M}\frac{\pi(\theta^\star + \frac{h}{\sqrt{n}})}{\pi(\theta^\star)} \vee \sup_{\|h\|\leq M}\frac{\pi(\theta^\star)}{\pi(\theta^\star + \frac{h}{\sqrt{n}})}.$$

On obtient alors ,

$$\begin{aligned} Z_n(M) &\leq \int_{\|h\|\leq M}\left|\delta_n^2\frac{\exp[\sup_{\|g\|\leq M}|R_n(g)|]}{\exp[-\sup_{\|g\|\leq M}|R_n(g)|]} - 1\right| Q_{n,M}(dh) \\ &= \left|\delta_n^2\frac{\exp[o_{\mathbb{P}_{\theta^\star}}(1)]}{\exp[-o_{\mathbb{P}_{\theta^\star}}(1)} - 1\right| = \delta_n^2(1 + o_{\mathbb{P}_{\theta^\star}}(1)) - 1. \end{aligned}$$

Comme π est continue en $\theta^\star$, δ_n tend vers 1 quand n tend vers l'infini, et l'on obtient bien que $Z_n(M) = o_{\mathbb{P}_{\theta^\star}}(1)$. Ceci est vrai pour tout $M > 0$, donc il existe une suite M_n qui tend vers l'infini telle que

$$Z_n(M_n) = o_{\mathbb{P}_{\theta^\star}}(1). \tag{5.4}$$

Maintenant,

$$\begin{aligned} &\left\|\mathcal{L}\left(\sqrt{n}(T-\theta^\star)|X_1,\ldots,X_n\right) - \mathcal{N}\left(\Delta_n(\theta^\star), I_{\theta^\star}^{-1}\right)\right\|_{VT} \\ &\qquad \leq \left\|\mathcal{L}\left(\sqrt{n}(T-\theta^\star)|X_1,\ldots,X_n\right) - L_{n,M_n}\right\|_{VT} \\ &\qquad\qquad + \left\|\mathcal{N}\left(\Delta_n(\theta^\star), I_{\theta^\star}^{-1}\right) - Q_{n,M_n}\right\|_{VT} + o_{\mathbb{P}_{\theta^\star}}(1). \end{aligned}$$

Maintenant, on a (calcul immédiat) :

$$\begin{aligned} \left\|\mathcal{N}\left(\Delta_n(\theta^\star), I_{\theta^\star}^{-1}\right) - Q_{n,M_n}\right\|_{VT} &= 2\mathcal{N}\left(\Delta_n(\theta^\star), I_{\theta^\star}^{-1}\right)[\|U\| \geq M_n] \\ &\leq 2\mathbf{1}_{\|\Delta_n\|\geq M_n/2} + C_1\int_{\|I_{\theta^\star}^{-1/2}v\|\geq M_n/2}\exp\left[-C_2\|v\|^2\right]du = o_{\mathbb{P}_{\theta^\star}}(1) + o(1) \end{aligned}$$

pour des constantes C_1 et C_2 qui dépendent uniquement de $I_{\theta^\star}$, donc

$$\left\|\mathcal{N}\left(\Delta_n(\theta^\star), I_{\theta^\star}^{-1}\right) - Q_{n,M_n}\right\|_{VT} = o_{\mathbb{P}_{\theta^\star}}(1).$$

De même,

$$\begin{aligned}\left\|\mathcal{L}\left(\sqrt{n}(T-\theta^\star)|X_1,\ldots,X_n\right)-L_{n,M_n}\right\|_{VT} &= 2\Pi\left(\|H\|\geq M_n|X_1,\ldots,X_n\right)\\ &= 2\Pi\left(\|H\|\geq M_n|X_1,\ldots,X_n\right)\psi_n + 2\Pi\left(\|H\|\geq M_n|X_1,\ldots,X_n\right)(1-\psi_n).\end{aligned}$$

On a

$$E_{\mathbb{P}_{\theta^\star}}\left[\Pi\left(\|H\|\geq M_n|X_1,\ldots,X_n\right)\psi_n\right]\leq \mathbb{P}_{\theta^\star}\psi_n = o(1),$$

donc

$$\Pi\left(\|H\|\geq M_n|X_1,\ldots,X_n\right)\psi_n = o_{\mathbb{P}_{\theta^\star}}(1).$$

Ensuite,

$$\Pi\left(\|H\|\geq M_n|X_1,\ldots,X_n\right)(1-\psi_n) = \frac{\int_{\|h\|\geq M_n}\pi(\theta^\star+\frac{h}{\sqrt{n}})\exp[\ell_n(\theta^\star+\frac{h}{\sqrt{n}})-\ell_n(\theta^\star)]dh}{\int\pi(\theta^\star+\frac{g}{\sqrt{n}})\exp[\ell_n(\theta^\star+\frac{g}{\sqrt{n}})-\ell_n(\theta^\star)]dg}(1-\psi_n).$$

On fixe $M>0$, et on a, pour des constantes $C_3>0$ et $C_4>0$,

$$\begin{aligned}\int\pi(\theta^\star+\frac{g}{\sqrt{n}})\exp[\ell_n(\theta^\star+\frac{g}{\sqrt{n}})-\ell_n(\theta^\star)]dg&\\ \geq \int_{\|g\|\leq M}\pi(\theta^\star+\frac{g}{\sqrt{n}})\exp[\ell_n(\theta^\star+\frac{g}{\sqrt{n}})-\ell_n(\theta^\star)]dg&\\ \geq C_3\exp\left[-C_2(M^2+\|\Delta_n(\theta^\star)\|^2)-\sup_{\|g\|\leq M}|R_n(g)|\right] &= \exp\left[-O_{\mathbb{P}_{\theta^\star}}(1)\right]\end{aligned}$$

donc il reste à montrer que

$$\int_{\|h\|\geq M_n}\pi(\theta^\star+\frac{h}{\sqrt{n}})\exp[\ell_n(\theta^\star+\frac{h}{\sqrt{n}})-\ell_n(\theta^\star)](1-\psi_n)\,dh = o_{\mathbb{P}_{\theta^\star}}(1).$$

Mais par Fubini

$$\begin{aligned}\mathbb{P}_{\theta^\star}\left[\int_{\|h\|\geq M_n}\pi(\theta^\star+\frac{h}{\sqrt{n}})\exp[\ell_n(\theta^\star+\frac{h}{\sqrt{n}})-\ell_n(\theta^\star)](1-\psi_n)\,dh\right]&\\ = \int_{\|h\|\geq M_n}\pi(\theta^\star+\frac{h}{\sqrt{n}})\mathbb{P}_{\theta^\star+\frac{h}{\sqrt{n}}}(1-\psi_n)\,dh&\\ \leq \int_{\|h\|\geq M_n}\pi(\theta^\star+\frac{h}{\sqrt{n}})e^{-cn(\|h/\sqrt{n}\|^2\wedge 1)}dh&\end{aligned}$$

par (5.3). Enfin, il existe $\delta>0$ et $C>0$ tels que pour $\|u\|\leq\delta$, $\pi(\theta^\star+u)\leq C$ (par continuité en $\theta^\star$) et on a

$$\begin{aligned}\int_{\|h\|\geq M_n}\pi(\theta^\star+\frac{h}{\sqrt{n}})e^{-cn(\|h/\sqrt{n}\|^2\wedge 1)}dh&\\ \leq C\int_{M_n\leq\|h\|\leq\delta\sqrt{n}}e^{-c\|h\|^2}dh + e^{-c\delta^2 n}\int_{\|h\|\geq\delta\sqrt{n}}\pi(\theta^\star+\frac{h}{\sqrt{n}})dh&\\ = o(1)+O\left((\sqrt{n})^k\right)e^{-c\delta^2 n} = o(1).&\end{aligned}$$

5.2.3 Conséquences du Théorème de Bernstein-von Mises

5.2.3.1 Ensembles de crédibilité

Dire que $E_{n,\alpha}$ est un ensemble de crédibilité pour $\theta^\star$ de niveau de crédibilité $1-\alpha$ signifie que $E_{n,\alpha}$ est un ensemble aléatoire (qui dépend des observations $X_1,\ldots,X_n$) tel que

$$\Pi\left(T \in E_{n,\alpha} | X_1,\ldots,X_n\right) \geq 1-\alpha.$$

Un tel ensemble est construit à partir du calcul de la loi a posteriori (par des algorithmes MCMC appropriés) et en choisissant un ensemble de couverture de cette loi (il est naturel de le choisir le plus concentré possible). Si l'on note

$$D_n = \left\| \mathcal{L}\left(\sqrt{n}(T-\theta^\star) | X_1,\ldots,X_n\right) - \mathcal{N}\left(\Delta_n\left(\theta^\star\right), I_{\theta^\star}^{-1}\right) \right\|_{VT},$$

on a par définition de la norme en variation totale et du fait qu'elle est invariante par translation et changement d'échelle

$$\left| \Pi\left(T \in E_{n,\alpha} | X_1,\ldots,X_n\right) - \mathcal{N}\left(\theta^\star + \frac{\Delta_n\left(\theta^\star\right)}{\sqrt{n}}, \frac{I_{\theta^\star}^{-1}}{n}\right)\left[E_{n,\alpha}\right] \right| \leq D_n.$$

Si par ailleurs $\widehat{\theta}_n$ est un estimateur (par exemple l'estimateur du maximum de vraisemblance, si les hypothèses assurent que) tel que

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^\star\right) = \Delta_n\left(\theta^\star\right) + w_n,$$

avec $w_n = o_{\mathbb{P}_{\theta^\star}}(1)$, on a que

$$E_{n,\alpha} = \widehat{\theta}_n - \frac{w_n}{\sqrt{n}} + \frac{I_{\theta^\star}^{-1/2}}{\sqrt{n}} A_{n,\alpha}$$

avec $A_{n,\alpha}$ tel que, si U suit la loi $\mathcal{N}_k\left(0, Id\right)$,

$$1-\alpha-D_n \leq P\left(U \in A_{n,\alpha}\right).$$

Par ailleurs, comme $\Delta_n\left(\theta^\star\right)$ converge en loi vers $\mathcal{N}\left(0, I_{\theta^\star}^{-1}\right)$, si l'on choisit A_α telle que

$$P\left(U \in A_\alpha\right) = 1-\alpha,$$

l'ensemble

$$C_{n,\alpha} = \widehat{\theta}_n + \frac{I_{\theta^\star}^{-1/2}}{\sqrt{n}} A_\alpha$$

est une région de confiance pour $\theta^\star$ asymptotiquement de niveau de confiance $1-\alpha$. On a ainsi

$$E_{n,\alpha} = C_{n,\alpha} - \frac{w_n}{\sqrt{n}} + \frac{I_{\theta^\star}^{-1/2}}{\sqrt{n}}\left(A_{n,\alpha} - A_\alpha\right) = C_{n,\alpha} + o_{\mathbb{P}_{\theta^\star}}(1/\sqrt{n}).$$

Le Théorème de Bernstein-von Mises implique que $P(U \in A_{n,\alpha}) = 1-\alpha+o_{\mathbb{P}_{\theta^\star}}(1)$. Donc si $A_{n,\alpha}$ est par exemple choisi de concentration maximum de la loi a posteriori, et que A_α est la boule centrée en 0 telle que $P(U \in A_\alpha) = 1-\alpha$, on a $A_{n,\alpha} = A_\alpha + o_{\mathbb{P}_{\theta^\star}}(1)$, et donc en ce cas

$$E_{n,\alpha} = C_{n,\alpha} + o_{\mathbb{P}_{\theta^\star}}(1),$$

et en ce cas, le Théorème de Bernstein-von Mises implique que, asymptotiquement, ensemble de crédibilité et région de confiance construite à partir de $\widehat{\theta}_n$ (par exemple l'estimateur du maximum de vraisemblance si les hypothèses permettent d'écrire son asymptotique) coincident en probabilité.

5.2.3.2 Estimateurs ponctuels bayésiens

Etant donnée une fonction de perte $L(\cdot,\cdot)$, on peut choisir Z_n qui minimise

$$z \mapsto E\left(L(z,T)|X_1,\ldots,X_n\right),$$

c'est l'estimateur bayésien relativement à L. Par exemple, pour la fonction de perte quadratique $L(z,t) = \|z-t\|^2$, Z_n est l'espérance de la loi a posteriori, et si $k=1$, pour la fonction de perte $L(z,t) = |z-t|$, Z_n est la médiane de la loi a posteriori. Dans ces deux cas, le Théorème de Bernstein-von Mises semble dire que $\sqrt{n}(Z_n-\theta^\star)$ est l'espérance (resp. la médiane) de la loi $\mathcal{N}\left(\Delta_n(\theta^\star), I_{\theta^\star}^{-1}\right)$, et donc que $\sqrt{n}(Z_n-\theta^\star) = \Delta_n(\theta^\star)+o_{\mathbb{P}_{\theta^\star}}(1)$. C'est facile à voir pour la médiane a posteriori.

Proposition 5.2.1. *On se place dans le cas $k=1$, et sous les hypothèses du Théorème de Bernstein-von Mises. Soit Z_n la médiane de la loi a posteriori. Alors*

$$\sqrt{n}(Z_n-\theta^\star) = \Delta_n(\theta^\star) + o_{\mathbb{P}_{\theta^\star}}(1)$$

et $\sqrt{n}(Z_n-\theta^\star)$ converge en loi vers $\mathcal{N}(0, 1/I_{\theta^\star})$.

Preuve. Par définition,

$$\Pi\left(T < Z_n|X_1,\ldots,X_n\right) \leq \frac{1}{2} \leq \Pi\left(T \leq Z_n|X_1,\ldots,X_n\right).$$

Par le Théorème de Bernstein-von Mises,

$$\Pi\left(T < Z_n|X_1,\ldots,X_n\right) = \mathcal{N}\left(\theta^\star + \frac{\Delta_n(\theta^\star)}{\sqrt{n}}, \frac{I_{\theta^\star}}{n}\right)\{]-\infty, Z_n[\} + o_{\mathbb{P}_{\theta^\star}}(1)$$

et

$$\Pi\left(T \leq Z_n|X_1,\ldots,X_n\right) = \mathcal{N}\left(\theta^\star + \frac{\Delta_n(\theta^\star)}{\sqrt{n}}, \frac{I_{\theta^\star}}{n}\right)\{]-\infty, Z_n]\} + o_{\mathbb{P}_{\theta^\star}}(1).$$

Comme la loi gaussienne ne charge aucun point, cela implique que

$$\mathcal{N}\left(\theta^\star + \frac{\Delta_n(\theta^\star)}{\sqrt{n}}, \frac{I_{\theta^\star}}{n}\right)\{]-\infty, Z_n]\} = \frac{1}{2} + o_{\mathbb{P}_{\theta^\star}}(1),$$

soit, en notant Φ la fonction de répartition de la loi gaussienne centrée réduite,

$$\Phi\left(\sqrt{I_{\theta^\star}}[\sqrt{n}(Z_n-\theta^\star)-\Delta_n(\theta^\star)]\right)=\frac{1}{2}+o_{\mathbb{P}_{\theta^\star}}(1).$$

Comme Φ est inversible d'inverse continue, le théorème de l'image continue implique que $\sqrt{I_{\theta^\star}}[\sqrt{n}(Z_n-\theta^\star)-\Delta_n(\theta^\star)]$ converge en $\mathbb{P}_{\theta^\star}$-probabilité vers $\Phi^{-1}(\frac{1}{2})=0$.

5.3 Exercices

Dans tous les exercices, on considère une famille de lois $(P_\theta)_{\theta\in\Theta}$ et une loi de probabilité "a priori" , loi d'une variable aléatoire T sur Θ. Pour tout entier n, la loi de $X_1,\dots,X_n$ conditionnellement à $T=\theta$ est $P_\theta^{\otimes n}$, c'est-à-dire que, conditionnellement à $T=\theta$, $X_1,\dots,X_n$ sont indépendantes et de même loi P_θ. La loi dite "a posteriori" est la loi de T conditionnellement à $(X_1,\dots,X_n)$.

Exercice 5.3.1. Soit Π la loi gaussienne sur $\mathbb{R}^k$ d'espérance τ et de variance Σ inversible. Soit P_θ la loi gaussienne sur $\mathbb{R}^k$ d'espérance θ et de variance identité.

1. Quelle est la loi a posteriori ?
2. On suppose que $(X_1,\dots,X_n)$ sont i.i.d. de loi P_{θ_0}
   a) On choisit comme estimateur $\widehat{\theta}_n$ de θ_0 l'espérance de la loi a posteriori. Quelles sont les propriétés asymptotiques de $\widehat{\theta}_n$?
   b) Vérifier directement que le théorème de Bernstein-von-Mises est vrai.

Exercice 5.3.2. Soit P_θ la loi de Bernoulli de paramètre θ et Π la loi a priori Beta de paramètre (α,β) ($\alpha>0$, $\beta>0$) c'est-à-dire de densité

$$\pi(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}_{]0,1[}(\theta),$$

où $\Gamma(\cdot)$ est la fonction donnée par $\Gamma(s)=\int_0^{+\infty}x^{s-1}e^{-x}dx$.

1. Quelle est la loi a posteriori ?
2. On suppose que $(X_1,\dots,X_n)$ sont i.i.d. de loi P_{θ_0} On choisit comme estimateur $\widehat{\theta}_n$ de θ_0 l'espérance de la loi a posteriori. Quelles sont les propriétés asymptotiques de $\widehat{\theta}_n$?

Exercice 5.3.3. Un exemple où l'estimateur du maximum de vraisemblance n'est pas consistant et la loi a posteriori est consistante.
$\Theta = \mathbb{N}^*$ (ensemble des entiers positifs non nuls). On fixe $C \in]0,1[$, et $h(x) = e^{1/x^2}$ pour tout $x > 0$. On définit $a_0 = 1$ et pour tout entier k, a_k par $\int_{a_k}^{a_{k-1}} (h(x) - C)dx = 1 - C$. Enfin, soit alors f_k la fonction définie sur $[0,1]$ par

$$f_k(x) = \begin{cases} h(x) & \text{si } a_k < x < a_{k-1} \\ C & \text{sinon} \end{cases}$$

On fixe Π une probabilité sur Θ telle que pour tout entier $j \geq 1$, $\Pi(\{j\}) > 0$.

1. Montrer que pour tout entier $k \geq 1$, $a_k \in]0,1[$, que $\lim_{k\to+\infty} a_k = 0$, et que f_k est une densité de probabilité.
2. Montrer que pour tout entier $j \geq 1$, $\Pi(\cdot|X_1, \ldots, X_n)$ est P_j-consistante.
3. Soit $j \geq 1$ fixé. On note $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance, $X_{(1)} = \min\{X_1, \ldots, X_n\}$, et on définit la variable aléatoire k_n par $k_n = k$ si et seulement si $X_{(1)} \in]a_k, a_{k-1}[$. On note $\ell_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$.
   a) Soit Y une variable aléatoire de loi uniforme sur $[0, 1/C]$. Montrer que pour tout réel $x > 0$, $P_j(X_1 > x) \leq P(Y > x)$. En déduire que $n(X_{(1)})^2$ tend vers 0 en $P_j^{\otimes n}$-probabilité.
   b) Montrer que pour tout $M > 0$, $P_j^{\otimes n}(\hat{\theta}_n \neq j) \geq P_j^{\otimes n}(\ell_n(k_n) - \ell_n(j) \geq M)$.
   c) Montrer que, si N_j^n est le nombre de X_i, $i = 1, \ldots, n$, tels que $X_i \in]a_j, a_{j-1}[$,

$$\ell_n(k_n) - \ell_n(j) \geq \log \frac{h\left(X_{(1)}\right)}{C} - N_j^n \log \frac{h(a_j)}{C}.$$

   d) En déduire que $\lim_{n\to+\infty} P_j^{\otimes n}(\hat{\theta}_n \neq j) = 1$.

Exercice 5.3.4. Un exemple où l'estimateur du maximum de vraisemblance est consistant et la loi a posteriori n'est pas consistante.
On pose $\Theta = [0,1] \cup [2,3]$, P_θ la loi uniforme sur $[0,\theta]$, et Π une probabilité sur Θ de densité π continue et strictement positive sur l'intérieur de Θ telle que $\pi(\theta) = e^{-1/(\theta-1)^2}$ si $\theta \in [0,1[$, ce qui est possible car $\int_0^1 e^{-1/(\theta-1)^2} d\theta < 1$. On pose $\theta_0 = 1$.

1. Montrer que l'estimateur du maximum de vraisemblance est consistant.
2. On va montrer que $\Pi([2,3]|X_1, \ldots, X_n)$ tend vers 1 en $P_{\theta_0}^{\otimes n}$-probabilité, ce qui suffira à déduire que la loi a posteriori n'est pas P_{θ_0}-consistante. On note $X_{(n)} = \max\{X_1, \ldots, X_n\}$
   a) Montrer que, $P_{\theta_0}^{\otimes n}$-p.s.,

$$\Pi([2,3]|X_1, \ldots, X_n) = \frac{1}{1 + \frac{\int_{X_{(n)}^n}^1 \theta^{-n}\pi(\theta)d\theta}{\int_2^3 \theta^{-n}\pi(\theta)d\theta}}.$$

b) Montrer que

$$-(\log 3)\Pi([2,3]) \leq \frac{1}{n}\log\int_2^3 \theta^{-n}\pi(\theta)d\theta \leq -(\log 2)\Pi([2,3]).$$

c) Montrer que

$$\frac{1}{n}\log\int_{X_{(n)}^n}^1 \theta^{-n}\pi(\theta)d\theta \leq -\left(1-\frac{1}{n}\right)\log X_{(n)}^n + \frac{1}{n}\log\left(1-X_{(n)}^{n-1}\right) + \frac{1}{n}\log\pi\left(X_{(n)}\right).$$

d) Conclure.

6 Sujets

6.1 Partiel de novembre 2011

Exercice 6.1.1. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles i.i.d. de loi P. On suppose que $\{h_\theta, \theta \in \Theta\}$ est un ensemble de fonctions réelles P-Glivenko-Cantelli, que $\theta \mapsto Ph_\theta$ est une fonction continue d'un ouvert $\Theta \subset \mathbb{R}$ dans $\mathbb{R}$, et que $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ est une suite d'estimateurs consistant de θ_0, $\theta_0 \in \Theta$.
Montrer qu'alors $\mathbb{P}_n h_{\widehat{\theta}_n}$ tend en probabilité vers Ph_{θ_0} quand n tend vers l'infini.
En déduire un estimateur consistant de $P(X_1 \leq E(X_1))$ lorsque X_1 admet une espérance, et que sa fonction de répartition est continue.

Exercice 6.1.2. Soit $\Theta \subset \mathbb{R}^k$, et $(P_\theta)_{\theta\in\Theta}$ un modèle différentiable en moyenne quadratique en θ_0, point intérieur à Θ, de score $\dot{\ell}_{\theta_0}$ et d'information de Fisher I_{θ_0} inversible. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d.

1. Donner des conditions suffisantes pour que l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ soit un estimateur consistant de θ_0.
2. Donner des conditions suffisantes pour que l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ vérifie

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}^{\otimes n}}(1). \tag{6.1}$$

3. Soit g une fonction continument différentiable de $\mathbb{R}^k$ dans $\mathbb{R}$. Montrer que si (6.1) est vérifiée, alors

$$\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta_0)\right) = \nabla g_{\theta_0}^T I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}^{\otimes n}}(1)$$

   où ∇g_{θ_0} est le vecteur gradient de g en θ_0.
4. En déduire que si (6.1) est vérifiée, alors $g(\widehat{\theta}_n)$ est un estimateur régulier de $g(\theta_0)$, asymptotiquement efficace au sens du théorème de convolution.

Exercice 6.1.3. Estimateur de James-Stein. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. à valeurs dans $\mathbb{R}^k$, $k \geq 3$. On considère la suite de modèles $[(P_m^{\otimes n})_{m\in\mathbb{R}^k}]_{n\geq 1}$, avec $P_m = \mathcal{N}_k(m, I)$ où I est la matrice identité. On note $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, et $\|\cdot\|$ la norme euclidienne. L'estimateur de James-Stein est défini par :

$$T_n = \left(1 - \frac{k-2}{n\|\overline{X}\|^2}\right)\overline{X}.$$

On admettra que si Z est une variable aléatoire dans $\mathbb{R}^k$ gaussienne centrée réduite, pour tout a de $\mathbb{R}^k$,

$$E\left(\frac{\langle Z, Z+a\rangle}{\|Z+a\|^2}\right) = (k-2)E\left(\frac{1}{\|Z+a\|^2}\right),$$

qui est finie dès que $k \geq 3$.

1. Montrer que le modèle $(P_m)_{m\in\mathbb{R}^k}$ est différentiable en moyenne quadratique en tout $m_0 \in \mathbb{R}^k$, et que l'information de Fisher est la matrice identité I.
2. Montrer que sous $P_m^{\otimes n}$,
$$\sqrt{n}\,(T_n - m) = Z - (k-2)\,\frac{Z + m\sqrt{n}}{\|Z + m\sqrt{n}\|^2}$$
où Z est une variable aléatoire dans $\mathbb{R}^k$ gaussienne centrée réduite.
3. Montrer que pour tout $m \in \mathbb{R}^k$, $E_m\|\sqrt{n}(\overline{X} - m)\|^2 = k$ et que
$$E_m\|\sqrt{n}(T_n - m)\|^2 = k - (k-2)^2 E\left(\frac{1}{\|Z + m\sqrt{n}\|^2}\right).$$
4. En déduire que pour tout $M > 0$,
$$\limsup_{n\to+\infty} \sup_{\|h\|\leq M} E_{\frac{h}{\sqrt{n}}}\|\sqrt{n}(T_n - \frac{h}{\sqrt{n}})\|^2 < k.$$
5. Montrer que, pour tout $m \in \mathbb{R}^k$, T_n est asymptotiquement efficace au sens du risque minimax local.
6. Montrer que si $m \neq 0$, T_n est un estimateur régulier efficace au sens du théorème de convolution.
7. Qu'en est-il en $m = 0$?
8. Commenter les résultats des questions 3 à 7.

6.2 Partiel de novembre 2012

Exercice 6.2.1. Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles i.i.d. de loi P admettant un moment d'ordre 4. On note $\sigma^2 = Var(X^2)$ où X est une variable aléatoire réelle de loi P . Soit

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2.$$

1. Montrer que $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n U_i^2 - \left(\frac{1}{n}\sum_{i=1}^n U_i\right)^2$ où l'on a posé $U_i = X_i - E(X)$.
2. Montrer que $\sqrt{n}(\widehat{\sigma}_n^2 - \sigma^2)$ converge en loi vers une variable gaussienne centrée et de variance
   $V = E[(X - E(X))^4] - (E[(X - E(X))^2])^2$.
3. Proposer un estimateur consistant de V et un intervalle de confiance pour σ^2 asymptotiquement de niveau α.

Exercice 6.2.2. Soit P_θ la probabilité sur $\mathbb{R}$ de densité par rapport à Lebesgue $p_\theta(x) = \frac{1}{\pi[1+(x-\theta)^2]}$.
On note $\mathcal{I} = \frac{4}{\pi}\int_{-\infty}^{+\infty} \frac{u^2}{(1+u^2)^3} du = \frac{1}{2}$. Le but de l'exercice est de construire un estimateur asymptotiquement efficace de θ pour tout réel θ. On note $\ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$.

1. Montrer que le modèle $(P_\theta)_{\theta\in\mathbb{R}}$ est différentiable en moyenne quadratique en tout réel θ, de score $\dot{\ell}_\theta(x) = \frac{-2(x-\theta)}{1+(x-\theta)^2}$ et d'information de Fisher $\mathcal{I}$ en θ.
2. Soit $\tilde{\theta}_n$ une médiane empirique qui vérifie
$$\sum_{i=1}^n \mathbb{1}_{X_i<\tilde{\theta}_n} = \sum_{i=1}^n \mathbb{1}_{X_i>\tilde{\theta}_n}$$
   a) Montrer que pour tout réel θ, $\tilde{\theta}_n$ converge en $P_\theta^{\otimes n}$-probabilité vers θ.
   b) Montrer que
$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) = \frac{\pi}{2\sqrt{n}}\sum_{i=1}^n \left(\mathbb{1}_{X_i>\theta} - \mathbb{1}_{X_i<\theta}\right) + o_{P_\theta^{\otimes n}}(1).$$
   c) $\tilde{\theta}_n$ est-il régulier ? $\tilde{\theta}_n$ est-il asymptotiquement efficace au sens du théorème de convolution ?
3. Soit $M > 0$ positif fixé. Soit $\widehat{\theta}_n(M)$ l'estimateur du maximum de vraisemblance sur $[-M, M]$, c'est-à-dire tel que $\widehat{\theta}_n(M) \in [-M, M]$ et
$$\ell_n\left(\widehat{\theta}_n(M)\right) \geq \sup_{|\theta'|\leq M} \ell_n\left(\theta'\right) - \frac{1}{n}.$$
   a) Montrer que $\{\log p_{\theta'}(x) \ : \ |\theta'| \leq M\}$ est P_θ-Glivenko-Cantelli pour tout réel θ.
   b) Montrer que si $\theta \in]-M; M[$, $\widehat{\theta}_n(M)$ converge en $P_\theta^{\otimes n}$-probabilité vers θ.
   c) Montrer que si $\theta \in]-M; M[$,
$$\sqrt{n}\left(\widehat{\theta}_n(M) - \theta\right) = \frac{1}{\mathcal{I}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_\theta(X_i) + o_{P_\theta^{\otimes n}}(1).$$

4. On souhaite maintenant s'affranchir de la connaissance a priori de M, et construire un estimateur ayant les mêmes propriétés asymptotiques. Pour cela, on utilise $\tilde{\theta}_n$ comme estimateur préliminaire, que l'on améliore par un pas de descente pour chercher un zéro de la dérivée de la vraisemblance. Autrement dit, si on note $Z_n(\theta) = \frac{1}{n}\sum_{i=1}^n \dot{\ell}_\theta(X_i)$ et $W_n(\theta) = \frac{1}{n}\sum_{i=1}^n \left[\dot{\ell}_\theta(X_i)\right]^2$, on pose

$$\widehat{\theta}_n = \tilde{\theta}_n + \frac{Z_n(\tilde{\theta}_n)}{W_n(\tilde{\theta}_n)}.$$

   a) Soit $h_\theta(x)$ la dérivée seconde de $\log p_\theta(x)$ par rapport à θ.
   Montrer que pour tout réel θ, $|h_\theta(x)| \leq 2$ pour tout réel x, puis que

$$\int_0^1 \left[\frac{1}{n}\sum_{i=1}^n h_{\theta+t(\tilde{\theta}_n-\theta)}(X_i)\right] dt = -\mathcal{I} + o_{P_\theta^{\otimes n}}(1),$$

   et en déduire que

$$\sqrt{n}\left[Z_n(\tilde{\theta}_n) - Z_n(\theta)\right] = -\mathcal{I}\sqrt{n}\left(\tilde{\theta}_n - \theta\right) + o_{P_\theta^{\otimes n}}(1).$$

   b) Montrer que pour tout réel θ, $W_n(\tilde{\theta}_n) = \mathcal{I} + o_{P_\theta^{\otimes n}}(1)$, puis que

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) = \frac{1}{\mathcal{I}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_\theta(X_i) + o_{P_\theta^{\otimes n}}(1).$$

   c) Commenter.

Exercice 6.2.3. Soit $\mathcal{P}$ l'ensemble des probabilités sur $\mathbb{R}$ de densité continue strictement positive. Pour tout $P \in \mathcal{P}$, on définit $\psi(P)$ comme étant la médiane de P, c'est à dire le nombre réel défini par $\psi(P) = F^{-1}(\frac{1}{2})$, F étant la fonction de répartition de P, qui est inversible puisque P a une densité strictement positive. Soit $P \in \mathcal{P}$ fixé de densité f.

1. Soit $\mathcal{G}$ l'ensemble des fonctions réelles g continues bornées et telles que $\int_{\mathbb{R}} g(x)f(x)dx = 0$ et
   $\int_{\mathbb{R}} g^2(x)f(x)dx < +\infty$. Montrer qu'il existe $\delta > 0$ tel que, si $|t| \leq \delta$, en posant $P_{t,g} = (1+tg)P$, on a $P_{t,g} \in \mathcal{P}$, puis que $\mathcal{G}$ est un ensemble tangent à $\mathcal{P}$ en P.
2. On fixe g dans $\mathcal{G}$. On note, pour tout $t \in]-\delta,\delta[$, $F(x,t) = \int_{-\infty}^x f(u)du + t\int_{-\infty}^x g(u)f(u)du$ la fonction de répartition de $P_{t,g}$. $\psi(P_{t,g})$ est alors l'unique réel tel que $F(\psi(P_{t,g}),t) = \frac{1}{2}$. En utilisant le théorème des fonctions implicites, montrer que $t \mapsto \psi(P_{t,g})$ est dérivable par rapport à t en $t = 0$ et que

$$\left.\frac{d\psi(P_{t,g})}{dt}\right|_{t=0} = -\frac{1}{2f(\psi(P))}\left[\int_{-\infty}^{\psi(P)} g(u)f(u)du - \int_{\psi(P)}^{+\infty} g(u)f(u)du\right].$$

3. Quelle est la fonction d'influence efficace de ψ au point P ?
4. Montrer que la médiane empirique est un estimateur efficace de $\psi(P)$ (On rappelle que si θ est la médiane de P, si $\widehat{\theta}_n$ est la médiane empirique, $\sqrt{n}(\widehat{\theta}_n - \theta) = \frac{1}{2f(\theta)\sqrt{n}} \sum_{i=1}^n (\mathbb{1}_{X_i>\theta} - \mathbb{1}_{X_i<\theta}) + o_P(1)$).

6.3 Partie de l'examen de janvier 2012

Exercice 6.3.1. Soit $\mathcal{P}$ l'ensemble des probabilités sur $\mathbb{R}$ de densité strictement positive. Pour tout $P \in \mathcal{P}$, on définit $\psi(P)$ comme étant la médiane de P, c'est à dire le nombre réel défini par $F(\psi(P)) = \frac{1}{2}$, F étant la fonction de répartition de P. Soit $P \in \mathcal{P}$ de densité f.

1. Soit $\mathcal{G}$ l'ensemble des fonctions réelles g bornées et telles que $\int_{\mathbb{R}} g(x)f(x)dx = 0$ et $\int_{\mathbb{R}} g^2(x)f(x)dx < +\infty$. Montrer qu'il existe $\delta > 0$ tel que, si $|t| \leq \delta$, en posant $P_{t,g} = (1 + tg)P$, on a $P_{t,g} \in \mathcal{P}$, puis que $\mathcal{G}$ est un ensemble tangent à $\mathcal{P}$ en P.
2. On fixe g dans $\mathcal{G}$. Montrer que si t tend vers 0, $\psi(P_{t,g})$ tend vers $\psi(P)$.
3. Montrer que
$$F\left(\psi(P_{t,g})\right) - F\left(\psi(P)\right) = -t \int_{-\infty}^{\psi(P_{t,g})} g(u)f(u)du.$$
4. En déduire que
$$\lim_{t\to 0} \frac{\psi(P_{t,g}) - \psi(P)}{t} = -\frac{1}{f(\psi(P)} \int_{-\infty}^{\psi(P)} g(u)f(u)du = -\frac{1}{2f(\psi(P)} \left[\int_{-\infty}^{\psi(P)} g(u)f(u)du - \int_{\psi(P)}^{+\infty} g(u)f(u)du\right],$$
Quelle est la fonction d'influence efficace de ψ au point P ?
5. Montrer que la médiane empirique est un estimateur efficace de $\psi(P)$ (On rappelle que si θ est la médiane de P, si $\widehat{\theta}$ est la médiane empirique, $\sqrt{n}(\widehat{\theta} - \theta) = \frac{1}{2f(\theta)\sqrt{n}} \sum_{i=1}^n (\mathbb{1}_{X_i>\theta} - \mathbb{1}_{X_i<\theta}) + o_P(1)$).

Exercice 6.3.2. Soit T une variable aléatoire de loi a priori Π sur $\mathbb{R}^+$ qui est la loi gamma de paramètres $a > 0$ et $b > 0$, notée $G(a; b)$, de densité
$$\frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \mathbb{1}_{\theta>0},$$
avec $\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}dx$. Pour tout entier n, les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes conditionnellement à $T = \theta$ et de loi de Poisson P_θ de paramètre θ, c'est à dire de densité
$$p_\theta(x) = \frac{\theta^x}{x!} e^{-\theta}$$

par rapport à la mesure de comptage μ (qui donne un poids égal à 1 à tout entier positif ou nul).

1. Montrer que la loi a posteriori est la loi gamma $G(\sum_{i=1}^n X_i + a; n + b)$.
2. Soit $\widehat{\theta}$ l'espérance de la loi a posteriori. Montrer que
$$\widehat{\theta} = \frac{\sum_{i=1}^n + a}{n + b}.$$
3. Soit $\theta_0 > 0$. Montrer que le modèle $(P_\theta)_{\theta>0}$ est différentiable en moyenne quadratique en θ_0, de score $\dot{\ell}_{\theta_0}(x) = \frac{x-\theta_0}{\theta_0}$ et d'information de Fisher $I_{\theta_0} = \frac{1}{\theta_0}$.
4. Montrer que si l'on note $\tilde{\theta}$ l'estimateur du maximum de vraisemblance,
$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = \sqrt{n}\left(\tilde{\theta} - \theta_0\right) + o_{P_{\theta_0}^{\otimes n}}(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \theta_0) + o_{P_{\theta_0}^{\otimes n}}(1).$$
5. Montrer que le théorème de Bernstein-von Mises est vérifié et l'énoncer.
6. Soit $\lambda > 0$. On veut tester "$\theta \leq \lambda$" contre "$\theta > \lambda$". Soit ϕ_n le test bayésien qui vaut 1 si
$$\Pi(T > \lambda | X_1, \ldots, X_n) \geq \Pi(T \leq \lambda | X_1, \ldots, X_n)$$
et 0 sinon. Déduire du théorème de Bernstein-von Mises que pour tout $\theta_0 < \lambda$,
$$\lim_{n\to+\infty} P_{\theta_0}^{\otimes n}(\phi_n = 1) = 0,$$
et que pour tout $\theta_0 > \lambda$,
$$\lim_{n\to+\infty} P_{\theta_0}^{\otimes n}(\phi_n = 0) = 0.$$

Bibliographie