Éléments de Statistique Asymptotique

Marie-Claude VIANO et Charles SUQUET

Master 2 Recherche de Mathématiques année 2010–2011

Table des matières

Introduction

Ce cours est d'abord une « visite guidée» de quelques points importants en statistique asymptotique : le rôle de la mesure empirique pour estimer la loi des variables, les δ-méthodes qui permettent d'obtenir les distributions limites d'estimateurs par des développements limités de la mesure empirique autour de cette loi, les notions de robustesse qui conduisent à la construction d'estimateurs non optimaux mais aux performances peu fragiles, les notions de contiguïté et de normalité asymptotique locale qui, par des méthodes de géométrie différentielle, fournissent une explication à certains « comportements invariants» en statistique classique comme la normalité asymptotique des estimateurs du maximum de vraisemblance et la vitesse en $\sqrt{n}$ obtenue dans la plupart des convergences en loi.

Le deuxième objectif de ce cours est de quitter le domaine des variables indépendantes, terrain de prédilection des statisticiens. Pour un statisticien, un «échantillon» est la réalisation d'un vecteur aléatoire à composantes indépendantes et de même loi. Depuis une cinquantaine d'années, on s'est intéressé à ce qu'il advient des résultats limites bien connus (loi des grands nombres, théorème central limite, loi du logarithme itéré, etc...) lorsque les variables sont dépendantes. On est arrivé dans bien des cas à évaluer l'impact de la perte d'indépendance. Dans la mesure du possible, chaque chapitre du cours consacre un paragraphe à cette question.

CHAPITRE 1

Rappels sur les notions de convergence

Dans ce chapitre, nous examinerons des questions de convergence de suites de variables aléatoires à valeurs dans des espaces métriques. Ces « variables aléatoires » seront définies sur un espace $(\Omega, \mathcal{F})$ et à valeurs dans un espace métrique E. Parler de variables aléatoires et pas seulement d'applications $\Omega \to E$, suppose une certaine mesurabilité. Pour cela, nous munirons généralement E de sa tribu *borélienne*, c'est à dire la tribu engendrée par les ouverts de E. On emploiera aussi l'expression « élément aléatoire de E » pour désigner de telles variables.

1. Distances entre mesures de probabilité

Entre autres distances, nous en retiendrons trois.

1.1. La distance de Prokhorov.

Définition 1.1 (distance de Prokhorov). *La distance de Prokhorov de deux mesures de probabilité P et Q sur la tribu borélienne $\mathcal{E}$ d'un espace métrique (E, d) est définie par :*

$$\pi(P,Q) := \inf\{\varepsilon > 0| \quad \forall A \in \mathcal{E},\ P(A) \leq Q(A^\varepsilon) + \varepsilon \text{ et } Q(A) \leq P(A^\varepsilon) + \varepsilon\},$$

où A^ε désigne l'ensemble des points de E dont la distance à A est strictement inférieure à ε.

Théorème 1.2.

(i) Dans la définition de π ci-dessus, on peut se restreindre à la famille des A fermés de E.

(ii) $\pi(P,Q) = \pi_1(P,Q) = \inf\{\varepsilon > 0| P(A) \leq Q(A^\varepsilon) + \varepsilon$ pour tout A fermé de $E\}$.

(iii) π est une distance sur l'ensemble des mesures de probabilité sur $\mathcal{E}$.

(iv) Si E est séparable[1]*, π métrise*[2] *la convergence en loi.*

Preuve.
Remarques préliminaires. Rappelons que si A est une partie d'un espace métrique (E, d), la distance d'un élément e de E à A est définie par $d(e, A) = \inf\{d(e, x);\ x \in A\}$. On vérifie facilement que $d(e, A) = 0$ si et seulement si e appartient à la fermeture $\overline{A}$ de A et que $d(e, A) = d(e, \overline{A})$. De plus, l'application $\varphi : E \to \mathbb{R}^+$, $x \mapsto d(x, A)$ est *continue* [23, p. 103], ce qui implique que $A^\varepsilon = \varphi^{-1}(]-\infty, \varepsilon[)$ est un ouvert de E.

1. C'est-à-dire s'il existe une partie dénombrable dense dans E.
2. Autrement dit, X_n converge en loi vers X si et seulement si $\pi(P_{X_n}, P_X)$ tend vers 0, où P_{X_n}, P_X désignent les lois des X_n et de X.

Pour toutes mesures de probabilités P et Q sur $\mathcal{E}$ et toute sous-famille $\mathcal{G}$ de $\mathcal{E}$, notons

$$I_{\mathcal{G}}(P,Q) := \{\varepsilon > 0 | \quad \forall A \in \mathcal{G},\ P(A) \leq Q(A^\varepsilon) + \varepsilon \text{ et } Q(A) \leq P(A^\varepsilon) + \varepsilon\}.$$

0n remarque alors que $I_{\mathcal{G}}(P,Q)$ est un *intervalle* de $\mathbb{R}$. En effet, si $\varepsilon \in I_{\mathcal{G}}(P,Q)$, pour tout $\varepsilon' > \varepsilon$, $Q(A^\varepsilon) + \varepsilon \leq Q(A^{\varepsilon'}) + \varepsilon'$ et de même avec P, d'où $\varepsilon' \in I_{\mathcal{G}}(P,Q)$. Ceci montre que si $I_{\mathcal{G}}(P,Q) \neq \emptyset$, c'est un intervalle de $\mathbb{R}$, de borne supérieure $+\infty$. En fait, $[1,+\infty[\subset I_{\mathcal{G}}(P,Q) \subset [0,+\infty[$, puisque P et Q étant des probabilités, il est clair que tout réel $\varepsilon \geq 1$ est dans $I_{\mathcal{G}}(P,Q)$. Ceci montre la première inclusion, la deuxième résultant de la condition « $\varepsilon > 0$ » dans la définition de $I_{\mathcal{G}}(P,Q)$.

Notons enfin que si $\mathcal{G} \subset \mathcal{G}'$, alors $I_{\mathcal{G}'}(P,Q) \subset I_{\mathcal{G}}(P,Q)$, d'où $\inf I_{\mathcal{G}}(P,Q) \leq \inf I_{\mathcal{G}'}(P,Q)$.

Preuve de (i). En notant $\mathcal{F}$ la famille des fermés de E et $\tilde{\pi}(P,Q)$ la borne inférieure de $I_{\mathcal{F}}(P,Q)$, nous avons à comparer cette borne avec la distance de Prokhorov $\pi(P,Q) = \inf I_{\mathcal{E}}(P,Q)$. D'abord puisque la famille $\mathcal{F}$ des fermés est incluse dans la tribu borélienne $\mathcal{E}$, la remarque précédente nous donne $\tilde{\pi}(P,Q) \leq \pi(P,Q)$ pour toutes probabilités P et Q sur $\mathcal{E}$. Supposons qu'il existe un couple de probabilités (P,Q) pour lequel cette inégalité soit stricte. Alors il existe un $\varepsilon \in]\tilde{\pi}(P,Q), \pi(P,Q)[$ et un borélien $A \in \mathcal{E}$ vérifiant l'une au moins des deux inégalités $P(A) > Q(A^\varepsilon) + \varepsilon$ ou $Q(A) > P(A^\varepsilon) + \varepsilon$. Pour fixer les idées, disons que la première est réalisée (le raisonnement qui suit reste valable en échangeant les rôles de P et Q). Alors comme $\varepsilon > \tilde{\pi}(P,Q)$, on a aussi $P(\overline{A}) \leq Q((\overline{A})^\varepsilon) + \varepsilon$. Comme $(\overline{A})^\varepsilon = A^\varepsilon$ et $P(\overline{A}) \geq P(A)$, ceci est contradictoire avec la minoration stricte de $P(A)$ supposée ci-dessus. Par conséquent, l'inégalité stricte $\tilde{\pi}(P,Q) < \pi(P,Q)$ ne peut avoir lieu pour aucun couple (P,Q), d'où $\tilde{\pi}(P,Q) \leq \pi(P,Q)$ pour toutes probabilités P et Q. On peut donc bien se restreindre aux A fermés pour calculer la distance de Prokhorov.

Preuve de (ii). Pour établir le point (ii), il suffit de vérifier que pour toutes mesures de probabilité P et Q sur $\mathcal{E}$, $\pi_1(P,Q) = \pi_1(Q,P)$.

Commençons par le cas particulier où $\pi_1(P,Q) = 0$. Dans ce cas on a pour tout A fermé et tout entier $n \geq 1$, $P(A) \leq Q(A^{1/n}) + \frac{1}{n}$. Comme A est fermé, $\cap_{n\geq 1} A^{1/n} = A$, d'où par continuité séquentielle décroissante[3] de Q, $P(A) \leq Q(A)$ en faisant tendre n vers l'infini. En passant aux complémentaires, on en déduit que pour tout ouvert B de E, $Q(B) \leq P(B)$. Maintenant, raisonnons par l'absurde en supposant que $\pi_1(Q,P) > 0$. Alors il existe un ε tel que $0 < \varepsilon < \pi_1(Q,P)$ et un fermé F tel que $Q(F) > P(F^\varepsilon) + \varepsilon$. Ainsi $Q(F^\varepsilon) \geq Q(F) > P(F^\varepsilon) + \varepsilon$, d'où $Q(F^\varepsilon) > P(F^\varepsilon)$. Mais comme F^ε est ouvert, ceci contredit la validité pour tout ouvert B de l'inégalité $Q(B) \leq P(B)$. On a donc bien $\pi_1(Q,P) = 0 = \pi_1(P,Q)$. En prime nous avons montré au passage que si $\pi_1(P,Q) = 0$, alors pour tout fermé A, on a à la fois $P(A) \leq Q(A)$ et $Q(A) \leq P(A)$, d'où $P(A) = Q(A)$. Ainsi les mesures finies P et Q coïncident sur la classe des fermés de E, qui est stable par intersections finies et engendrent la tribu borélienne $\mathcal{E}$, donc elles coïncident sur $\mathcal{E}$, d'où $P = Q$.

Traitons maintenant le cas $\pi_1(P,Q) > 0$. Alors pour tout ε tel que $0 < \varepsilon < \pi_1(P,Q)$, il existe un fermé A tel que $P(A) > Q(A^\varepsilon) + \varepsilon$, d'où en passant aux complémentaires $Q(E \setminus A^\varepsilon) > P(E \setminus A) + \varepsilon$. Remarquons que si $d(x,y) < \varepsilon$ et $x \in A$, alors $y \in A^\varepsilon$.

3. $A^{1/n} \downarrow A \Rightarrow Q(A^{1/n}) \downarrow A$.

On en déduit que si $d(x,y) < \varepsilon$ et $y \in E \setminus A^\varepsilon$, alors $x \notin A$, ce qui établit l'inclusion $(E \setminus A^\varepsilon)^\varepsilon \subset E \setminus A$. On a donc $Q(E \setminus A^\varepsilon) > P(E \setminus A) + \varepsilon \geq P\Big((E \setminus A^\varepsilon)^\varepsilon\Big) + \varepsilon$. On obtient ainsi l'existence d'un fermé $F = E \setminus A^\varepsilon$ tel que $Q(F) > P(F^\varepsilon) + \varepsilon$, ce qui entraîne que $\varepsilon \leq \pi_1(Q,P)$. Nous venons ainsi de vérifier que pour tout $0 < \varepsilon < \pi_1(P,Q)$, $\varepsilon \leq \pi_1(Q,P)$. On en déduit que $\pi_1(P,Q) \leq \pi_1(Q,P)$. En tenant compte du cas particulier $\pi_1(P,Q) = 0$ étudié ci-dessus, nous avons maintenant établi que pour *tout* couple (P,Q) de probabilités, $\pi_1(P,Q) \leq \pi_1(Q,P)$. En échangeant P et Q l'inégalité inverse est donc aussi vraie et finalement $\pi_1(P,Q) = \pi_1(Q,P)$.

Preuve de (iii). Vérifions que π est une distance sur l'ensemble de smesures de probabilité sur $\mathcal{E}$. La symétrie découle immédiatement de la définition de π. D'autre part si $P = Q$, tout $\varepsilon > 0$ vérifie les inégalités figurant dans la définition de π, d'où $\pi(P,P) = \inf]0,+\infty[= 0$. Réciproquement, on a vu comme sous-produit de la preuve de (ii) que l'égalité $\pi(P,Q) = 0$ implique $\pi_1(P,Q) = 0$ et que ceci implique l'égalité de P et Q.

Il ne nous reste plus qu'à montrer que π vérifie l'inégalité triangulaire, ou plus simplement que π_1 la vérifie : $\pi_1(P,Q) \leq \pi_1(P,R) + \pi_1(R,Q)$ pour P,Q,R probabilités quelconques sur $\mathcal{E}$. On remarque pour cela que pour tous réels x et y tels que $\pi_1(P,R) < x$ et $\pi_1(R,Q) < y$ et tout borélien A de E,

$$P(A) \leq R(A^x) + x \leq Q\Big((A^x)^y\Big) + y + x \leq Q(A^{x+y}) + x + y,$$

d'où $\pi_1(P,Q) \leq x + y$. Avec $x \downarrow \pi_1(P,R)$ et $y \downarrow \pi_1(R,Q)$, il vient $\pi_1(P,Q) \leq \pi_1(P,R) + \pi_1(R,Q)$.

Preuve partielle de (iv). Nous nous contenterons de démontrer que si les X_n, X sont des éléments aléatoires de E de loi respective P_n, P et si $\pi_1(P_n,P)$ tend vers 0, alors X_n converge en loi dans E vers X. Pour cela, en vertu du *portmanteau theorem* (Th. 1.11 ci-dessous), il suffit de montrer que si A est un borélien tel que $P(\partial A) = 0$, $P_n(A)$ converge vers $P(A)$. Rappelons ici que l'intérieur de A est le plus grand ouvert contenu dans A et l'extérieur de A le plus grand ouvert contenu dans son complémentaire. La frontière ∂A de A est l'ensemble des points qui ne sont ni intérieurs ni extérieurs à A. Comme l'intérieur de A est l'extérieur de son complémentaire et vice-versa, A et $E \setminus A$ ont même frontière. Fixons $\varepsilon > 0$ quelconque. Notons que pour tout entier $k \geq 1$,

$$(A^{1/k} \setminus A) \subset (A^{1/k} \setminus \mathring{A}) \downarrow (\overline{A} \setminus \mathring{A}) = \partial A.$$

Comme $P(\partial A) = 0$, on peut trouver $0 < \delta < \varepsilon$ suffisamment petit pour que $P(A^\delta \setminus A) < \varepsilon$ et $P((E \setminus A)^\delta \setminus (E \setminus A)) < \varepsilon$. Par hypothèse, $\pi_1(P_n,P)$ tend vers 0, donc on peut trouver un n_0 dépendant de ε tel que pour tout $n \geq n_0$, $\pi_1(P_n,P) < \delta$. On en déduit les inégalités $P_n(A) \leq P(A^\delta) + \delta \leq P(A) + 2\varepsilon$ et de même $P_n(E \setminus A) \leq P(E \setminus A) + 2\varepsilon$. En réécrivant cette dernière sous la forme $1 - P_n(A) \leq 1 - P(A) + 2\varepsilon$, on obtient $P(A) \leq P_n(A) + 2\varepsilon$ et finalement $-2\varepsilon \leq P_n(A) - P(A) \leq 2\varepsilon$. Comme ε était quelconque, la convergence de $P_n(A)$ vers $P(A)$ est établie.

$\square$

Exercice 1.1. On suppose $Q = (1-\alpha)P + \alpha R$ où $\alpha \in [0,1]$ et R est une mesure de probabilité. Montrer que $\pi(P,Q) \le \alpha$. ◁

Exercice 1.2. Montrer que pour tous $x, y \in E$, $\pi(\delta_x, \delta_y) = \min(1, d(x,y))$. ◁

Exercice 1.3. On définit la distance de Paul Lévy de deux lois de probabilité P et Q sur $\mathbb{R}$ par

$$L(P,Q) = \inf\{\varepsilon > 0 | G(x-\varepsilon) - \varepsilon \le F(x) \le G(x+\varepsilon) + \varepsilon,\ \forall x\},$$

où F et G sont les fonctions de répartition respectives de P et Q.

a) Vérifier que L est une distance.

b) Justifiez l'inégalité $L(P,Q) \le \pi(P,Q)$ et en déduire que si la suite de variables aléatoires réelles $(X_n)_{n\ge 1}$ converge en loi vers X, $L(P_{X_n}, P_X)$ converge vers 0, où la notation P_Y désigne la loi de Y.

c) Montrez que si $L(P_{X_n}, P_X)$ converge vers 0, alors X_n converge en loi vers X. Ainsi L métrise la convergence en loi.

d) Trouver deux suites de lois $(P_n)_{n\ge 1}$ et $(Q_n)_{n\ge 1}$ telles que $L(P_n, Q_n)$ converge vers 0, mais pas $\pi(P_n, Q_n)$. En déduire que les distances π et L ne sont pas équivalentes.

◁

1.2. La distance en variation totale. Ici, $(E, \mathcal{E})$ est un espace mesurable quelconque.

$$\|P - Q\| = \sup_{A\in\mathcal{E}}\{|P(A) - Q(A)|\}.$$

Cette distance est issue de la norme en variation totale définie sur l'ensemble des mesures signées, ensemble dont on ne parlera pas ici. Sur l'ensemble des mesures positives, cette norme est tout simplement la masse totale de la mesure.

Exercice 1.4. Montrer que $\pi(P,Q) \le \|P - Q\|$ et trouver un exemple où l'inégalité est stricte. ◁

Exercice 1.5. (Théorème de Scheffé). Si P et Q ont respectivement pour densité f et g par rapport à une même mesure μ,

$$\|P - Q\| = \frac{1}{2}\int |f(x) - g(x)|\,\mathrm{d}\mu(x) = 1 - \int \min\{f(x), g(x)\}\,\mathrm{d}\mu(x). \tag{1.1}$$

Montrer que l'hypothèse d'existence de densités n'est pas restrictive. ◁

Ce dernier résultat rend la distance en variation totale souvent facile à manipuler, même si la distance de Prohorov, puisqu'elle métrise la convergence en loi, est plus adaptée à beaucoup de problèmes statistiques.

On a l'habitude de noter $P \wedge Q$ la mesure qui a pour densité $\min\{f(x), g(x)\}$. L'égalité (1.1) s'écrit

$$\|P - Q\| = 1 - \|P \wedge Q\|.$$

Il est à noter que la distance en variation totale a une signification statistique, comme on le voit dans l'exercice qui suit.

Exercice 1.6. Supposons qu'on désire tester l'hypothèse que la loi d'une variable X est P, contre l'hypothèse que sa loi est Q. Pour toute région critique C on regarde la somme des deux erreurs $P(C) + 1 - Q(C)$. Montrer que la valeur minimale de cette somme est atteinte et qu'elle vaut $\|P \wedge Q\|$. ◁

1.3. La distance d'Hellinger. Avec les notations de l'exercice 1.5, on pose

$$H^2(P,Q) = \int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \mathrm{d}\mu(x).$$

La distance $H = \sqrt{H^2}$ ainsi définie est invariante par changement de mesure de référence.

On appelle *affinité d'Hellinger* la quantité

$$A(P,Q) = \int \sqrt{f(x)g(x)}\, \mathrm{d}\mu(x).$$

On a évidemment

$$H^2(P,Q) = 2(1 - A(P,Q)).$$

On remarque que, si P_1 et Q_1 (resp. P_2 et Q_2) sont deux mesures sur E_1 (resp. E_2) à densité par rapport à μ_1 (resp. μ_2) on a

$$A(P_1 \otimes P_2, Q_1 \otimes Q_2) = A(P_1, Q_1)A(P_2, Q_2),$$

ce qui montre que la distance d'Hellinger est de manipulation particulièrement aisée pour des mesures produit. C'est la situation rencontrée en statistique lorsqu'on a affaire à des variables indépendantes. De plus, la distance d'Hellinger se compare bien à la distance en variation totale.

Exercice 1.7. Montrer que

$$\frac{1}{2}H^2(P,Q) \le \|P - Q\| \le \min\Big(H(P,Q); 2 - A^2(P,Q)\Big).$$

◁

Exercice 1.8. Calculer la distance en variation totale et la distance d'Hellinger entre la gaussienne standard et la gaussienne d'espérance m et de variance 1. ◁

Exercice 1.9. Calculer $H^2(P_{\theta_0}, P_\theta)$ lorsque P_θ est la loi uniforme sur $[0, \theta]$. ◁

Exercice 1.10. On considère P_θ la loi triangulaire centrée en θ : sa densité est nulle en dehors de $[\theta - 1, \theta + 1]$ et sur cet intervalle elle est égale à $1 - (x - \theta)\mathrm{sgn}(x - \theta)$. Montrer que

$$H^2(P_0, P_\theta) = -\frac{\theta^2}{2}\ln\theta + o(\theta^2) \quad \text{lorsque } \theta \to 0.$$

◁

2. Convergence de variables aléatoires à valeurs dans des espaces métriques

2.1. Rappels. Pour les convergences de suites de variables aléatoires à valeurs dans $\mathbb{R}^k$, le lecteur peut se reporter au chapitre 2 de [24], ou encore à [15]. On se contente ici de deux rappels et d'un complément. Dans la suite de ce cours, ces résultats pourront utilisés dans le contexte plus général où les X_n sont définies sur des ensembles probabilisés $(\Omega_n, \mathcal{A}_n, P_n)$ différents selon les valeurs de n. Au lecteur de vérifier que les définitions, les notations aussi bien que les résultats qui suivent se transposent sans difficulté dans ce contexte.

Définition 1.3. *Considérons, sur l'espace $(\Omega, \mathcal{A}, P)$, une suite $(X_n)_{n\geq 1}$ de vecteurs aléatoires. Cette suite est dite bornée en probabilité si pour tout $\varepsilon > 0$ il existe M tel que*

$$\sup_{n\geq 1} P(\|X_n\| > M) \leq \varepsilon.$$

Pour des raisons évidentes, on écrit aussi $X_n = O_P(1)$.

Tout naturellement les notions de o et O se transposent pour les suites aléatoires. Dans la définition suivante, $(X_n)_{n\geq 1}$ est une suite de vecteurs aléatoires de $\mathbb{R}^k$ et $(R_n)_{n\geq 1}$, $(Y_n)_{n\geq 1}$ deux suites de variables aléatoires réelles positives.

Définition 1.4.

$$X_n = o_P(R_n) \qquad \textit{signifie que} \quad \|X_n\| = Y_n R_n \textit{ où } \quad Y_n \xrightarrow[n\to+\infty]{\mathrm{P}} 0$$

$$X_n = O_P(R_n) \qquad \textit{signifie que} \quad \|X_n\| = Y_n R_n \textit{ où } \quad Y_n = O_P(1).$$

Le théorème de Prokhorov étend le théorème de Heine-Borel aux suites bornées en probabilité .

Théorème 1.5 (Théorème de Prohorov)**.** *Soit une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^p$.*

(1) Si la suite (X_n) de vecteurs aléatoires converge en loi, elle est bornée en probabilité.

(2) Si la suite (X_n) est bornée en probabilité, il existe une sous suite $\varphi(n)$ telle que $(X_{\varphi(n)})$ a une limite en loi.

Preuve. Supposons que la suite (X_n) a une limite en loi X. Soit $\varepsilon > 0$ fixé. D'après le (4) du théorème 1.11, étant donné $M > 0$, on a $P(\|X_n\| \geq M) \leq P(\|X\| \geq M) + \varepsilon$ pour $n \geq n_0$. Choisissant M de sorte que $P(\|X\| \geq M) < \varepsilon$ on obtient

$$P(\|X_n\| \geq M) \leq 2\varepsilon \quad \forall n \geq n_0.$$

Enfin, quitte à modifier M, l'inégalité qui précède est valable aussi pour $n < n_0$ ce qui prouve le (1). Prouvons le (2). Pour simplifier la preuve, on suppose que $p = 1$. Considérons la suite (F_n) des fonctions de répartitions des X_n. On sait (lemme de Helly, voir par exemple [24], page 9) que de cette suite on peut extraire une sous suite $(F_{\varphi(n)})$ et une fonction croissante positive F telle $F_{\varphi(n)}(x) \to F(x)$ en tout point de continuité de F. Il ne reste plus qu'à montrer que F est une fonction de répartition. C'est à dire que $F(x)$ tend vers 1 (resp 0) lorsque lorsque $x \to \infty$ (resp. $-\infty$). Pour cela on utilise le fait que $X_n = O_P(1)$.

En effet il existe alors M tel que $F_n(x) \geq 1 - \varepsilon$ (resp. $F_n(x) \leq \varepsilon$) pour tout n et tout $x \geq M$ (resp. $x \leq -M$). On peut toujours supposer que M (resp. $-M$) est un point de continuité de F (il suffit éventuellement de le remplacer par une valeur plus grande). Ceci suffit à prouver que $F(x) \geq 1 - \varepsilon$ pour tout $x \geq M$ et $F(x) \leq \varepsilon$ pour tout $x \leq -M$, et le résultat est obtenu.

□

Terminons en signalant que la distance en variation totale définie au paragraphe 1.2 donne lieu à une notion de convergence

Définition 1.6. *On dit que X_n converge en variation totale (ou, plus brièvement « en variation») si*

$$\|P_{X_n} - P_X\| \to 0 \quad \textit{quand} \quad n \to \infty.$$

C'est une convergence de la loi de X_n vers celle de X. Elle est plus forte que la convergence en loi (à vous de trouver un exemple). Grâce au théorème de Scheffé (exercice 1.5), il est parfois facile de prouver cette convergence.

Proposition 1.7. *Si X_n (pour tout n) et X ont toutes une densité (notées f_n et f) par rapport à une même mesure μ telles que, lorsque n tend vers l'infini*

$$f_n \to f \quad \mu\textit{-presque-partout}$$

alors X_n converge en variation vers X.

Preuve. D'après (1.1), il suffit de montrer que $\int \min(f_n, f)\,\mathrm{d}\mu$ tend vers 1. La convergence μ-p.p. de f_n vers f entraîne celle de $\min(f_n, f)$ vers $\min(f, f) = f$. Cette dernière convergence est dominée par la fonction μ intégrable f, donc par le théorème de Lebesgue, on a

$$\lim_{n\to\infty} \int \min(f_n, f)\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu = 1,$$

puisque f est une densité de probabilité relativement à μ. □

2.2. Variables à valeurs dans des espaces métriques généraux. Les variables de ce paragraphe sont à valeurs dans un espace métrique E muni de la tribu borélienne $\mathcal{E}$ associée à sa distance d, c'est à dire la plus petite tribu qui contient les ouverts pour d. Mis à part les exemples simples que sont $\mathbb{R}^k$, ou encore la droite complétée $[-\infty, \infty]$ (munie de la distance $d(x, y) = |\Phi(x) - \Phi(y)|$ où Φ est n'importe quelle fonction strictement croissante de 0 à 1), on peut citer :

Exemple 1.1. L'espace $l^\infty(T)$ des fonctions bornées d'un ensemble T dans $\mathbb{R}$, muni de la norme $\sup_{t\in T} |g(t)|$ et de la distance associée. ◁

Exemple 1.2. L'espace $C[a, b]$ des fonctions continues de $[a, b]$ dans $\mathbb{R}$, muni de la distance

$$d(f, g) = \sup_{x\in[a,b]} |f(x) - g(x)|.$$

Ici il est possible de prendre $a = -\infty$ ou (et) $b = +\infty$, mais alors les fonctions sont supposées avoir une limite en ces points. ◁

Exemple 1.3. L'espace $D[a,b]$ des fonctions càd làg de $[a,b]$ dans $\mathbb{R}$, muni de la même distance. On a bien sûr

$$C[a,b] \subset D[a,b] \subset l^\infty([a,b]).$$

$\triangleleft$

2.2.1. *Convergences presque sûre et en probabilité.* Les définitions sont les mêmes que pour les variables à valeurs dans $\mathbb{R}^k$, mise à part l'introduction d'une difficulté nouvelle dûe au fait que $d(X,Y)$ n'est pas nécessairement mesurable même si X et Y le sont. Voyons ce point en détail. Si X et Y sont des éléments aléatoires de E définis sur le même espace probabilisable $(\Omega,\mathcal{F})$, ils sont mesurables $\mathcal{F}$–$\mathcal{E}$, où $\mathcal{E}$ est la tribu borélienne de E. Il en résulte que le couple (X,Y) est mesurable $\mathcal{F}$–$\mathcal{E}\otimes\mathcal{E}$, où $\mathcal{E}\otimes\mathcal{E}$ est la tribu produit engendrée par les produits cartésiens $A\times B$, A, $B\in\mathcal{E}$. D'autre part on peut munir $E\times E$ de la topologie produit $\mathcal{T}$, dont les ouverts sont les *unions quelconques* de « rectangles ouverts » $U\times V$, U et V ouverts de E. La tribu $\sigma(\mathcal{T})$ engendrée par $\mathcal{T}$ contient toujours la tribu $\mathcal{E}\otimes\mathcal{E}$ et l'inclusion inverse est *vraie lorsque E est séparable* [14, Prop. 4.1.7]. Dans ce cas, d étant continue relativement à la topologie produit $\mathcal{T}$, donc mesurable $\sigma(\mathcal{T})-\mathrm{Bor}(\mathbb{R})$, l'inclusion $\sigma(\mathcal{T})\subset\mathcal{E}\otimes\mathcal{E}$ nous permet de conclure que l'application $d(X,Y)=d\circ(X,Y)$: $\omega\mapsto d(X(\omega),Y(\omega))$ est $\mathcal{F}$–$\mathrm{Bor}(\mathbb{R})$ mesurable. En particulier $\{\omega\in\Omega;\ d(X(\omega),Y(\omega))>\varepsilon\}$ est bien un élément de $\mathcal{F}$ et sa probabilité est bien définie. Par contre, si E n'est pas séparable, on ne peut affirmer que $\sigma(\mathcal{T})=\mathcal{E}\otimes\mathcal{E}$ et en cas d'inclusion stricte $\mathcal{E}\otimes\mathcal{E}\subsetneq\sigma(\mathcal{T})$, la mesurabilité de $d\circ(X,Y)$ n'est plus garantie.

Pour la convergence en probabilité cette difficulté de mesurabilité de d se résout en introduisant la probabilité extérieure P^* (voir par exemple [24]), définie sur la famille $\mathcal{P}(E)$ de toutes les parties de E par :

$$\forall A\in\mathcal{P}(E),\quad P^*(A)=\inf\{P(B);\ B\supset A,\ B\in\mathcal{E}\}. \tag{1.2}$$

Définition 1.8. *On dit que la suite X_n converge en probabilité vers X, élément aléatoire à valeurs dans E si, lorsque n tend vers l'infini,*

$$P^*\left(d(X_n,X)\geq\varepsilon\right)\to 0\quad\forall\varepsilon>0. \tag{1.3}$$

Lorsque E est séparable, on peut remplacer P^* par P dans (1.3).

Pour la convergence presque sûre, on demande à la distance d'être majorée par une suite de variables aléatoires positives convergeant p.s. vers 0.

Définition 1.9. *On dit que la suite X_n converge presque sûrement vers X, élément aléatoire à valeurs dans E si il existe une suite Δ_n de variables aléatoires positives telle que*

$$d(X_n,X)\leq\Delta_n\quad\forall n\qquad \textit{et}\quad P(\Delta_n\to 0)=1.$$

Dans le cas E séparable, on peut prendre $\Delta_n=d(X_n,X)$, ce qui revient à la définition classique de la convergence presque sûre : $P(d(X_n,X)\to 0)=1$.

2.2.2. *Convergence en loi.* Soit (X_n) une suite d'éléments aléatoires définis sur une suite d'espaces probabilisés $(\Omega_n, \mathcal{A}_n, P_n)$ et à valeurs dans le même espace métrique E muni de sa tribu borélienne. Soit X un élément aléatoire défini sur un espace probabilisé $(\Omega, \mathcal{A}, P)$ et à valeurs dans le même espace métrique E muni de sa tribu borélienne. On note $\mathbb{E}_n$ l'espérance sous P_n et $\mathbb{E}$ l'espérance sous P.

Définition 1.10. *On dit que la suite X_n converge en loi vers X, élément aléatoire à valeurs dans E si, lorsque n tend vers l'infini,*

$$\mathbb{E}_n(f(X_n)) \to \mathbb{E}(f(X))$$

pour toute application continue et bornée de E dans $\mathbb{R}$.

Le théorème suivant de caractérisation de la convergence en loi étend les résultats déjà connus pour les suites de vecteurs aléatoires de $\mathbb{R}^d$ avec une preuve analogue. Dans la littérature anglaise, il est connu sous le sobriquet de *portmanteau theorem*, qu'il serait erroné de traduire par « théorème du portemanteau », puisque le faux-ami *portmanteau* désigne en fait une grande valise à deux compartiments [4]. Le premier compartiment (1)–(2) contient des caractérisations fonctionnelles de la convergence en loi le deuxième (3)–(5) contient des caractérisations ensemblistes.

Théorème 1.11. *Sont équivalentes à la convergence en loi de X_n vers X les cinq propositions suivantes :*

(1) Pour toute application bornée et lipschitzienne de E dans $\mathbb{R}$,

$$\mathbb{E}_n(f(X_n)) \to \mathbb{E}(f(X)).$$

(2) Pour toute application continue de E dans $\mathbb{R}^+$,

$$\liminf \mathbb{E}_n(f(X_n)) \geq \mathbb{E}(f(X)).$$

(3) Pour tout ouvert O,

$$\liminf P_n(X_n \in O) \geq P(X \in O).$$

(4) Pour tout fermé F,

$$\limsup P_n(X_n \in F) \leq P(X \in F).$$

(5) Pour tout borélien A tel que $P(X \in \partial A) = 0$,

$$P_n(X_n \in A) \to P(X \in A).$$

4. D'après le Collins Cobuild : **1** *A portmanteau is a large travelling case which opens out into two equal compartments ; an old fashioned use.* **2** *Portmanteau is used to describe* **2.1** *a word that combines parts of the forms and meanings of two other words. For example, 'brunch' is formed from 'breakfast' and 'lunch'...* **2.2** *someone or something that combines many different features or uses.*

Exercice 1.11. Dans le cas $E = \mathbb{R}$, trouvez des exemples simples montrant que les inégalités de ce théorème peuvent être strictes. $\triangleleft$

Comme pour le cas de variables à valeurs dans $\mathbb{R}^k$, la convergence est conservée par transformation continue : si X_n converge en loi vers X alors $g(X_n)$ converge en loi vers $g(X)$ dès que g est continue. Le théorème qui suit (voir [24] pour la preuve) est un raffinement de ce résultat, qui sera utilisé au chapitre 4. La suite $g(X_n)$ est remplacée par $g_n(X_n)$ où les g_n peuvent ne pas être définies sur E tout entier.

Théorème 1.12. *Soit $(E_n)_{n\geq 0}$ une suite de sous ensembles de E et $(g_n)_{n\geq 0}$ une suite d'applications de E_n dans un espace métrique G. Si pour toute suite $x_n \in E_n$ on a $g_{\varphi(n)}(x_{\varphi(n)}) \to g_0(x)$ dès que $x_{\varphi(n)} \to x \in E_0$ le long de la sous suite $\varphi(n)$ alors, si la suite X_n d'éléments aléatoires de E_n converge en loi vers X_0, élément aléatoire de E_0, on a*

$$g_n(X_n) \xrightarrow[n\to+\infty]{\text{loi}} g_0(X_0).$$

Terminons en ajoutant que les liens bien connus entre convergence en loi, convergence en probabilité et convergence presque sûre sont conservés, nous ne les rappellerons pas.

CHAPITRE 2

Théorèmes classiques : rappels et compléments

1. Théorème de Glivenko-Cantelli. Ergodicité. Mesure empirique

1.1. Le théorème de Glivenko-Cantelli. Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs réelles (en fait on peut très bien énoncer ce qui suit pour des variables à valeurs dans $\mathbb{R}^k$, mais on suppose $k=1$ pour simplifier). On pose pour tout $x \in \mathbb{R}$

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{]-\infty,x]}(X_j). \tag{2.1}$$

La fonction aléatoire F_n est appelée fonction de répartition empirique.

Si les variables ont toutes même loi de fonction de répartition F, et si pour tout $x \in \mathbb{R}$, la loi forte des grands nombres s'applique aux variables aléatoires $\mathbb{I}_{]-\infty,x]}(X_j)$, $F_n(x)$ converge presque sûrement vers $F(x)$.

Le théorème suivant montre que sans hypothèse supplémentaire, il existe un évènement de probabilité 1, sur lequel la convergence de F_n vers F est uniforme (relativement à x) sur $\mathbb{R}$.

Théorème 2.1 (Théorème de Glivenko Cantelli)**.** *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires de même loi, de fonction de répartition F, et telle que pour tout intervalle A*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_A(X_i) \xrightarrow{\text{p.s.}} P(X \in A) = \int_A dF. \tag{2.2}$$

Sous ces conditions,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{p.s.}} 0.$$

Preuve. Pour tout $x \in \mathbb{R}$, en prenant $A =]-\infty, x]$ puis $]-\infty, x[$ dans (2.2), on obtient

$$F_n(x) \xrightarrow{\text{p.s.}} F(x), \qquad F_n(x^-) \xrightarrow{\text{p.s.}} F(x^-). \tag{2.3}$$

On rappelle maintenant que l'inverse généralisé de F (ou fonction quantile) est défini sur $]0,1[$ par

$$\forall u \in]0,1[, \quad F^{-1}(u) := \inf\{t \in \mathbb{R};\ F(t) \geq u\}. \tag{2.4}$$

En raison de la continuité à droite de F, on a toujours $F(F^{-1}(u)) \geq u$. D'autre part, il résulte de la définition de $F^{-1}(u)$ comme borne inférieure des t tels que $F(t) \geq u$ que pour tout $s < F^{-1}(u)$, $F(s) < u$. Comme F a une limite à gauche en $F^{-1}(u)$, notée $F(F^{-1}(u)^-)$, on en déduit que $F(F^{-1}(u)^-) \leq u$. Nous avons donc l'encadrement :

$$\forall u \in]0,1[, \quad F(F^{-1}(u)^-) \leq u \leq F(F^{-1}(u)). \tag{2.5}$$

Fixons $N \in \mathbb{N}$ et, pour tout $j \in \{0, 1, \dots, N\}$, posons $x_{j,N} = \inf\left\{x : F(x) \geq \frac{j}{N}\right\}$ avec la convention habituelle $\inf(\emptyset) = +\infty$. Ainsi, $x_{0,N} = -\infty$, $x_{j,N} = F^{-1}(j/N)$ pour $1 \leq j < N$ et $x_{N,N}$ est un réel fini s'il existe $t \in \mathbb{R}$ tels que $F(t) = 1$, sinon $x_{N,N} = +\infty$. Pour N fixé, la suite finie $(x_{j,N})_{0 \leq j \leq N}$ est croissante. Pour être sûr de recouvrir dans tous les cas $\mathbb{R}$ par la réunion d'intervalles $[x_{j,N}, x_{j+1,N}[$, on pose $x_{N+1,N} := +\infty$. Il est commode dans ce qui suit de prolonger toutes les fonctions de répartition en leur attribuant la valeur 0 en $-\infty$ et la valeur 1 en $+\infty$. Notons que pour tout $N \geq 1$, $F(x_{N,N}) = 1 = F(x_{N+1,N}^-) = F(x_{N+1,N})$.

Avec ces notations, et compte-tenu de (2.5), on a pour tout $j = 0, 1, \dots, N$,

$$F(x_{j,N}^-) \leq \frac{j}{N} \leq F(x_{j,N}), \tag{2.6}$$

ce qui implique que pour $0 \leq j < N$,

$$F(x_{j,N}) + \frac{1}{N} \geq F(x_{j+1,N}^-). \tag{2.7}$$

Cette inégalité est aussi vérifiée trivialement pour $j = N$. Maintenant, définissons les intervalles

$$\Delta_{0,N} =]-\infty, x_{1,N}[, \quad \Delta_{j,N} = [x_{j,N}, x_{j+1,N}[, \quad 1 \leq j \leq N.$$

Lorsque l'intervalle $\Delta_{j,N}$ n'est pas vide, par croissance de F_n et de F sur cet intervalle, on a l'encadrement :

$$\forall x \in \Delta_{j,N}, \quad F_n(x_{j,N}) - F(x_{j+1,N}^-) \leq F_n(x) - F(x) \leq F_n(x_{j+1,N}^-) - F(x_{j,N}).$$

En utilisant (2.7), on en déduit facilement l'encadrement :

$$\forall x \in \Delta_{j,N}, \quad F_n(x_{j,N}) - F(x_{j,N}) - \frac{1}{N} \leq F_n(x) - F(x) \leq F_n(x_{j+1,N}^-) - F(x_{j+1,N}^-) + \frac{1}{N}.$$

Posons pour alléger les écritures

$$D_{n,j,N} := |F_n(x_{j,N}) - F(x_{j,N})|, \quad D'_{n,j,N} := \left|F_n(x_{j+1,N}^-) - F(x_{j+1,N}^-)\right|,$$

$$D_{n,N} := \max_{0 \leq j \leq N} \max(D_{n,j,N}, D'_{n,j,N}).$$

On déduit alors de l'encadrement précédent que

$$\forall j = 0, 1, \dots, N, \ \forall x \in \Delta_{j,N}, \quad |F_n(x) - F(x)| \leq \max(D_{n,j,N}, D'_{n,j,N}) + \frac{1}{N}.$$

Comme la réunion des $\Delta_{j,N}$, $0 \leq j \leq N$ est $\mathbb{R}$,

$$\forall x \in \mathbb{R}, \quad |F_n(x) - F(x)| \leq D_{n,N} + \frac{1}{N}.$$

d'où

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq D_{n,N} + \frac{1}{N}.$$

Dans cette inégalité, le membre de gauche est un supremum d'un ensemble non dénombrable de variables aléatoires positives. Il n'est donc pas acquis qu'il soit mesurable. Par contre, le majorant est mesurable comme max d'un nombre fini de variables aléatoires positives. D'après (2.3), chacune de ces variables tend p.s. vers 0 quand n tend vers l'infini [1] et il en va de même pour $D_{n,N}$ (rappelons que N est fixé pour l'instant). Il existe donc un évènement $\Omega_N \in \mathcal{F}$, de probabilité 1 et tel que :

$$\forall \omega \in \Omega_N, \quad \limsup_{n\to+\infty} \sup_{x\in\mathbb{R}} |F_n(\omega,x) - F(x)| \leq \frac{1}{N}.$$

Cette inégalité étant vraie pour tout $N \in \mathbb{N}^*$, notons Ω' l'intersection pour $N \in \mathbb{N}^*$ des Ω_N. Cette intersection est encore un évènement de probabilité 1 *qui ne dépend pas de* N et l'on a

$$\forall \omega \in \Omega', \forall N \in \mathbb{N}^*, \quad \limsup_{n\to+\infty} \sup_{x\in\mathbb{R}} |F_n(\omega,x) - F(x)| \leq \frac{1}{N}.$$

Comme le premier membre de cette inégalité ne dépend pas de N, on en déduit qu'il est nul. En conclusion, nous avons montré l'existence d'un $\Omega' \in \mathcal{F}$, de probabilité 1 tel que pour tout $\omega \in \Omega'$, $\sup_{x\in\mathbb{R}} |F_n(\omega,x) - F(x)|$ converge vers 0 quand n tend vers l'infini, ce qui à un détail près est la conclusion annoncée dans l'énoncé du théorème. Ce détail est que la convergence presque-sûre est définie pour des suites de variables aléatoires, or on ne sait pas encore si $\sup_{x\in\mathbb{R}} |F_n - F|$ est une variable aléatoire. Ce point fait l'objet du lemme suivant. □

Lemme 2.2. *La f.d.r. empirique F_n bâtie sur les variables aléatoires $X_1, \dots, X_n$, de même loi X_i de f.d.r. F vérifie :*

$$\forall \omega \in \Omega, \quad \sup_{x\in\mathbb{R}} |F_n(\omega,x) - F(x)| = \sup_{x\in\mathbb{Q}} |F_n(\omega,x) - F(x)|. \tag{2.8}$$

En conséquence, $\|F_n - F\|_\infty$ est une variable aléatoire réelle.

Preuve du lemme 2.2. Fixons ω quelconque dans Ω et notons pour alléger

$$\tau := \|F_n(\omega,.) - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(\omega,x) - F(x)|.$$

Comme $F_n(\omega,x)$ et $F(x)$ sont toujours deux réels de $[0,1]$, ce supremum τ est fini. Alors pour tout $\varepsilon > 0$, il existe un x_ε tel que

$$\tau - \varepsilon < |F_n(\omega,x_\varepsilon) - F(x_\varepsilon)| \leq \tau.$$

Les f.d.r. $F_n(\omega,.)$ et F étant continues à droite au point x_ε, la valeur absolue de leur différence l'est aussi. Il existe donc un $\delta > 0$ tel que

$$\forall t \in]x_\varepsilon, x_\varepsilon + \delta[, \quad |F_n(\omega,t) - F(t)| > |F_n(\omega,x_\varepsilon) - F(x_\varepsilon)| - \varepsilon$$

Dans l'intervalle $]x_\varepsilon, x_\varepsilon + \delta[$, il y a au moins un nombre *rationnel* t. Ce rationnel vérifiant l'inégalité ci-dessus, on en déduit

$$\sup_{r\in\mathbb{Q}} |F_n(\omega,r) - F(r)| > \tau - 2\varepsilon.$$

1. Si $x_{j,N} = +\infty$ (cas $j = N+1$ et peut-être $j = N$), cela ne résulte pas de (2.3), mais c'est évident directement puisqu'alors F_n et F valent 1 en $x_{j,N}$.

Le premier membre ne dépendant pas de ε et $\varepsilon > 0$ étant arbitraire, on en déduit :

$$\sup_{r\in\mathbb{Q}} |F_n(\omega, r) - F(r)| \geq \tau = \sup_{r\in\mathbb{R}} |F_n(\omega, r) - F(r)| .$$

Puisque ω était quelconque, ceci vaut pour tout $\omega \in \Omega$. L'inégalité dans l'autre sens est évidente, donc l'égalité (2.8) est démontrée pour tout $\omega \in \Omega$. Elle permet de voir l'application $\|F_n - F\|_\infty : \Omega \to \mathbb{R}$, $\omega \mapsto \|F_n(\omega, .) - F\|_\infty$ comme le sup d'une famille *dénombrable* de variables aléatoires réelles. Cette application $\|F_n - F\|_\infty$ hérite ainsi de la mesurabilité de ces v.a., c'est donc elle-même une variable aléatoire. □

Exercice 2.1. Montrez que si G et F sont deux fonctions de répartition [2],

$$\sup_{x\in\mathbb{R}} |G(x) - F(x)| = \sup_{x\in\mathbb{Q}} |G(x) - F(x)| .$$

◁

Exercice 2.2. On considère une fonction ψ, continue à droite, croissante et bornée sur $\mathbb{R}$. Montrer que si, pour tout t, $n^{-1}\sum_{j=1}^n \psi(X_j - t)$ converge presque sûrement vers $\mathbb{E}(\psi(X_1 - t))$, la convergence est presque sûrement uniforme en t. ◁

1.2. La mesure empirique. On appelle ainsi la mesure (aléatoire) P_n dont la fonction de répartition est définie en (2.1). C'est une somme pondérée de masses de Dirac aux observations X_j :

$$P_n = \frac{1}{n}\sum_{j=1}^n \delta_{X_j}. \tag{2.9}$$

Son rôle est important car la plupart du temps une statistique $T_n(X_1, \dots, X_n)$ peut s'exprimer comme la valeur en P_n d'une fonctionnelle définie sur un ensemble plus vaste de mesures de probabilité.

$$T_n(X_1, \dots, X_n) = T(P_n).$$

On notera indifféremment cette fonctionnelle $T(P_n)$ ou $T(F_n)$. La moyenne correspond à la fonctionnelle $T(F) = \int x dF(x)$ et la variance empirique à $T(F) = \int x^2 dF(x) - (\int x dF(x))^2$. On verra d'autres exemples plus loin.

Cette remarque prend tout son intérêt à la lumière des théorèmes de convergence de ce chapitre. Le théorème de Glivenko-Cantelli assure en effet que, pour une large classe de suite de variables aléatoires (voir la sous-section 1.3), la suite des mesures empiriques P_n converge faiblement vers P et ceci avec une probabilité 1. Ce résultat permet de démontrer facilement la convergence presque sûre de beaucoup de statistiques.

Exercice 2.3. *Médianes d'un échantillon observé*

1) Soit $x_1, \dots, x_n$ une suite finie de réels, pas forcément distincts. On appelle *médiane* de cette suite tout réel m tel qu'au moins la moitié des x_i vérifie $x_i \geq m$ et au moins la moitié des x_i vérifie $x_i \leq m$:

$$\operatorname{card}\Big\{i \in \{1, \dots, n\};\ x_i \geq m\Big\} \geq \frac{n}{2} \quad \text{et} \quad \operatorname{card}\Big\{i \in \{1, \dots, n\};\ x_i \leq m\Big\} \geq \frac{n}{2}. \tag{2.10}$$

2. C'est juste pour voir si vous avez bien lu la preuve du lemme.

Trouver toutes les médianes des suites finies suivantes :
a) 2 1 3 ; b) 4 2 5 7 ; c) 4 2 4 7.

2) Exprimer la double condition (2.10) à l'aide de la fonction de répartition F_n de la mesure $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ (c'est la fonction de répartition empirique construite sur l'échantillon observé $x_1, \dots, x_n$). Illustrer graphiquement la recherche de la ou des médianes à partir du graphe de F_n, dans le cas où les $x_1, \dots, x_n$ sont tous distincts, avec n impair puis avec n pair. $\triangleleft$

Exercice 2.4. *Médianes d'une loi*

Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé et X une v.a. réelle définie sur cet espace. On appelle *médiane de* X, tout réel m vérifiant :

$$P(X \geq m) \geq \frac{1}{2} \quad \text{et} \quad P(X \leq m) \geq \frac{1}{2}. \tag{2.11}$$

Cette double condition pouvant se réécrire $P_X([m, +\infty[) \geq \frac{1}{2}$ et $P_X(]-\infty, m]) \geq \frac{1}{2}$, il est clair que cette médiane si elle existe, ne dépend que de la loi de X sous P. Il serait donc plus correct de parler de médiane de la loi de X sous P. Mais nous commettrons l'abus de langage habituel (comme pour la f.d.r., l'espérance, les moments,...) en parlant de médiane(s) de X.

1) Montrer que l'ensemble des médianes de X est un intervalle[3] de $\mathbb{R}$. On vérifiera que si $m' < m''$ sont deux médianes de X, tout $m \in [m', m'']$ est une médiane de X.

2) Montrer que x est une médiane de X si et seulement si

$$P(X < x) \leq \frac{1}{2} \leq P(X \leq x), \tag{2.12}$$

ce qui peut s'écrire à l'aide de la f.d.r. F de X sous la forme

$$F(x-) \leq \frac{1}{2} \leq F(x). \tag{2.13}$$

En déduire que si F est continue sur $\mathbb{R}$, toute médiane m vérifie $F(m) = \frac{1}{2}$ et que s'il existe un intervalle médian $[a, b]$, c'est l'ensemble des solutions de l'équation $F(x) = \frac{1}{2}$.

3) Montrer que $F^{-1}(1/2)$ est toujours une médiane de X et que c'est la plus petite des médianes de X (F^{-1} désigne ici la fonction quantile définie par (2.4)). $\triangleleft$

Exercice 2.5. *Médiane et meilleure approximation L^1*

Montrer que si m est une médiane de la v.a. intégrable X,

$$\forall c \in \mathbb{R}, \quad \mathbb{E}\,|X - c| \geq \mathbb{E}\,|X - m|\,. \tag{2.14}$$

L'inégalité ci-dessus est triviale si $\mathbb{E}\,|X| = +\infty$. Montrer que dans ce cas $|X - c| - |X - m|$ est intégrable et vérifie $\mathbb{E}(|X - c| - |X - m|) \geq 0$ pour tout $c \in \mathbb{R}$. $\triangleleft$

Exercice 2.6. *Estimation de médiane*

Sur l'espace probabilisé $(\Omega, \mathcal{F}, P)$, soit $(X_i)_{i \geq 1}$ une suite de variables aléatoires réelles de même loi de f.d.r. F et vérifiant le théorème de Glivenko-Cantelli. On suppose que F^{-1}

3. L'ensemble vide est un intervalle, mais on verra à la question 3) que l'ensemble des médianes de X n'est jamais vide.

définie par (2.4) est continue au point $\frac{1}{2}$. Donner une condition suffisante sur F pour qu'il en soit ainsi.

On se propose d'estimer la plus petite médiane de la loi des X_i, donc $F^{-1}(\frac{1}{2})$ à partir de l'observation d'un échantillon de grande taille. On note F_n la f.d.r. empirique construite sur l'échantillon $X_1, \dots, X_n$ et on pose

$$M_n := F_n^{-1}\Big(\frac{1}{2}\Big) = \inf\Big\{t \in \mathbb{R};\ F_n(t) \geq \frac{1}{2}\Big\}.$$

Le but de cet exercice est de prouver que M_n converge presque-sûrement vers $F^{-1}(\frac{1}{2})$ lorsque n tend vers $+\infty$.

1) En écrivant $F(M_n) = F_n(M_n) + \big(F(M_n) - F_n(M_n)\big)$ et en comparant $F_n(M_n)$ avec $\frac{1}{2}$, montrer grâce au théorème de Glivenko-Cantelli, qu'il existe un $\Omega_1 \in \mathcal{F}$ tel que $P(\Omega_1) = 1$ et

$$\forall \omega \in \Omega_1,\ \forall \varepsilon > 0,\ \exists N_1 = N_1(\omega, \varepsilon),\ \forall n \geq N_1, \quad F(M_n(\omega)) > \frac{1}{2} - \varepsilon. \tag{2.15}$$

2) De même en écrivant $F(M_n - \varepsilon) = F_n(M_n - \varepsilon) + \big(F(M_n - \varepsilon) - F_n(M_n - \varepsilon)\big)$, pour tout $\varepsilon > 0$, et en comparant $F_n(M_n - \varepsilon)$ avec $\frac{1}{2}$, montrer qu'il existe un $\Omega_2 \in \mathcal{F}$ tel que $P(\Omega_2) = 1$ et

$$\forall \omega \in \Omega_2,\ \forall \varepsilon > 0,\ \exists N_2 = N_2(\omega, \varepsilon),\ \forall n \geq N_2, \quad F(M_n(\omega) - \varepsilon) < \frac{1}{2} + \varepsilon. \tag{2.16}$$

3) Déduire de ce qui précède l'existence d'un $\Omega_0 \in \mathcal{F}$ de probabilité 1 tel que

$$\forall \omega \in \Omega_0,\ \forall \varepsilon > 0,\ \exists N_0 = N_0(\omega, \varepsilon),\ \forall n \geq N_0, \quad F^{-1}\Big(\frac{1}{2} - \varepsilon\Big) < M_n(\omega) < F^{-1}\Big(\frac{1}{2} + \varepsilon\Big) + \varepsilon.$$

Conclure sur la convergence p.s. de M_n.

4) À quel endroit a-t-on vraiment besoin de la convergence *uniforme* presque sûre du théorème de Glivenko-Cantelli ? ◁

1.3. Loi des grands nombres, ergodicité. On rappelle que la loi forte des grands nombres de Kolmogorov-Khintchine s'énonce de la façon suivante.

Théorème 2.3. *Si les variables X_j sont i.i.d., leur moyenne arithmétique*

$$M_n = \overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$$

a une limite au sens presque sûr si et seulement si $\mathbb{E}\,|X_1| < \infty$. La limite est alors $\mathbb{E}(X_1)$.

Ce résultat établit que la moyenne empirique converge vers l'espérance. En l'appliquant aux indicatrices et en utilisant le théorème de Glivenko-Cantelli on obtient comme il a été dit plus haut la convergence faible presque sûre de la suite des mesures empiriques pour toute suite de variables i.i.d.

En fait la convergence presque sûre de la moyenne arithmétique est largement valable en dehors de l'hypothèse d'indépendance. Notamment on peut énoncer :

Proposition 2.4. *Soit une suite de variables aléatoires $(X_n)_{n\geq 1}$ de carré intégrable, ayant une même espérance $\mathbb{E}(X_n) = \mu$. On suppose de plus que $\sup_{k\geq 1} \operatorname{Var} X_k < +\infty$ et qu'il existe une constante c telle que :*

$$\forall n \in \mathbb{N}^*, \quad \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \leq cn. \tag{2.17}$$

Alors la moyenne arithmétique $M_n = n^{-1}\sum_{j=1}^{n} X_j$ converge presque sûrement vers μ.

Preuve. Notons $S_n := \sum_{k=1}^{n} X_k$. Puisque les X_k ont même espérance,

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{1}{n}\,\mathbb{E}\,S_n = \frac{1}{n}(n\,\mathbb{E}\,X_1) = \mathbb{E}\,X_1.$$

Posons $X'_k := X_k - \mathbb{E}\,X_k$ et $S'_n := \sum_{k=1}^{n} X'_k = S_n - n\,\mathbb{E}\,X_1$. Alors $S'_n/n = S_n/n - \mathbb{E}\,X_1$, donc la convergence p.s. de S_n/n vers $\mathbb{E}\,X_1$ équivaut à la convergence p.s. de S'_n/n vers 0. Notons que $\sum_{i=1}^{n} X'_i$ ayant même variance que $\sum_{i=1}^{n} X_i$, vérifie aussi (2.17). On ne perd donc pas de généralité en supposant désormais pour le confort d'écriture que $\mu = 0$. Il s'agit maintenant de prouver que

$$M_n := \frac{1}{n} S_n \xrightarrow[n\to+\infty]{\text{p.s.}} 0. \tag{2.18}$$

En introduisant la suite $(U_n)_{n\geq 1}$ définie par

$$U_n := \max_{n^2 \leq k < (n+1)^2} |M_k|$$

et en remarquant que pour tout $n \geq 1$, $|M_n| \leq U_{[\sqrt{n}\,]}$, il suffit pour établir (2.18) de prouver que $(U_n)_{n\geq 1}$ converge presque-sûrement vers zéro [4].

En remarquant que pour $n^2 \leq k < (n+1)^2$,

$$|M_k| = \frac{1}{k}\,|S_k| \leq \frac{1}{n^2}\Big(\,|S_{n^2}| + |S_k - S_{n^2}|\,\Big) \leq M_{n^2} + \frac{1}{n^2}\sum_{n^2 \leq j < (n+1)^2} |X_j| =: M_{n^2} + T_n,$$

d'où $U_n \leq M_{n^2} + T_n$, on est ramenés à prouver séparément la convergence presque-sûre vers 0 de M_{n^2} et de T_n.

Pour la première convergence, l'inégalité de Bienaymé-Tchebycheff et l'hypothèse (2.17) nous donnent :

$$\forall \varepsilon > 0, \quad P(|M_{n^2}| \geq \varepsilon) = P(|S_{n^2}| \geq n^2\varepsilon) \leq \frac{\operatorname{Var}(S_{n^2})}{\varepsilon^2 n^4} \leq \frac{c}{\varepsilon^2 n^2}.$$

On en déduit que

$$\forall \varepsilon > 0, \quad \sum_{n=1}^{+\infty} P(|M_{n^2}| \geq \varepsilon) < +\infty,$$

4. La notation $[\sqrt{n}\,]$ désigne ici la partie entière de $\sqrt{n}$, d'où $[\sqrt{n}\,]^2 \leq n < ([\sqrt{n}\,]+1)^2$. La suite d'entiers $([\sqrt{n}\,])_{n\geq 1}$ tend vers l'infini comme $\sqrt{n}$, mais avec des blocs de valeurs répétées de plus en plus longs. La suite $(U_{[\sqrt{n}\,]})_{n\geq 1}$ n'est donc pas à proprement parler une sous-suite de $(U_n)_{n\geq 1}$. À vous de compléter les détails de la réduction à la convergence vers 0 de $(U_n)_{n\geq 1}$.

ce qui implique

$$M_{n^2} \xrightarrow[n\to+\infty]{\text{p.s.}} 0. \tag{2.19}$$

Pour montrer la convergence de $(T_n)_{n\geq 1}$, on commence par majorer $\mathbb{E}\,T_n^2$ en utilisant l'inégalité de Cauchy-Schwarz dans $L^2(\Omega)$, $\mathbb{E}\,|X_i|\,|X_j| \leq (\mathbb{E}\,X_i^2)^{1/2}(\mathbb{E}\,X_j^2)^{1/2} \leq b$, où $b := \sup_{k\geq 1} \operatorname{Var} X_k = \sup_{k\geq 1} \mathbb{E}\,X_k^2$. On obtient ainsi :

$$\mathbb{E}\,T_n^2 = \frac{1}{n^4} \sum_{n^2\leq i,j<(n+1)^2} \mathbb{E}\,|X_i|\,|X_j| \leq \frac{((n+1)^2-n^2)^2}{n^4} b = \frac{(2n+1)^2 b}{n^4}.$$

L'inégalité de Markov avec moment d'ordre 2 nous donne alors

$$\forall \varepsilon > 0, \quad P(|T_n| > \varepsilon) \leq \frac{(2n+1)^2 b}{n^4 \varepsilon^2}.$$

Ce majorant étant le terme général d'une série convergente, on en déduit :

$$T_n \xrightarrow[n\to+\infty]{\text{p.s.}} 0. \tag{2.20}$$

□

Plus généralement, on aimerait savoir à quelle condition on est assuré de la convergence presque sûre de $n^{-1}\sum_{j=1}^n g(X_j)$ pour toute fonction g telle que $g(X_n)$ soit intégrable et d'intégrale indépendante de n. Pour énoncer cela de façon plus précise, on a besoin de la notion de processus stationnaire.

1.3.1. *Généralités sur les processus.*

Définition 2.5. *Soit T un ensemble, $(\Omega, \mathcal{A}, P)$ un espace probabilisé, et $(\Xi, \mathcal{X})$ un espace mesurable. Un processus stochastique à temps T et à valeurs dans Ξ est une famille $(X_t)_{t\in T}$ de variables aléatoires à valeurs dans Ξ.*

Dans ce paragraphe, $\Xi = \mathbb{R}$, mais il pourrait être pris égal à $\mathbb{N}$, ou encore $\mathbb{C}$ (on parle dans tous ces cas de processus *univarié*), ou $\mathbb{R}^p$, $p > 1$ (on parle alors de processus *multivarié*). Le temps T sera dans ce paragraphe $\mathbb{N}$ (processus à *temps discret*), mais il pourrait être $\mathbb{R}$ ou un intervalle de $\mathbb{R}$ (on parle alors de processus à *temps continu unidimensionnel*) ou $\mathbb{N}^p$, $p = 2, 3$ ou $\mathbb{Z}^p$, $p = 2, 3$ (ces processus à *temps multidimensionnel* servent pour le traitement d'image).

Dans le cas particulier qui nous intéresse, un processus $(X_n)_{n\in\mathbb{N}}$ est simplement une suite réelle aléatoire, c'est à dire une variable aléatoire dans un espace de suites réelles. On dit que $(X_n(\omega))_n$ (ω fixé) est une *trajectoire* ou une réalisation du processus. Pour $(X_n(\omega))_{1\leq n\leq N}$ on parle de *trajectoire de longueur N*.

1.3.2. *Rôle des lois fini-dimensionnelles.* Faisons d'abord quelques remarques de mesurabilité. Soit sur Ξ^T la *tribu cylindrique* engendrée par les cylindres finis $\prod_{t\in T} A_t$ de $\mathcal{X}^T$ tels que $A_t = \Xi$ sauf pour un nombre fini d'indices. On notera $\mathcal{X}^{\otimes T}$ cette tribu. L'espace $(\Xi^T, \mathcal{X}^{\otimes T})$ est aussi noté $(\Xi, \mathcal{X})^T$. Il est évident que l'application X qui à ω associe $(X_t(\omega))_{t\in T}$ est mesurable de $(\Omega, \mathcal{A})$ dans $(\Xi, \mathcal{X})^T$. Dans le cas particulier des processus à temps discret et à valeurs réelles, l'espace qu'on vient de considérer est $\mathbb{R}^{\mathbb{N}}$ (ou $\mathbb{C}^{\mathbb{N}}$) muni de la tribu engendrée par les cylindres.

Comme conséquence du théorème de Dynkin, la loi image P^X de P par cette application, c'est à dire la loi du processus, est caractérisée par les mesures qu'elle attribue aux cylindres finis, c'est à dire par les lois des vecteurs $(X_{t_1}, \dots, X_{t_k})$.

Définition 2.6. *On appelle lois fini-dimensionnelles du processus les lois des vecteurs*

$$(X_{t_1}, \dots, X_{t_k}) \quad k \in \mathbb{N}^* \quad t_1, \dots, t_k \in T.$$

Ces lois sont notées $P^X_{t_1,\dots,t_k}$.

On vient de voir que la famille des lois fini-dimensionnelles suffit à déterminer la loi du processus. On peut aussi se demander si, étant donnée une famille

$$\mathcal{P} = \{P_{t_1,\dots,t_k}, \quad k \in \mathbb{N}^*, t_1, \dots, t_k \in T\},$$

où $P_{t_1,\dots,t_k}$ est une loi sur $(\Xi^k, \mathcal{X}^{\otimes k})$, il existe un processus à valeurs dans $(\Xi, \mathcal{X})$ et à temps T dont ce soit précisément la famille lois fini-dimensionnelles. La réponse est dans le théorème de Kolmogorov que nous nous contenterons d'énoncer (la preuve est faite, par exemple, dans [6]).

Théorème 2.7 (Kolmogorov)**.** *Si Ξ est métrique complet et séparable. Supposons que la famille $\mathcal{P}$ vérifie la condition de compatibilité*

$$\begin{cases} P_{t_1,\dots,t_k,\dots,t_{k+h}}(A_1 \times \cdots \times A_k \times \Xi \times \cdots \times \Xi) = P_{t_1,\dots,t_k}(A_1 \times \cdots \times A_k) \\ P_{t_{\sigma(1)},\dots,t_{\sigma(k)}}(A_{\sigma(1)} \times \cdots \times A_{\sigma(k)}) = P_{t_1,\dots,t_k}(A_1 \times \cdots \times A_k) \end{cases}$$

pour tous k entier et $A_1, \dots, A_k$ dans $\mathcal{X}$ et pour toute permutation σ des indices $\{1, \dots, k\}$. Alors il existe un processus à temps T et à valeurs dans Ξ qui admet $\mathcal{P}$ comme famille de lois fini-dimensionnelles.

Un peu de vocabulaire. On dit que deux processus X et Y sont équivalents s'ils ont mêmes lois fini-dimensionnelles (donc s'ils induisent les deux mêmes lois P^X et P^Y).
On dit que Y est une modification de X si

$$\forall t \in T, \qquad X_t = Y_t, \text{ p.s.} \tag{2.21}$$

On note que, dans ce cas, X et Y sont équivalents. D'autre part, *si T est dénombrable*, ce qui est le cas pour les processus à temps discret, (2.21) équivaut à

$$\text{p.s.}, \qquad X_t = Y_t, \ \forall t \in T,$$

qui signifie que les deux processus sont presque sûrement les mêmes.

1.3.3. *Processus stationnaires.* Un processus est stationnaire si ses lois fini-dimensionnelles sont invariantes par translation du temps.

Définition 2.8. *La suite $(X_n)_{n\in\mathbb{N}}$ est dite stationnaire si pour tous $k, h \in \mathbb{N}^*$ et pour tous $n_1 < \cdots < n_k$, les lois de $(X_{n_1}, \dots, X_{n_k})$ et de $(X_{n_1+h}, \dots, X_{n_k+h})$ sont les mêmes. Cette définition se généralise aux processus indexés par $\mathbb{Z}$.*

Par exemple, une suite de variables i.i.d est stationnaire. C'est en cela que les processus stationnaires sont une généralisation des suites i.i.d. dans la direction de la non-indépendance.

On parle parfois plutôt de *stricte stationnarité* pour bien faire la différence avec la notion de stationnarité d'ordre deux qui ne porte que sur les moments jusqu'à l'ordre deux.

Pour alléger les écritures, désignons par $\mathcal{C}$ la tribu cylindrique de $\mathbb{R}^{\mathbb{N}}$, au lieu de $\mathrm{Bor}(\mathbb{R})^{\otimes\mathbb{N}}$ selon les notations vues à la sous-section 1.3.2. Soit $X = (X_k)_{k\in\mathbb{N}}$ un processus stationnaire défini sur $(\Omega, \mathcal{F}, P)$, les X_k étant à valeurs dans $\mathbb{R}$. Ainsi X est $\mathcal{F} - \mathcal{C}$ mesurable. Notons $\mathcal{F}_X$ la sous tribu de $\mathcal{F}$ engendrée par X, c'est à dire l'ensemble des évènements $A \in \mathcal{F}$, de la forme $A = X^{-1}(B) = \{X \in B\}$, pour $B \in \mathcal{C}$. Notons enfin S, le « shift » ou décalage sur $\mathbb{R}^{\mathbb{N}}$ défini par :

$$\mathsf{S} : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}, \quad x = (x_k)_{k\in\mathbb{N}} \longmapsto \mathsf{S}x := (x_{k+1})_{k\in\mathbb{N}}.$$

Exercice 2.7. Montrez que le processus $X = (X_k)_{k\in\mathbb{N}}$ est stationnaire si et seulement si X et $\mathsf{S}X$ considérés comme éléments aléatoires à valeurs dans $\mathbb{R}^{\mathbb{N}}$ ont même loi. ◁

Exercice 2.8. Les variables $(\varepsilon_j)_{j\in\mathbb{Z}}$ étant i.i.d. et de carré intégrable, les X_j sont définies par

$$X_j = \sum_{l=0}^{\infty} a^l \varepsilon_{j-l},$$

où $a \in]-1, 1[$. Étudiez les modes de convergence de la série et montrez que le processus $(X_n)_{n\in\mathbb{N}}$ est stationnaire. ◁

Exercice 2.9. Soient $(X_n)_{n\in\mathbb{Z}}$ un procesus stationnaire et $f : \mathbb{R}^{2k+1} \to \mathbb{R}$ une fonction mesurable (borélienne), montrer que le processus défini par $Y_n = f(X_{n-k}, \dots, X_n, \dots, X_{n+k})$ est stationnaire. Même question avec $Z_n = g(X_n, X_{n+1}, \dots) = g(\mathsf{S}^n X)$, où $g : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ est borélienne. ◁

1.3.4. *Ergodicité et loi des grands nombres.* Pour les processus stationnaires, on peut énoncer une loi des grands nombres qui n'est que la traduction du théorème ergodique dans le cas des processus stationnaires. Dans le cadre de ce cours, nous ne présenterons pas le théorème ergodique dans sa forme générale, nous contentant d'une approche minimaliste adaptée aux processus stationnaires. On peut trouver de multiples présentations du théorème ergodique dans la littérature, par exemple [5].

Définition 2.9. *Un évènement $A \in \mathcal{F}_X$ est dit invariant pour le processus $X = (X_k)_{k\in\mathbb{N}}$, s'il existe un $B \in \mathcal{C}$ tel que pour tout $n \in \mathbb{N}$, $A = \{\mathsf{S}^n X \in B\}$, où S^0 est l'identité sur $\mathbb{R}^{\mathbb{N}}$.*

Exercice 2.10. Montrer que l'événement $\{n^{-1}\sum_{j=1}^{n} X_j \quad \text{converge}\}$ est invariant. ◁

Proposition 2.10. *L'ensemble des évènements invariants pour le processus $X = (X_k)_{k\in\mathbb{N}}$ constitue une tribu $\mathcal{I}$, appelée tribu invariante du processus.*

La vérification est élémentaire et laissée au lecteur.

Le théorème ergodique de Birkhoff, dont la preuve se trouve par exemple dans [4], [5] ou [6], établit la convergence de la moyenne arithmétique vers une variable aléatoire $\mathcal{I}$-mesurable.

Théorème 2.11 (Birkhoff)**.** *Si le processus $(X_n)_{n\in\mathbb{N}}$ est stationnaire, pour toute fonction mesurable f telle que $\mathbb{E}\,|f(X_1)| < +\infty$ on a,*

$$\frac{1}{n}\sum_{j=1}^{n} f(X_j) \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbb{E}(f(X_1)\,|\,\mathcal{I}).$$

La loi des grands nombres énoncée pour des variables i.i.d. assure la convergence de la moyenne arithmétique vers l'espérance de la variable. Le théorème ergodique, faute d'hypothèses supplémentaires, se contente d'assurer la convergence vers une espérance conditionnelle. D'où l'intérêt de la notion d'ergodicité.

Définition 2.12. *Le processus stationnaire $(X_n)_{n\in\mathbb{N}}$ est dit ergodique si sa tribu des invariants $\mathcal{I}$ est triviale, c'est à dire ne contient que des événements de probabilité 0 ou 1.*

Le caractère trivial de la tribu $\mathcal{I}$ équivaut aussi au fait que toute variable $\mathcal{I}$-mesurable est presque sûrement constante. Comme $\mathbb{E}(f(X_1)\,|\,\mathcal{I})$ est $\mathcal{I}$-mesurable, elle est presque sûrement constante, donc presque sûrement égale à son espérance qui est $\mathbb{E}(f(X_1))$ et on peut énoncer la loi des grands nombres pour un processus stationnaire ergodique

Corollaire 2.13. *Si $(X_n)_{n\in\mathbb{N}}$ est stationnaire et ergodique, pour toute f telle que $\mathbb{E}\,|f(X_1)| < +\infty$, on a, lorsque n tend vers l'infini,*

$$\frac{1}{n}\sum_{j=1}^{n} f(X_j) \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbb{E}(f(X_1)). \tag{2.22}$$

Puis, en prenant $f = \mathbb{I}_{]-\infty,x]}$ on obtient la convergence presque sûre de la fonction de répartition empirique $F_n(x)$ vers $F(x)$. Donc, le théorème de Glivenko-Cantelli s'applique.

Corollaire 2.14. *Si $(X_n)_{n\in\mathbb{N}}$ est stationnaire et ergodique,*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to+\infty]{\text{p.s.}} 0.$$

Évidemment, la question importante est de savoir à quelle condition un processus stationnaire est ergodique. Commençons par l'exemple le plus simple.

Proposition 2.15. *Toute suite i.i.d. $X = (X_k)_{k\in\mathbb{N}}$ est un processus stationnaire ergodique.*

Preuve. On commence par noter que la tribu invariante $\mathcal{I}$ associée à X est une sous-tribu de la « tribu de queue » ou « tribu asymptotique » associée à X et définie comme $\cap_{n\geq 1}\mathcal{F}_{\mathsf{S}^n X}$. Or la loi du zéro-un de Kolmogorov nous dit que tout élément de la tribu asymptotique d'une suite i.i.d. a pour probabilité 0 ou 1. □

Exercice 2.11. Soit $X = (X_n)_{n\in\mathbb{N}}$ un processus stationnaire ergodique. Définissons le processus $Z = (X_n)_{n\in\mathbb{N}}$, par $Z_n = g(X_n, X_{n+1}, \dots) = g(\mathsf{S}^n X)$, où $g : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ est borélienne. Alors Z est lui aussi stationnaire ergodique. ◁

En combinant cet exercice avec la proposition 2.15, on peut construire une large classe de processus ergodiques.

Voici maintenant un résultat sur les processus gaussiens que l'on admettra (voir par exemple [6]).

Proposition 2.16. *Un processus gaussien stationnaire est ergodique si et seulement si la covariance* $\mathrm{Cov}(X_1, X_n)$ *tend vers zéro au sens de Cesàro quand* $n \to \infty$*.*

Comme exemple d'application de ce qui précède, on note que le processus de l'exercice 2.8 est ergodique (dans le cas où les ε_i sont gaussiennes). Donc, pour ce processus, on a (2.22) pour toute f telle que $\mathbb{E}\,|f(X_1)| < +\infty$.

Exemple 2.1. Un processus régulier d'innovation gaussienne est ergodique (consulter un cours de prévision). En particulier, un processus ARMA de représentation

$$X_n + \sum_{j=1}^{p} a_j X_{n-j} = \varepsilon_n + \sum_{j=1}^{q} b_j \varepsilon_{n-j} \qquad \forall n \in \mathbb{Z}, \tag{2.23}$$

où $A(z) = 1 + \sum_{j=1}^{p} a_j z^j$ n'a pas de racine de module 1 et où le bruit ε_n est gaussien, est ergodique. Comme cas particulier on retrouve le processus de l'exercice 2.8. ◁

2. Théorème central limite. Convergence du Processus empirique

2.1. Variables i.i.d. C'est le contexte habituel où sont énoncés les théorèmes limites.

2.1.1. *Théorème central limite pour les suites i.i.d.* Nous l'énonçons sans le démontrer, sous sa forme multivariée.

Théorème 2.17 (Théorème central limite)**.** *Soit* $(X_j)_{j\in\mathbb{N}}$ *une suite de vecteurs aléatoires de* $\mathbb{R}^d$*, i.i.d., de carré intégrable (*$\mathbb{E}\,\|X_1\|^2 < +\infty$*). La suite* $(S_n^*)_{n\geq 1}$ *définie par*

$$S_n^* := \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j - \mathbb{E}(X_j))$$

converge en loi lorsque n *tend vers l'infini vers un vecteur gaussien centré ayant même matrice de covariance que* X_1*.*

Si la loi des X_j est suffisamment régulière, on obtient une convergence en variation.

Théorème 2.18. *Soit* $(X_n)_{n\geq 1}$ *une suite de vecteurs aléatoires de* $\mathbb{R}^d$*, i.i.d., de carré intégrable, ayant une fonction caractéristique intégrable. Lorsque* n *tend vers l'infini, la loi de* $n^{-1/2}\sum_{j=1}^{n}(X_j - \mathbb{E}(X_j))$ *converge en variation vers la loi gaussienne indiquée dans le théorème 2.17.*

Preuve. Pour simplifier on suppose que la dimension est $d = 1$ et sans perdre de généralité on se ramène au cas où $\mathbb{E}(X_1) = 0$ et $\mathrm{Var}(X_1) = 1$. Comme la fonction caractéristique ψ de X_n est intégrable, la loi de cette variable a une densité qui s'écrit

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-itx} \psi(t)\,\mathrm{d}t.$$

La somme normalisée $n^{-1/2}\sum_{j=1}^n X_j$ a comme densité

$$f_n(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-itx}\left(\psi\left(\frac{t}{\sqrt{n}}\right)\right)^n \mathrm{d}t. \tag{2.24}$$

D'après le théorème 1.7, il suffit de prouver la convergence presque partout de f_n vers la densité gaussienne ϕ qui peut s'écrire elle aussi comme transformée de Fourier inverse :

$$\phi(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-itx}\mathrm{e}^{-t^2/2}\,\mathrm{d}t.$$

Comme $\mathbb{E}\,X_1 = 0$ et $\mathbb{E}\,X_1^2 = 1$, on a au voisinage de 0 le développement limité

$$\psi(u) = 1 - \frac{u^2}{2} + o(u^2).$$

Il existe donc un $\varepsilon \in]0, 2[$ tel que :

$$\forall u \in [-\varepsilon, \varepsilon], \quad |\psi(u)| \leq 1 - \frac{u^2}{4}. \tag{2.25}$$

Découpant maintenant l'intégrale dans (2.24) en $\int_{|t|\leq\varepsilon\sqrt{n}} + \int_{|t|>\varepsilon\sqrt{n}}$, on va montrer que ces deux intégrales convergent respectivement vers $\phi(x)$ et vers 0 quand n tend vers l'infini.

La première intégrale s'écrit $\int_{\mathbb{R}} g_n(t)\,\mathrm{d}t$, où l'on a posé

$$g_n(t) = \mathrm{e}^{-itx}\left(\psi\left(\frac{t}{\sqrt{n}}\right)\right)^n \mathbb{I}_{[0,\varepsilon\sqrt{n}]}(|t|).$$

On remarque d'abord que pour tout $t \in \mathbb{R}$,

$$g_n(t) \xrightarrow[n\to+\infty]{} \mathrm{e}^{-itx}\mathrm{e}^{-t^2/2}.$$

Pour pouvoir appliquer le théorème de convergence dominée, on cherche un majorant de $|g_n|$. Grâce à (2.25), on a

$$\forall t \in [-\varepsilon\sqrt{n}, \varepsilon\sqrt{n}], \quad |g_n(t)| \leq \left(1 - \frac{t^2}{4n}\right)^n,$$

en rappelant que $\varepsilon < 2$, donc ici $1 - \frac{t^2}{4n} > 0$. En utilisant l'inégalité de concavité $\ln(1+v) \leq v$ valable pour tout $v > -1$, avec $v = -\frac{t^2}{4n}$, il vient :

$$\forall t \in [-\varepsilon\sqrt{n}, \varepsilon\sqrt{n}], \quad |g_n(t)| \leq \mathrm{e}^{-t^2/4}.$$

Ce majorant est aussi valable pour $|t| > \varepsilon\sqrt{n}$, puisqu'alors $|g_n(t)| = 0$. Par convergence dominée, on a donc

$$\frac{1}{2\pi}\int_{\mathbb{R}} g_n(t)\,\mathrm{d}t \xrightarrow[n\to+\infty]{} \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-itx}\mathrm{e}^{-t^2/2}\,\mathrm{d}t = \phi(x)$$

et cette convergence a lieu pour tout $x \in \mathbb{R}$.

Pour prouver la convergence vers 0 de l'intégrale $\int_{|t|>\varepsilon\sqrt{n}}$, nous allons exploiter le lemme suivant dont la preuve est légèrement différée.

Lemme 2.19. *Si ψ est la fonction caractéristique d'une loi à densité, $|\psi(u)|$ est* strictement *inférieur à* 1 *pour tout $u \in \mathbb{R}^*$.*

Comme ψ est la fonction caractéristique d'une loi à densité, le théorème de Riemann-Lebesgue nous assure que $\psi(u)$ tend vers 0 quant u tend vers $\pm\infty$. On peut donc trouver un réel $a > \varepsilon$ tel que pour tout u hors de l'intervalle $[-a, a]$, $|\psi(u)| \leq 1/2$. Ensuite, par continuité de ψ sur le compact $[-a, -\varepsilon] \cup [\varepsilon, a]$, $|\psi|$ atteint son maximum M sur ce compact et en raison du lemme 2.19, $M < 1$. En posant $c = \max(M, 1/2)$, on a donc

$$\sup_{|u| \geq \varepsilon} |\psi(u)| \leq c < 1.$$

On en déduit la majoration

$$\left| \int_{|t| > \varepsilon\sqrt{n}} \mathrm{e}^{-itx} \left(\psi \left(\frac{t}{\sqrt{n}} \right) \right)^n \mathrm{d}t \right| \leq \sqrt{n} c^{n-1} \int_{\mathbb{R}} |\psi(u)| \, \mathrm{d}u,$$

en rappelant que ψ est intégrable sur $\mathbb{R}$. Comme $0 < c < 1$, ce majorant tend vers 0 quand n tend vers l'infini et ceci est valable pour tout $x \in \mathbb{R}$. □

Preuve du lemme 2.19. On raisonne par l'absurde en supposant l'existence d'un réel $s \neq 0$ tel que $|\psi(s)| = 1$. Alors il existe un réel b tel que

$$\psi(s) = \mathbb{E} \exp(isX) = \mathrm{e}^{ib}, \quad \text{d'où} \quad \mathbb{E} \exp(i(sX - b)) = 1.$$

Ainsi la variable aléatoire $Y = sX - b$ est à densité (parce que X l'est et $s \neq 0$) et vérifie $\mathbb{E} \exp(iY) = 1$, donc aussi $\mathbb{E} \cos Y = 1$. Ceci implique que toute la masse de la loi de Y est portée par l'ensemble $2\pi\mathbb{Z} = \{x \in \mathbb{R};\ \cos x = 1\}$. En effet supposons qu'il existe un $c < 1$ tel que $P(\cos Y \leq c) > 0$, alors

$$\mathbb{E} \cos Y \leq cP(\cos Y \leq c) + 1 \times P(\cos Y > c) < 1.$$

Donc pour tout $c < 1$, $P(\cos Y \leq c) = 0$, d'où $P(\cos Y = 1) = 1$. Ainsi $P(Y \in 2\pi\mathbb{Z}) = 1$ et comme $2\pi\mathbb{Z}$ est dénombrable, il existe au moins un $k \in \mathbb{Z}$ tel que $P(Y = 2k\pi) > 0$, ce qui est impossible puisque Y est à densité. Par conséquent, il ne peut exister de $s \neq 0$ tel que $|\psi(s)| = 1$. □

Remarquons au passage que dans la preuve du lemme l'hypothèse d'existence d'une densité pour la loi de X sert seulement à montrer que pour $s \neq 0$, la loi de $Y = sX - b$ n'a pas de masse ponctuelle ($\forall y \in \mathbb{R},\ P(Y = y) = 0$). Il est clair que le même argument vaut plus généralement si la loi de X n'a pas de masse ponctuelle. Le lemme est donc vrai dès que la loi de X est *continue*, c'est-à-dire à f.d.r. continue.

2.1.2. *Le processus empirique en variables i.i.d.* Une application particulièrement intéressante du T.C.L. concerne le processus empirique. Pour simplifier supposons les variables X_j à valeurs dans $\mathbb{R}$.

Définition 2.20. *On appelle processus empirique le processus indexé par* $\mathbb{R}$

$$\tilde{F}_n(x) := \sqrt{n} \left(F_n(x) - F(x) \right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\mathbb{I}_{]-\infty, x]}(X_j) - F(x) \right).$$

Il s'agit à proprement parler d'une suite (en n) de fonctions aléatoires de la variable x, donc de processus indexés par $\mathbb{R}$. Si les variables X_j sont i.i.d., non seulement la suite $\sqrt{n}(F_n(x) - F(x))$ converge, à x fixé, vers une variable gaussienne centrée et de variance $F(x)(1-F(x))$, mais on a le résultat plus riche suivant, directement issu du théorème 2.17 (prouvez le). Considérons un processus gaussien $W(x)$ à temps $x \in \mathbb{R}$, centré, de covariance

$$\mathrm{Cov}(W(s), W(t)) = F(\min(s,t)) - F(s)F(t),$$

alors la suite de processus $\sqrt{n}(F_n(x) - F(x))$ converge vers $W(x)$ au sens que les lois fini-dimensionnelles du premier convergent vers celles du second.

En réalité, on a un résultat encore plus intéressant pour les applications statistiques si on considère le processus empirique comme une suite de fonctions aléatoires, c'est à dire une suite d'éléments aléatoires d'un espace de fonctions. Son énoncé suppose un petit investissement topologique. Le processus empirique est une suite de fonctions aléatoires càdlàg qui s'annulent en $\pm\infty$. Appelons $D_0(\mathbb{R})$ l'ensemble des fonctions ayant ces propriétés. Munissons le de la topologie uniforme.

Théorème 2.21. *Si les variables X_j sont i.i.d. de fonction de répartition F, la suite $\sqrt{n}(F_n(x) - F(x))$ d'éléments aléatoires de $D_0(\mathbb{R})$ muni de la topologie uniforme converge en loi vers $W(x)$.*

Nous ne démontrerons pas ce théorème ici (voir par exemple [3]). Il implique la convergence en loi de $g(\sqrt{n}(F_n(.) - F(.)))$ vers $g(W(.))$ pour toute g continue pour la topologie uniforme. Notamment

$$\sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to+\infty]{\text{loi}} \sup_{x\in\mathbb{R}} |W(x)|,$$

ce qui donne la loi limite de la statistique de Kolmogorov-Smirnov, utilisée tester que la loi marginale des variables X_j est F. Une autre statistique classique dans le problème de test d'adéquation à une loi donnée est celle de Cramér-von Mises qui s'écrit :

$$\left\|\tilde{F}_n\right\|_2^2 = n \int_{\mathbb{R}} (F_n(x) - F(x))^2 \,\mathrm{d}x.$$

On aimerait pouvoir dire que sa loi limite est celle de $\|W\|_2^2 = \int_{\mathbb{R}} W(x)^2 \,\mathrm{d}x$, mais cela ne résulte pas en général du théorème 2.21, puisque $g \mapsto \|g\|_2^2$ n'est pas une fonctionnelle continue sur $D_0(\mathbb{R})$ (elle n'est même pas définie sur tout l'espace). On peut obtenir cette convergence dans le cas où la loi F est à support borné (exercice). Mais dans le cas général, il est préférable d'utiliser directement la topologie de $L^2(\mathbb{R})$. En fait, le résultat optimal s'obtient en utilisant le théorème limite central dans l'espace de Hilbert séparable $L^2(\mathbb{R})$. L'énoncé de ce théorème est le même qu'en dimension finie.

Théorème 2.22. *Soit $(Y_i)_{i\geq 1}$ une suite i.i.d. d'éléments aléatoires d'un espace de Hilbert séparable H, centrés. Alors*

$$S_n^* := n^{-1/2} \sum_{i=1}^{n} Y_i \xrightarrow[n\to+\infty]{\text{loi}} Z$$

si et seulement si $\mathbb{E}\,\|Y_1\|_H^2 < +\infty$. *Dans le cas où cette convergence a lieu, l'élément aléatoire limite* Z *est gaussien, de même structure de covariance que* Y_1.

On dit que Y_1 est centré si pour tout $h \in H$, $\mathbb{E}\,\langle h, Y_1\rangle = 0$. L'élément aléatoire Z de H est dit gaussien si pour tout $h \in H$, la variable aléatoire réelle $\langle h, Z\rangle$ est gaussienne. La covariance de Z est la forme bilinéaire continue définie sur H^2 par $\mathrm{Cov}(Z)(h_1, h_2) = \mathrm{Cov}(\langle h_1, Z\rangle, \langle h_2, Z\rangle)$. On pourra trouver ce théorème dans [20]. On l'applique ici avec $H = L^2(\mathbb{R})$ et

$$Y_i : \mathbb{R} \to \mathbb{R}, \quad x \longmapsto \mathbb{I}_{]-\infty,x]}(X_i) - F(x) = \mathbb{I}_{[X_i,+\infty[}(x) - F(x), \tag{2.26}$$

en notant que $S_n^* = \tilde{F}_n$. Il faut bien sûr commencer par regarder à quelles conditions Y_i est bien un élément aléatoire de $L^2(\mathbb{R})$. Comme

$$\int_{\mathbb{R}} Y_i(x)^2\,\mathrm{d}x = \int_{-\infty}^{X_i} F(x)^2\,\mathrm{d}x + \int_{X_i}^{+\infty} (1 - F(x))^2\,\mathrm{d}x,$$

on voit que Y_i est élément aléatoire de $L^2(\mathbb{R})$ si et seulement si :

$$\int_{-\infty}^{0} F(x)^2\,\mathrm{d}x < +\infty \quad \text{et} \quad \int_{0}^{+\infty} (1 - F(x))^2\,\mathrm{d}x < +\infty. \tag{2.27}$$

Regardons maintenant ce que donne la condition nécessaire et suffisante du théorème 2.22 :

$$\mathbb{E}\,\|Y_1\|_{L^2(\mathbb{R})}^2 = \mathbb{E}\int_{\mathbb{R}} Y_i(x)^2\,\mathrm{d}x = \int_{\mathbb{R}} \mathbb{E}(Y_i(x)^2)\,\mathrm{d}x = \int_{\mathbb{R}} \mathrm{Var}(Y_1(x))\,\mathrm{d}x = \int_{\mathbb{R}} F(x)(1 - F(x))\,\mathrm{d}x.$$

L'intégrabilité sur $\mathbb{R}$ de $F(1-F)$ équivaut à celles de F au voisinage de $-\infty$ et de $1-F$ au voisinage de $+\infty$ et ces deux conditions impliquent les conditions 2.27, puisque $F(x)$ et $1 - F(x)$ étant dans $[0,1]$ majorent leur carré. On peut finalement énoncer.

Corollaire 2.23. *Le processus empirique* $\tilde{F}_n$ *converge en loi dans* $L^2(\mathbb{R})$ *vers* W *si et seulement si*

$$\int_{\mathbb{R}} F(x)(1 - F(x))\,\mathrm{d}x < +\infty. \tag{2.28}$$

Sous cette condition, on a la convergence en loi de la statistique de Cramér-von Mises :

$$n\int_{\mathbb{R}} (F_n(x) - F(x))^2\,\mathrm{d}x \xrightarrow[n\to+\infty]{\text{loi}} \int_{\mathbb{R}} W^2(x)\,\mathrm{d}x.$$

Exercice 2.12. Z étant la limite de S_n^* obtenue lorsque les Y_i sont définis en (2.26), vérifiez que W et Z ont même loi. On admettra que la loi d'un élément aléatoire gaussien centré de $L^2(\mathbb{R})$ est déterminée uniquement par sa covariance. ◁

Exercice 2.13. Montrez que la condition (2.28) équivaut à $\mathbb{E}\,|X_1| < +\infty$. ◁

2.2. Variables indépendantes mais non équi-distribuées. Nous rappelons ici une version du théorème central limite pour des variables aléatoires qui ne sont pas de même loi. On se donne un tableau triangulaire

$$X_{n,1}, \dots, X_{n,k_n} \qquad n = 1, \dots \quad k_n \in \mathbb{N}$$

de variables aléatoires de L^2. On pose $D_n^2 = \sum_{j=1}^{k_n} \mathbb{E}\,(X_{n,j} - \mathbb{E}(X_{n,j}))^2$ et on suppose satisfaites les conditions $(\mathcal{L})$:

($\mathcal{L}$.1) à n fixé, les variables $X_{n,1}, \ldots, X_{n,k_n}$ sont indépendantes ;

($\mathcal{L}$.2) pour tout $t > 0$,

$$\frac{1}{D_n^2} \sum_{j=1}^{k_n} \mathbb{E}\left((X_{n,j} - \mathbb{E}(X_{n,j}))^2 \, \mathbb{I}_{|X_{n,j} - \mathbb{E}(X_{n,j}| > tD_n} \right) \xrightarrow[n \to +\infty]{} 0. \tag{2.29}$$

Théorème 2.24 (Lindeberg-Feller)**.** *Sous les conditions* ($\mathcal{L}$)*, on a*

$$\frac{1}{D_n} \sum_{i=1}^{k_n} (X_{n,j} - \mathbb{E}(X_{n,j})) \xrightarrow[n \to +\infty]{\text{loi}} \mathfrak{N}(0,1). \tag{2.30}$$

Nous ne démontrerons pas ce théorème dont la preuve se trouve dans la plupart des cours de maîtrise.

Remarque 2.1. Si l'on se place dans le contexte du théorème central limite classique, $k_n = n$, $X_{n,j} = Y_j$ où (Y_j) est une suite de variables aléatoires L^2 et i.i.d. et $D_n^2 = n\sigma^2 = \mathrm{Var}(Y_1)$. La condition (2.29) devient

$$\mathbb{E}\left((Y_1 - \mathbb{E}(Y_1))^2 \, \mathbb{I}_{|Y_1 - \mathbb{E}(Y_1)| > t\sigma\sqrt{n}} \right) \xrightarrow[n \to +\infty]{} 0,$$

qui est vérifiée. ◁

Remarque 2.2. La condition (2.29) signifie que les variables re-normalisées par D_n sont en quelque sorte « uniformément petites ». Plus exactement :

$$\sup_{1 \le j \le k_n} \frac{1}{D_n^2} \mathbb{E}\left(X_{n,j} - \mathbb{E}(X_{n,j}) \right)^2 \xrightarrow[n \to +\infty]{} 0. \tag{2.31}$$

◁

Il est évident que ce théorème, appliqué aux indicatrices $\mathbb{I}_{]-\infty,s]}(X_j)$, doit pouvoir conduire à une convergence du processus empirique, à condition que les lois des variables varient suffisamment peu pour qu'on puisse donner un sens à $F(x)$. Nous ne développerons pas ici cette idée.

2.3. Variables non indépendantes. Ce qu'il advient le théorème central limite lorsque les variables X_j ne sont pas indépendantes est un des champs de recherches les plus actifs depuis une cinquantaine d'années. En gros on peut distinguer deux larges types de dépendances : la faible dépendance et la forte dépendance, la frontière entre les deux n'étant pas clairement définie. La première recouvre les situations où le théorème central limite reste valable, bien qu'avec une variance limite modifiée, la seconde les cas où la normalité asymptotique du processus empirique est conservée, mais avec une normalisation qui n'est plus $\sqrt{n}$. Nous ne donnerons ici que deux exemples.

2.3.1. *Variables faiblement dépendantes.* On va supposer ici que le processus $(X_n)_{n \ge 1}$ est stationnaire, mais que les variables ne sont plus indépendantes. En réalité, si la dépendance est assez faible, on conserve un théorème central limite. Le théorème qui suit énonce un T.C.L. pour les processus linéaires.

Théorème 2.25. *Soit* $(\varepsilon_n)_{n\in\mathbb{Z}}$ *une suite i.i.d. centrée de* L^2 *et soit* $(c_n)_{n\in\mathbb{Z}}$ *une suite sommable. Considérons le processus* $X_n = \sum_{k=-\infty}^{\infty} c_k \varepsilon_{n-k}$. *Supposons que* $\operatorname{Var} \varepsilon_n > 0$. *Lorsque* n *tend vers l'infini,*

$$n^{1/2}\,\overline{X}_n \xrightarrow[n\to+\infty]{\text{loi}} \mathfrak{N}\left(0, \sum_{-\infty}^{\infty} \Gamma(j)\right),$$

où $\Gamma(j) = \operatorname{Cov}(X_0, X_j)$.

Regardons maintenant ce qui se passe pour le processus empirique et pour cela commençons par évaluer sa covariance.

$$\begin{aligned}
&\operatorname{Cov}\left(\sqrt{n}(F_n(s)-F(s)), (\sqrt{n}(F_n(t)-F(t))\right) = \frac{1}{n}\sum_{j,k=1}^{n} \operatorname{Cov}\left(\mathbb{I}_{]-\infty,s]}(X_j), \mathbb{I}_{]-\infty,t]}(X_k)\right)\\
= \;& \operatorname{Cov}(\mathbb{I}_{]-\infty,s]}(X_1), \mathbb{I}_{]-\infty,t]}(X_1)) +\\
+ \;& \sum_{l=1}^{n-1}(1-\frac{l}{n})\left(\operatorname{Cov}(\mathbb{I}_{]-\infty,s]}(X_1), \mathbb{I}_{]-\infty,t]}(X_{1+l})) + \operatorname{Cov}(\mathbb{I}_{]-\infty,t]}(X_1), \mathbb{I}_{]-\infty,s]}(X_{1+l}))\right).
\end{aligned}$$

Supposons que la série de terme général $P(X_1 \le s, X_h \le t) - F(s)F(t)$ converge absolument. Alors, si le processus empirique converge en loi vers un processus gaussien Z, celui-ci a nécessairement la limite de la covariance trouvée, c'est à dire

$$\operatorname{Cov}(Z(s), Z(t)) = \sum_{h=-\infty}^{\infty} (P(X_1 \le s, X_h \le t) - F(s)F(t)) =: \Gamma(s,t), \tag{2.32}$$

ce qui, dans le cas de l'indépendance, redonne le résultat énoncé dans le théorème 2.21. Nombreuses sont les conditions que l'on peut mettre pour assurer la convergence du processus empirique vers un processus gaussien. Elles impliquent toutes la convergence de la série (2.32). Consulter par exemple [10] pour une liste assez exhaustive des résultats existants.

Nous donnons ici deux résultats, le premier concerne les processus linéaires dits «causaux»

$$X_n = \sum_{j=0}^{\infty} a_j \varepsilon_{n-j},$$

où $\sum |a_j| < \infty$ et où la suite $(\varepsilon_n)_n$ est i.i.d. centrée de variance 1. Pour de tels processus, pourvu que la loi de ε_0 soit assez régulière, le processus empirique a la limite attendue. Ce théorème est démontré dans [13].

Théorème 2.26. *Supposons que* $\mathbb{E}(\varepsilon_0^4) < \infty$ *et que pour tout* u

$$|\,\mathbb{E}\exp(iu\varepsilon_0)| \le \frac{C}{(1+|u|)^{\delta}} \tag{2.33}$$

où $1/2 < \delta \le 1$.

Alors,

$$\sum_{h=0}^{\infty} |P(X_1 \le s, X_h \le t) - F(s)F(t)| < \infty$$

et le processus empirique $\sqrt{n}\,(F_n(x) - F(x))$ *converge en loi dans* $D_0(\mathbb{R})$ *(muni de la topologie uniforme) vers un processus gaussien centré dont la covariance est donnée par (2.32).*

Remarquons que la condition (2.33) est une condition de régularité de la loi du bruit : elle implique, si F_ε désigne la fonction de répartition de cette loi

$$|F_\varepsilon(x) - F_\varepsilon(y)| < C|x-y| \quad \forall x, y \in \mathbb{R}.$$

Le théorème qui suit est démontré dans [3] (ch. 22). Il est basé sur la notion de ϕ-mélange qui s'énonce de la façon suivante : soit $(X_n)_{n\in\mathbb{Z}}$ un processus stationnaire. Notons $\mathcal{M}_{-\infty}^0$ et $\mathcal{M}_n^\infty$ les tribus engendrées respectivement par $(\ldots, X_{-1}, X_0)$ et $(X_n, X_{n+1}, \ldots)$. Le coefficient de ϕ-mélange est l'indice de dépendance défini par

$$\phi(n) = \sup\{|P(E_2|E_1) - P(E_2)|; E_1 \in \mathcal{M}_{-\infty}^0, E_2 \in \mathcal{M}_n^\infty, P(E_1) \neq 0\}.$$

Le processus est dit ϕ-mélangeant si ce coefficient tend vers zéro lorsque n tend vers l'infini.

Théorème 2.27. *Soit* $(X_n)_{n\in\mathbb{Z}}$ *un processus stationnaire ϕ-mélangeant. Supposons que* $\sum n^2\sqrt{\phi(n)} < \infty$ *et que la loi marginale a une fonction de répartition F continue. Alors la série de terme général* $P(X_1 \le s, X_h \le t) - F(s)F(t)$ *converge absolument et* $\sqrt{n}\,(F_n(x) - F(x))$ *converge en loi dans* $D_0(\mathbb{R})$ *vers un processus gaussien centré de covariance (2.32).*

2.3.2. *Variables fortement dépendantes.* Que se passe-t'il lorsque la série des covariances diverge absolument ? On n'a encore, dans ce cas de « dépendance forte », que des résultats partiels. On pourra consulter [10] pour un tour d'horizon. En ce qui concerne le processus empirique la normalité asymptotique est en général conservée, mais la normalisation en $\sqrt{n}$ est à remplacer par une normalisation d'ordre plus petit d'une part, et d'autre part la structure du processus limite est radicalement différente. Voici en particulier un résultat valable pour les processus gaussiens (voir [10] page 82).

Théorème 2.28. *Supposons que* $(X_n)_{n\geq 1}$ *est un processus stationnaire gaussien, centré, tel que*

$$\mathrm{Cov}(X_1, X_j) \sim j^{-\alpha} \qquad \text{si } j \to \infty$$

où $\alpha \in]0,1[$. *Alors, lorsque n tend vers l'infini,*

$$\sqrt{\frac{2n^\alpha}{(1-\alpha)(2-\alpha)}}(F_n(x) - F(x)) \xrightarrow[n\to+\infty]{\text{loi}} Z(x) = f(x)Z,$$

où Z est une variable gaussienne standard, $f(x)$ la densité gaussienne des variables X_j, et où la convergence a lieu dans l'espace $D_0(\mathbb{R})$ *muni de la topologie uniforme.*

Ce résultat appelle un commentaire : contrairement à ce qui se passe en faible dépendance, le processus gaussien limite $Z(x)$ est dégénéré dans la mesure où la variable x ne figure que dans le facteur déterministe $f(x)$ tandis que c'est la même variable gaussienne Z qui apparaît pour toutes les valeurs de x.

CHAPITRE 3

Méthodes du maximum de vraisemblance et M-estimateurs

1. Introduction

On appelle habituellement M-estimateur une statistique obtenue en minimisant [1] par rapport à t un critère du type

$$M_n(t, X_1, \ldots, X_n) = \sum_{j=1}^{n} m_t(X_j). \tag{3.1}$$

Si m_t est une fonction suffisamment régulière de t, l'estimateur s'obtient aussi comme solution [2] d'un système d'équations de type

$$\sum_{j=1}^{n} \psi_{t,k}(X_j) = 0, \quad k = 1, \ldots, q. \tag{3.2}$$

Sous des conditions convenables, la loi des grands nombres assure la convergence de $n^{-1}M_n(t, X_1, \ldots, X_n)$ vers $\mathbb{E}(m_t(X_1))$, ce qui laisse présager la convergence du M-estimateur vers la valeur θ de t, si elle existe, qui minimise cette espérance. Ainsi, lorsqu'on veut estimer un paramètre θ, une bonne procédure consiste à minimiser sur l'échantillon une fonction de t dont l'espérance est rendue minimale, sous la loi P_θ, en prenant $t = \theta$.

Exemple 3.1. *Le maximum de vraisemblance*

Les variables X_j sont i.i.d selon une loi P_θ, de densité f_θ par rapport à une mesure μ, la même pour tous les P_θ, et θ est un paramètre inconnu qu'il faut estimer. La méthode du maximum de vraisemblance consiste à maximiser en t la vraisemblance ou, lorsque la densité ne s'annule jamais, son logarithme

$$L_n(t, X_1, \ldots, X_n) = \sum_{j=1}^{n} \ln f_t(X_j). \tag{3.3}$$

La méthode du maximum de vraisemblance a été et reste très populaire car elle permet la construction effective d'estimateurs. D'autre part, comme on le verra, elle conduit la plupart du temps à des procédures efficaces. Elle peut évidemment être transposée en dehors du cas de la situation i.i.d, mais alors la forme canonique (3.3) est perdue, et la maximisation peut poser de gros problèmes de mise en oeuvre. ◁

1. Ou en maximisant, ce qui revient au même en changeant le signe du critère. Ainsi le M dans « M-estimateur » évoque aussi bien une minimisation ou une maximisation.

2. Dans ce cas, on parle aussi de Z-estimateur, où le Z évoque un zéro d'une fonction.

Exercice 3.1. Trouvez l'estimateur du maximum de vraisemblance de θ pour la loi uniforme sur $]0, \theta]$, pour une loi gaussienne de paramètres $\theta = (\mu, \sigma^2)$ et pour la loi double exponentielle centrée en θ de densité $\frac{1}{2}\exp(-|x-\theta|)$. ◁

Exercice 3.2. On reprend l'exercice 2.8. Le paramètre a est inconnu. Vous n'observez que $(X_1, \dots, X_n)$. Estimez le paramètre a par la méthode du maximum de vraisemblance. Inutile de chercher à terminer les calculs… ◁

Exemple 3.2. *Estimation des moindres carrés.* On appelle ainsi toute procédure qui consiste à minimiser, par rapport à t,

$$\sum_{j=1}^{n} (X_j - f_t(Z_j))^2. \tag{3.4}$$

Suivant les situations, les Z_j sont considérés comme aléatoires ou non. Dans la seconde situation l'échantillon est constitué des n couples (X_j, Z_j). Cet exemple recouvre notamment l'estimation des paramètres d'une régression linéaire ou non linéaire : on suppose que, les Z_j étant des quantités connues éventuellement vectorielles et (ε_n) une suite de variables aléatoires centrées,

$$X_j = f_\theta(Z_j) + \varepsilon_j, \quad \forall j = 1, \dots, n.$$

On prend pour estimation de θ la valeur de t qui minimise (3.4). ◁

Exercice 3.3. Reprenez l'exercice 3.2. Vous constatez que la méthode du maximum de vraisemblance est proche d'une méthode des moindres carrés. Donnez la solution des moindres carrés. La variance des ε_j est en fait σ^2, nouveau paramètre inconnu. Comment l'estimeriez vous ? ◁

Exemple 3.3. *Estimateur du minimum de contraste.* La notion de minimum contraste est une généralisation de celle du maximum de vraisemblance et des moindres carrés à des fonctions $M_n(t, X_1, \dots, X_n)$ plus générales. On pourra consulter [7]. ◁

Reprenons (3.1). Un cas particulier très étudié est celui où on minimise (ou maximise)

$$M_n(t) = \sum_{j=1}^{n} m(X_j - t). \tag{3.5}$$

Les estimateurs définis par cette équation peuvent être considérés comme estimateurs d'un paramètre de position puisqu'ils ont la propriété « d'équivariance par translation ». Ceci signifie que lorsque les données X_i sont translatées d'une quantité fixe α, il en va de même pour l'estimateur : si $\hat{\theta}_n$ minimise (resp. maximise) M_n définie par (3.5), $\hat{\theta}_n + \alpha$ est évidemment minimiseur (resp. maximiseur) de $\sum_{j=1}^{n} m(X_j + \alpha - t)$. C'est le cas de l'estimateur du maximum de vraisemblance d'un paramètre de position : étant donnée une loi de densité f, les variables X_j sont indépendantes et distribuées selon la loi de densité $f(x-\theta)$. La méthode du maximum de vraisemblance conduit à maximiser en t la fonction $\sum_{j=1}^{n} \ln f(X_j - t)$.

Exercice 3.4. Quelles statistiques bien connues s'obtiennent en minimisant la fonction $M_n(t)$ lorsque $m(x) = x^2$? Lorsque $m(x) = |x|$? Qu'estiment-elles ? ◁

Lorsque m est dérivable de dérivée $m' = \psi$, minimiser ou maximiser $M_n(t)$ revient, dans le cas d'un paramètre uni-dimensionnel, à chercher les valeurs de t pour lesquelles $\sum_{j=1}^n \psi(X_j - t)$ s'annule *en changeant de signe*. Comme on le voit dans l'exemple de la médiane, il se peut que cette équation n'ait pas une solution unique. Dans ce cas on pourra, par exemple si ψ est croissante, choisir

$$\hat{\theta}_n = \inf \Big\{ t | \sum_{j=1}^n \psi(X_j - t) \le 0 \Big\}. \tag{3.6}$$

Notez que pour l'estimation de l'espérance de la loi des X_i par la moyenne empirique, $\psi(x) = x$, et que pour l'estimation d'une médiane de la loi des X_i par une médiane de l'échantillon, $\psi(x) = \mathrm{sign}(x)$, où $\mathrm{sign}(x) = -1$ si $x < 0$, 0 si $x = 0$, $+1$ si $x > 0$.

Une famille de Z-estimateurs généralisant les deux précédents est celle des estimateurs de Huber correspondant aux fonctions :

$$\psi_k(x) := \begin{cases} -k & \text{si } x < -k, \\ x & \text{si } |x| \le k, \\ k & \text{si } x > k, \end{cases} \qquad k \in \mathbb{N}^*.$$

L'introduction de ces estimateurs est liée à une question de *robustesse* (cf. la section 3 ci-dessous). Il s'agit de l'influence des valeurs extrêmes de l'échantillon sur l'estimateur. La plus gande et la plus petite valeur de l'échantillon ont relativement peu d'influence sur la médiane tandis qu'elles peuvent influencer très fortement la moyenne en cas de valeurs « aberrantes ». On dit que la moyenne empirique n'est pas robuste contre les valeurs aberrantes. Si on considère les valeurs extrêmes de l'échantillon comme peu fiables, on a certainement intérêt à limiter leur influence sur l'estimateur. D'un autre côté, on peut considérer que la médiane réussit trop bien dans cet objectif. Le paramètre k permet en quelque sorte d'interpoler entre médiane et moyenne, pour le Z-estimateur de Huber. Le comportement de cet estimateur est similaire à celui de la moyenne pour les grandes valeurs de k et à celui de la médiane pour les petites valeurs de k.

2. Propriétés asymptotiques

Les résultats ne seront donnés que pour quelques classes d'exemples, car il n'existe pas de résultat très général. Dans un premier paragraphe on estime un paramètre de translation θ par un M-estimateur défini en (3.6) en prenant une fonction ψ croissante et assez régulière. Ces hypothèses mises sur la fonction ψ ne sont pas toujours réalisées pour l'estimateur du maximum de vraisemblance. L'étude de celui-ci occupera la sous-section 2.2.

2.1. Les M-estimateurs. On considère uniquement le cas de l'estimation d'un paramètre de translation uni-dimensionnel. P_0 est une loi fixée de fonction de répartition $F_0 = F$ et P_θ est la translatée de P_0, de fonction de répartition $F_\theta(x) = F(x - \theta)$. Si X est une variable aléatoire et h une fonction borélienne P_θ-intégrable sur $\mathbb{R}$, $\mathbb{E}_\theta\, h(X)$ désigne l'espérance de $h(X)$ sous la loi P_θ (c'est-à-dire lorsque la loi de X est P_θ). Cette espérance peut s'écrire comme intégrale de Lebesgue relativement à P_θ ou comme intégrale de Stieltjes relativement à F_θ :

$$\mathbb{E}_\theta\, h(X) = \int_{\mathbb{R}} h(x)\,\mathrm{d}P_\theta(x) = \int_{\mathbb{R}} h(x)\,\mathrm{d}F_\theta(x).$$

Lorsque h est de la forme $g(.-t)$, en appliquant le changement de variable [3] $x = y+\theta$ et en remarquant que la mesure image de P_θ par ce changement de variable est P_0 puisque $P_\theta(X-\theta \leq s) = F_\theta(s+\theta) = F(s)$, on obtient

$$\mathbb{E}_\theta\, g(X-t) = \int_{\mathbb{R}} g(x-t)\,\mathrm{d}P_\theta(x) = \int_{\mathbb{R}} g(y+\theta-t)\,\mathrm{d}P_0(y) = \mathbb{E}_0\, g(X+\theta-t). \tag{3.7}$$

Ainsi on a, lorsque l'intégrale est définie,

$$\mathbb{E}_\theta\, \psi(X-t) = \lambda(t-\theta), \tag{3.8}$$

en définissant la fonction λ par

$$\lambda(s) := \mathbb{E}_0\, \psi(X-s).$$

2.1.1. *Consistance des M-estimateurs.* Lorsque la suite $(X_n)_{n\in\mathbb{N}}$ est i.i.d. et plus généralement lorsqu'elle est stationnaire et ergodique, si la loi commune des X_j est P_θ et si

$$\forall t \in \mathbb{R}, \quad \mathbb{E}_\theta\, |\psi(X-t)| < \infty, \tag{3.9}$$

alors, sous P_θ,

$$\frac{1}{n}\sum_{j=1}^{n} \psi(X_j - t) \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} \lambda(t-\theta), \quad \forall t \in \mathbb{R}. \tag{3.10}$$

On peut donc espérer que le point d'annulation du premier membre converge vers celui du second membre. En particulier, si P est une loi paire, et si ψ est impaire on aura $\lambda(0) = 0$ et l'estimateur défini en (3.6) convergera vers θ. Le théorème suivant s'applique à la médiane, à la moyenne, et à la plupart des M-estimateurs associés à une ψ croissante.

Théorème 3.1. *Si ψ est croissante, si les X_j sont de même loi P_θ, si* (3.9) *et* (3.10) *sont vérifiées et si $\lambda(t) > 0$ lorsque $t < 0$ et $\lambda(t) < 0$ lorsque $t > 0$ alors l'estimateur défini en* (3.6) *converge presque sûrement vers θ lorsque $n \to +\infty$.*

Preuve. Posons pour alléger les écritures :

$$\varphi_n(t) := \frac{1}{n}\sum_{j=1}^{n} \psi\left(X_j - t\right)$$

et notons qu'en raison de la croissance de ψ, φ_n est décroissante. Pour $k \in \mathbb{N}^*$ fixé, en appliquant (3.10) avec $t = \theta + 1/k$, on a

$$\varphi_n\left(\theta + \frac{1}{k}\right) = \frac{1}{n}\sum_{j=1}^{n} \psi\left(X_j - \theta - \frac{1}{k}\right) \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} \lambda\left(\frac{1}{k}\right) < 0.$$

3. pour les intégrales au sens de Lebesgue.

On a donc $P_\theta -$ p.s., $\varphi_n(\theta+1/k) < 0$ à partir d'un rang n_0 (aléatoire). On déduit alors de la définition de $\hat{\theta}_n$ que $\hat{\theta}_n \leq \theta + 1/k$ pour $n \geq n_0$ (notez que la croissance de ψ n'intervient pas ici). De même, avec $t = \theta - 1/k$,

$$\frac{1}{n}\sum_{j=1}^{n} \psi\left(X_j - \theta + \frac{1}{k}\right) \xrightarrow[n\to+\infty]{P_\theta - \text{p.s.}} \lambda\left(-\frac{1}{k}\right) > 0,$$

d'où $\varphi_n(\theta - 1/k) > 0$ à partir d'un rang n_1 (aléatoire). Comme φ_n est décroissante et $\hat{\theta}_n = \inf\{s \in \mathbb{R};\ \varphi_n(t) \leq 0\}$, l'inégalité $\varphi_n(t) > 0$, implique $t \leq \hat{\theta}_n$, donc ici $\theta - 1/k \leq \hat{\theta}_n$.

Avec une P_θ-probabilité 1 on a donc $\hat{\theta}_n \in [\theta - \frac{1}{k}, \theta + \frac{1}{k}]$ pour n assez grand. On en déduit que pour tout $k \in \mathbb{N}^*$, l'évènement

$$\Omega_k := \left\{\theta - \frac{1}{k} \leq \liminf_{n\to+\infty} \hat{\theta}_n \leq \limsup_{n\to+\infty} \hat{\theta}_n \leq \theta + \frac{1}{k}\right\}$$

est de P_θ-probabilité 1. L'évènement $\Omega' := \cap_{k\geq 1}\Omega_k$ est donc lui aussi de P_θ-probabilité 1. Sur Ω', on a clairement $\liminf \hat{\theta}_n \leq \limsup \hat{\theta}_n = \theta$, ce qui établit la convergence $P_\theta -$ p.s. de $\hat{\theta}_n$ vers θ. □

On déduit des résultats du chapitre 2 que la convergence presque sûre de $\hat{\theta}_n$ vers θ a lieu sous (3.9) dès que le processus $(X_j)_{j\in\mathbb{N}}$ est stationnaire et ergodique.

2.1.2. *Normalité asymptotique.* Dans ce paragraphe on se place dans le même contexte et on considère encore la suite des estimateurs (3.6).

Théorème 3.2. *On suppose que ψ est croissante, bornée et P_0 presque partout continue, ou bien qu'elle est croissante, uniformément continue et que $\mathbb{E}_0(\psi(X-t)^2) < +\infty$ pour tout t. On suppose en outre que la fonction $\lambda(t)$ est nulle en 0, dérivable en 0 et que $\lambda'(0) \neq 0$ (donc $\lambda'(0) < 0$).*

Dans ces conditions, les X_j étant i.i.d. de loi P_θ,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi}} U,$$

où U est une variable aléatoire gaussienne centrée dont la variance

$$v_\infty = \frac{\mathbb{E}_0\left(\psi(X)^2\right)}{(\lambda'(0))^2} \tag{3.11}$$

ne dépend pas de θ.

Preuve. Puisque φ_n est décroissante et $\hat{\theta}_n = \inf\{s \in \mathbb{R};\ \varphi_n(s) \leq 0\}$, les inégalités $\hat{\theta}_n \leq t$ et $n^{1/2}\varphi_n(t) \leq 0$ sont équivalentes, d'où avec $t = \theta + xn^{-1/2}$,

$$\hat{\theta}_n - \theta \leq \frac{x}{\sqrt{n}} \quad\Longleftrightarrow\quad \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \psi\left(X_j - \theta - \frac{x}{\sqrt{n}}\right) \leq 0.$$

On utilise ensuite le développement suivant

$$\psi\left(X_j-\theta-\frac{x}{\sqrt{n}}\right)=\psi(X_j-\theta)+\lambda\left(\frac{x}{\sqrt{n}}\right)+\underbrace{\psi\left(X_j-\theta-\frac{x}{\sqrt{n}}\right)-\psi(X_j-\theta)-\lambda\left(\frac{x}{\sqrt{n}}\right)}_{=:R_{n,j}}.$$

Le théorème central limite s'applique aux $\psi(X_j-\theta)$, ce qui implique, puisque $\mathbb{E}_\theta(\psi(X-\theta)=\lambda(0)=0$, que sous P_θ,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\psi(X_j-\theta)\xrightarrow[n\to+\infty]{\text{loi}}V, \tag{3.12}$$

où V est une variable gaussienne centrée de variance

$$\mathbb{E}_\theta\left(\psi(X_1-\theta)\right)^2=\mathbb{E}_0\left(\psi(X_1)^2\right),$$

en appliquant (3.7) avec $g=\psi^2$.

Puisque λ est dérivable en zéro et nulle en zéro,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\lambda\left(\frac{x}{\sqrt{n}}\right)=\sqrt{n}\lambda\left(\frac{x}{\sqrt{n}}\right)\xrightarrow[n\to+\infty]{}x\lambda'(0). \tag{3.13}$$

Enfin, pour le terme de reste $R_n:=n^{-1/2}\sum_{j=1}^{n}R_{n,j}$, on vérifie que

$$R_n:=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\psi\left(X_j-\theta-\frac{x}{\sqrt{n}}\right)-\psi(X_j-\theta)-\lambda\left(\frac{x}{\sqrt{n}}\right)\right)\xrightarrow[n\to+\infty]{L^2}0. \tag{3.14}$$

Pour cela, on remarque d'abord que le terme général de la somme ci-dessus est une variable aléatoire centrée puisque $\lambda(n^{-1/2}x)=\mathbb{E}_0\,\psi(X_j-n^{-1/2}x)=\mathbb{E}_\theta\,\psi(X_j-\theta-n^{-1/2}x)$ et $\mathbb{E}_\theta\,\psi(X_j-\theta)=\mathbb{E}_0\,\psi(X_j)=\lambda(0)=0$. On peut donc interpréter $\mathbb{E}_\theta\,R_n^2$ comme la variance d'une somme de variables aléatoires i.i.d. (sous P_θ), d'où :

$$\mathbb{E}_\theta\,R_n^2=\operatorname{Var}_\theta\left(\psi\left(X_1-\theta-\frac{x}{\sqrt{n}}\right)-\psi(X_1-\theta)\right),$$

puis en utilisant l'inégalité élémentaire $\operatorname{Var}Y\leq\mathbb{E}\,Y^2$ et en appliquant (3.7) avec $g=(\psi(\,.-n^{-1/2}x)-\psi)^2$,

$$\mathbb{E}_\theta\,R_n^2\leq\int_{\mathbb{R}}\left|\psi\left(y-\frac{x}{\sqrt{n}}\right)-\psi(y)\right|^2\,\mathrm{d}F(y).$$

Cette dernière intégrale tend vers zéro soit parce que ψ est bornée et P_0 presque partout continue (par convergence dominée), soit parce qu'elle est uniformément continue (vérification élémentaire).

En rassemblant (3.12), (3.13) et (3.14) et en utilisant le lemme de Slutsky, on obtient

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\psi\left(X_j-\theta-\frac{x}{\sqrt{n}}\right)\xrightarrow[n\to+\infty]{\text{loi}}V+x\lambda'(0)$$

et comme la f.d.r. Ψ de la v.a. gaussienne $V + x\lambda'(0)$ est continue au point 0, on en déduit que

$$P_\theta\left(\hat{\theta}_n - \theta \leq \frac{x}{\sqrt{n}}\right) = P_\theta\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\psi\left(X_j - \theta - \frac{x}{\sqrt{n}}\right) \leq 0\right) \xrightarrow[n\to+\infty]{} \Psi(0).$$

La variable gaussienne $V+x\lambda'(0)$ a pour espérance $x\lambda'(0)$ et pour écart type $\sigma = \mathbb{E}_0^{1/2}\left(\psi(X_1)^2\right)$, En notant Φ la f.d.r. de la loi normale $\mathfrak{N}(0,1)$, on a $\Psi(t) = \Phi((t - x\lambda'(0))/\sigma)$, d'où pour tout x réel

$$P_\theta\left(\sqrt{n}(\hat{\theta}_n - \theta) \leq x\right) \xrightarrow[n\to+\infty]{} \Phi\left(\frac{-\lambda'(0)}{\sigma}x\right) = \Phi\left(\frac{|\lambda'(0)|}{\sigma}x\right),$$

en rappelant que $\lambda'(0) < 0$. La f.d.r. limite est bien celle d'une gaussienne centrée de variance

$$\frac{\sigma^2}{\lambda'(0)^2} = \frac{\mathbb{E}_0\left(\psi(X_1)^2\right)}{\lambda'(0)^2} = v_\infty$$

indépendante de θ, ce qui achève la preuve. □

Remarque 3.1. Si $\lambda'(0) = 0$, on a

$$\forall x \in \mathbb{R}, \quad P_\theta\left(\sqrt{n}(\hat{\theta}_n - \theta) \leq x\right) \xrightarrow[n\to+\infty]{} \Phi(0) = \frac{1}{2}.$$

La limite n'étant pas une fonction de répartition, il n'y a pas convergence en loi. ◁

Exercice 3.5. Appliquez le théorème 3.2 à la médiane. ◁

Exercice 3.6. Appliquez le théorème 3.2 au cas où ψ est strictement croissante, impaire, bornée. ◁

On reviendra au chapitre 4 sur la convergence des M-estimateurs pour des suites de variables non indépendantes.

2.2. L'estimateur du maximum de vraisemblance. Classiquement, la convergence en loi de l'estimateur du maximum de vraisemblance du paramètre θ est démontrée sous des conditions de dérivabilité d'ordre deux par rapport à θ de la fonction $f(x,\theta)$ (voir par exemple [**15**]). On verra ici qu'on peut alléger cette hypothèse.

2.2.1. *Convergence presque sûre.* Pour simplifier, nous nous plaçons dans le cas où l'ensemble Θ des paramètres est un ouvert de $\mathbb{R}$. Le modèle $(P_\theta)_{\theta\in\Theta}$ est supposé régulier c'est à dire qu'il est dominé (les P_θ sont à densité $f(\,.\,,\theta)$ par rapport à une même mesure de probabilité μ) et les densités $f(\,.\,,\theta)$ vérifient les conditions suivantes.

R.1. $f(x,\theta) > 0$ pour tout (x,θ).

R.2. Pour tout x, la fonction $f(x,\,.\,)$ est partout dérivable par rapport à θ.

R.3. Pour tout $\theta \in \Theta$, la dérivabilité sous le signe intégrale est valable :

$$\int_{\mathbb{R}}\left|\frac{\partial f(x,\theta)}{\partial\theta}\right| \mathrm{d}\mu(x) < \infty \text{ et } \int_{\mathbb{R}}\frac{\partial f(x,\theta)}{\partial\theta}\mathrm{d}\mu(x) = \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f(x,\theta)\,\mathrm{d}\mu(x) = 0. \tag{3.15}$$

R.4. Pour tout $\theta \in \Theta$, *l'information de Fisher* $I(\theta)$ est non nulle :

$$I(\theta) := \mathbb{E}_\theta \left(\frac{\partial \ln f(X,\theta)}{\partial \theta} \right)^2 > 0. \tag{3.16}$$

En raison de la régularité du modèle $(P_\theta)_{\theta\in\Theta}$, la fonction $t \longmapsto M_n(t, X_1, \ldots, X_n) = \sum_{j=1}^n \ln f(X_j, t)$ est dérivable et tout estimateur du maximum de vraisemblance est solution de l'équation de vraisemblance

$$\sum_{j=1}^n \frac{\partial \ln f(X_j, t)}{\partial t} = 0. \tag{3.17}$$

Théorème 3.3. *Supposons que les variables X_j sont P_θ-i.i.d. pour tout $\theta \in \Theta$, que le modèle $(P_\theta)_{\theta\in\Theta}$ est régulier et que l'application qui à t associe P_t est injective. Alors, pour tout $\theta \in \Theta$ et pour P_θ presque tout échantillon infini $(x_j)_{j\geq 1} = (X_j(\omega))_{j\geq 1}$, il existe une suite $(\hat\theta_n)$ convergente vers θ, telle que $\hat\theta_n$ soit solution de l'équation de vraisemblance (3.17) pour tout n suffisamment grand.*

Preuve. Fixons θ. On peut toujours, dans les équations de vraisemblance (3.17), remplacer $f(X_j, t)$ par $f(X_j, t) f(X_j, \theta)^{-1}$. Considérons donc la suite

$$\frac{1}{n} \sum_{j=1}^n \ln \frac{f(X_j, t)}{f(X_j, \theta)}. \tag{3.18}$$

Les variables $Y_j = \ln(f(X_j, t) f(X_j, \theta)^{-1})$ sont i.i.d. et vérifient $\mathbb{E}_\theta Y_j^+ < +\infty$. Pour le voir, il suffit d'appliquer l'inégalité $\ln x \leq x$:

$$\mathbb{E}_\theta Y_j^+ = \mathbb{E}_\theta \left(Y_j \mathbb{I}_{Y_j\geq 0} \right) \leq \mathbb{E}_\theta \left(\frac{f(X_j,t)}{f(X_j,\theta)} \mathbb{I}_{Y_j \geq 0} \right) \leq \mathbb{E}_\theta \frac{f(X_j,t)}{f(X_j,\theta)} = \int_{\mathbb{R}} \frac{f(x,t)}{f(x,\theta)} f(x,\theta) \,\mathrm{d}\mu(x) = 1.$$

Leur l'espérance est négative (mais éventuellement égale à $-\infty$) car, d'après l'inégalité de Jensen :

$$\mathbb{E}_\theta \left(\ln \frac{f(X_j,t)}{f(X_j,\theta)} \right) < \ln \mathbb{E}_\theta \left(\frac{f(X_j,t)}{f(X_j,\theta)} \right) = 0.$$

L'inégalité est stricte pour $t \neq \theta$: en effet, la fonction ln étant strictement concave on a

$$\mathbb{E}_\theta \left(\ln \frac{f(X,t)}{f(X,\theta)} \right) = \ln \mathbb{E}_\theta \left(\frac{f(X,t)}{f(X,\theta)} \right) \iff \frac{f(X,t)}{f(X,\theta)} = c \quad \mu - \text{p.s.}$$

Comme pour tout t, $f(\,.\,, t)$ est une densité, la constante c est nécessairement égale à 1. Mais $\frac{f(X,t)}{f(X,\theta)} = 1$ μ-p.s. pour un $t \neq \theta$ contredit l'injectivité de l'application $t \mapsto P_t$.

La loi des grands nombres permet de conclure que, pour tout $t \neq \theta$, P_θ-presque sûrement,

$$\frac{1}{n} \sum_{j=1}^n \ln \frac{f(X_j,t)}{f(X_j,\theta)} \to \mathbb{E}_\theta \left(\ln \frac{f(X,t)}{f(X,\theta)} \right) < 0, \tag{3.19}$$

le résultat étant valable même si $\mathbb{E}_\theta \left(\ln \frac{f(X_j,t)}{f(X_j,\theta)} \right) = -\infty$, cf. exercice 3.7.

Considérons maintenant l'ensemble dénombrable

$$\Theta_0 = \left\{ t \in \Theta | t = \theta \pm \frac{1}{k},\ k \in \mathbb{N}^* \right\},$$

et, pour tout $t \in \Theta_0$, appelons Ω_t l'ensemble des $\omega \in \Omega$ tels que (3.19) a lieu, $P_\theta(\Omega_t) = 1$. Évidemment, puisque Θ_0 est dénombrable, l'intersection Ω'_θ des Ω_t est de probabilité P_θ égale à 1. Soit maintenant un ω appartenant à Ω'_θ. D'après la définition des Ω_t, pour tout $t \in \Theta_0$, il existe une constante $-\infty \le \ell(t) < 0$ telle que,

$$\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega), t)}{f(X_j(\omega), \theta)} \xrightarrow[n\to+\infty]{} \ell(t). \tag{3.20}$$

Pour tout $k \in \mathbb{N}^*$, posons $t_k = \theta - \frac{1}{k}$ et $t'_k = \theta + \frac{1}{k}$. D'après (3.20), il existe un entier $n_k(\omega)$ tel que, pour tout $n \ge n_k(\omega)$,

$$\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega), t_k)}{f(X_j(\omega), \theta)} < 0 \quad \text{et} \quad \frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega), t'_k)}{f(X_j(\omega), \theta)} < 0.$$

Ceci nous laisse la possibilité de prendre la suite $(n_k(\omega))_{k\ge1}$ *strictement* croissante, choix que nous adoptons désormais.

Considérons la fonction $\frac{1}{n}\sum_{j=1}^{n} \ln \frac{f(X_j(\omega),t)}{f(X_j(\omega),\theta)}$. Elle est nulle pour $t = \theta$ et strictement négative pour $t = t_k$ et $t = t'_k$. Comme elle est partout dérivable, il existe un point de $]t_k, t'_k[$ qui annule sa dérivée. Pour $n_k(\omega) \le n < n_{k+1}(\omega)$, notons $\hat{\theta}_n(\omega)$ ce point. Avec cette construction, la suite des $\hat{\theta}_n(\omega)$ est définie pour $n \ge n_1(\omega)$. On complète sa définiton pour $n < n_1(\omega)$ par un choix arbitraire. Par construction cette suite vérifie :

$$\forall k \in \mathbb{N}^*, \quad \theta - \frac{1}{k} \le \liminf_{n\to+\infty} \hat{\theta}_n(\omega) \le \limsup_{n\to+\infty} \hat{\theta}_n(\omega) \le \theta + \frac{1}{k}.$$

En faisant tendre k vers l'infini, on en déduit la convergence de $\hat{\theta}_n(\omega)$ vers θ. □

Exercice 3.7. *Une loi forte des grands nombres avec limite infinie.*
1) Soit $(X_i)_{i\ge1}$ une suite i.i.d. de variables aléatoires positives telle que $\mathbb{E}\,X_1 = +\infty$. Montrer que

$$\frac{1}{n}\sum_{i=1} X_i \xrightarrow[n\to+\infty]{\text{p.s.}} +\infty.$$

2) Généraliser ce résultat au cas où les X_i sans être positives vérifient l'une des deux conditions

i) $\mathbb{E}\,X_1^+ = +\infty$ et $\mathbb{E}\,X_1^- < +\infty$;
ii) $\mathbb{E}\,X_1^+ < +\infty$ et $\mathbb{E}\,X_1^- = +\infty$.

◁

Le théorème 3.3 n'est pas tout à fait un résultat de consistance forte de l'estimateur du maximum de vraisemblance (voyez-vous pourquoi ?). Par ailleurs, les hypothèses du théorème ne sont pas du tout nécessaires.

Exercice 3.8. Trouver la limite de l'estimateur par maximum de vraisemblance de θ lorsque P_θ est la loi uniforme sur $[0, \theta]$ (dans ce cas, le modèle n'est pas régulier.) ◁

2.2.2. *Convergence en loi.* Nous plaçant dans les conditions du théorème 3.3, nous allons maintenant regarder l'écart renormalisé $\sqrt{n}(\hat{\theta}_n - \theta)$.

Théorème 3.4. *Supposons vérifiées les hypothèses du théorème 3.3 ainsi que les hypothèses suivantes.*

- *Quel que soit x, $f(x,\theta)$ est partout deux fois dérivable par rapport à θ ;*
- *on peut dériver deux fois par rapport à θ sous le signe somme dans $\int_{\mathbb{R}} f(x,\theta)\,\mathrm{d}\mu(x)$;*
- *la dérivée seconde $D_2(x,\theta) := \frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2}$ est partout continue par rapport à θ, et cette continuité est uniforme en x au sens suivant : $\sup_x |D_2(x,t) - D_2(x,\theta)|$ tend vers 0 quand t tend vers θ ;*
- *en tout θ, l'information de Fisher $I(\theta)$, définie en (3.16) est finie.*

Alors, pour toute suite $(\hat{\theta}_n)$ de solutions des équations de vraisemblance qui converge presque sûrement vers θ, on a sous la loi P_θ,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to+\infty]{\text{loi}} \mathfrak{N}\Big(0, \frac{1}{I(\theta)}\Big).$$

Preuve. Posons

$$h_n(X_1, \cdots, X_n, \theta) := \frac{1}{n}\sum_{j=1}^{n} \frac{\partial \ln f(X_j,\theta)}{\partial\theta}.$$

Sous les hypothèses faites, $(\frac{\partial \ln f(X_j,\theta)}{\partial\theta})_{n\geq 1}$ est une suite i.i.d. de variables aléatoires centrées et de variance $I(\theta)$. On applique le théorème central limite : sous la loi P_θ,

$$\sqrt{n}h_n(X_1, \cdots, X_n, \theta) \xrightarrow[n\to+\infty]{\text{loi}} \mathfrak{N}\Big(0, I(\theta)\Big). \tag{3.21}$$

Puis, développons h_n en utilisant le théorème des accroissements finis :

$$h_n(X_1, \cdots, X_n, t) = h_n(X_1, \cdots, X_n, \theta) + (t-\theta)K_n(X_1, \cdots, X_n, \theta^*), \tag{3.22}$$

où $\min(t,\theta) < \theta^* < \max(t,\theta)$ et

$$K_n(X_1, \cdots, X_n, u) := \frac{1}{n}\sum_{j=1}^{n} D_2(X_j, u).$$

Puis, en remplaçant t par $\hat{\theta}_n$ on a avec $\min(\hat{\theta}_n,\theta) < \theta_n^* < \max(\hat{\theta}_n,\theta)$,

$$0 = h_n(X_1, \cdots, X_n, \theta) + (\hat{\theta}_n - \theta)K_n(X_1, \cdots, X_n, \theta_n^*). \tag{3.23}$$

Le comportement asymptotique de $K_n(X_1, \cdots, X_n, \theta)$, somme des variables i.i.d. $D_2(X_j,\theta)$, est régi par la loi forte des grands nombres, pourvu que $\mathbb{E}\,|D_2(X_1,\theta)|$ soit fini. Vérifions ce point.

$$\begin{aligned}\mathbb{E}_\theta \left|\frac{\partial^2 \ln f(X_1,\theta)}{\partial\theta^2}\right| &= \int_{\mathbb{R}} \left|\frac{f(x,\theta)\frac{\partial^2 f}{\partial\theta^2}(x,\theta) - \left(\frac{\partial f}{\partial\theta}(x,\theta)\right)^2}{f(x,\theta)^2}\right| f(x,\theta)\,\mathrm{d}\mu(x)\\ &\leq \int_{\mathbb{R}} \left|\frac{\partial^2 f}{\partial\theta^2}(x,\theta)\right|\,\mathrm{d}\mu(x) + \mathbb{E}_\theta\left|\frac{\partial \ln f}{\partial\theta}(X_1,\theta)\right|^2 < +\infty,\end{aligned}$$

car dans ce majorant la finitude du premier terme vient de l'hypothèse de double dérivabilité sous le signe somme de $\int_{\mathbb{R}} f(x,\theta)\,\mathrm{d}\mu(x)$ et le deuxième terme n'est autre que $I(\theta)$. Par conséquent,

$$\frac{1}{n}\sum_{j=1}^{n} D_2(X_j,\theta) \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} \mathbb{E}_\theta \frac{\partial^2 \ln f(X_1,\theta)}{\partial\theta^2}$$

et en reprenant le calcul ci-dessus sans les valeurs absolues, on voit que cette limite est égale à $-I(\theta)$, en notant que la nullité de $\int_{\mathbb{R}} \frac{\partial^2 f}{\partial\theta^2}(x,\theta)\,\mathrm{d}\mu(x)$ s'obtient en dérivant 2 fois sous le signe somme l'identité $\int_{\mathbb{R}} f(x,\theta)\,\mathrm{d}\mu(x) = 1$. Finalement,

$$K_n(X_1,\cdots,X_n,\theta) \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} -I(\theta). \tag{3.24}$$

Par ailleurs,

$$\begin{aligned} |K_n(X_1,\cdots,X_n,\theta_n^*) - K_n(X_1,\cdots,X_n,\theta)| &\le \frac{1}{n}\sum_{j=1}^{n} |D_2(X_j,\theta_n^*) - D_2(X_j,\theta)| \\ &\le \sup_x |D_2(x,\theta_n^*) - D_2(x,\theta)|. \end{aligned} \tag{3.25}$$

Comme $\hat\theta_n$ converge P_θ-presque sûrement vers θ, il en va de même pour θ_n^* qui est dans $]\min(\hat\theta_n,\theta),\max(\hat\theta_n,\theta)[$. Grâce à la continuité uniforme en x de la dérivée seconde $D_2(x,\theta)$, il résulte de (3.24) et (3.25) que,

$$K_n(X_1,\cdots,X_n,\theta_n^*) \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} -I(\theta). \tag{3.26}$$

Comme $I(\theta)$ n'est pas nul, l'évènement $A_n := \{K_n(X_1,\cdots,X_n,\theta_n^*) \neq 0\}$ a donc une probabilité qui tend vers 1 quand n tend vers l'infini. *Sur cet évènement*, on peut réécrire (3.23) sous la forme :

$$\sqrt{n}(\hat\theta_n - \theta) = \frac{-1}{K_n(X_1,\cdots,X_n,\theta_n^*)}\sqrt{n}h_n(X_1,\cdots,X_n,\theta) = R_n\sqrt{n}h_n(X_1,\cdots,X_n,\theta),$$

en prenant $R_n := K_n(X_1,\cdots,X_n,\theta_n^*)^{-1}$ sur A_n et $R_n := I(\theta)^{-1}$ sur A_n^c. Ainsi le produit $R_n\sqrt{n}h_n(X_1,\cdots,X_n,\theta)$ est bien défini sur tout Ω et converge en loi par (3.26), (3.21) et le lemme de Slutsky vers une gaussienne centrée de variance $I(\theta)/I(\theta)^2 = 1/I(\theta)$. Comme $P(A_n)$ tend vers 1, cette convergence en loi a lieu aussi pour $\sqrt{n}(\hat\theta_n - \theta)$, avec la même limite (voir l'exercice suivant). □

Exercice 3.9. Soient $(A_n)_{n\ge1}$ une suite d'évènements dont la probabilité tend vers 1, $(Y_n)_{n\ge1}$ et $(Z_n)_{n\ge1}$ deux suites de variables aléatoires telles que la première converge en loi vers Y et que sur A_n, $Y_n = Z_n$. Alors Z_n converge en loi vers Y. ◁

Comme pour le théorème précédent, le résultat reste valable sous des hypothèses plus faibles que celles sous-lesquelles il est démontré. On reprendra notamment l'exemple de l'estimateur du maximum de vraisemblance du paramètre de translation de la loi double exponentielle (exercice 3.1), via l'exercice 3.5.

On trouvera dans [24] la preuve du théorème suivant, qui ne demande pas la régularité du modèle.

Théorème 3.5. *Supposons que Θ est un ouvert de $\mathbb{R}$, que le modèle $(P_\theta)_{\theta\in\Theta}$ est dominé et dérivable en moyenne quadratique au point $\theta_0 \in \Theta$, c'est à dire qu'il existe une application mesurable $\dot\ell$ telle que*

$$\int \left|\sqrt{f(x,\theta)} - \sqrt{f(x,\theta_0)} - \frac{1}{2}(\theta-\theta_0)\dot\ell(x)\sqrt{f(x,\theta_0)}\right|^2 \mathrm{d}\mu(x) = o\left(|\theta-\theta_0|^2\right), \tag{3.27}$$

et telle que

$$0 < I(\theta_0) = \mathbb{E}_{\theta_0}(\dot\ell^2(X)) < \infty.$$

On suppose de plus qu'il existe une application mesurable v telle que $\mathbb{E}_{\theta_0}(v^2) < \infty$ et telle que pour tous θ_1 et θ_2 dans un voisinage de θ_0,

$$|\ln f(x,\theta_1) - \ln f(x,\theta_2)| \leq v(x)|\theta_1-\theta_2|.$$

Alors, si, sous la loi P_{θ_0} la suite $\hat\theta_n$ d'estimateurs du maximum de vraisemblance converge presque sûrement vers θ_0, on a

$$\sqrt{n}(\hat\theta_n - \theta_0) = I(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{j=1}^n \dot\ell(X_j) + o_{P_{\theta_0}}(1). \tag{3.28}$$

La décomposition (3.28) implique la convergence en loi de $\sqrt{n}(\hat\theta_n-\theta_0)$ vers une gaussienne centrée de variance $I(\theta_0)^{-1}$ pourvu que $\mathbb{E}_{\theta_0}(\dot\ell(X)) = 0$. Vérifions cette égalité.

Vérification. En fixant h et en prenant $\theta = \theta_0 + n^{-1/2}h$, on déduit de (3.27) que

$$\int \left|\sqrt{f(x,\theta_0+n^{-1/2}h)} - \sqrt{f(x,\theta_0)} - \frac{h\dot\ell(x)}{2\sqrt{n}}\sqrt{f(x,\theta_0)}\right|^2 \mathrm{d}\mu(x) = \frac{h^2\varepsilon_n}{n}, \quad \varepsilon_n \to 0. \tag{3.29}$$

Comme $\int \dot\ell(x)^2 f(x,\theta_0)\,\mathrm{d}\mu(x) = I(\theta_0) < \infty$, on obtient

$$\sqrt{f(\,.\,,\theta_0+n^{-1/2}h)} \xrightarrow[n\to+\infty]{L^2(\mu)} \sqrt{f(\,.\,,\theta_0)}. \tag{3.30}$$

Écrivons maintenant $\mathbb{E}_{\theta_0}(\dot\ell(X)) = \int \dot\ell(x)f(x,\theta_0)\,\mathrm{d}\mu(x)$ sous la forme :

$$\mathbb{E}_{\theta_0}(\dot\ell(X)) = \int \frac{\dot\ell(x)}{2}\sqrt{f(x,\theta_0)}(2\sqrt{f(x,\theta_0)})\,\mathrm{d}\mu(x). \tag{3.31}$$

On remarque que d'après (3.29),

$$\alpha_n := n^{1/2}\left(\sqrt{f(\,.\,,\theta_0+n^{-1/2}h)} - \sqrt{f(\,.\,,\theta_0)}\right) \xrightarrow[n\to+\infty]{L^2(\mu)} \frac{h\dot\ell}{2}\sqrt{f(\,.\,,\theta_0)}$$

et d'après (3.30),

$$\beta_n := \sqrt{f(\,.\,,\theta_0+n^{-1/2}h)} + \sqrt{f(\,.\,,\theta_0)} \xrightarrow[n\to+\infty]{L^2(\mu)} 2\sqrt{f(\,.\,,\theta_0)}.$$

Ces deux convergences $L^2(\mu)$ impliquent la convergence du produit scalaire $\langle \alpha_n, \beta_n \rangle_{L^2(\mu)}$, ce qui nous donne par report dans (3.31) :

$$\mathbb{E}_{\theta_0}(\dot{\ell}(X)) =$$
$$\lim_{n\to\infty} \sqrt{n} \int \left(\sqrt{f(x, \theta_0 + n^{-1/2}h)} - \sqrt{f(x, \theta_0)}\right) \left(\sqrt{f(x, \theta_0 + n^{-1/2}h)} + \sqrt{f(x, \theta_0)}\right) \mathrm{d}\mu(x) =$$
$$\lim_{n\to\infty} \sqrt{n} \int \left(f(x, \theta_0 + n^{-1/2}h) - f(x, \theta_0)\right) \mathrm{d}\mu(x) = 0.$$

□

Exercice 3.10. Comment définiriez vous la dérivabilité en moyenne quadratique de $f(x, \theta)$? Pourquoi l'énoncé parle-t-il de dérivabilité en moyenne quadratique à propos de la condition (3.27) ? La notation $I(\theta_0) = \mathbb{E}_{\theta_0}(\dot{\ell}^2(X))$ semble suggérer un lien entre cette quantité et l'information de Fisher. Pourriez vous clarifier cette question ? ◁

Le théorème 3.5 ne demande pas la dérivabilité (à l'ordre 1 et 2) de $f(x, \theta)$ par rapport à θ en tout x. Il s'applique en particulier à la loi double exponentielle.

Exercice 3.11. Montrer que la condition de dérivabilité en moyenne quadratique implique pour la distance d'Hellinger (voir la section 1.3 du chapitre 1) $H^2(P_\theta, P_{\theta_0}) = O((\theta - \theta_0)^2)$. En déduire que la famille des lois uniformes sur $[0, \theta]$ n'est pas dérivable en moyenne quadratique, et qu'il en est de même pour la famille de lois triangulaires de l'exercice 1.10. ◁

Exercice 3.12. Reprendre l'exercice 3.8 et trouver la loi limite de l'estimateur. ◁

2.3. Méthodes alternatives au maximum de vraisemblance. On verra dans la section 4 que la variance limite $1/I(\theta)$ obtenue dans le théorème 3.4 est, pour les modèles réguliers, la plus petite possible. C'est un argument solide en faveur de la méthode du maximum de vraisemblance.

Cependant, la section précédente laisse ouvertes plusieurs questions. Dans le cas favorable où chaque équation de vraisemblance (3.17) a une solution unique, les théorèmes 3.3 et 3.4 donnent des conditions pour la convergence de la suite des estimateurs ainsi obtenus. Il ne reste qu'à vérifier que cet estimateur maximise bien la vraisemblance. Mais que se passe-t-il lorsqu'il n'y a pas unicité ?

– Parmi les solutions des équations de vraisemblance, quelles sont celles qui correspondent à un maximum ?
– Comment s'assurer que la suite de solutions choisie a bien la limite en loi fournie par le théorème 3.4 ?

Et, bien sûr, il ne faut pas oublier que les solutions des équations de vraisemblance ne sont pas toujours calculables.

Le résultat qui suit donne un moyen simple pour construire, à partir d'un estimateur convergent, un autre estimateur convergent qui atteint la variance limite minimale du théorème 3.4. Sur ce sujet, le lecteur trouvera des compléments et des exemples dans le chapitre 4 de [21], le chapitre 6 de [22] ou encore dans [15].

On utilisera ici une version simplifiée des notations introduites dans la preuve du théorème 3.4 :

$$\begin{aligned} h_n(t) &= h_n(X_1,\dots,X_n,t) = \frac{1}{n}\sum_{j=1}^n \frac{\partial \ln f(X_j,t)}{\partial t} \\ K_n(t) &= K_n(X_1,\dots,X_n,t) = \frac{1}{n}\sum_{j=1}^n D_2(X_j,t) \end{aligned}$$

Théorème 3.6. *Supposons vérifiées les hypothèses des théorèmes 3.3 et 3.4. Soit $\tilde\theta_n$ une suite telle que, sous la loi P_θ, $\tilde\theta_n$ converge presque sûrement vers θ et*

$$\sqrt{n}(\tilde\theta_n - \theta) = O_P(1),$$

alors, l'estimateur défini par

$$\tilde\theta_n^1 = \tilde\theta_n - \frac{h_n(\tilde\theta_n)}{K_n(\tilde\theta_n)}$$

a, sous P_θ, la loi limite obtenue dans le théorème 3.4 :

$$\sqrt{n}(\tilde\theta_n^1 - \theta) \xrightarrow[n\to+\infty]{\text{loi}} \mathfrak{N}\Big(0, \frac{1}{I(\theta)}\Big).$$

Preuve. On ne fait que l'esquisser, laissant au lecteur le soin de la terminer. On reprend le développement (3.22) de la fonction h_n autour du point $t=\theta$

$$h_n(\tilde\theta_n) = h_n(\theta) + (\tilde\theta_n - \theta)K_n(\theta_n^*)$$

où θ_n^* est entre θ et $\tilde\theta_n$. Reportant dans l'expression de $\tilde\theta_n^1$ on obtient

$$\sqrt{n}(\tilde\theta_n^1 - \theta) = -\sqrt{n}\frac{h_n(\theta)}{K_n(\tilde\theta_n)} + \sqrt{n}(\tilde\theta_n - \theta)\left(1 - \frac{K_n(\theta_n^*)}{K_n(\tilde\theta_n)}\right)$$

Les arguments de convergence de $K_n(t)$ et de $\sqrt{n}h_n(\theta)$ déjà utilisés dans la preuve du théorème 3.4, joints au fait que $\sqrt{n}(\tilde\theta_n-\theta)$ est borné en probabilité, permettent de conclure. □

3. Robustesse

3.1. Généralités. La notion de M-estimateur a été introduite dans les années 60 par P. J. Huber pour répondre à des préoccupations de robustesse. Mais certains des estimateurs robustes étudiés dans la littérature depuis cette époque sont en fait bien plus anciens. L'idée est la suivante : on travaille en général avec des hypothèses plus ou moins explicites. Par exemple les variables qui modélisent le phénomène observé ont toutes même loi. Ou bien elles ont toutes une loi gaussienne. Ou encore, elles sont mutuellement indépendantes. En réalité il ne s'agit là que d'hypothèses qui facilitent l'étude et, dans bien des cas, elles sont tout simplement peu réalistes. Quelle est la distorsion subie par l'estimateur qu'on a choisi si une des hypothèses sous lesquelles il était censé bien se comporter n'est pas réalisée ? Cette question en appelle une autre : pourrait-on construire des estimateurs robustes, c'est à dire des estimateurs qui soient peu sensibles à un écart aux hypothèses de base. En réalité, on n'a que peu de réponses à la deuxième question pour les écarts à l'indépendance, qui peuvent occasionner des distorsions importantes, et on en a beaucoup pour les écarts à une loi spécifiée. C'est dans ce cadre que nous allons nous placer.

3.2. Quelques exemples. Un problème que rencontrent tous les statisticiens est celui des données aberrantes (« outliers » en anglais). La présence de valeurs aberrantes peut indiquer soit que quelques unes des variables ont une loi décentrée par rapport à celle des autres, ou bien que leur variance est beaucoup plus grande que celle des autres, soit que toutes les variables ont bien même loi, mais que cette loi n'est pas celle que l'on pensait. Quoi qu'il en soit, on peut avoir des attitudes différentes dans cette situation : supprimer les valeurs aberrantes, ou bien en tenir compte et refaire une modélisation, ou avoir recours à un estimateur qui soit peu sensible à leur présence. Tout le monde sait, par exemple, que la moyenne est très sensible aux valeurs aberrantes alors que la médiane ne l'est pas. La moyenne α-tronquée (en anglais « α-trimmed mean ») qui est la moyenne de ce qui reste lorsqu'on a enlevé à l'échantillon un pourcentage α des valeurs les plus grandes et le même pourcentage des valeurs les plus petites, a la même qualité. Un autre exemple est l'écart type empirique qui est plus sensible aux valeurs aberrantes que la médiane des écarts à la médiane.

Les statisticiens se sont attachés à construire des indices pour quantifier ces notions. Nous parlerons ici de la fonction d'influence et du point de rupture. Pour des compléments, se reporter à [17].

3.3. Fonction d'influence. On suppose ici que, comme c'était le cas dans les exemples précédents, l'estimateur est la valeur que prend une fonctionnelle $T(P)$ définie sur un ensemble de mesures de probabilités lorsque $P = P_n$.

Définition 3.7. *On appelle fonction d'influence de T au point P la limite*

$$\lim_{u\to 0^+} \frac{T\left((1-u)P + u\delta_x\right) - T(P)}{u} =: I_P(x), \tag{3.32}$$

définie sur l'ensemble des x où cette limite existe.

Comme dans un certain nombre de cas l'expression de $T(P)$ fait intervenir F, la fonction de répartition de P, on écrira indifféremment $T(F)$ pour $T(P)$ et $I_F(x)$ pour $I_P(x)$.

L'existence de la limite dans (3.32) est la dérivabilité en zéro de $T\left((1-u)P + u\delta_x\right)$ considérée comme une fonction de u.

Exercice 3.13. Trouver la fonction d'influence de $T(P) = \int_{\mathbb{R}} g \,\mathrm{d}P$. Le cas particulier où g est l'identité fournit la fonction d'influence de la moyenne. ◁

Exercice 3.14. La moyenne α-tronquée correspond à la fonctionnelle

$$T(P) = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y\, dF(y) = \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(s)ds.$$

Ce n'est pas un M-estimateur. Quelle est sa fonction d'influence ? ◁

Au chapitre 4 on obtiendra l'expression de la fonction d'influence de toute une classe de M-estimateurs comme conséquence de la dérivabilité de la fonctionnelle correspondante, et on verra son rôle dans le développement au premier ordre de la fonctionnelle et dans la variance limite de l'estimateur.

La fonction d'influence peut être vue comme une évaluation quantitative de l'action sur la fonctionnelle T d'une faible contamination de la loi P par une masse de Dirac. Sa forme est en général indépendante de P.

Il est naturel de préférer une statistique à fonction d'influence bornée si on désire qu'elle résiste bien aux données aberrantes. Pour une telle statistique la fonction d'influence calculée lorsque P est remplacée par la mesure empirique P_n, indique l'action sur l'estimateur du déplacement en une valeur x fixée d'une faible proportion des observations.

3.4. Point de rupture. C'est un autre indice de robustesse. Considérons un estimateur $T(P_n)$. Le point de rupture de T en P indique jusqu'à quel point la mesure Q peut s'éloigner de la mesure de référence P en distance de Prokhorov sans que, sous la loi Q, la suite $T(P_n)$ cesse d'être bornée. Plus exactement

$$\varepsilon^* = \sup\{\varepsilon \leq 1 | \exists r_\varepsilon,\ \pi(P,Q) \leq \varepsilon \Rightarrow Q(|T(P_n)| \leq r_\varepsilon) \to 1 \text{ quand } n \to \infty\}.$$

Dans cette définition, π désigne la distance de Prokhorov et la notation $Q(|T(P_n)|)$ sous-entend que la probabilité est calculée sous l'hypothèse où les variables sont i.i.d selon la loi Q. En général, la valeur du point de rupture ne dépend pas de la mesure P au voisinage de laquelle on se place.

Deux exemples sont instructifs.

Exemple 3.4. *La moyenne.* Son point de rupture est nul en toute loi P telle que $\int |x| \, \mathrm{d}P(x) < +\infty$. Pour le voir, soit R la loi d'une variable aléatoire positive telle que $\int x \, \mathrm{d}R(x) = +\infty$. Considérons la loi $Q = (1-\varepsilon)P + \varepsilon R$. D'une part on a $\pi(P,Q) \leq \varepsilon$ et d'autre part, sous la loi Q, la suite $T(P_n) = \overline{X}_n$ converge presque sûrement vers $+\infty$. La conclusion est alors facile à tirer. $\triangleleft$

Exemple 3.5. *La médiane.* Le point de rupture de la médiane est égal à 1/2 en toute loi P dont la fonction de répartition F est bijective.

Soit d'abord $\varepsilon < 1/2$. Prenons

$$x_1 < F^{-1}\left(\frac{1}{2} - \varepsilon\right) - \varepsilon \quad \text{et} \quad x_2 > F^{-1}\left(\frac{1}{2} + \varepsilon\right) + \varepsilon.$$

Si $\pi(P,Q) \leq \varepsilon$, en notant G la fonction de répartition de Q, on a pour tout x

$$F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon$$

si bien que

$$G(x_1) < \frac{1}{2} \quad \text{et} \quad G(x_2) > \frac{1}{2}.$$

D'après le théorème de Glivenko-Cantelli, si les variables sont i.i.d selon la loi G, F_n converge (presque sûrement) uniformément vers G, et on a donc, pour n assez grand, presque sûrement,

$$F_n(x_1) < \frac{1}{2} \quad \text{et} \quad F_n(x_2) > \frac{1}{2}.$$

De ce fait la médiane de F_n est dans $[x_1, x_2]$. Ceci prouve que $\varepsilon^* \geq 1/2$.

Ensuite, prenons $\varepsilon > 1/2$ et considérons dans la boule de rayon ε les contaminées $G = (1-\varepsilon)P + \varepsilon R$. Pour tout x, $(1-\varepsilon)F(x) < 1/2$. Pour r fixé, choisissons $R = \delta_{2r}$. La médiane de G est égale à $2r$, ce qui veut dire que pour tout r on peut trouver dans la boule une loi dont la médiane est strictement supérieure à r. Il est alors aisé de conclure que $\varepsilon^* \leq 1/2$ ◁

Exercice 3.15. Montrer que le point de rupture de la moyenne α-tronquée est α. ◁

Exercice 3.16. Le point de rupture du M-estimateur de Huber correspondant à ψ_1 est $1/2$. ◁

4. Robustesse contre efficacité

Lorsque les variables sont i.i.d., et sous certaines conditions de régularité, on sait que la variance d'un estimateur sans biais de θ est limitée inférieurement par une quantité non nulle. Rappelons le théorème de Cramér-Rao (voir par exemple [15])

Théorème 3.8. *Supposons que le modèle est régulier. Soit $\hat{g}_n = \hat{g}_n(X_1, \cdots, X_n)$ un estimateur sans biais de $g(\theta)$. Supposons que g est dérivable sur Θ et que (dérivabilité sous le signe intégrale)*

$$\forall \theta \in \Theta, \quad \frac{\partial\, \mathbb{E}_\theta(\hat{g}_n)}{\partial \theta} = \frac{\partial\, \mathbb{E}_\mu(\hat{g}_n V_\theta)}{\partial \theta} = \mathbb{E}_\mu\left(\hat{g}_n \frac{\partial V_\theta}{\partial \theta}\right), \tag{3.33}$$

où l'on note $V_\theta := \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$. On dit qu'un tel estimateur est régulier.

Sous les conditions précédentes,

$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta\left(\hat{g}_n - g(\theta)\right)^2 \geq \frac{g'(\theta)^2}{nI(\theta)}. \tag{3.34}$$

Le membre de droite dans l'inégalité (3.34) s'appelle borne de Cramér-Rao. Un estimateur est dit efficace si sa variance est égale à cette borne. On dit qu'il est asymptotiquement efficace si $\sqrt{n}(\hat{g}_n - g(\theta))$ a une loi limite de variance $g'(\theta)^2 I(\theta)^{-1}$.

4.1. Efficacité asymptotique de l'estimateur du maximum de vraisemblance. Un corollaire immédiat du théorème 3.4 est que l'estimateur du maximum de vraisemblance est asymptotiquement efficace (ici $g(\theta) = \theta$), et qu'il en est de même pour les estimateurs introduits en section 2.3. L'estimateur de vraisemblance peut d'ailleurs être efficace à distance finie comme c'est le cas pour la moyenne arithmétique lorsque les variables sont i.i.d. selon une loi gaussienne dont on estime l'espérance.

4.2. Perte d'efficacité des estimateurs robustes. Posons le problème dans le cas particulier du paramètre de translation. La loi des variables X_n a pour fonction de répartition $F(x-\theta)$, et on estime θ. On a vu que, si les variables sont indépendantes, et sous quelques conditions de régularité, l'estimateur de maximum de vraisemblance est asymptotiquement efficace. Mais il n'est la plupart du temps pas robuste vis à vis d'écarts à la loi spécifiée. On peut se rappeler l'exemple de la moyenne dont le point de rupture est nul.

En revanche, les estimateurs robustes ne sont en général pas efficaces. Autrement dit, en passant de l'estimateur du maximum de vraisemblance à un estimateur robuste on troque une sorte d'optimalité sous la loi spécifiée contre une performance moins bonne mais plus résistante puisque valable même en dehors du modèle spécifié. Au statisticien de décider ce qu'il préfère.

4.2.1. *Perte d'efficacité en variables indépendantes.* En variables indépendantes, les M-estimateurs ne sont en général pas asymptotiquement efficaces. La perte d'efficacité peut s'évaluer par le quotient de la variance du M-estimateur (cf. th. 3.2) par la variance asymptotique de l'estimateur de θ par maximum de vraisemblance (cf. th. 3.4)) :

$$\text{Perte d'efficacité} = \frac{n^{-1}v_\infty}{n^{-1}I(\theta)^{-1}} = I(\theta)v_\infty$$

quantité qui, on va le voir dans la Proposition 3.9, est toujours supérieure ou égale à 1. On peut vérifier qu'ici $I(\theta)$ ne dépend pas de θ. En effet, la f.d.r. de P_θ est $F(\,.\,-\theta)$ et en supposant P_0 absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$, P_θ est de densité $f(x,\theta) = f(x-\theta)$. On a alors

$$\begin{aligned}I(\theta) = \mathbb{E}_\theta\left(\frac{\frac{\partial f}{\partial\theta}(X,\theta)}{f(X,\theta)}\right)^2 &= \int_{\mathbb{R}}\left(\frac{\partial}{\partial\theta}f(x-\theta)\right)^2\frac{1}{f(x-\theta)^2}f(x-\theta)\,\mathrm{d}x\\ &= \int_{\mathbb{R}}\frac{(-f'(x-\theta))^2}{f(x-\theta)^2}f(x-\theta)\,\mathrm{d}x\\ &= \int_{\mathbb{R}}\frac{(-f'(y))^2}{f(y)^2}f(y)\,\mathrm{d}y = \int_{\mathbb{R}}\frac{f'(y)^2}{f(y)}\,\mathrm{d}y.\end{aligned}$$

Après ce calcul, on peut réécrire la perte d'efficacité d'un M-estimateur d'un paramètre de position sous la forme :

$$\text{Perte d'efficacité} = \int\frac{f'(x)^2}{f^2(x)}\,\mathrm{d}F(x)\int\frac{\psi(x)^2}{\lambda'(0)^2}\,\mathrm{d}F(x).$$

Proposition 3.9. *On suppose que ψ est croissante et bornée sur $\mathbb{R}$, que la densité f est absolument continue et que le modèle est régulier. Alors*

$$v_\infty = \frac{\int\psi^2(s)\,\mathrm{d}F(s)}{{\lambda'_F}^2(0)} \geq I(\theta)^{-1}, \tag{3.35}$$

et l'égalité n'est atteinte que si la fonction ψ est proportionnelle à $f'f^{-1}$.

Preuve. La clé de la preuve est l'égalité (justifiée ci-dessous)

$$\lambda'(0) = \int_{\mathbb{R}}\psi(y)f'(y)\,\mathrm{d}y, \tag{3.36}$$

qui réécrite sous la forme $\int_{\mathbb{R}} f'(y)\psi(y)(\lambda'(0))^{-1}\,\mathrm{d}y = 1$, nous permet d'obtenir en appliquant l'inégalité de Cauchy-Schwarz dans $L^2(f(y)\,\mathrm{d}y)$:

$$1 = \left\{\int_{\mathbb{R}}\frac{f'(y)}{f(y)}\frac{\psi(y)}{\lambda'(0)}f(y)\,\mathrm{d}y\right\}^2 \leq \left\{\int_{\mathbb{R}}\frac{f'(y)^2}{f(y)^2}f(y)\,\mathrm{d}y\right\}\left\{\int_{\mathbb{R}}\frac{\psi(y)^2}{\lambda'(0)^2}f(y)\,\mathrm{d}y\right\} = I(\theta)v_\infty,$$

ce qui nous donne (3.35) et son cas d'égalité.

Pour vérifier (3.36), on commence par rappeler que :

$$\lambda(s) = \mathbb{E}_0\, \psi(X-s) = \int_{\mathbb{R}} \psi(x-s) f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \psi(y) f(y+s)\,\mathrm{d}y,$$

d'où

$$\forall s>0, \quad \frac{\lambda(s)-\lambda(0)}{s} = \int_R \psi(y) \frac{f(y+s)-f(y)}{s}\,\mathrm{d}y.$$

Comme λ est dérivable en zéro, sa dérivée en zéro s'obtient en faisant tendre s vers zéro (à droite) dans l'égalité ci-dessus. Or, f étant *absolument continue*, elle est dérivable presque partout (relativement à la mesure de Lebesgue sur $\mathbb{R}$) et vérifie $\int_{\mathbb{R}} |f(t)|\,\mathrm{d}t < +\infty$ et $f(y+s)-f(y) = \int_y^{y+s} f'(t)\,\mathrm{d}t$. On en déduit par une application immédiate du théorème de Fubini :

$$\begin{aligned}\frac{\lambda(s)-\lambda(0)}{s} &= \int_{\mathbb{R}^2} \psi(y) f'(t)\, \mathbb{I}_{[y,y+s]}(t)\,\mathrm{d}t\,\mathrm{d}y \\ &= \int_{\mathbb{R}} f'(t) \left\{ \int_{\mathbb{R}} \frac{1}{s} \mathbb{I}_{[t-s,t]}(y)\psi(y)\,\mathrm{d}y \right\} \mathrm{d}t \\ &= \int_{\mathbb{R}} f'(t) \left\{ \frac{1}{s} \int_{t-s}^{t} \psi(y)\,\mathrm{d}y \right\} \mathrm{d}t.\end{aligned}$$

Posons $\Psi(s,t) := \frac{1}{s}\int_{t-s}^t \psi(y)\,\mathrm{d}y$. Comme ψ est monotone, elle est continue presque partout sur $\mathbb{R}$, il est alors élémentaire de vérifier que pour presque tout t, $\lim_{s\to 0} \Psi(s,t) = \psi(t)$. En rappelant que $\int_R |f'(t)|\,\mathrm{d}t < +\infty$, une utilisation de la convergence dominée, dans l'intégrale relative à t, avec fonction dominante $|f'|\,\|\psi\|_\infty$ nous permet d'obtenir (3.36). $\square$

Exercice 3.17. La suite $(X_j)_{j\geq 1}$ étant i.i.d. gaussienne, d'espérance θ, calculer la perte d'efficacité subie lorsqu'on estime θ par la médiane. $\triangleleft$

Remarque 3.2. Bien sûr, la fonction $f'(x)f(x)^{-1}$ ne vérifie pas nécessairement les conditions du théorème 3.2, qui, soulignons le à nouveau, ne sont pas nécessaires. Par exemple, si f est la densité de la gaussienne standard, l'efficacité n'est obtenue que si $\psi(x) = \lambda x$ qui n'est pas bornée. On remarque cependant que l'estimateur lié à cette fonction est la moyenne dont on connaît les propriétés sans avoir recours au théorème 4.8.

4.2.2. *Perte d'efficacité en dépendance faible.* Dès que les variables ne sont plus indépendantes, la question est plus difficile puisqu'on ne connaît pas de borne inférieure pour la variance des estimateurs sans biais. C'est pourquoi nous nous contenterons d'effleurer l'exemple de l'estimation de l'espérance θ d'une loi gaussienne de variance 1. Il est assez difficile de savoir si la moyenne arithmétique est efficace. Par contre on sait en général comparer sa variance asymptotique à celle du « meilleur » estimateur linéaire sans biais (en anglais abrégé « B.L.U.E. » : best linear unbiased estimator), c'est à dire au barycentre des observations qui a la plus petite variance. Si la densité spectrale du processus (X_n) est continue et ne s'annule jamais, la moyenne arithmétique a la même variance asymptotique que le B.L.U.E. (voir par exemple [16]). Dans le cas qui nous interesse, le processus est gaussien. Sous les hypothèses du théorème 2.27, il est ϕ-mélangeant. Il en résulte que les variables sont m-dépendantes (voir en particulier [12]), ce qui implique que la densité spectrale est continue. Il est donc légitime, comme dans le cas indépendant, de comparer à 1 la variance limite $\operatorname{Var} X_0 + 2\sum_{j\geq 1} \operatorname{Cov}(X_0, X_j)$, mais la comparaison reste à faire.

4.2.3. *Perte d'efficacité en forte dépendance.* Sous les conditions du théorème 2.28, si la fonction Ψ satisfait aux conditions du théorème 4.8 la variance de la limite de l'estimateur normalisé $\sqrt{\frac{2n^\alpha}{(1-\alpha)(2-\alpha)}}(T(F_n) - T(F))$ est 1, donc il n'y a pas de perte d'efficacité des M-estimateurs par rapport à la moyenne arithmétique. Ce fait est étonnant. Par ailleurs, bien qu'en forte dépendance la moyenne arithmétique soit asymptotiquement moins efficace que le B.L.U.E, des travaux (voir par exemple [2]) montrent que la perte d'efficacité est assez faible. Si l'on s'en tient à ces résultats, on constate que , contrairement à ce qui se passe si les variables sont indépendantes, on ne perd pas beaucoup d'efficacité en utilisant un M-estimateur pour un paramètre de translation lorsque les variables sont fortement dépendantes. Il faut cependant noter que ce qui précède suppose que le paramètre α, caractéristique de la longue mémoire, est connu du statisticien, faute de quoi la statistique normalisée $\sqrt{\frac{2n^\alpha}{(1-\alpha)(2-\alpha)}}(T(F_n) - T(F))$ est incalculable. Et si on travaille avec $\sqrt{n}(T(F_n) - T(F))$, la perte d'efficacité peut être considérable.

CHAPITRE 4

Les delta-méthodes

1. Introduction

Partons de deux exemples.

Exemple 4.1. On dispose d'un n-échantillon d'une variable de Bernoulli de paramètre inconnu $p \in]0,1[$. On note S_n le nombre de 1 obtenus. On sait que, lorsque n tend vers l'infini, $\sqrt{n}\Big(S_n/n - p\Big)$ converge en loi vers une gaussienne centrée de variance $p(1-p)$. Comme, à son tour, cette variance peut être estimée de façon consistante par $\frac{S_n}{n}\Big(1 - \frac{S_n}{n}\Big)$, il est évidemment intéressant de connaître la loi limite de

$$\sqrt{n}\left(\frac{S_n}{n}\left(1 - \frac{S_n}{n}\right) - p(1-p)\right).$$

◁

Exemple 4.2. On considère une suite $(X_n)_{n\geq 1}$ de variables aléatoires i.i.d. Soit p un nombre entier tel que $\mathbb{E}\, X_1^{2p} < +\infty$. Il est facile de vérifier que $n^{-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^p$ converge presque sûrement vers $\mathbb{E}\Big(X_1 - \mathbb{E}(X_1)\Big)^p$. Comment trouver la limite en loi de la suite

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Big((X_j - \overline{X}_n)^p - \mathbb{E}\Big(X_1 - \mathbb{E}(X_1)\Big)^p\Big)?$$

◁

Les questions posées dans ces deux exemples se présentent sous la même forme. Trouver la loi limite de $\sqrt{n}\Big(g(U_n) - g(\mu)\Big)$ lorsqu'on connaît celle de $\sqrt{n}\Big(U_n - \mu\Big)$. Dans le deuxième exemple, on prendra pour U_n le vecteur de coordonnées $V_{k,n} = n^{-1}\sum_{j=1}^{n} X_j^k$, $k \in \{0, \dots, p\}$, $\mu = \mathbb{E}\, U_n$ et $g(t_0, t_1, \dots, t_p) = \sum_{k=0}^{p} C_p^k t_k t_1^{p-k}$ (vérifiez !).

Le problème est vite résolu si la fonction g est suffisamment régulière, car alors on peut approcher $g(U_n)$ par $g(\mu) + V(U_n - \mu)$ où V est la matrice jacobienne de g au point U. Le théorème qui suit précise les conditions d'application.

On note X^T la transposée du vecteur X.

Théorème 4.1 (Théorème de Cramér)**.** *Soit g de $\mathbb{R}^d$ dans $\mathbb{R}^k$ une fonction de classe $\mathcal{C}^1$ au voisinage de μ. Soit $(U_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires dans $\mathbb{R}^d$ telle que*

$$\sqrt{n}(U_n - \mu) \xrightarrow[n\to+\infty]{\text{loi}} U. \tag{4.1}$$

Alors

$$\sqrt{n}(g(U_n) - g(\mu)) \xrightarrow[n\to+\infty]{\text{loi}} (\nabla g(\mu))^T\, U, \tag{4.2}$$

où

$$\nabla g = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_k}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial g_1}{\partial x_d} & \cdots & \frac{\partial g_k}{\partial x_d} \end{pmatrix}.$$

L'examen de la preuve ci-dessous devrait convaincre le lecteur que la normalisation $\sqrt{n}$ dans l'hypothèse (4.2) pourrait être remplacée par n'importe quelle suite r_n de constantes strictement positives tendant vers l'infini.

Preuve. Dans cette démonstration, nous utiliserons 3 normes : pour l'espace de départ $\mathbb{R}^d$, l'espace d'arrivée $\mathbb{R}^k$ et la norme opérateur pour $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^k)$ l'espace des applications linéaires (automatiquement continues puisque la dimension est finie) de $\mathbb{R}^d \to \mathbb{R}^k$. Nous les noterons toutes $\|x\|$, la nature de x permettant de dissiper toute ambiguité quant à la norme concernée. La matrice jacobienne ∇g sera considérée comme application linéaire de $\mathbb{R}^d \to \mathbb{R}^k$.

D'abord on remarque que

$$U_n - \mu = \sqrt{n}(U_n - \mu)\frac{1}{\sqrt{n}} \xrightarrow[n\to+\infty]{\mathrm{P}} 0,$$

en effet la suite $(\sqrt{n}(U_n - \mu))_{n\geq 1}$ convergeant en loi est équitendue et donc en particulier stochastiquement bornée. Pour tout $\varepsilon > 0$, il existe un $r(\varepsilon) > 0$ tel que pour tout $n \geq 1$, $\mathbf{P}(n^{1/2}\|U_n - \mu\| > r(\varepsilon)) < \varepsilon$. Alors pour tout $\delta > 0$, $\mathbf{P}(\|U_n - \mu\| > \delta) = \mathbf{P}(n^{1/2}\|U_n - \mu\| > n^{1/2}\delta) < \varepsilon$ dès que $n^{1/2}\delta > r(\varepsilon)$.

Comme g est de classe $\mathcal{C}^1$ au voisinage de μ, il existe $\eta > 0$ tel que si $\|x - \mu\| \leq \eta$,

$$g(x) = g(\mu) + \left(\int_0^1 \nabla g(\mu + \theta(x-\mu))^T \,\mathrm{d}\theta\right)(x - \mu).$$

Sur l'ensemble $\{\|U_n - \mu\| \leq \eta\}$, on a donc

$$\sqrt{n}(g(U_n) - g(\mu)) = \left(\int_0^1 \nabla g(\mu + \theta(U_n-\mu))^T \,\mathrm{d}\theta\right)\sqrt{n}(U_n - \mu),$$

d'où la décomposition

$$\sqrt{n}(g(U_n) - g(\mu)) = \varphi_n\left(\sqrt{n}(U_n - \mu)\right) + R_n\,\mathbb{I}_{\|U_n-\mu\|>\eta}, \tag{4.3}$$

où φ_n est l'application linéaire aléatoire $\mathbb{R}^d \to \mathbb{R}^k$ définie par :

$$\varphi_n := \begin{cases} \displaystyle\int_0^1 \nabla g(\mu + \theta(U_n-\mu))^T \,\mathrm{d}\theta & \text{sur} \quad \{\|U_n - \mu\| \leq \eta\} \\ \nabla g(\mu)^T & \text{sur} \quad \{\|U_n - \mu\| > \eta\}. \end{cases}$$

et

$$R_n = \sqrt{n}(g(U_n) - g(\mu)) - \varphi_n\left(\sqrt{n}(U_n - \mu)\right).$$

La fonction

$$y \longmapsto \int_0^1 \nabla g(\mu + \theta y)^T \,\mathrm{d}\theta$$

est continue en zéro (par convergence dominée car g est de classe $\mathcal{C}^1$, donc ∇g est continue et bornée sur une boule fermée donc compacte de centre μ) et U_n converge en probabilité vers μ, donc

$$\varphi_n \xrightarrow[n\to+\infty]{\mathrm{P}} \nabla g(\mu)^T.$$

Notez que cette convergence en probabilité a lieu dans $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^k)$. Comme $\sqrt{n}(U_n - \mu)$ converge en loi vers U, le premier terme de (4.3) converge en loi vers $\nabla g(\mu)^T U$, en appliquant le résultat de l'exercice 4.1 ci-dessous.

Pour le second terme de (4.3), on vérifie sa convergence en probabilité vers 0 comme suit. Pour tout $\varepsilon > 0$,

$$\begin{aligned}\mathbf{P}\Big(\|R_n\| \ \mathbb{I}_{\|U_n-\mu\|\geq\eta} \geq \varepsilon\Big) &= \mathbf{P}\Big(\{\|R_n\| \geq \varepsilon\} \cap \{\|U_n - \mu\| \geq \eta\}\Big)\\ &\leq \mathbf{P}\Big(\|U_n - \mu\| \geq \eta\Big) \xrightarrow[n\to+\infty]{} 0,\end{aligned}$$

ce qui termine la preuve.

□

Exercice 4.1. Soient E et F deux espaces vectoriels normés, $\mathcal{L}(E, F)$ l'espace des applications linéaires *continues* $E \to F$, muni de la norme opérateur : $\|\varphi\| = \sup\{\|\varphi(x)\| ; \|x\| \leq 1\}$. On suppose que φ_n est une suite d'éléments aléatoires de $\mathcal{L}(E, F)$ qui converge en probabilité vers φ élément non aléatoire de $\mathcal{L}(E, F)$ et que la suite de vecteurs aléatoires $(Z_n)_{n\geq1}$ converge en loi vers Z dans E. Montrez qu'alors $\varphi(Z_n)$ converge en loi vers $\varphi(Z)$ dans F. ◁

Exercice 4.2. Si $(X_n)_{n\in\mathbb{N}}$ est une suite i.i.d de L^2, quelle est la loi limite de la suite $\sqrt{n}(\bar{X}_n^2 - \mathbb{E}(X_1)^2)$? ◁

Exercice 4.3. Traiter les exemples 4.1 et 4.2. ◁

2. Notions de dérivabilité directionnelle

Dans le contexte de la statistique classique où l'on ne s'intéresse qu'à la loi marginale des variables, une statistique peut souvent être vue comme la valeur pour la mesure empirique P_n (ou sa fonction de répartition F_n, ce qui revient au même) d'une fonctionnelle définie sur un ensemble de mesures. On l'écrira indifféremment $T(F_n)$ ou $T(P_n)$. Ainsi, l'application $T(P) = \int x \,\mathrm{d}P(x)$ définit la moyenne empirique $\overline{X}_n = n^{-1}\sum_{j=1}^n X_j$ tandis que la médiane empirique est définie par $T(P) = \inf\{x|F(x) \geq 1/2\}$.

Il est donc naturel de penser que, sous des conditions convenables de régularité sur T, la normalité asymptotique de $\sqrt{n}(F_n - F)$ va induire celle de $\sqrt{n}(T(F_n) - T(F))$.

Parmi les notions de dérivabilité que l'on rencontre dans la littérature, citons en trois. Soient D et E deux espaces vectoriels normés et T une application de $D_T \subset D$ dans E.

Définition 4.2. Dérivabilité au sens de Gateaux.

T est dérivable au sens de Gateaux au point $F \in D_T$ si pour tout $h \in D$ pour lequel $F + th$ appartient à D_T si $t \in \mathbb{R}$ est assez petit, il existe $\psi_h \in E$ tel que

$$\left\| \frac{T(F + th) - T(F)}{t} - \psi_h \right\|_E \xrightarrow[t \to 0]{} 0.$$

On peut remarquer que cette définition n'est autre que celle de la dérivabilité à l'origine de l'application de $\mathbb{R}$ dans E qui à t associe $T(F + th)$, h étant fixé. En fait, si $D = \mathbb{R}$, c'est la notion habituelle de dérivabilité. En dehors de ce cas, c'est une notion de dérivabilité directionnelle, qui, comme on le sait, n'implique même pas la continuité de T. On le voit par exemple en considérant la fonction

$$\Phi(x, y) = \frac{x^2 + y^2}{x} \text{ si } x \neq 0, \quad \Phi(0, y) := 0.$$

Définition 4.3. Dérivabilité au sens de Fréchet.

L'application T est dérivable au sens de Fréchet au point $F \in D_T$ si $F + h \in D_T$ pour $\| h \|_D$ assez petit et si il existe une application linéaire continue *T'_F telle que*

$$\|T(F + h) - T(F) - T'_F(h)\|_E = o(\| h \|_D).$$

Cette notion est plus forte que la précédente, mais souvent trop forte pour être vérifiée par les fonctionnelles utiles en statistique, comme on le verra plus loin.

Définition 4.4. Dérivabilité au sens d'Hadamard.

L'application T est dérivable au sens d'Hadamard au point $F \in D_T$ si il existe une application linéaire continue T'_F de D dans E telle que, pour tout $h \in D$ et pour toutes suites $h_n \in D$ et $u_n \to 0$ telles que $h_n \to h$ et $F + u_n h_n \in D_T$ pour n assez grand, on ait

$$\left\| \frac{T(F + u_n h_n) - T(F)}{u_n} - T'_F(h) \right\|_E \xrightarrow[n \to +\infty]{} 0.$$

Cette définition, plus forte que la dérivabilité au sens de Gateaux, est plus faible que celle au sens de Fréchet. Elle lui est en fait équivalente lorsque dans D la boule unité fermée est compacte[1]. Par exemple si $D = \mathbb{R}^k$, elles sont toutes deux équivalentes à la différentiabilité usuelle.

Parfois on peut se contenter d'une définition un peu plus faible qui est la dérivabilité tangentiellement à un sous espace.

Définition 4.5. *L'application T est dérivable au sens d'Hadamard au point $F \in D_T$, tangentiellement à un sous-ensemble D_0 de D si dans la définition qui précède on suppose seulement $h \in D_0$, la dérivée $T'_F(h)$ pouvant alors n'être définie que sur cet ensemble.*

On verra quelques exemples dans la suite de ce chapitre. La notion de dérivabilité au sens d'Hadamard est d'autant plus intéressante que, bien que relativement faible, elle maintient tout de même la validité du théorème de dérivation en chaîne, ce que le lecteur vérifiera lui même.

1. Autrement dit, lorsque D est de dimension finie.

Proposition 4.6. *Soit T_1 une application de D_{T_1} dans E_1 et T_2 une application de $D_{T_2} \subset E_1$ dans E_2. Supposons T_1 Hadamard-dérivable au point $F \in D_{T_1}$ tangentiellement à D_0 et notons $T'_{1,F}$ sa dérivée. Si T_2 est Hadamard-dérivable au point $T_1(F)$ tangentiellement à $T'_{1,F}(D_0)$, sa dérivée étant notée $T'_{2,T_1(F)}$, alors $T_2 \circ T_1$ est Hadamard-dérivable au point F, tangentiellement à D_0, et sa dérivée est*

$$(T_2 \circ T_1)'_F = T'_{2,T_1(F)} \circ T'_{1,F}.$$

3. La delta-méthode fonctionnelle

Le résultat qui suit est la généralisation naturelle du théorème 4.1. Soient D et E deux espaces vectoriels normés et T une application de $D_T \subset D$ dans E.

Théorème 4.7. *Soit $\Phi \in D_T$, (Φ_n) une suite d'éléments aléatoires à valeurs dans D_T et une suite $r_n \to \infty$ telles que la suite $r_n(\Phi_n - \Phi)$ converge en loi vers Z, élément aléatoire à valeurs dans $D_0 \subset D$. Si l'application T est dérivable au sens d'Hadamard au point Φ tangentiellement à D_0, alors la suite*

$$r_n\Big(T(\Phi_n) - T(\Phi)\Big)$$

converge en loi vers $T'_\Phi(Z)$.

Preuve. Pour tout n, la fonction $g_n(h) = r_n\left(T(\Phi + h/r_n) - T(\Phi)\right)$, est définie sur $D_n = \{h | \Phi + h/r_n \in D_T\}$. Par définition de la dérivabilité au sens d'Hadamard, pour toute suite $h_n \in D_n$ qui converge vers un élément h de D_0, $g_n(h_n) \to T'_\Phi(h)$. Appliquant alors le théorème 1.12 de transport de la convergence en loi, on obtient le résultat annoncé en choisissant h_n et h *aléatoires* comme suit : $h_n = r_n(\Phi_n - \Phi)$ et $h = Z$ puisqu'alors $\Phi_n = \Phi + h_n/r_n$ et $r_n\Big(T(\Phi_n) - T(\Phi)\Big) = g_n(h_n)$. □

4. Application aux M-estimateurs

Le cas le plus simple est celui de l'estimation d'un paramètre de position unidimensionnel. On suppose que les variables $(X_j)_{j\geq 1}$ ont la même loi, de fonction de répartition $F(x-\theta)$, et on estime θ par (3.6). On a vu que cet estimateur est « raisonnable » lorsque $\mathbb{E}_\theta(\psi(X-\theta)) = 0$. On commence par étudier la dérivabilité de la fonctionnelle qui définit le M-estimateur, puis, comme on l'a fait pour les quantiles, on en déduit grâce au théorème 4.7 la convergence en loi du M-estimateur dans des situations qui ne recouvrent pas tout à fait celles du chapitre 2.

4.1. Dérivabilité de la fonctionnelle. Considérons l'opérateur T défini par

$$G \longmapsto T(G) = \inf\{t |\ \lambda_G(t) \leq 0\}, \quad \text{où} \quad \lambda_G(t) = \mathbb{E}_G\, \psi(X-t). \tag{4.4}$$

On commence par préciser les hypothèses permettant de définir $T(G)$ pour toute fonction de répartition G. La définition ci-dessus nous permet seulement de considérer $T(G)$ comme un élément de $\overline{\mathbb{R}}$. Une condition suffisante pour que $T(G) \in \mathbb{R}$ est que λ_G prenne des valeurs strictement négatives (ce qui implique $T(G) < +\infty$ et des valeurs strictement positives (ce qui implique $T(G) > -\infty$). Supposons que ψ est croissante et bornée sur $\mathbb{R}$, qu'elle prend des valeurs strictement négatives et des valeurs strictement positives. On a alors

$$-\infty < \lim_{s\to-\infty} \psi(s) =: \ell^- < 0 \quad \text{et} \quad 0 < \lim_{s\to+\infty} \psi(s) =: \ell^+ < +\infty.$$

Comme ψ est bornée, $\mathbb{E}_G\,\psi(X-t)$ est bien défini (comme nombre réel) pour tout $t \in \mathbb{R}$ et ce, quelle que soit la loi G de la variable aléatoire X. C'est une fonction décroissante de t. Par convergence monotone (ou dominée par $\|\psi\|_\infty$), on a

$$\lim_{t\to+\infty} \mathbb{E}_G\,\psi(X-t) = \mathbb{E}_G\,\ell^- = \ell^- < 0 \quad \text{et} \quad \lim_{t\to-\infty} \mathbb{E}_G\,\psi(X-t) = \mathbb{E}_G\,\ell^+ = \ell^+ > 0.$$

Ceci montre en particulier que l'infimum $T(G)$ est fini.

Pour étudier la dérivabilité de T en tout point F, où F est une fonction de répartition, il nous faut d'abord vérifier qu'on peut prolonger T en une fonctionnelle définie sur une boule de centre F de rayon assez petit dans l'espace $D(\mathbb{R})$ des fonctions càdlàg bornées, muni de la norme $\| \ \|_\infty$. Pour cela, nous commençons par une formule qui nous permettra de définir λ_G pour toute $G \in D(\mathbb{R})$.

Supposons désormais que ψ est en plus *absolument continue* sur $\mathbb{R}$ et montrons qu'alors pour toute fonction de répartition G,

$$\lambda_G(t) := \mathbb{E}_G\,\psi(X-t) = \ell^+ - \int_{\mathbb{R}} \psi'(x-t)G(x)\,\mathrm{d}x. \tag{4.5}$$

Puisque ψ est absolument continue, elle continue sur $\mathbb{R}$, presque-partout dérivable [2] sur $\mathbb{R}$, sa dérivée ψ' est Lebesgue intégrable sur $\mathbb{R}$ ($\int_{\mathbb{R}} |\psi'(y)|\,\mathrm{d}y < +\infty$) et pour tous réels $u < v$, $\psi(v) - \psi(u) = \int_{]u,v]} \psi'(y)\,\mathrm{d}y$. En particulier, on dispose de la représentation : $\psi(x-t) = \ell^- + \int_{]-\infty,x]} \psi'(y-t)\,\mathrm{d}y$. La vérification de la formule d'intégration par parties (4.5) repose alors sur le théorème de Fubini.

$$\begin{aligned}
\mathbb{E}_G\,\psi(X-t) = \int_{\mathbb{R}} \psi(x-t)\,\mathrm{d}G(x) &= \int_{\mathbb{R}} \left\{ \ell^- + \int_{]-\infty,x]} \psi'(y-t)\,\mathrm{d}y \right\} \mathrm{d}G(x) \\
&= \ell^- + \int_{\mathbb{R}} \int_{\mathbb{R}} \psi'(y-t)\,\mathbb{I}_{]-\infty,x]}(y)\,\mathrm{d}y\,\mathrm{d}G(x) \\
&= \ell^- + \int_{\mathbb{R}} \psi'(y-t) \left\{ \int_{\mathbb{R}} \mathbb{I}_{[y,+\infty[}(x)\,\mathrm{d}G(x) \right\} \mathrm{d}y \\
&= \ell^- + \int_{\mathbb{R}} \psi'(y-t)\Big(1 - G(y^-)\Big)\,\mathrm{d}y \\
&= \ell^- + (\ell^+ - \ell^-) - \int_{\mathbb{R}} \psi'(y-t)G(y^-)\,\mathrm{d}y \\
&= \ell^+ - \int_{\mathbb{R}} \psi'(y-t)G(y)\,\mathrm{d}y,
\end{aligned}$$

en notant que l'ensemble des discontinuités de la f.d.r. G est au plus dénombrable et donc $G(y^-) = G(y)$ presque-partout.

2. En fait toute fonction monotone est presque-partout dérivable sur $\mathbb{R}$ d'après un théorème de Lebesgue.

La formule (4.5) nous permet de *redéfinir* λ_G *pour toute fonction* $G \in D(\mathbb{R})$ par

$$\lambda_G(t) := \ell^+ - \int_{\mathbb{R}} \psi'(x-t)G(x)\,\mathrm{d}x, \tag{4.6}$$

cette définition étant compatible grâce à (4.5) avec la définition initiale de $\lambda_G(t)$ par $\mathbb{E}_G\,\psi(X-t)$ lorsque G est une f.d.r. En effet puisque $\psi' \in L^1(\mathbb{R},\,\mathrm{d}x)$ et G est bornée sur $\mathbb{R}$ comme élément de $D(\mathbb{R})$, l'intégrale figurant dans (4.6) est bien définie. Notons que cette nouvelle fonction λ_G reste *continue*. Pour le vérifier, rappelons que les translations opèrent continûment sur $L^1(\mathbb{R},\,\mathrm{d}x)$: ce qui signifie que si $f \in L^1(\mathbb{R},\,\mathrm{d}x)$, $\lim_{u\to 0}\int |f(x+u)-f(x)|\,\mathrm{d}x \to 0$. On applique ceci avec $f = \psi'(\,.\,-t)$ et $u = s-t$, après avoir noté que

$$|\lambda_G(t) - \lambda_G(s)| \le \|G\|_\infty \int_{\mathbb{R}} |\psi'(x-t) - \psi'(x-s)|\,\mathrm{d}x.$$

Par contre, il n'y a aucune raison que λ_G conserve sa monotonie du cas où G est une f.d.r.

Passons au prolongement de la fonctionnelle T, obtenu en posant :

$$\begin{aligned}\forall G \in D(\mathbb{R}),\quad T(G) &= \inf\{t\in\mathbb{R}|\ \lambda_G(t)\le 0\}\\ &= \inf\left\{t\in\mathbb{R}|\ \int_{\mathbb{R}}\psi'(x-t)G(x)\,\mathrm{d}x \ge \ell^+\right\}.\end{aligned} \tag{4.7}$$

Pour l'instant, ceci ne définit $T(G)$ que comme un élément de $\overline{\mathbb{R}}$. Une condition suffisante pour que $T(G)$ soit un réel est que λ_G prenne des valeurs strictement positives et des valeurs strictement négatives. Nous allons voir que cette condition est réalisée dans toute boule centrée sur une f.d.r. F et de rayon assez petit. En effet,

$$\begin{aligned}\lambda_{F+h}(t) :=& \ell^+ - \int \psi'(x-t)\left(F(x)+h(x)\right)\,\mathrm{d}x\\ =&\lambda_F(t) - \int \psi'(x-t)h(x)\,\mathrm{d}x\\ =&\lambda_F(t) + O(\|h\|_\infty),\end{aligned} \tag{4.8}$$

car $\int_{\mathbb{R}} |\psi'(x-t)|\,\mathrm{d}x = \int_{\mathbb{R}} |\psi'(x)|\,\mathrm{d}x < +\infty$. Ceci nous montre que si

$$\|h\|_\infty < r := \frac{\min(\ell^+, |\ell^-|)}{2\,\|\psi'\|_1}, \tag{4.9}$$

la fonction λ_{F+h} prend des valeurs strictement positives et des valeurs strictement négatives. Ainsi $T(G)\in\mathbb{R}$, pour tout $G \in B(F,r)$. Enfin, la continuité de λ_G nous permet de voir que pour toute $G \in B(F,r)$, $\lambda_G(T(G)) = 0$. En effet, en notant que $T(G) > -\infty$, pour tout $t < T(G)$, $\lambda_G(t) > 0$, donc $\lambda_G(T(G)) = \lim_{t\to T(G)^-}\lambda_G(t) \ge 0$; d'autre part puisque $T(G) < +\infty$, il existe une suite $t_n \downarrow T(G)$ telle que pour tout n, $\lambda_G(t_n) \le 0$, d'où $\lambda_G(T(G)) = \lim_{n\to+\infty}\lambda_G(t_n) \le 0$.

Théorème 4.8. *Supposons* ψ *absolument continue, bornée, croissante et impaire. Soit* F *une fonction de répartition telle que* λ_F *ait une dérivée strictement négative au point*

$t = T(F)$. L'application T définie par (4.7) sur une boule de $D(\mathbb{R})$ centrée en F est Fréchet-dérivable au point F et sa dérivée est :

$$T'_F(h) = \frac{\int_{\mathbb{R}} \psi'(y - T(F))h(y)\,\mathrm{d}y}{\lambda'_F(T(F))}.$$

Preuve. Remarquons d'abord que T'_F définie ci-dessus est une forme linéaire continue sur $L^\infty(\mathbb{R})$ en raison de l'appartenance à L^1 de ψ'. Il s'agit de prouver que

$$|T(F+h) - T(F) - T'_F(h)| = o(\|h\|_\infty).$$

On impose d'emblée $\|h\|_\infty < r$, où r est donné par (4.9), ce qui implique que $\lambda_F(T(F))$ et $\lambda_{F+h}(T(F+h))$ sont nuls. Compte-tenu de la nullité de $\lambda_F(T(F))$, la dérivabilité de λ_F au point $T(F)$ peut s'écrire sous la forme :

$$\forall t \in \mathbb{R}, \quad \lambda_F(t) = (t - T(F))\Big(\lambda'_F(T(F)) + \eta(t - T(F))\Big), \quad \text{avec} \quad \lim_{u\to 0} \eta(u) = 0.$$

En appliquant ceci avec $t = T(F+h)$, en reportant dans (4.8), on obtient :

$$0 = \lambda_{F+h}(T(F+h)) = (T(F+h) - T(F))\Big((\lambda'_F(T(F)) + \eta(T(F+h) - T(F))\Big) + O(\|h\|_\infty)$$

Comme de plus $\lambda'_F(T(F)) \neq 0$, on en déduit que

$$T(F+h) - T(F) = O(\|h\|_\infty), \tag{4.10}$$

dès que l'on sait que T est continue au point F. Pour vérifier cette continuité, on déduit de (4.8) que $\lambda_F(T(F+h)) = O(\|h\|_\infty)$ et la continuité et la monotonie de λ_F obligent $T(F+h)$ à tendre vers $T(F)$, *unique zéro*[3] de λ_F, quand h tend vers 0 (écrivez les détails).

Par définition de T, pour tout $\varepsilon > 0$,

$$\lambda_{F+h}(T(F+h) - \varepsilon) > 0 \quad \text{et} \tag{4.11}$$

$$\exists \varepsilon' \in]0, \varepsilon[, \quad \lambda_{F+h}(T(F+h) + \varepsilon') \leq 0. \tag{4.12}$$

En rappelant que $\lambda_{F+h}(t) = \lambda_F(t) - \int_{\mathbb{R}} \psi'(x-t)h(x)\,\mathrm{d}x$, on déduit de (4.11) que

$$\lambda_F(T(F+h) - \varepsilon) > \int_{\mathbb{R}} \psi'(x - T(F+h) + \varepsilon)h(x)\,\mathrm{d}x,$$

puis en utilisant la dérivabilité de λ_F au point $T(F)$:

$$\Big(T(F+h) - \varepsilon - T(F)\Big)\Big(\lambda'_F(T(F)) + \eta(T(F+h) - \varepsilon - T(F))\Big) > \int_{\mathbb{R}} \psi'(x - T(F+h) + \varepsilon)h(x)\,\mathrm{d}x.$$

Choisissant désormais $\varepsilon = o(\|h\|_\infty)$ on en déduit grâce à (4.10) :

$$\lambda'_F(T(F))\Big(T(F+h) - \varepsilon - T(F)\Big) > \int_{\mathbb{R}} \psi'(x - T(F+h) + \varepsilon)h(x)\,\mathrm{d}x + o(\|h\|_\infty), \tag{4.13}$$

puis comme $\lambda'_F(T(F)) < 0$,

$$T(F+h) - T(F) < \frac{\int_{\mathbb{R}} \psi'(x - T(F+h) + \varepsilon)h(x)\,\mathrm{d}x}{\lambda'_F(T(F))} + o(\|h\|_\infty). \tag{4.14}$$

3. L'unicité résulte de la monotonie de λ_F combinée avec la non nullité de $\lambda'_F(T(F))$.

En utilisant la continuité L^1 de ψ',

$$\left|\int h(x)\psi'(x-T(F+h)+\varepsilon)\,\mathrm{d}x - \int h(x)\psi'(x-T(F))\,\mathrm{d}x\right|$$
$$\leq \|h\|_\infty \int |\psi'(x-T(F+h)+\varepsilon)-\psi'(x-T(F))|\ \mathrm{d}x = o(\|h\|_\infty).$$

En reportant cette estimation dans (4.14),

$$T(F+h)-T(F) < \frac{\int_{\mathbb{R}} \psi'(x-T(F))h(x)\,\mathrm{d}x}{\lambda_F'(T(F))} + o(\|h\|_\infty)$$

D'où

$$T(F+h)-T(F)-T_F'(h) \leq o(\|h\|_\infty).$$

Reprenant la même ligne d'arguments, à partir de (4.12), en notant que le choix $\varepsilon = o(\|h\|_\infty)$ fait ci-dessus entraîne $\varepsilon' = o(\|h\|_\infty)$, on aboutit à l'inégalité dans l'autre sens, ce qui complète la preuve. □

Remarque 4.1. Il est possible parfois de s'affranchir de l'hypothèse de monotonie de ψ. On utilise ainsi la fonction ψ impaire définie sur $\mathbb{R}^+$ par

$$\psi(x) = \left\{ \begin{array}{ll} x & \text{si } x \in [0,1] \\ 1 & \text{si } x \in [1,2] \\ 3-x & \text{si } x \in [2,3] \\ 0 & \text{si } x > 3 \end{array} \right\}.$$

L'estimateur $T(F_n)$ (dit de Hampel) construit sur cette fonction, est défini comme étant le t_0 le plus proche de zéro tel que $\mathbb{E}_{F_n}(\psi(X-t)) = 0$. Si F est la fonction de répartition d'une loi paire, $T(F) = 0$. Si en plus $\mathbb{E}_F(\psi(X-t))$ est strictement monotone au voisinage de zéro, il n'est pas difficile de voir que $T(F+h)$ est parfaitement défini pour $\|h\|_\infty$ assez petit. La preuve du théorème 4.8 s'adapte ensuite assez facilement à la nouvelle fonctionnelle. ◁

4.2. Indépendance et dépendance faible. Le théorème 4.8 permet d'obtenir la limite en loi des M-estimateurs. Sous les conditions du théorème 2.27, en supposant pour simplifier que $T(F) = 0$,

$$\sqrt{n}\,T(F_n) \xrightarrow[n\to+\infty]{\text{loi}} \frac{1}{\lambda_F'(0)} \int_{\mathbb{R}} \psi'(s)Z(s)\,\mathrm{d}s, \tag{4.15}$$

où Z est un processus gaussien centré dont la covariance $\Gamma(s,t)$ est donnée par (2.32).

Il est facile de voir que l'intégrale dans (4.15) est une variable gaussienne centrée. Sa variance est

$$\begin{aligned} \mathbb{E}\left(\int \psi'(s)Z(s)\,\mathrm{d}s\right)^2 &= \mathbb{E}\left(\int \psi'(s)\psi'(t)Z(s)Z(t)\,\mathrm{d}s\mathrm{d}t\right) = \int \psi'(s)\psi'(t)\,\mathbb{E}(Z(s)Z(t))\,\mathrm{d}s\mathrm{d}t \\ &= \int \psi'(s)\psi'(t)\Gamma(s,t)\,\mathrm{d}s\mathrm{d}t. \end{aligned}$$

Notamment, si les variables sont indépendantes,

$$\operatorname{Var}\Big(\int \psi'(s)Z(s)\,\mathrm{d}s\Big) = \int \psi'(s)\psi'(t)\Big(F(\min(s,t)) - F(s)F(t)\Big)\,\mathrm{d}s\mathrm{d}t.$$

Pour retrouver l'expression (3.11) de la variance, remarquons que

$$\begin{aligned}\int \psi'(s)\psi'(t)F(\min(s,t))\,\mathrm{d}s\mathrm{d}t &= 2\int \psi'(s)F(s)\left(\int_{s<t}\psi'(t)\,\mathrm{d}t\right)\mathrm{d}s\\ &= \int(\ell^+ - \psi(s))^2 dF(s) + \Big[(\ell^+ - \psi(s))^2F(s)\Big]_{-\infty}^{\infty}\\ &= \int \psi(s)^2 dF(s) + (\ell^+)^2.\end{aligned}$$

Puis

$$\int \psi'(s)\psi'(t)F(s)F(t)\,\mathrm{d}s\mathrm{d}t = \Big(\int \psi'(s)F(s)\,\mathrm{d}s\Big)^2 = (\ell^+)^2,$$

ce qui conduit immédiatement à (3.11), c'est à dire à

$$\operatorname{Var}\left(\frac{\int \psi'(s)Z(s)\,\mathrm{d}s}{\lambda'_F(0)}\right) = \frac{\int \psi^2(s)\,\mathrm{d}F(s)}{{\lambda'_F}^2(0)}. \tag{4.16}$$

4.3. Retour sur la fonction d'influence. Il est facile de voir que la dérivabilité au sens d'Hadamard de la fonctionnelle T au point F implique que sa fonction d'influence en ce point existe et est égale à

$$I_F(x) = T'_F(\mathbb{I}_{[x,\infty[} - F). \tag{4.17}$$

Exercice 4.4. En supposant que ψ est partout continue à droite et en supposant pour simplifier $T(F) = 0$ montrer que la fonction d'influence du M-estimateur lié à ψ est

$$I_F(x) = -\frac{\psi(x)}{\lambda'_F(0)},$$

et que, de ce fait la dérivabilité Frechet de T s'écrit aussi, lorsque G est une fonction de répartition,

$$T(G) - T(F) = \int I_F(x)\,\mathrm{d}(G-F)(x)\,\mathrm{d}x + o(\|G-F\|_\infty), \tag{4.18}$$

ce qui met bien en évidence le rôle de la fonction d'influence dans le développement à l'ordre 1 de la fonctionnelle. ◁

Par ailleurs, il est important de noter que, dans les cas les plus réguliers, la variance limite s'exprime à travers cette fonction d'influence :

$$v_\infty = \int I_F(x)^2\,\mathrm{d}F(x). \tag{4.19}$$

On le vérifie aisément pour les M-estimateurs. Ce sera aussi vrai pour les quantiles et la moyenne α-tronquée qui seront étudiés au chapitre 5.

L'expression (4.19) s'explique heuristiquement (et sans trop de souci de justification) de la façon suivante : l'expression (4.18) découle par intégration de (4.17) en remarquant que

$$G(y) = \int_{\mathbb{R}} \mathbb{I}_{[x,\infty[}(y)\,\mathrm{d}G(x) \text{ et } \int I_F(x) dF(x) = T_F'(0) = 0.$$

En remplaçant G par

$$\hat{F}_n = F + \sqrt{n}\frac{\hat{F}_n - F}{\sqrt{n}} = F + \frac{\tilde{F}_n}{\sqrt{n}},$$

on obtient

$$\sqrt{n}(T(\hat{F}_n) - T(F)) = \int I_F(x)\,\mathrm{d}\tilde{F}_n(x) + \text{reste} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} I_F(X_j) + \text{reste},$$

et sous l'hypothèse de variables i.i.d $\frac{1}{\sqrt{n}}\sum_{j=1}^{n} I_F(X_j)$, converge vers une gaussienne centrée de variance (4.19). On voit que la delta-méthode constitue simplement une justification rigoureuse de ces considérations imprécises. Mais tout imprécises qu'elles soient, elles sont utiles car elles mettent en lumière le rôle de la fonction d'influence dans le développement de la fonctionnelle et dans la variance asymptotique de l'estimateur.

CHAPITRE 5

Les quantiles

La delta-méthode fonctionnelle est particulièrement adaptée à l'étude asymptotique des quantiles et des statistiques dérivées de la fonction quantile.

1. Définition

Pour $p \in]0,1[$, le quantile d'ordre p de la loi P dont F est la fonction de répartition est défini par

$$F^{-1}(p) = \inf\{x | F(x) \geq p\}. \tag{5.1}$$

Le quantile d'ordre 1/2 est bien entendu la (plus petite des) médiane(s).

2. Quelques propriétés élémentaires

Leur preuve est laissée au lecteur.

Tout d'abord, quelques propriétés de l'inverse généralisée F^{-1}, fonction définie sur $]0,1[$ par (5.1), en tout point continue à gauche et à valeurs dans le support de F.

(1) $F^{-1}(s) \leq x$ ssi $s \leq F(x)$
(2) $F \circ F^{-1}(s) \geq s$, avec égalité si et seulement si F est continue au point $F^{-1}(s)$
(3) $F_{-} \circ F^{-1}(s) \leq s$
(4) $F^{-1} \circ F(x) \leq x$
(5) $F^{-1} \circ F \circ F^{-1} = F^{-1}$ et $F \circ F^{-1} \circ F = F$
(6) $(F \circ G)^{-1} = G^{-1} \circ F^{-1}$

Puis, présentons la transformation quantile qui permet de passer de la loi uniforme à une loi donnée de fonction de répartition F.

(1) Si la variable U a une loi uniforme sur $[0,1]$ alors la variable $F^{-1}(U)$ a comme loi F.
(2) Si F est partout continue, $F(X)$ est uniformément distribuée sur $[0,1]$ si et seulement si X a comme loi F.

Exercice 5.1. Soit F une fonction de répartition. Déduire de ce qui précède que pour toute g bornée

$$\int_0^1 g \circ F^{-1}(s)\,\mathrm{d}s = \int_{\mathbb{R}} g(x)\,\mathrm{d}F(x)$$

◁

Soit maintenant G_n une suite de fonctions de repartition. En utilisant la transformation quantile, on montre facilement que G_n converge vers la fonction de répartition G en tout point de continuité de cette dernière si et seulement si G_n^{-1} converge vers G^{-1} de la même façon. En appliquant le théorème de Glivenko-Cantelli, on déduit que, si la suite strictement stationnaire X_n est ergodique, le quantile empirique $F_n^{-1}(p)$ converge presque sûrement vers $F^{-1}(p)$ dès que F n'a pas de palier d'ordonnée p (comparer avec le théorème 3.1).

3. Propriétés asymptotiques des quantiles.

Pour étudier la convergence en loi des quantiles, étudions l'application correspondante. On remarque que, si G est une fonction de répartition, le quantile d'ordre p est bien défini pour toute fonction H telle que $\|G-H\|_\infty < \min\{p, 1-p\}$. En effet, pour une telle fonction il existe x_0 et x_1 tels que

$$H(x) < p \text{ si } x \le x_0 \quad \text{et} \quad H(x) > p \text{ si } x \ge x_1,$$

ce qui implique que le quantile d'ordre p de H existe et est situé dans $[x_0, x_1]$. On notera D_p l'ensemble des fonctions de $D(\mathbb{R})$ à distance (pour la norme du sup) strictement inférieure à $\min\{p, 1-p\}$ d'une fonction de répartition.

Théorème 5.1. *Si F est une fonction de répartition prenant la valeur p en un point unique $\xi_0 := F^{-1}(p)$ où elle est dérivable et de dérivée non nulle, alors l'application quantile T_p définie sur D_p est Hadamard-dérivable au point F tangentiellement au sous-ensemble C de $D_0(\mathbb{R})$ constitué des fonctions h continues au point $F^{-1}(p)$. Sa dérivée est*

$$T'_{p,F}(h) = -\frac{h(F^{-1}(p))}{F'(F^{-1}(p))}.$$

Preuve. Soit u_n une suite convergeant vers zéro. Remarquons que, puisque $\|h_n - h\|_\infty \to 0$, la suite h_n est uniformément bornée et la fonction $G_n := F + u_n h_n$ est dans D_p pour n assez grand. Soit ξ_n son quantile d'ordre p. Pour tout $\varepsilon > 0$

$$G_n(\xi_n - \varepsilon) < p \le G_n(\xi_n). \tag{5.2}$$

Donc, pour tout $\varepsilon > 0$

$$F(\xi_n - \varepsilon) - c|u_n| < p \le F(\xi_n) + c|u_n|, \tag{5.3}$$

Soit

$$F(\xi_n - \varepsilon) < p + c|u_n| \quad \text{et} \quad F(\xi_n) \ge p - c|u_n|,$$

qui, grâce à la propriété (1) de la fonction F^{-1} entraîne

$$F^{-1}(p - c|u_n|) \le \xi_n \le \varepsilon + F^{-1}(p + c|u_n|)$$

Mais F^{-1} est développable au point p, ce qui donne

$$\xi_0 - c\frac{|u_n|}{F'(\xi_0)} + o(u_n) \le \xi_n \le \varepsilon + \xi_0 + c\frac{|u_n|}{F'(\xi_0)} + o(u_n).$$

Il suffit alors de prendre $\varepsilon = O(u_n)$ pour voir que $\xi_n - \xi_0 = O(u_n)$. On reprend alors (5.3), en développant F au voisinage de ξ_0, et on obtient, puisque $F(\xi_0) = p$,

$$(\xi_n-\xi_0-\varepsilon)F'(\xi_0)+o(\xi_n-\xi_0-\varepsilon)+u_nh_n(\xi_n-\varepsilon) < 0 \leq (\xi_n-\xi_0-\varepsilon)F'(\xi_0)+o(\xi_n-\xi_0)+u_nh_n(\xi_n).$$

Pour finir, $h_n(x) = h(x) + R_n(x)$, où $R_n(x)$ tend vers zéro uniformément en x. En prenant maintenant $\varepsilon = o(u_n)$ on obtient

$$(\xi_n - \xi_0)F'(\xi_0) + u_nh(\xi_n) + o(u_n) < 0 \leq (\xi_n - \xi_0)F'(\xi_0) + u_nh(\xi_n) + o(u_n),$$

ce qui prouve le résultat en utilisant la continuité de h au point ξ_0. □

Il est intéressant de noter que l'application quantile n'est pas Fréchet-dérivable au point F dans les conditions du théorème précédent. En effet, si elle l'était, on aurait

$$T_p(G) - T_p(F) + \frac{(G-F)(T(F))}{F'(T(F))} = o(\|G - F\|_\infty). \tag{5.4}$$

Prenons comme exemple $p = 1/2$ et supposons que $F(0) = 1/2$. Étant donnée une suite a_n positive tendant vers zéro, prenons $G_n(x) = F(x)$ si $x < 0$ ou si $x \geq a_n$ et $G_n(x) = F(a_n)$ pour $x \in [0, a_n[$. On a $\|G_n - F\|_\infty = F(a_n) - 1/2$ et $T_{1/2}(G_n) = T_{1/2}(F) = 0$. Comme $G_n(0) - F(0) = F(a_n) - 1/2$, (5.4) n'est évidemment pas vérifiée.

La dérivabilité au sens d'Hadamard suffit, comme on l'a vu, pour trouver la limite en loi des quantiles, et des théorèmes 4.7 et 5.1 on déduit

Corollaire 5.2. *Soit r_n une suite tendant vers l'infini. Si $r_n(F_n(x) - F(x))$ converge en loi dans $D_0(\mathbb{R})$ (muni de la topologie du* sup*) vers un processus $Z(x)$ à trajectoires presque sûrement continues, alors sous les conditions du théorème 5.1, $r_n(F_n^{-1}(p)-F^{-1}(p))$ converge en loi vers $-\frac{Z(F^{-1}(p))}{F'(F^{-1}(p))}$.*

On est en mesure maintenant de trouver la loi limite des quantiles dans diverses situations plus générales que celles du théorème 3.2 puisque la validité de ce théorème repose sur la convergence du processus empirique qui, comme on l'a vu au chapitre 2, est largement vérifiée en dehors du cas où les variables sont indépendantes.

Exercice 5.2. Calculer la variance limite du quantile empirique centré et normalisé $\sqrt{n}(F_n^{-1}(p)-F^{-1}(p))$ lorsque les X_j sont i.i.d de loi F. ◁

Exercice 5.3. Trouver la fonction d'influence des quantiles, et vérifier que l'égalité (4.19) est valable. ◁

4. La fonction quantile

Dans ce qui précède on a travaillé à p fixé. Il est facile d'améliorer les résultats en considérant la fonction quantile $\sqrt{n}(F_n^{-1}(p) - F^{-1}(p))$ pour $p \in [p_1, p_2]$ $(0 < p_1 < p_2)$. Une lecture attentive de la preuve du théorème 5.1 montre que les o() sont uniformes en p pourvu qu'il reste dans $[p_1, p_2]$ strictement inclus dans $]0, 1[$. On obtient alors

Proposition 5.3. *Si F est une fonction de répartition strictement croisante et partout dérivable alors l'application Q_p qui à G associe la fonction quantile $G^{-1}(s)$ restreinte à $[p, 1-p]$ est Hadamard-dérivable au point F tangentiellement au sous-ensemble C de $D_0(\mathbb{R})$ constitué des fonctions h partout continues. Sa dérivée est*

$$Q'_{p,F}(h) = -\frac{h \circ F^{-1}}{F' \circ F^{-1}}.$$

Exercice 5.4. On désigne par $\mathrm{med}_{1\leq j\leq n} U_j$ la médiane de $(U_1, \ldots, U_n)$.

On appelle écart absolu médian du n-uplet $(X_1, \ldots, X_n)$ la quantité

$$\mathrm{MAD}(X_1, \ldots, X_n) = \mathrm{MAD}(\hat{F}_n) = \mathrm{med}_{1\leq j\leq n} |X_j - \mathrm{med}_{1\leq j\leq n} X_j|.$$

1) Quel peut être l'usage de cette statistique, et dans quel(s) cas peut elle être préférée à certaines statistiques plus familières ?

2) Soit X une variable aléatoire de fonction de répartition F. Quelle est la fonction de répartition de $|X|$? On suppose maintenant que la loi de X est symétrique, c'est à dire que, pour tout x, $F(x) = 1 - F(-x)$. Exprimer $\mathrm{MAD}(F)$.

3) Etudier la dérivabilité au sens d'Hadamard de la fonctionnelle MAD.

4) Quelle est la loi limite de $\mathrm{MAD}(X_1, \ldots, X_n)$ pour une suite de variables i.i.d dont la loi est comme à la question 2 ?

◁

Exercice 5.5. Les variables X_j sont i.i.d. Pour estimer la dispersion de leur loi, on choisit l'écart interquartile

$$T(\hat{F}_n) = \hat{F}_n^{-1}(3/4) - \hat{F}_n^{-1}(1/4) \qquad (5.5)$$

1) Pourquoi $T(G)$ peut-il être pris comme indice de dispersion de la loi G ? Quel(s) avantage(s) a-t-il par rapport à l'écart-type ?

2) La fonction de répartition F des X_j est C^1 et sa dérivée f est partout non nulle. Montrer que $\sqrt{n}\Big(T(\hat{F}_n) - T(F)\Big)$ converge en loi. Explicitez les caractéristiques de le loi limite.

3) On suppose (pour simplifier) que f est paire. Donnez l'expression de la fonction d'influence de T au point F. Que remarquez vous ?

4) L'estimateur défini en (5.5) est robuste par rapport certaines perturbations des données, mais sensible à d'autres. Lesquelles ? ◁

5. Une application : l'écart absolu médian

Si $\mathrm{med}_{1\leq j\leq n} U_j$ la médiane de $(U_1, \ldots, U_n)$. On appelle écart absolu médian du n-uplet $(X_1, \ldots, X_n)$ la quantité

$$\mathrm{MAD}(X_1, \ldots, X_n) = \mathrm{MAD}(\hat{F}_n) = \mathrm{med}_{1\leq j\leq n} |X_j - \mathrm{med}_{1\leq j\leq n} X_j|.$$

C'est une mesure de la dispersion de l'échantillon plus robuste que la variance et un peu moins grossière que l'écart interquartile puisque calée sur la médiane.

La fonctionnelle associée est la composée de 4 fontionnelles

$$T = T_4 \circ T_3 \circ T_2 \circ T_1$$

où $T_1(G) = (G, G^{-1}(1/2)$, $T_2(G, a) = G(. - a)$, T_3 est la fonction de répartition de X, si X a comme loi G (autrement dit $T_3(G)(x) = G(x) - G^{-}(-x)$, et $T_4(G) = G^{-1}(1/2)$.

Il est facile de voir que si F est la fonction de répartition d'une loi paire, $T(F)$ est le quartile supérieur : $T(F) = F^{-1}(3/4)$.

Exercice 5.6.

1) Etudier la dérivabilité au sens d'Hadamard de la fonctionnelle MAD.

2) Quelle est la loi limite de $\mathrm{MAD}(X_1, \dots, X_n)$ pour une suite de variables i.i.d dont la loi est comme à la question 2 ? ◁

6. Une application : la moyenne α-tronquée

On reprend l'exercice 3.14 et la fonctionnelle

$$T_\alpha(P) = \frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y \, \mathrm{d}F(y) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(s) \, \mathrm{d}s.$$

On voit bien qu'en appliquant la proposition 5.3 et la règle de dérivation en chaine, on va obtenir la dérivabilité de T_α. En effet $T_\alpha = \mathcal{I} \circ Q_\alpha$, où $\mathcal{I}$ est l'application qui à ϕ associe

$$\mathcal{I}(\phi) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} \phi(s) \, \mathrm{d}s.$$

Proposition 5.4. *La fonctionnelle T_α est Hadamard-dérivable tangentiellement au sous-ensemble C de $D_0(\mathbb{R})$ constitué des fonctions h partout continues, en toute F continument dérivable et de dérivée f strictement positive. Sa dérivée est*

$$\begin{aligned} T'_{\alpha,F}(h) &= -\frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} \frac{h \circ F^{-1}(s)}{f \circ F^{-1}(s)} \, \mathrm{d}s = -\frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \frac{h(x)}{f(x)} \, \mathrm{d}F(x) \\ &= -\frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} h(x) \, \mathrm{d}x. \end{aligned}$$

Exercice 5.7. Utiliser ce qui précède pour trouver la loi limite de la moyenne α-tronquée pour une suite de variables i.i.d dont la loi F est conforme aux hypothèses de la proposition 5.4. ◁

Exercice 5.8. Retrouver la fonction d'influence de la moyenne α-tronquée, que vous aviez calculée directement dans l'exercice 3.14. Cette fonction d'influence ressemble beaucoup à celle du M-estimateur de Huber défini par

$$\Psi(x) = \begin{cases} -1 & \text{si } x < -1 \\ x & \text{si } -1 \le x \le 1 \\ 1 & \text{si } x > 1 \end{cases}$$

Pourtant les estimateurs correspondants sont très différents. Expliquez. ◁

ANNEXE A

Solutions d'exercices

EXERCICE 1.1. *On suppose* $Q = (1-\alpha)P + \alpha R$ *où* $\alpha \in [0,1]$ *et* R *est une mesure de probabilité. Montrer que* $\pi(P,Q) \leq \alpha$.

SOLUTION. Pour prouver que $\pi_1(P,Q) \leq \alpha$, on vérifie pour tout fermé A l'inégalité $P(A) \leq Q(A^\alpha) + \alpha$ en écrivant simplement :

$$\begin{aligned} Q(A^\alpha) + \alpha &= (1-\alpha)P(A^\alpha) + \alpha R(A^\alpha) + \alpha \\ &\geq (1-\alpha)P(A) + 0 + \alpha P(A) = P(A). \end{aligned}$$

Et on conclut en rappelant que $\pi_1(P,Q) = \pi(P,Q)$. ◀

EXERCICE 1.2. *Montrer que pour tous* $x, y \in E$, $\pi(\delta_x, \delta_y) = \min(1, d(x,y))$.

SOLUTION. Commençons par remarquer que la distance Prokhorov de deux probabilités P et Q quelconques est au plus 1 car l'inégalité $P(F) \leq Q(F^1) + 1$ est trivialement vérifiée par tout fermé F de E.

Posons $r = d(x,y)$ et montrons d'abord que $\pi_1(\delta_x, \delta_y) \leq r$. Pour cela, on vérifie que pour tout $s > r$, pour tout fermé F de E,

$$\delta_x(F) \leq \delta_y(F^s) + s. \tag{A.1}$$

Cette inégalité est évidente si $y \in F^s$. Sinon, $d(y,F) \geq s$, donc pour tout $z \in F$, $d(y,z) \geq s > r$, ce qui interdit l'appartenance de x à F puisque $d(x,y) = r$, d'où $\delta_x(F) = 0$ et l'inégalité (A.1) est vérifiée. À ce stade, nous avons montré que $\pi_1(\delta_x, \delta_y) \leq \min(1, d(x,y))$ est vraie pour tous $x, y \in E$.

Montrons maintenant que $\pi_1(\delta_x, \delta_y) \geq \min(1, d(x,y))$. Pour cela on raisonne par l'absurde en supposant $\pi_1(\delta_x, \delta_y) < \min(1,r)$, ce qui implique l'existence d'un $s < \min(1,r)$ tel que (A.1) soit vérifiée pour tout fermé F. En prenant $F = \{x\}$, on a $F^s = B(x,s)$ et comme $d(x,y) = r > s$, $y \notin F^s$ donc $\delta_y(F^s) = 0$ et comme $s < 1$ et $\delta_x(A) = 1$, (A.1) ne peut être vraie pour ce choix de F. On a donc bien $\pi_1(\delta_x, \delta_y) \geq \min(1, d(x,y))$. ◀

EXERCICE 1.4. Comparaison entre distances de Prokhorov et en variation. *Montrer que* $\pi(P,Q) \leq \|P - Q\|$ *et trouver un exemple où l'inégalité est stricte.*

SOLUTION. Pour montrer l'inégalité, $\pi(P,Q) \leq \|P-Q\|$, il suffit de remarquer que, pour tout borélien A de E, $|P(A) - Q(A)| \leq \|P-Q\|$, ce qui implique

$$P(A) \leq Q(A) + \|P-Q\| \leq Q(A^{\|P-Q\|}) + \|P-Q\|,$$

d'où le résultat par définition de $\pi(P,Q)$.

Voici un exemple montrant que l'inégalité peut être stricte. On prend $E = [0, 1/2]$, $P = \delta_0$, $Q = \delta_{1/2}$. Il est clair que $\|P - Q\| = 1$. Notons d'autre part que si $\emptyset \neq A \subset E$, $A^{1/2} = E$. On a alors $P(A) \leq Q(A^{1/2}) + 1/2 = 3/2$ pour tout $A \neq \emptyset$ (si $A = \emptyset$, $P(A) = 0 \leq 0 + 1/2$). Donc $\pi_1(P, Q) \leq 1/2 < \|P - Q\|$. ◀

Exercice 1.5. Théorème de Scheffé. *Si P et Q ont respectivement pour densité f et g par rapport à une même mesure μ,*

$$\|P - Q\| = \frac{1}{2}\int |f(x) - g(x)| \,\mathrm{d}\mu(x) = 1 - \int \min\{f(x), g(x)\} \,\mathrm{d}\mu(x). \tag{A.2}$$

Montrer que l'hypothèse d'existence de densités n'est pas restrictive.

Solution. Remarquons d'emblée qu'il existe toujours une mesure μ dominant P et Q, c'est-à-dire telle que P et Q aient une densité relativement à μ. Il suffit de prendre $\mu = P + Q$. En effet $\mu(A) = 0$ implique $P(A) = Q(A) = 0$.

La deuxième égalité dans (A.2) est une conséquence de l'égalité élémentaire

$$|a - b| = a + b - 2\min(a, b)$$

vérifiée pour tous réels a et b.

Pour montrer la première égalité, on commence par remarquer que

$$\|P - Q\| = \sup_{A \in \mathcal{E}} |P(A) - Q(A)| = \sup_{A \in \mathcal{E}} \left| \int_A (f - g) \,\mathrm{d}\mu \right|.$$

En posant $E^+ := \{x \in E;\ f(x) > g(x)\} = \{f > g\}$ et $E^- := \{f < g\}$, on voit que

$$\begin{aligned}\|P - Q\| &= \sup_{A \in \mathcal{E}} \left| \int_{A \cap E^+} |f - g| \,\mathrm{d}\mu - \int_{A \cap E^-} |f - g| \,\mathrm{d}\mu \right| \\ &\leq \max\left(\int_{E^+} |f - g| \,\mathrm{d}\mu, \int_{E^-} |f - g| \,\mathrm{d}\mu \right).\end{aligned}$$

Cette inégalité est en fait une égalité (faire $A = E^+$ et $A = E^-$ pour minorer le sup).

Pour finir, il ne nous reste plus qu'à vérifier que $\int_{E^+} |f - g| \,\mathrm{d}\mu = \int_{E^-} |f - g| \,\mathrm{d}\mu$ et comme leur somme vaut $\int_E |f - g| \,\mathrm{d}\mu$, on en déduira que $\|P - Q\| = \int_{E^+} |f - g| \,\mathrm{d}\mu = \frac{1}{2} \int |f(x) - g(x)| \,\mathrm{d}\mu$.

$$\begin{aligned}\int_{E^+} |f - g| \,\mathrm{d}\mu = \int_{E^+} (f - g) \,\mathrm{d}\mu &= \int_E (f - g) \,\mathrm{d}\mu - \int_{E^-} (f - g) \,\mathrm{d}\mu \\ &= 1 - 1 - \int_{E^-} (f - g) \,\mathrm{d}\mu \\ &= \int_{E^-} (g - f) \,\mathrm{d}\mu = \int_{E^-} |f - g| \,\mathrm{d}\mu.\end{aligned}$$

La première égalité dans (A.2) est ainsi démontrée. ◀

Exercice 1.6. *Supposons qu'on désire tester l'hypothèse que la loi d'une variable X est P, contre l'hypothèse que sa loi est Q. Pour toute région critique C on regarde la somme des deux erreurs $P(C) + 1 - Q(C)$. Montrer que la valeur minimale de cette somme est atteinte et qu'elle vaut $\|P \wedge Q\|$.*

Solution. En notant μ une mesure dominant P et Q, par exemple $\mu = P + Q$, et f, g les densités respectives de P et Q par rapport à μ, on voit que

$$P(C) + 1 - Q(C) = 1 - \int_C (g - f)\,\mathrm{d}\mu$$

est minimal lorsque $\int_C (g - f)\,\mathrm{d}\mu$ est maximal. Ce maximum est clairement atteint pour $C = \{g > f\}$. Ainsi

$$\begin{aligned}\min_{C\in\mathcal{E}}\Big(P(C) + 1 - Q(C)\Big) &= 1 - \int_{\{g>f\}} (g - f)\,\mathrm{d}\mu \\ &= \int_E \Big(g - (g-f)\,\mathbb{I}_{\{g>f\}}\Big)\,\mathrm{d}\mu \\ &= \int_E \Big(g\,\mathbb{I}_{\{g\le f\}} + f\,\mathbb{I}_{\{g>f\}}\Big)\,\mathrm{d}\mu \\ &= \int_E \min(f,g)\,\mathrm{d}\mu = \|P \wedge Q\|\,.\end{aligned}$$

◀

Exercice 1.7. *Montrer que*

$$\frac{1}{2}H^2(P,Q) \le \|P - Q\| \le \min\Big(H(P,Q); 1 - \frac{1}{2}A^2(P,Q)\Big).$$

Solution. Soit μ une mesure dominant P et Q et f, g les densités respectives de P, Q par rapport à μ.

Pour la première inégalité, il suffit de remarquer que $\min(f,g) \le \sqrt{fg}$, d'où

$$\frac{1}{2}H^2(P,Q) = 1 - \int \sqrt{fg}\,\mathrm{d}\mu \le 1 - \int \min(f,g)\,\mathrm{d}\mu = \|P - Q\|\,.$$

Le coefficient $\frac{1}{2}$ est optimal car si on prend P et Q de supports disjoints (par exemple $P = \delta_x$ et $Q = \delta_y$), on voit facilement que $\pi(P,Q) = 1$ et $H^2(P,Q) = 2$.

Montrons maintenant l'inégalité $\|P - Q\| \le H(P,Q)$. Pour cela, on utilise l'inégalité de Cauchy-Schwarz dans $L^2(\mu)$ comme suit.

$$\begin{aligned}2\,\|P - Q\| = \int |f - g|\ \mathrm{d}\mu &= \int \Big|(f^{1/2})^2 - (g^{1/2})^2\Big|\ \mathrm{d}\mu \\ &= \int \Big|f^{1/2} - g^{1/2}\Big|\Big|f^{1/2} + g^{1/2}\Big|\ \mathrm{d}\mu \\ &\le \left\{\int \Big(f^{1/2} - g^{1/2}\Big)^2\,\mathrm{d}\mu\right\}^{1/2}\left\{\int \Big(f^{1/2} + g^{1/2}\Big)^2\,\mathrm{d}\mu\right\}^{1/2} \\ &= H(P,Q)\left\{\int \Big(f^{1/2} + g^{1/2}\Big)^2\,\mathrm{d}\mu\right\}^{1/2}.\end{aligned}$$

En utilisant maintenant l'inégalité élémentaire $(a+b)^2 \le 2(a^2 + b^2)$, on obtient :

$$2\,\|P - Q\| \le H(P,Q)\left\{\int 2(f+g)\,\mathrm{d}\mu\right\}^{1/2} = 2H(P,Q).$$

Il nous reste à montrer l'inégalité $\|P-Q\| \leq 1 - \frac{1}{2}A^2(P,Q)$. Pour cela, nous allons comparer $(\int \sqrt{fg}\,\mathrm{d}\mu)^2$ et $\int \min(f,g)\,\mathrm{d}\mu$. On remarque alors que

$$fg = \max(f,g)\min(f,g) \leq (f+g)\min(f,g),$$

d'où en appliquant l'inégalité de Cauchy-Schwarz dans $L^2(\mu)$:

$$A^2(P,Q) = \left\{\int \sqrt{fg}\,\mathrm{d}\mu\right\}^2 \leq \left\{\int (f+g)^{1/2}\min(f,g)^{1/2}\,\mathrm{d}\mu\right\}^2 \leq 2\int \min(f,g)\,\mathrm{d}\mu.$$

Par conséquent

$$\|P-Q\| = 1 - \int \min(f,g)\,\mathrm{d}\mu \leq 1 - \frac{1}{2}A^2(P,Q).$$

◀

Exercice 1.11. *Dans le cas $E=\mathbb{R}$, trouvez des exemples simples montrant que les inégalités dans le « portmanteau theorem » peuvent être strictes.*

Solution.

Voici un exemple où $\liminf \mathbb{E}(f(X_n)) > \mathbb{E}(f(X))$ avec f continue $\mathbb{R} \to \mathbb{R}_+$. Nécessairement on devra prendre f non bornée. Soit $(Y_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d. de densité $ct^{-4}\mathbb{I}_{\{|t|\geq 1\}}$ et $X_n := n^{-1/2}(Y_1 + \cdots + Y_n)$. Pour tout $i \geq 1$, $\mathbb{E}\,Y_i = 0$ et $\mathbb{E}\,Y_i^2 =: \sigma^2 < +\infty$, donc par le TLC, X_n converge en loi vers X de loi $\mathfrak{N}(0,\sigma^2)$. On choisit maintenant $f(t) = \exp(t)$. Alors $\mathbb{E}\,f(X)$ est fini. Mais comme $\mathbb{E}\,f(X_n) = +\infty$ pour tout $n \geq 1$ (voir justification ci-après), sa limite inférieure vaut aussi $+\infty$ et on a donc bien $\liminf \mathbb{E}(f(X_n)) > \mathbb{E}(f(X))$. Pour voir que $\mathbb{E}\,f(X_n) = +\infty$ pour tout n, on remarque que par indépendance et équidistribution des Y_i, $\mathbb{E}\,f(X_n) = \Big(\mathbb{E}\exp(n^{-1/2}Y_1)\Big)^n$ et que :

$$\mathbb{E}\exp\left(\frac{Y_1}{\sqrt{n}}\right) = \int_{\{|t|\geq 1\}} \exp\left(\frac{t}{\sqrt{n}}\right)\frac{c}{t^4}\,\mathrm{d}t > \int_1^{+\infty} \exp\left(\frac{t}{\sqrt{n}}\right)\frac{c}{t^4}\,\mathrm{d}t = +\infty.$$

Voyons maintenant un exemple où pour un ouvert O, $\liminf P(X_n \in O) > P(X \in O)$. On prend X_n de loi uniforme sur $[-1/n, 0]$ et X de loi δ_0. Il est clair que X_n converge en loi vers X (dessinez les f.d.r.). Avec $O =]-\infty, 0[$, on a pour tout $n \geq 1$, $P(X_n \in O) = 1$ et comme $P(X \in O) = 0$, on a bien $\liminf P(X_n \in O) > P(X \in O)$.

La même suite X_n peut servir pour trouver un fermé F tel que $\limsup P(X_n \in F) = 0 < P(X \in F) = 1$ en prenant $F = [0, +\infty[$ par exemple.

Et avec la même suite X_n, pour $A = \{0\}$, on a $P(X \in \partial A) = P(X = 0) = 1$ et $P(X_n \in A) = 0$ ne converge pas vers $P(X \in A) = 1$.

◀

Exercice 2.5. *Médiane et meilleure approximation L^1.*
Montrer que si m est une médiane de la v.a. intégrable X,

$$\forall c \in \mathbb{R}, \quad \mathbb{E}\,|X-c| \geq \mathbb{E}\,|X-m|\,. \tag{A.3}$$

L'inégalité ci-dessus est triviale si $\mathbb{E}\,|X| = +\infty$. Montrer que dans ce cas $|X-c| - |X-m|$ est intégrable et vérifie $\mathbb{E}(|X-c| - |X-m|) \geq 0$ pour tout $c \in \mathbb{R}$.

Solution. L'intégrabilité de X équivaut à celle de $X-c$ pour toute constante c. Exprimons $|X-a|-|X-b|$ sans valeur absolue, pour $a<b$ réels quelconques

$$|X-a|-|X-b| = \begin{cases} a-b & \text{si } X \le a, \\ 2X-a-b & \text{si } a \le X \le b, \\ b-a & \text{si } X \ge b. \end{cases} \tag{A.4}$$

On en déduit que

$$\begin{aligned}\mathbb{E}(|X-a|-|X-b|) &= (a-b)P(X<a) + \int_{\{a\le X<b\}} (2X-a-b)\,\mathrm{d}P + (b-a)P(X\ge b) \\ &\ge (a-b)P(X<a) + (a-b)P(a\le X<b) + (b-a)P(X\ge b) \\ &\ge (a-b)P(X<b) + (b-a)P(X\ge b),\end{aligned}$$

d'où

$$\mathbb{E}(|X-a|-|X-b|) \ge (b-a)\Big(P(X\ge b)-P(X<b)\Big), \quad a<b. \tag{A.5}$$

De manière analogue :

$$\begin{aligned}\mathbb{E}(|X-b|-|X-a|) &= (b-a)P(X\le a) + \int_{\{a<X\le b\}} (a+b-2X)\,\mathrm{d}P + (a-b)P(X>b) \\ &\ge (b-a)P(X\le a) + (a-b)P(a<X\le b) + (a-b)P(X>b) \\ &\ge (b-a)P(X\le a) + (a-b)P(X>a),\end{aligned}$$

d'où

$$\mathbb{E}(|X-b|-|X-a|) \ge (b-a)\Big(P(X\le a)-P(X>a)\Big), \quad a<b. \tag{A.6}$$

Soit maintenant m une médiane de X et $c<m$. En appliquant (A.5) avec $a=c$ et $b=m$, on obtient

$$\mathbb{E}(|X-c|-|X-m|) \ge (m-c)\Big(2P(X\ge m)-1\Big) \ge 0,$$

puisque $P(X\ge m)\ge \frac{1}{2}$. De même, pour $c>m$, en appliquantt (A.6) avec $a=m$ et $b=c$, il vient :

$$\mathbb{E}(|X-c|-|X-m|) \ge (c-m)\Big(2P(X\le m)-1\Big) \ge 0,$$

puisque $P(X\le m)\ge \frac{1}{2}$. Si X est intégrable, ceci complète la vérification de (A.3).

Si X n'est pas intégrable, la seule différence est qu'on n'a plus le droit d'écrire $\mathbb{E}(|X-c|-|X-m|) = \mathbb{E}\,|X-c| - \mathbb{E}\,|X-m|$ comme on vient de le faire implicitement pour déduire (A.3) de la positivité de $\mathbb{E}(|X-c|-|X-m|)$. Par contre, la variable aléatoire $|X-c|-|X-m|$ est intégrable puisque bornée en valeur absolue par $|c-m|$ et on peut quand même dire que son espérance est positive et finie. ◀

Exercice 3.7. *Une loi forte des grands nombres avec limite infinie.*
1) *Soit $(X_i)_{i\ge1}$ une suite i.i.d. de variables aléatoires positives telle que $\mathbb{E}\,X_1 = +\infty$. Montrer que*

$$\frac{1}{n}\sum_{i=1} X_i \xrightarrow[n\to+\infty]{\text{p.s.}} +\infty. \tag{A.7}$$

2) *Généraliser ce résultat au cas où les X_i sans être positives vérifient l'une des deux conditions*

i) $\mathbb{E}\,X_1^+ = +\infty$ *et* $\mathbb{E}\,X_1^- < +\infty$ *;*

ii) $\mathbb{E}\,X_1^+ < +\infty$ *et* $\mathbb{E}\,X_1^- = +\infty$*.*

Solution.

1) Posons pour $k \in \mathbb{N}^*$, $X_{i,k} := \max(X_i, k)$. Pour chaque k, les variables aléatoires positives $(X_{i,k})_{i\geq 1}$ sont i.i.d. et bornées donc vérifient la loi forte des grands nombres. En faisant tendre n vers l'infini dans l'inégalité

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq \frac{1}{n}\sum_{i=1}^{n} X_{i,k},$$

on obtient donc

$$\text{p.s.} \quad \liminf_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \geq \mathbb{E}\,X_{1,k}.$$

Cette inégalité dans $\overline{\mathbb{R}}_+$ étant vérifiée pour tout $k \in \mathbb{N}^*$, on peut faire tendre k vers $+\infty$ en notant qu'alors $X_{1,k} \uparrow X_1$ et que d'après le théorème de convergence croissante (Beppo Levi), $\mathbb{E}\,X_{1,k} \uparrow \mathbb{E}\,X_1$, d'où

$$\text{p.s.} \quad \liminf_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \geq \mathbb{E}\,X_1 = +\infty,$$

ce qui établit (A.7).

2) Si on remplace l'hypothèse de positivité des X_i par $\mathbb{E}\,X_1^+ = +\infty$ et $\mathbb{E}\,X_1^- < +\infty$, en appliquant le résultat de la question préccédente à la suite $(X_i^+)_{i\geq 1}$ et la loi forte des grands nombres classique à la suite $(X_i^-)_{i\geq 1}$, on obtient :

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} X_i^+ - \frac{1}{n}\sum_{i=1}^{n} X_i^- \xrightarrow[n\to+\infty]{\text{p.s.}} +\infty - \mathbb{E}\,X_1^- = +\infty.$$

Si $\mathbb{E}\,X_1^+ < +\infty$ et $\mathbb{E}\,X_1^- = +\infty$, on obtient de même :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\text{p.s.}} \mathbb{E}\,X_1^+ - \infty = -\infty.$$

◀

Exercice 4.1. *Soient E et F deux espaces vectoriels normés, $\mathcal{L}(E,F)$ l'espace des applications linéaires* continues $E \to F$*, muni de la norme opérateur :* $\|\varphi\| = \sup\{\|\varphi(x)\|\,;\ \|x\| \leq 1\}$*. On suppose que φ_n est une suite d'éléments aléatoires de $\mathcal{L}(E,F)$ qui converge en probabilité vers φ élément non aléatoire de $\mathcal{L}(E,F)$ et que la suite de vecteurs aléatoires $(Z_n)_{n\geq 1}$ converge en loi vers Z dans E. Montrez qu'alors $\varphi(Z_n)$ converge en loi vers $\varphi(Z)$ dans F.*

Solution. On part de la décomposition

$$\varphi_n(Z_n) = \varphi(Z_n) + \big(\varphi_n(Z_n) - \varphi(Z_n)\big).$$

Le premier terme $\varphi(Z_n)$ converge en loi dans F vers $\varphi(Z)$ par conservation de la convergence en loi par l'application continue φ.

On montre ensuite que le terme de reste converge en probabilité vers 0 en notant que

$$\|\varphi_n(Z_n) - \varphi(Z_n)\|_F \leq \|\varphi_n - \varphi\|_{\mathcal{L}(E,F)} \|Z_n\|_E \, .$$

On utilise alors la convergence en probabilité de φ_n vers φ et le fait que (Z_n) est stochastiquement bornée dans E pour conclure. ◀

Bibliographie

[1] Azencott R., Dacunha-Castelle D. (1984). Séries d'observations irrégulières. Masson.

[2] Beran, J. (1994). Statistics for long-memory processes. Monographs on Stat. and Appl. Prob. 61, Chapman-Hall.

[3] Billingsley P. (1968). Convergence of probability measures. Wiley.

[4] Billingsley P. (1995). Probability and measure. Wiley

[5] Breiman L. (1968), Probability, Addison-Wesley.

[6] Cramer H., Leadbetter R. (1967). Stationary and related stochastic processes. Wiley.

[7] Dacunha-Castelle D., Duflo M. (1983). Probabilités et Statistiques : vol. 1. Problèmes à temps fixe. Masson.

[8] Dacunha-Castelle D., Duflo M. (1983). Probabilités et Statistiques : vol. 1. Problèmes à temps fixe. Exercices. Masson.

[9] Dacunha-Castelle D., Duflo M. (1983). Probabilités et Statistiques : vol. 2. Problèmes à temps mobile. Masson.

[10] Dehling H., Mikosh T., Sørensen M. (Eds) (2002). Empirical process techniques for dependent data. Birkhauser.

[11] Doob J.L. (1953). Stochastic processes. Wiley.

[12] Doukhan P.(1994) mixing : Properties and examples. LNS. Springer.

[13] Doukhan P., Surgailis D. (1998). Functional central limit theorem for the empirical process of short memory linear processes. C.R. Acad. Sci. Paris, t. 326, Série I, 87–92.

[14] Dudley R.M. (2002). Real analysis and probability. Cambridge University Press.

[15] Ferguson T. S. (1996). A course in large sample theory. Chapman & Hall, Texts in Statistical Science.

[16] Grenander U., Rosenblatt M., (1957). Statistical analysis of stationary time series. Wiley.

[17] Hampel F.R., Ronchetti E.M., Rousseeuw P.J. Stahel W.A. (1986). Robust Statistics. The approach based on influence functions. Wiley.

[18] Hannan E. J. (1970). Multiple time series. Wiley.

[19] Ibragimov I., Linnik

[20] Ledoux M., Talagrand M. (1991). Probability in Banach Spaces. Springer.

[21] Lehmann E. L. (2004). Elements of large sample theory. Springer texts in statistics. Springer.

[22] Lehmann E. L., Casella G. (2001). Theory of point estimation. Springer texts in statistics. Springer.

[23] Schwartz L. (1970). Topologie générale et analyse fonctionnelle. Hermann.

[24] van der Vaart A. W. (2000). Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics.

[25] van der Vaart A. W., Wellner J. A. (1996). Weak convergence and empirical processes. Springer.