Sous-échantillonnage pour la Régression Logistique : Approches théoriques et Applications

Travail d'étude et de recherche

Direction de A. Guyader & M. Sangnier

T. Fassina & Y. Ghariani

Mai 2024

Table des matières

1 Introduction **3**

2 Les théorèmes asymptotiques **4**
2.1 Introduction 4
2.2 Le théorème centrale limite classique 4
2.3 Conditions de Lindeberg-Feller et Lyapunov 5
2.4 Symboles o et O stochastiques 7

3 La régression linéaire et logistique **8**
3.1 La régression linéaire 8
3.1.1 Régression linéaire 8
3.1.2 Régression linéaire dans le cas gaussien et propriétés de l'estimateur MLE 9
3.1.3 Application du théorème de Lindeberg Feller à la regression linéaire pour la normalité asymptotique 10
3.1.4 Regression lineaire généralisée 10
3.2 La régression logistique 11
3.2.1 Existence et consistance de estimateur du maximum de vraisemblance 13
3.2.2 La normalité asymptotique de l'estimateur du maximum de vraisemblance 15

4 Propriétés conditionnelles **16**
4.1 Symboles o et O conditionnels 16
4.2 Convergence en loi conditionnelle et quelques premières propriétés 16
4.3 Théorème de Lévy conditionnel 17
4.4 Théorème de Lindeberg-Feller conditionnel 18

5 Subsampling and Optimal subsampling **19**
5.1 Introduction de l'estimateur par sous-échantillonage 19
5.2 La consistance 19
5.3 La normalité asymptotique 20
5.4 Les poids optimaux 20
5.5 L'algorithme en deux pas 21

6 Les simulations numériques **23**
6.1 Introduction 23
6.2 La consistance 23
6.3 La normalité asymptotique 23
6.4 L'erreur quadratique 24

7 Conclusion **26**

8 Annexes **27**
8.1 Théorème asymptotiques 27
8.2 La régression linéaire et logistique 34
8.3 Propriétés conditionnelles 38
8.4 Subsampling et Optimal Subsampling 43

1 Introduction

Dans le cadre de l'analyse de données massives pour l'estimation des paramètres de modèles statistiques ou, plus généralement, pour l'étude de relation entre événements et phénomènes du monde, le manque de données peut constituer un obstacle important dans la mise en oeuvre des méthodologies statistiques connues à ce jour. A ce problème s'ajoute malheureusement le problème exactement inverse, à savoir l'impossibilité, principalement due à des limitations computationnelles, d'analyser des échantillons trop importants.

Le texte qui suit s'inscrit précisément dans le contexte défini par le problème de l'identification de solutions d'analyse efficaces lorsque l'échantillon observé est trop grand. En particulier, ce qui est proposé dans ce qui suit est une voie théorique qui peut nous permettre de comprendre comment, dans le cas de la régression logistique, il est possible d'effectuer une estimation efficace des paramètres de régression en utilisant un sous-échantillon convenablement choisi de l'échantillon de grande taille à notre disposition.

Tout notre parcours vise à démontrer la consistance et la normalité asymptotique d'un estimateur des paramètres de régression construit par sous-échantillonnage. Ces deux résultats sont présents dans l'article de 2018 "Optimal Subsampling for Large Sample Logistic Regression" de HaiYing Wang, Rong Zhu et Ping Ma [7]. Pour comprendre ces résultats, il est cependant nécessaire de suivre un long chemin théorique.

Dans le chapitre 2, nous nous familiariserons avec les principaux résultats de convergence asymptotique, notamment le Théorème de Lindeberg-Feller.

Dans le chapitre 3, nous aborderons la définition de la régression linéaire et de la régression logistique. L'introduction de la régression linéaire vise simplement à montrer ce que signifie étudier un estimateur de fonction de régression sans se limiter au cas de la régression logistique. Par rapport à la régression logistique, nous n'en resterons pas à l'introduction du modèle de régression, mais nous approfondirons l'existence et les propriétés de l'estimateur de maximum de vraisemblance dans ce modèle.

Une fois ces outils introduits, nous explorerons, dans le chapitre 4, le concept de conditionnement. Dans les résultats de l'article, la notion d'espérance conditionnelle, de convergence en loi conditionnelle, de $o_{\mathbb{P}|\mathcal{F}_n}(1)$ et de $O_{\mathbb{P}|\mathcal{F}_n}(1)$ sont essentielles pour comprendre et démontrer les énoncés des théorèmes. Nous énoncerons et démontrerons un théorème de Lévy conditionnel ainsi qu'un théorème de Lindeberg-Feller conditionnel.

Avec une connaissance approfondie des notions et propriétés nécessaires ainsi que des propriétés de l'estimateur de maximum de vraisemblance, nous pourrons enfin, dans le chapitre 5 aborder le problème de l'estimateur défini par sous-échantillonnage. En particulier, nous montrerons un résultat de consistance et de normalité asymptotique de cet estimateur. Une fois le parcours théorique achevé, nous passerons à des simulations numériques. Nous prendrons en compte cinq lois différentes, générerons des échantillons et présenterons graphiquement les résultats de consistance et de normalité asymptotique. En particulier, nous énoncerons quelles sont les probabilités optimales à mettre sur l'échantillon pour obtenir un estimateur par sous-échantillonnage aussi efficace que possible.

Le chemin que nous avons suivi nous permettra de mieux appréhender les implications pratiques et théoriques de ces résultats et d'explorer comment le sous-échantillonnage peut améliorer la performance des estimations dans le contexte de la régression.

2 Les théorèmes asymptotiques

2.1 Introduction

Nous commençons notre voyage par une discussion approfondie du théorème de la limite centrale et de ses généralisations. Le théorème de la limite centrale dans sa variante standard à une dimension est un théorème de convergence en loi très important pour la variable aléatoire

$$\bar{X}_n = \frac{\sum_i^n X_i}{n}$$

à la loi d'une variable aléatoire de type gaussien. L'histoire du théorème central limite (TCL), où l'adjectif central fait référence au "théorème" et non à "limite", est une histoire qui remonte à 1733, lorsque de Moivre réussit à prouver

$$\sqrt{n}(\bar{X}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$$

dans le cas où $\{X_i\}$ soient des variables aléatoires de type Bernoulli indépendantes et paramètre p.

Nous aborderons ensuite l'énoncé et la preuve du théorème central limite dans sa version standard et la plus connue, à la fois dans le cas univarié et multivarié (vecteurs aléatoires X_i).

Dans les sections suivantes, nous nous consacrerons ensuite à l'énoncé et à la démonstration des deux généralisations les plus importantes du TCL : le TCL de Lindeberg-Feller, que nous prouverons dans le cas univarié et multivarié. Nous verrons qu'il ne sera heureusement pas nécessaire de proposer une preuve supplémentaire pour le TCL de Lyapunov puisque cette version du TCL est une conséquence assez simple du TCL de Lindeberg-Feller.

Avant de commencer notre parcours, on énonce un théorème très important et célèbre, qu'on utilisera largement dans les sections suivantes, notamment le théorème de convergence de Lévy.

Théorème 2.1 (Théorème de convergence de Lévy)**.**
Soit $\{X_n\}$ suite de variables (resp. vecteurs) aléatoires dans $\mathbb{R}$ (resp. $\mathbb{R}^d$) et $\{\Phi_n(t)\}$ leurs fonctions caractéristiques. On a l'équivalence entre les deux propriétés suivantes :

1. X_n converge en loi vers X.
2. $\Phi_n(t)$ converge vers $\phi_X(t)$ pour tout t dans $\mathbb{R}$ (resp. $\mathbb{R}^d$).

2.2 Le théorème centrale limite classique

Le théorème centrale limite classique est, avec la Loi Forte des Grands Nombres, le résultat fondamental le plus important, parce que il constitue un résultat très général qui met en relation une grand partie de variables aléatoires réelles avec la loi gaussienne et parce qu'il est constamment utilisé en probabilité et statistique. Dans le lignes qui suivent on va énoncer le théorème central limite dans le cas unidimensionnel, le prouver, énoncer le théorème central limite dans le cas multidimensionnel et montrer sa preuve, qui, on verra, découle de la preuve du cas univarié.

Théorème 2.2 (Théorème Central Limite Univarié)**.**
Soit $\{X_n\}$ une suite de variables aléatoires réelles iid admettant une moyenne, notée μ et une variance, notée σ^2. Alors, en notant

$$\bar{X}_n = \frac{\sum_i^n X_i}{n},$$

on a

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

ou, de manière équivalente,

$$\frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

(Preuve en annexes).

On énonce dans les prochaines lignes le théorème central limite dans le cas multivarié, dont la preuve repose sur la preuve dans le cas univarié et sur le théorème de Lévy.

Théorème 2.3 (Théorème Central Limite Multivarié)**.**
Soit $\{\bar{X}^i = (X_n^1, ..., X_n^d)'\}$ une suite de vecteurs aléatoires iid admettant une moyenne, notée μ, et une matrice de covariance Σ. Notant

$$\bar{X}_n = \frac{\sum_i^n X_i}{n},$$

on a

$$\sqrt{n}\Big(\bar{X}_n - \mu\Big) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

ou, de manière équivalente,

$$\frac{\sum_i^n X_i - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

L'analogie avec le cas univarié est plus que forte (Preuve en annexes).

2.3 Conditions de Lindeberg-Feller et Lyapunov

Les extensions les plus classiques du TCL sont celles de Lindeberg-Feller et de Lyapunov. On parle ici d'extension car nous n'avons plus besoin que les variables soient identiquement distribuées. A la place, nous imposons une condition sur les moments d'ordre 2 des variables aléatoires. Ce résultat est une généralisation du TCL, dans le sens où les hypothèses du TCL impliquent les hypothèses de Lindeberg-Feller. Nous verrons aussi que la condition de Lyapunov est en réalité une condition plus forte que celle de Lindeberg-Feller mais qui a l'avantage d'être plus facile à vérifier en pratique.

Nous avons besoin de ces deux lemmes pour prouver le TCL de Lindeberg-Feller

Lemme 2.1.

$$\left| e^{ix} - \sum_{k=0}^{n-1} \frac{(ix)^k}{k!} \right| \leq min\left\{ \frac{|x|^n}{n!}, 2\frac{|x|^{n-1}}{(n-1)!} \right\}$$

pour tout $x \in \mathbb{R}$ et $n \in \mathbb{N}^*$. (Preuve en annexes).

Lemme 2.2.
Soient $\{a_n\}$ et $\{b_n\}$ deux suites réelles t.q. pour tout i dans $\mathbb{N}$ $|a_i| \leq 1$ et $|b_i| \leq 1$. Alors

$$\left| \prod_{i=i}^{n} a_i - \prod_{i=i}^{n} b_i \right| \leq \sum_{i}^{n} |a_i - b_i|$$

pour tout $n \in \mathbb{N}$

Théorème 2.4 (TCL condition de Lindeberg)**.**
Soit $\{X_{ni} : 1 \leq i \leq r_n\}_{n \geq 1}$ un tableau triangulaire de variables aléatoires tel que :

- $X_{n1}, \cdots X_{nr_n}$ sont indépendantes pour tout $n \geq 1$,
- $\mathbb{E}[X_{ni}] = 0, 0 < \mathbb{E}[X_{ni}^2] := \sigma_{ni}^2 < +\infty$ pour tout $1 \leq i \leq r_n$ et $n \geq 1$,
- $\lim_{n \to +\infty} \frac{1}{s_n^2} \sum_{i=1}^{r_n} \mathbb{E}[X_{ni}^2 \mathbb{1}_{\{|X_{ni} > \epsilon s_n\}}] = 0$ avec $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ pour tout $n \geq 1$, pour tout $\epsilon > 0$.

Alors

$$\frac{S_n}{s_n} \xrightarrow[n \to +\infty]{d} \mathcal{N}(0, 1)$$

avec $S_n = \sum_{i=1}^{r_n} X_{ni}$ (Preuve en annexes).

Nous venons de démontrer que la condition de Lindeberg était suffisante pour avoir la normalité asymptotique. Sous certaines conditions de régularité sur le tableau de variables, il s'avère que cette condition est nécessaire. C'est ce qu'a démontré William Feller en 1935.

Théorème 2.5 (Feller)**.**
Soit $\{X_{ni} : 1 \leq i \leq r_n\}_{n\geq 1}$ un tableau triangulaire de variables aléatoires indépendantes centrées et L^2 telles que

$$\lim_{n\to+\infty} \max_{1\leq i\leq r_n} \frac{\sigma_{ni}^2}{s_n^2} = 0 \text{ et } \frac{S_n}{s_n} \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

Alors $\{X_{ni} : 1 \leq i \leq r_n\}_{n\geq 1}$ satisfait la condition de Lindeberg
(En notant $s_n^2 = \sum_{i=1}^{r_n} \mathbb{E}[X_{ni}^2]$ et $S_n = \sum_{i=1}^{r_n} X_{ni}$) (Preuve en annexes).

La condition de Lindeberg peut parfois être difficile à vérifier, une condition plus simple est celle de Lyapunov

Théorème 2.6 (TCL condition de Lyapunov)**.**
Soit $\{X_{ni} : 1 \leq i \leq r_n\}_{n\geq 1}$ un tableau triangulaire de variables indépendantes centrées et de carré intégrable. On note $\sigma_{nj}^2 := \mathbb{E}[X_{nj}^2]$, $s_n^2 := \sum_{j=1}^{r_n} \sigma_{nj}^2$, $S_n := \sum_{j=1}^{r_n} X_{nj}$ pour tout $n \geq 0, 1 \leq j \leq r_n$ Si il existe $\delta > 0$ tel que

$$\frac{1}{s_n^{2+\delta}} \sum_{j=1}^{r_n} \mathbb{E}[|X_{nj}^{2+\delta}|] \xrightarrow[n\to+\infty]{} 0$$

Alors

$$\frac{S_n}{s_n} \xrightarrow[n\to+\infty]{} \mathcal{N}(0,1)$$

(Preuve en annexes).

Un outil très efficace pour l'extension au cas multivarié est le théorème suivant :

Théorème 2.7 (Cramer-Wold)**.**

Soit $(X_n)_{n\geq 0}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ et soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$. Alors

$$X_n \xrightarrow[n\to+\infty]{d} X \iff \forall a \in \mathbb{R}^d \quad \langle a, X_n\rangle \xrightarrow[n\to+\infty]{d} \langle a, X\rangle.$$

Démonstration.
"$\Longleftarrow$" Soit $a \in \mathbb{R}^d$

$$\phi_{X_n}(a) = \phi_{\langle a, X_n\rangle}(1) \xrightarrow[n\to+\infty]{} \phi_{\langle a,X\rangle}(1) = \phi_X(a)$$

L'autre sens est une conséquence direct du théorème de Lévy. □

Remarque : Il est à noter que l'on peut se restreindre aux vecteurs de norme 1.

Théorème 2.8 (TCL Lindeberg-Feller Multivarié)**.**
Soit $\{X_{ni} : 1 \leq i \leq r_n\}_{n\geq 1}$ un tableau triangulaire de variables aléatoires à valeurs dans $\mathbb{R}^d$ tel que $\mathbb{E}[X_{nj}] = 0$ et $\frac{1}{r_n} \sum_{j=1}^{r_n} \mathrm{Cov}(X_{nj}) = \mathrm{Id}$ pour tout $n \geq 1, 1 \leq j \leq r_n$.
Si pour tout $\epsilon > 0$, on a

$$\lim_{n\to+\infty} \frac{1}{r_n} \sum_{j=1}^{r_n} \mathbb{E}[\|X_{nj}\|^2 \mathbb{1}_{\|X_{nj}\| > \epsilon\sqrt{r_n}}] = 0,$$

alors

$$\frac{\sum_{j=1}^{r_n} X_{nj}}{\sqrt{r_n}} \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, I_d)$$

(Preuve en annexes).

Remarque : Quitte à normaliser et centrer les variables, les hypothèses $\mathbb{E}[X_{nj}] = 0$ et $\frac{1}{r_n}\sum_{j=1}^{r_n} \mathrm{Cov}(X_{nj}) =$ Id ne sont en réalité pas restrictives.

2.4 Symboles o et O stochastiques

On rappelle ici deux définitions qui nous seront utiles pendant la suite et on introduit aussi un théorème concernant ces définitions.

Definition 2.1 ($o_{\mathbb{P}}(1)$)**.**
On dit que une suite $\{X_n\}_n$ de variables aléatoires est égale à un $o_{\mathbb{P}}(1)$ si $X_n \xrightarrow{d} 0$.

Definition 2.2 ($O_{\mathbb{P}}(1)$)**.**
On dit que une suite $\{X_n\}_n$ des variables aléatoires réelles ou vecteurs aléatoires réels est égale à un $O_{\mathbb{P}}(1)$ si la suite est tendue, c'est à dire $\forall \epsilon > 0$ il existe un compact K t.q. $\mathbb{P}(X_n \notin K) < \epsilon \forall n \geq 1$.

On a une propriété importante qu'on va démontrer et qu'on utilisera dans la suite qui lie strictement les suites tendues et la convergence en loi.

Théorème 2.9.
Soit $\{X_n\}_n$ une suites de variables(resp. vecteurs) aléatoires réelles et X une variable (resp. vecteur) aléatoire réelle. Alors

1. Si $X_n \xrightarrow{d} X$, alors $X_n = O_{\mathbb{P}}(1)$.
2. Si $X_n = O_{\mathbb{P}}(1)$, alors $\exists \{X_{n_k}\}_k$ sous suite de $\{X_n\}_n$ t.q. $X_{n_k} \xrightarrow{d} X$.

3 La régression linéaire et logistique

3.1 La régression linéaire

L'objectif de ce sous-chapitre est de présenter brièvement les outils fondamentaux utilisés en apprentissage supervisé. Nous introduirons donc, la notation, la signification et quelques résultats sur la régression linéaire, auxquels on fera suivre une discussion approfondie sur la régression logistique.

3.1.1 Régression linéaire

La régression linéaire est l'un des outils statistiques les plus utilisés et les plus anciens pour étudier la relation entre les variables aléatoires et les événements, et sa première utilisation historiquement établie remonte à 1805 et aux travaux d'Adrien-Marie Legendre.

La régression linéaire consiste essentiellement à

— Étudier la relation qu'il y a entre une variable Y et un ensemble de variables $X^1, ..., X^p$.

— Supposer que la relation entre Y et $X^1, ..., X^p$ est une relation essentiellement linéaire, perturbé par un petit bruit :

$$Y = \beta_1 X^1 + ... + \beta_p X^p + \epsilon$$

où ϵ représente le bruit. Il est évident que l'hypothèse de linéarité de la relation entre les variables X et Y n'est pas nécessairement raisonnable et appropriée, comme le sont de nombreuses autres hypothèses. Il est évidemment nécessaire de faire des hypothèses sur la relation entre les variables, et il n'est pas possible de connaître la relation exacte qui lie $X^1, ..., X^p$. Au contraire, c'est l'une des tâches principales de la statistique d'être capable de produire des connaissances sur des phénomènes pour lesquels il devient complexe de construire une représentation explicative.

— Étudier un ensemble de variables du type $(Y_i, X_i^1, ..., X_i^p)$ considérées comme iid, avec $i = 1, ..., n$, et d'essayer d'estimer à partir des observations empiriques de vecteurs aléatoires de ce type les paramètres β_i et les caractéristiques du bruit ϵ.

Plus formellement, la régression linéaire consiste à poser :

— Une matrice (n, p) $X = \{X_i^j\}_{i=1,...,n;j=0,...,p}$ de valeurs connues où $\forall i = 1, ..., n$ $X_i^0 = 1$, supposée de rang maximal.

— Un vecteur aléatoire $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ de variables aléatoires centrées iid et de variance σ^2.

— Un vecteur aléatoire $Y = (Y_1, ..., Y_n)'$ tel que

$$Y = X\beta + \epsilon$$

où $\beta = (\beta_1, ..., \beta_p)'$ est le vecteur des paramètres à estimer.

Une fois que nous avons introduit le modèle que nous souhaitons étudier et dont nous souhaitons estimer les paramètres, il ne reste plus qu'à introduire des méthodes pour estimer effectivement ces paramètres et faire des prédictions sur les prochaines occurrences de Y, en partant de la connaissance des valeurs des variables $X^1, ..., X^p$.

De nombreuses méthodes peuvent être utilisées pour estimer les paramètres β_i, méthodes dont la qualité est essentiellement jugée sur la base de la simplicité de calcul et des propriétés de l'estimateur obtenu. Dans le cas de la régression linéaire, l'estimateur du vecteur β est choisi de la manière suivante :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \ \|Y - X\beta\|^2 = \sum_{i=1}^{n} (Y_i - X_i\beta)^2.$$

D'un point de vue géométrique, compte tenu de l'hypothèse du rang maximal de la matrice X, la norme que nous venons d'écrire est minimisée lorsque $X\beta$ représente la projection (unique) de Y sur l'espace vectoriel généré par les colonnes de la matrice X. Le calcul explicite, utilisant purement des théorèmes et des concepts d'algèbre linéaire, conduit à la conclusion que

$$\hat{\beta} = (X'X)^{-1}X'Y$$

et que la matrice de projection sur l'espace engendré par X est donnée par

$$P_X = X(X'X)^{-1}X'.$$

Par un calcul direct, nous avons deux propriétés importantes concernant $\hat{\beta}$:

$$\mathbb{E}[\hat{\beta}] = \beta$$
$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

Il est important de pouvoir estimer la variance σ^2 des variables aléatoires indépendantes ϵ_i. Une façon naturelle d'estimer les valeurs prises par les variables aléatoires ϵ_i dans l'échantillon que nous pouvons observer est de définir un vecteur

$$\hat{\epsilon} = Y - X\hat{\beta} = Y - P_X Y = (I - P_X)Y = P_X^{\perp} Y.$$

L'intérêt d'introduire un tel vecteur d'estimation de ϵ est lié à la possibilité d'étudier le comportement de notre bruit. En effet, nous rappelons que, par hypothèse, le vecteur ϵ est un vecteur de variables aléatoires iid et représente donc un véritable échantillon de cette variable aléatoire que nous avons intuitivement appelée bruit. En particulier, la possibilité d'estimer ϵ nous permet d'estimer la variance σ^2 qui, comme nous l'avons déjà mentionné, implique également la loi de l'estimateur $\hat{\beta}$ de β.

En particulier, en définissant

$$\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|^2}{n-p} = \frac{\|P_X^{\perp} Y\|}{n-p},$$

nous obtenons un estimateur non biaisé de σ^2

3.1.2 Régression linéaire dans le cas gaussien et propriétés de l'estimateur MLE

Un cas particulièrement intéressant est celui où l'on suppose que le vecteur ϵ est un vecteur de gaussiennes centrées mutuellement indépendantes, c'est-à-dire que le bruit est distribué de manière gaussienne avec une certaine variance σ^2.

En conservant les hypothèses d'indépendance et de rang que nous avons faites dans le cas plus général, la régression linéaire dans le cas gaussien est essentiellement la régression linéaire dans laquelle les variables $\epsilon_1, ..., \epsilon_n$ sont supposées être indépendantes et de loi gaussienne de variance σ^2.

Dans ce cas, puisque nous sommes sous des hypothèses plus fortes, nous avons toujours l'unicité de la solution $\hat{\beta}$ du problème de minimisation vu précédemment et encore $\hat{\beta}$ et $\hat{\epsilon}$ sont deux estimateurs non biaisés de β et ϵ. En outre :

— Étant donné $Y \sim \mathcal{N}(\beta, \sigma^2)$, nous pouvons facilement prouver que

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1}).$$

— Grâce au théorème de Cochran, nous pouvons déduire que $\frac{(n-p)\hat{\sigma}^2}{\sigma} \sim \chi^2_{n-p}$, où χ^2_{n-p} désigne une variable chi-2 avec $n-p$ degrés de liberté.

— Encore une fois, grâce au théorème de Cochran et au fait que $\hat{\sigma^2}$ est une fonction de $P_X^{\perp} Y$ et que $\hat{\beta}$ est une fonction de $P_X Y$, nous savons que $\hat{\sigma}^2$ et $\hat{\beta}$ sont deux variables aléatoires indépendantes.

Ces propriétés importantes des deux estimateurs que nous étudions nous permettent de construire des intervalles et, plus généralement, des régions de confiance pour les estimateurs, de faire des estimations ponctuelles et d'étudier le comportement de la variance de ces estimateurs.

Nous avons obtenu des résultats très rapides dans ces pages concernant la régression linéaire et nous avons rapidement fourni les idées et les perspectives qui sous-tendent la régression linéaire et la régression linéaire dans le cas gaussien.

Le but ultime de la régression linéaire n'est bien sûr pas seulement l'estimation des paramètres. Certes, il peut être utile de connaître la valeur des paramètres de la régression linéaire, mais l'application naturelle de notre estimation est de pouvoir faire des prédictions dans le futur.

Si l'on considère, dans le cas gaussien, que nous disposons d'un autre vecteur d'observations $X_{n+1} = X_{n+1}^0, ..., X_{n+1}^p$ et si l'on suppose à nouveau

$$Y_{n+1} = X_{n+1}\beta + \epsilon_{n+1}$$

où ϵ_{n+1} a la même distribution d'éléments que le vecteur ϵ, on peut facilement essayer de prévoir Y_{n+1} via

$$\hat{Y}_{n+1} = X_{n+1}\hat{\beta}$$

qui, toujours dans le cas gaussien, respecte de très bonnes propriétés. Il est en effet vrai que

$$Y_{n+1} - \hat{Y}_{n+1} = \mathcal{N}(0, \sigma^2(1 + X'_{n+1}(X'X)^{-1}X_{n+1})).$$

Nous ne procédons pas à la construction explicite des intervalles et des régions de prévision parce qu'ils découlent, avec un peu de travail, des propriétés posées et parce que, dans ce texte, nous ne nous intéressons qu'au rappel et à la réintroduction de certains des outils utilisés dans l'apprentissage. Pour plus d'informations, veuillez consulter le livre [5]

3.1.3 Application du théorème de Lindeberg Feller à la regression linéaire pour la normalité asymptotique

Nous avons vu dans le cadre particulier de la régressien linéaire gaussienne que l'on a la normalité de l'estimateur $\hat{\beta}$. Cependant cette normalité s'appuyait sur l'hypothèse faite sur la loi du bruit. Nous pouvons utiliser le TCL de Lindeberg-Feller multivarié pour établir une normalité asymptotique de $\hat{\beta}$ dans des cadres plus généraux.
Par les notations utilisées dans la section 3.1.1, on a :

$$\hat{\beta} = (X'X)^{-1}X'Y \quad P_X = X(X'X)^{-1}X \quad \text{Var}(\epsilon) = \sigma^2 I_p$$

Pour obtenir la normalité asymptotique, nous devons rajouter une condition supplémentaire sur les X_i. On note $(h_{ii})_{i=1}^n$ les coefficients diagonaux de P_X.

Théorème 3.1.
Si

$$\max_{1\leq i\leq n} h_{ii} \underset{n\to+\infty}{\longrightarrow} 0,$$

alors on a

$$\sigma^{-1}(X'X)^{\frac{1}{2}}(\hat{\beta} - \beta) \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}(0, I_p).$$

(Preuve en annexe)

3.1.4 Regression lineaire généralisée

La régression linéaire présentée jusqu'à présent, bien que très utile dans certaines situations, présente certains défauts. Par exemple, elle n'est pas adéquate quand la variable réponse est censée être bornée ou qu'elle prend ses valeurs dans $\mathbb{N}$ par exemple. Les modèles linéaires généralisés (GLMs) ont été introduits par Nelder et Wedderburn en 1972 et constituent une classe de modèle avec une applicabilité relativement large et des propriétés statistiques intéressantes.
Il est utile aussi de souligner que dans le modèle linéaire que nous avons présenté jusqu'ici, les X_i étaient supposées déterministes. Dans tout ce qui suit, nous ne faisons plus cette hypothèse et nous passons à un modèle où les X_i sont aléatoires.

Dans un GLM, on pose :

• Les réponses Y_i suivent une loi de la famille exponentielle, de fonction de densité de la forme

$$f_\theta(y|x) = a(y_i)b(\theta)\exp(\eta(\theta)\cdot T(y))$$

où a et b sont des fonctions réelles et θ un paramètre. On appelle η le paramètre naturel de la loi.
Cette formulation inclut la plupart des lois usuelles comportant un ou deux paramètres : gaussienne, gamma, Poisson, binomiale ...

• Une fonction de lien $g(\cdot)$ qui exprime une relation fonctionnelle entre la fonction moyenne $\mathbb{E}[Y_i|X_i] := \mu_i$ et le prédicteur linéaire $X_i\beta$ par

$$g(\mu_i) = X_i\beta$$

où X_i est le vecteur de variables explicatives et β est le vecteur des paramètres. La fonction de lien qui associe la moyenne μ_i au paramètre naturel est appelée fonction de lien canonique. Dans ce cas,

$$g(\mu_i) = \eta(\theta) = X_i\beta.$$

Les liens canoniques résultent en des simplifications de calcul. On les préfèrera souvent aux autres fonctions de lien.
Le tableau ci-dessous exhibe la décomposition de certaines loi usuelles en leur forme exponentielles

Distribution	$\mathcal{N}(\mu,\sigma^2)$	Poisson(μ)	Bernoulli(p)	$\mathcal{E}(\mu)$
Paramètre	μ (σ supposé connu)	μ	p	μ
Paramètre naturel η	$\frac{\mu}{\sigma}$	$\ln(\mu)$	$\ln(\frac{p}{1-p})$	$-\mu$
Fonction réciproque de η	$\sigma\eta$	$\exp(\eta)$	$\frac{e^\eta}{1+e^\eta}$	$-\eta$

La fonction réciproque de η est donc la fonction de lien canonique. Nous allons d'ailleurs utiliser cette fonction de lien dans le cas de la régression logistique (réponse bernoulli) dans la section suivante.

En résumé, on fait l'hypothèse d'une relation linéaire sur une transformation de la moyenne (qui offre des avantages calculatoires et statistiques) et une famille exponentielle de la distribution de la réponse. Cette théorie permet d'englober plusieurs autres modèles (linéaire gaussien, logit, log-linéaire). Plusieurs des méthodes utilisés dans les autres modèles peuvent aussi être utilisées en GLM, comme par exemple l'estimation des paramètres β_i par maximum de vraisemblance ou méthode des moindres carrés.

3.2 La régression logistique

La régression logistique, comme la régression linéaire, est un outil important pour étudier la relation de dépendance entre une variable $(X^1, ..., X^p)$ et une variable dépendante (supposée) Y. Dans le cas de la régression linéaire, la variable Y est considérée comme une variable aléatoire qui prend ses valeurs dans tout $\mathbb{R}$ Dans le cas où Y est au contraire une variable aléatoire dichotomique, c'est-à-dire capable de prendre des valeurs du type $0, 1$, ou capable de prendre un nombre fini de valeurs, la régression linéaire ne semble pas être un instrument approprié. Il existe de nombreux cas dans lesquels il peut être utile d'étudier des variables dichotomiques. Parmi tous les cas, celui où il est nécessaire de comprendre si un courrier électronique présentant certaines caractéristiques, correspondant aux valeurs de $(X^1, ..., X^p)'$, doit être placé dans la boîte à spam ($Y = 1$) ou non, et celui où il est possible de prédire le diagnostic d'une certaine maladie ($Y = 1$) à partir de certains paramètres biométriques, correspondant eux aussi aux valeurs de $(X^1, ..., X^p)'$, sont exemplaires et importants.

Dans cette brève discussion, nous n'examinerons que le cas où la variable aléatoire Y est une variable dichotomique. Dans ce cas, l'hypothèse du modèle concernant la relation entre les variables indépendantes et la variable dépendante Y est la suivante :

$$\log\left(\frac{\mathbb{P}(Y=1|X^1,...,X^p)}{\mathbb{P}(Y=0|X^1,...,X^p)}\right) = \beta_1 X^1 + ... + \beta_p X^p.$$

Il est important de noter que ce modèle n'a de sens que si l'on exclut les cas pathologiques pour lesquels $\mathbb{P}(Y=1|X^1,...,X^p)$ peut valoir 0 ou 1.

Comme il s'agit d'une variable dichotomique, $\mathbb{P}(Y=1|X^1,...,X^p) = 1-\mathbb{P}(Y=0|X^1,...,X^p)$ et en notant pour simplicité

$$\pi = \mathbb{P}(Y=1|X^1,...,X^p)$$

nous pouvons dire que supposer le modèle que nous venons de proposer équivaut à affirmer que

$$\pi = \frac{e^{\beta_1 X^1+...+\beta_p X^p}}{1+e^{\beta_1 X^1+...+\beta_n X^p}}.$$

Cette deuxième forme nous permet de voir plus explicitement la dépendance directe entre Y et $X^1,...,X^p$.

Le modèle que nous venons de proposer est appelé modèle logistique précisément parce qu'il utilise la fonction logit

$$\text{logit}(x) = \log\frac{x}{1-x}$$

introduite pour la première fois par le statisticien et médecin américain Joseph Berkson en 1944 [2].

Tout simplement, comme dans le cas de la régression linéaire, une fois le modèle posé, il ne reste plus qu'à essayer d'estimer les paramètres du vecteur du modèle $\beta = (\beta_1,...,\beta_p)'$, la dernière étape pour comprendre la relation entre le vecteur $X = (X^1,...,X^p)$ et Y. Comme dans le cas de la relation de régression linéaire, il est donc important d'identifier une méthode d'estimation qui soit à la fois relativement facile à calculer et qui produise un estimateur $\hat{\beta}$ présentant des propriétés intéressantes.

Contrairement au cas précédent, où, pour trouver $\hat{\beta}$, on essayait de minimiser un risque empirique, une méthode très courante pour estimer β consiste à rechercher un estimateur du maximum de vraisemblance.

Étant donné le vecteur aléatoire $(X^1,...,X^p,Y)$, nous pouvons affirmer que la loi de la variable aléatoire $Y|X_1,...,X_p$ admet une densité par rapport à la mesure $\delta_0+\delta_1$ et que cette densité $f(y)$ peut être écrite sous la forme

$$f(y) = \pi^y(1-\pi)^{(1-y)}.$$

Dans ce cas, en considérant un échantillon $\{(X_i^1,...,X_i^p,Y_i)\}_{i=1,...,n}$ tel que l'échantillon de variables aléatoires $\{Y_i|X_i^1,...,X_i^p\}_{i=1,...,n}$ est un échantillon de variables aléatoires iid, dont la densité jointe évaluée sur l'échantillon s'écrit comme le produit des densités :

$$f(Y_1,...,Y_n|X_1,...,X_n) = \prod_{i=1}^n f(Y_i) = \prod_{i=1}^n\left(\pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)}\right)$$

En supposant que π_i est différent de 0 et 1, la maximisation de cette fonction en fonction de π est équivalente à la maximisation de la fonction

$$\log\left(\prod_{i=1}^n \pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)}\right) = \sum_{i=1}^n\left(Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i)\right)$$

Nous pouvons encore réécrire cette fonction, en tant que fonction de $\beta_0,...,\beta_p$ comme suit :

$$\begin{aligned}\log f_\beta(Y_1,...,Y_n|X_1,...,X_n) &= \sum_{i=1}^n\left(y_i\log\left(\frac{e^{\beta_1 X_i^1+...+\beta_p X_i^p}}{1+e^{\beta_1 X_i^1+...+\beta_n X_i^p}}\right) + (1-Y_i)\log\left(1-\frac{e^{\beta_1 X_i^1+...+\beta_p X_i^p}}{1+e^{\beta_1 X_i^1+...+\beta_n X_i^p}}\right)\right)\\ &= \sum_{i=1}^n\left(Y_i\log\left(\frac{e^{\beta_1 X_i^1+...+\beta_p X_i^p}}{1+e^{\beta_1 X_i^1+...+\beta_n X_i^p}}\right) + (1-Y_i)\log\left(\frac{1}{1+e^{\beta_1 X_i^1+...+\beta_n X_i^p}}\right)\right).\end{aligned}$$

Maximiser cette dernière fonction signifie, à un niveau moral, rechercher le β capable de rendre notre échantillon observé aussi probable que possible. Inversement, donc, plus un paramètre est capable de rendre un échantillon probable, plus il est probable qu'on se trouve devant le vrai paramètre.

En notant pour commodité $X_i\beta$ la somme $\beta_1 X_i^1 + ... + \beta_p X_i^p$, on peut réécrire la log-vraisemblance dessus :

$$\log f_\beta(Y_1, ..., Y_n | X_1, ..., X_p) = \sum_{\{i|Y_i=1\}} \log\left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right) + \sum_{\{i|Y_i=0\}} \log\left(\frac{1}{1+e^{X_i\beta}}\right).$$

3.2.1 Existence et consistance de estimateur du maximum de vraisemblance

Dans le cas de la régression logistique, on peut apprécier la complexité de la fonction de vraisemblance qu'il faut maximiser et, surtout, la complexité dans la compréhension même de l'existence et unicité d'un point de maximum de $\log f_\beta$. Un autre gros problème qu'il faut aborder, une fois qu'on a assuré l'existence et unicité d'un estimateur MLE, est le problème de la détermination du point de maximum et de ses propriétés asymptotiques.

Le parcours qu'on propose part de l'établissement d'un théorème d'existence et unicité de la solution à n fixé qui permet d'identifier des conditions d'existence suffisantes et nécessaires. Ensuite nous allons énoncer et démontrer un important théorème qui donne des conditions suffisantes pour avoir consistance forte et existence asymptotique de l'estimateur MLE dans le cas de la régression logistique. Enfin on va s'intéresser à la convergence en loi asymptotique de notre estimateur. Avant de commencer la discussion, il faut seulement dire que, comme dans la régression linéaire on va noter comme $\hat{\beta}$ l'estimateur MLE :

$$\hat{\beta}_n = \arg\max_{\beta \in \mathbb{R}^p} l_n(\beta, X, Y)$$

où $l_n(\beta, X, Y) = \log f_\beta(Y_1, ..., Y_n | X_1, ..., X_n)$ est la log-vraisemblance, Y est le vecteur colonne $(Y_1, ..., Y_n)$ et X la matrice n x p composée des lignes $X_1, ..., X_n$.

Dans toute la suite de la section on prendra en considération seulement les cas où $n \geq p$ et on va supposer que pour tout n la matrice X est une matrice de rang p presque sûrement.

Avant de donner les résultats d'existence et unicité, c'est utile d'introduire ces définitions.

Definition 3.1 (Séparation complète)**.** On dit que un ensemble de points $\{(x_i, y_i)\}_{i=1,...,n}$ est complètement séparé si $\exists\beta$ tel que

1. $x_i\beta > 0$ $\forall i$ t.q. $y_i = 1$
2. $x_i\beta < 0$ $\forall i$ t.q. $y_i = 0$

Definition 3.2 (Séparation quasi-complète)**.** On dit que un ensemble de points $\{(x_i, y_i)\}_{i=1,...,n}$ est quasi-complètement séparé si $\exists\beta$ tel que

1. $x_i\beta \geq 0$ $\forall i$ i t.q. $y_i = 1$
2. $x_i\beta \leq 0$ $\forall i$ t.q. $y_i = 0$
3. $\exists$ i t.q. $x_i\beta = 0$ (la troisième condition sert à distinguer la complétude et la quasi-complétude dans tous cas)

Definition 3.3 (Overlap)**.** On dit que dans l'ensemble de points $\{(x_i, y_i)\}_{i=1,...,n}$ il y a overlap si l'ensemble n'est pas complètement ou quasi-complètement séparable, c'est à dire si $\forall\beta$ il existe $i \in \{1, ..., n\}$ tel que

$$(y_i = 1 \text{ et } x_i\beta < 0) \text{ ou } (y_i = 0 \text{ et } x_i\beta > 0)$$

Une fois qu'on a donné cette définition on peut énoncer un théorème important, inclus et démontré dans l'article de Albert et Anderson (1984) [1] qui nous clarifie la situation par rapport à l'existence et unicité.

Théorème 3.2 (Existence et unicité du maximum de vraisemblance dans la régression logistique binaire)**.** Soit $\{(X_i, Y_i)(\omega)\}_{i=1,...,n}$ une réalisation de l'échantillon des vecteurs indépendants $\{(X_i, Y_i)\}_{i=1,...,n}$. Soit $\omega \in \Omega$:

1. $\{(X_i, Y_i)\}_{i=1,...,n}$ respecte les hypothèses du modèle présenté dans ce qui précède.
2. Pour tout $i = 1, ..., n$ $X_i(\omega)$ vecteur aléatoire de dimension p et $Y_i(\omega)$ variables aléatoire binaire à valeurs dans $\{0, 1\}$.
3. $n \geq p$ et les vecteurs $\{X_i(\omega)\}_{i=1,...,n}$ sont linéairement indépendants.

Alors :

1. Si $\{(X_i, Y_i)(\omega)\}_{i=1,...,n}$ est complètement séparé ou quasi-complètement séparé, il n'existe pas un vecteur $\hat{\beta} = \arg\max_{\mathbb{R}^p} \log(f_\beta(Y_1(\omega), ..., Y_n(\omega)|X_1(\omega), ..., X_n(\omega)))$.
2. Si pour l'ensemble $\{(X_i, Y_i)(\omega)\}_{i=1,...,n}$ il y a overlap, il existe une solution unique, c'est à dire il existe un unique $\hat{\beta}$ t.q. $\hat{\beta} = \arg\max_{\mathbb{R}^p} \log(f_\beta(Y_1(\omega), ..., Y_n(\omega)|X_1(\omega), ..., X_n(\omega)))$.

Le fait que dans les cas de séparation on ne peut pas affirmer l'existence d'un unique estimateur est interprétable de manière géométrique. En effet si il existe un β qui fait le travail de séparation, on a une droite qui est capable de diviser l'espace $\mathbb{R}^p$ en deux semi-espaces qui contiennent respectivement le nuage de points $X_i(\omega)$ t.q. $Y_i(\omega) = 1$ et le nuage de points $X_i(\omega)$ t.q. $Y_i(\omega) = 0$. Dans ce cas on peut voir avec un simple dessin (par exemple avec $p = 1$) que on n'a pas l'existence d'un maximum.

Évidemment, notre théorème ne nous donne pas des conditions effectives pour assurer l'existence d'un estimateur de MLE, c'est à dire que si on observe un échantillon $\{(X_i, Y_i)(\omega)\}_{1,...,n}$ qui respecte la propriété de overlapping on peut assurer l'existence d'un point de maximum de la log-vraisemblance, mais on ne sait rien sur l'existence de cet estimateur avant de l'observation de l'échantillon. Que veut dire cette limite en terme pratique ? En terme pratique, on ne peut pas garantir qu'une collecte de données du type $(X^1, ..., X^p, Y)$ nous conduira vers l'observation d'un échantillon sur le quel on peut faire de l'optimisation (c'est à dire on peut estimer β par un $\hat{\beta}$). Supposons d'être des scientifiques qui étudient le rapport entre certaines caractéristiques physiologique et la probabilité de diagnostique d'une maladie. Si nous ne pouvons pas établir des conditions plus générales d'existence et d'unicité, nous serions obligés de nous lancer dans une expérimentation pour estimer le paramètre β sans savoir si l'expérience conduira effectivement à un estimateur de β. En outre si on n'est pas capable de garantir l'unicité du MLE (c'est à dire on a plusieurs points de maximum $\hat{\beta}_1, ..., \hat{\beta}_m$) on peut avoir beaucoup de difficulté dans la compréhension de quel $\hat{\beta}_i$ utiliser dans notre travail de prévision future.

Pour pouvoir travailler avec la régression logistique, il faut donc être capable de trouver des conditions qui nous permettent d'affirmer l'existence de l'estimateur MLE dans tous, ou dans la plus part, des échantillons du type $\{(X_i, Y_i)\}_{i=1,...,n}$. Heureusement, on a un résultat qui établi en 1981 par Gourieroux et Monfort [4] qui des conditions suffisantes pour avoir existence asymptotique presque sûre et consistance de l'estimateur MLE.

Dans la suite, on notera $F(x) = \frac{1}{1+e^{-x}}$ et $f(x)$ sa dérivée. On peut noter qu'on a

$$F(X_i\beta) = p_i \text{ et aussi } f(X_i\beta) = p_i(1 - p_i)$$

On peut donc écrire

$$l_n(\beta, X, Y) = \sum_{i=1}^{n} Y_i(X_i\beta) + \sum_{i=1}^{n} \log(1 - F(X_i\beta))$$

et que avec quelques calculs on peut trouver que

$$\nabla\ l_n(\beta, X, Y) = \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} F(X_i\beta)X_i,$$

$$\mathcal{H}\ l_n(\beta, X, Y) = -\sum_{i=1}^{n} X_i' X_i f(X_i\beta).$$

Théorème 3.3 (Existence asymptotique et consistance forte de l'estimateur MLE)**.**

Soit $\{(X_i, Y_i)\}_{i\in\mathbb{N}}$ une suite de vecteurs aléatoires de dimension $p+1$ indépendants i.i.d tels que pour tout $i \in \mathbb{N}$, $\mathbb{P}(Y_i = 1|X_i) = F(X_i\beta)$. On suppose que presque sûrement pour tout $n \geq p$ la matrice X composée des lignes X_i est de rang p.

Dans ce cas la matrice $\mathcal{H}\ l_n(\beta, X, Y)$ est définie négative.

En outre si

— Il existe $M_0 > 0$ t.q. presque sûrement pour tout $i \in \mathbb{N}$ et pour tout $j = 1, ..., p\ |X_i^j| < M_0$

— Il existe $M_1 > 0$ t.q. presque sûrement pour tout $n \in \mathbb{N}$ la plus grande valeur propre (λ_{pn}) et la plus petite valeur propre (λ_{1n}) de $-\mathcal{H}\ l_n(\beta, X, Y)$ sont t.q. $\lambda_{pn}/\lambda_{1n} < M_1$.

Alors L'estimateur $\hat{\beta}_n$ existe presque sûrement asymptotiquement (i.e. pour tout ω dans un ensemble de mesure 1 pour tous les n assez grands il existe un point de maximum de la fonction $l_n(\beta, X, Y)(\omega)$) et $\hat{\beta}_n$ est fortement consistant si et seulement si

$$[-\mathcal{H}\ l_n(\beta_0, X, Y)]^{-1}\nabla\ l_n(\beta_0, X, Y) \xrightarrow[n\to+\infty]{p.s.} 0.$$

(Preuve en annexe)

3.2.2 La normalité asymptotique de l'estimateur du maximum de vraisemblance

Théorème 3.4.

Sous les hypothèses d'existence asymptotique et si $\mathbb{E}[f(X_1\beta)X_1'X_1]$ (qu'on notera A) est définie positive (l'espérance est prise composante par composante), alors on a

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, A^{-1}BA^{-1})$$

avec $B = \mathbb{E}[(Y_1 - F(X_1\beta))^2X_1'X_1]$

(Preuve en annexe)

4 Propriétés conditionnelles

Avant d'introduire effectivement le contenu de l'article, nous nous consacrons à la définition de certains objets et à l'énoncé de quelques théorèmes qui seront utiles dans la suite du texte. Beaucoup des éléments introduits peuvent être retrouvés dans (Xiong et Li, 2008) [8].

4.1 Symboles o et O conditionnels

On avait introduit la notion de $o_{\mathbb{P}}(1)$ et de $O_{\mathbb{P}}(1)$. Maintenant on introduit deux notions similaires qui sont strictement liées aux notions précédemment définies.

Definition 4.1.
Soit (X_n) une suite de v.a.r et $(\mathcal{F}_n)$ une suite de tribu On dit que :

— $X_n = o_{\mathbb{P}|\mathcal{F}_n}(1)$ si $\mathbb{P}(|X_n| > \epsilon|\mathcal{F}_n) \xrightarrow[n\to+\infty]{\mathbb{P}} 0$ pour tout $\epsilon > 0$, c'est à dire pour tout $\epsilon, \delta > 0$

$$\mathbb{P}(\mathbb{P}(|X_n| > \epsilon|\mathcal{F}_n) > \delta) \underset{n\to+\infty}{\longrightarrow} 0.$$

— $X_n = O_{\mathbb{P}|\mathcal{F}_n}(1)$ si pour tout $\delta > 0$, $\sup_n \mathbb{P}(\mathbb{P}(|X_n| > M|\mathcal{F}_n) > \delta) \underset{M\to+\infty}{\longrightarrow} 0$.

— On écrira $X_n = O_{\mathbb{P}|\mathcal{F}_n}(f(n))$ et $X_n = o_{\mathbb{P}}(f(n))$ pour indiquer respectivement $\frac{X_n}{f(n)} = O_{\mathbb{P}|\mathcal{F}_n}(1)$ ou $\frac{X_n}{f(n)} = o_{\mathbb{P}}(1)$, où $f(n)$ indique simplement une fonction déterministe de n.

Ces mêmes définitions sont valables pour les vecteurs aléatoires en prenant la norme et non pas la valeur absolue.

4.2 Convergence en loi conditionnelle et quelques premières propriétés

Definition 4.2.
Soit (X_n) une suite de variable aléatoires dans $\mathbb{R}^p$ et X une variable aléatoire dans $\mathbb{R}^p$. On considère $(\mathcal{F}_n)$ une suite de tribus. On dit que X_n converge en loi conditionnelle vers X (et on écrit) $X_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} X$ si pour toute fonction continue $\phi : \mathbb{R}^p \to \mathbb{R}$:

$$\mathbb{E}(\phi(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi(X)) = o_{\mathbb{P}}(1).$$

Théorème 4.1.
On a les propriétés suivantes. Soit X_n suite des v.a.r. . Pour tout tribu $\mathcal{F}_n$:

1. $X_n = o_{\mathbb{P}|\mathcal{F}_n}(1)$ ssi $X_n = o_{\mathbb{P}}(1)$.
2. $X_n = O_{\mathbb{P}|\mathcal{F}_n}(1)$ ssi $X_n = O_{\mathbb{P}}(1)$.
3. $X_n = o_{\mathbb{P}}(1)$ ssi $X_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} 0$.

(Preuve des deux premiers points en annexe)

En outre, il est possible de montrer que :

— $O_{\mathbb{P}}(1) + O_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$.
— $O_{\mathbb{P}}(1)O_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$.
— $O_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$.
— $|X_n| \leq |Y_n|$ et $Y_n = O_{\mathbb{P}}(1)$ alors $X_n = O_{\mathbb{P}}(1)$.
— $|X_n| \leq |Y_n|$ et $Y_n = o_{\mathbb{P}}(1)$ alors $X_n = o_{\mathbb{P}}(1)$.
— Si $X_n = O_{\mathbb{P}}(1)$ alors $(X_n)^k = O_{\mathbb{P}}(1)$ pour tout $k > 0$.

Théorème 4.2. (Slutsky conditionnel)
Soient (X_n) et (Y_n) deux suite de variables aléatoires dans $\mathbb{R}^p$, (A_n) suite de matrices aléatoires avec p

colonnes et $(\mathcal{F}_n)$ une suite de tribus. Si $X_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} X$, $A_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} A$ (avec A déterministe) et $Y_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} c$ ($c \in \mathbb{R}^p$ déterministe), alors :

$$A_n X_n + Y_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} AX + c.$$

On s'intéresse maintenant à reconstruire un partie de la théorie des probabilités pour arriver à démontrer un théorème de Lindeberg-Feller conditionnel. Dans l'ordre suivant on va énoncer et démontrer :

4.3 Théorème de Lévy conditionnel

Definition 4.3 (Uniforme equicontinuité en probabilité)**.**
Soient $f_n : \mathbb{R}^p \times \Omega \to \mathbb{R}$ une suite de fonctions aléatoires. On dit que la suite (f_n) est uniformément equicontinue en probabilité si

$$\forall \epsilon, \delta > 0, \text{ il existe } \delta_\epsilon > 0 \text{ t.q., pour tout } x, y \in \mathbb{R}^p \text{ t.q. } \|x - y\| < \delta_\epsilon, \text{ on a}$$

$$\sup_n \mathbb{P}(|f_n(x,.) - f_n(y,.)| > \delta) < \epsilon.$$

Definition 4.4 (Convergence uniforme en probabilité)**.**
Soient $f_n, f : \mathbb{R}^p \times \Omega \to \mathbb{R}$ des fonctions aléatoires et $A \subset \mathbb{R}^p$. On dit que (f_n) converge uniformément sur A vers f en probabilité si

$$sup_{x\in A}|f_n(x,.), f(x,.)| = o_{\mathbb{P}}(1)$$

Théorème 4.3. (Théorème de convergence des intégrales en probabilité)
Soient $f_n, f : \mathbb{R}^p \times \Omega \to \mathbb{R}$ des fonctions aléatoires. On suppose que (f_n) converge uniformément en probabilité vers f sur tout compact K et que

$$|f_n(x,\omega)|, |f(x,\omega)| \leq h(x) \text{ p.s., avec } h \in L^1(\mathbb{R}^p)$$

Alors

$$\int_{\mathbb{R}^p} |f_n(x,.) - f(x,.)|dx = o_{\mathbb{P}}(1)$$

Théorème 4.4.
La convergence en loi conditionnel est équivalente à la convergence des espérances des fonctions $C_c^2(\mathbb{R}^p)$, comme dans le cas non conditionnel. Plus formellement, si X_n, X sont dans $\mathbb{R}^p$, on à l'équivalence :

1. $X_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} X$
2. $\mathbb{E}(\phi(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi(X)) = o_{\mathbb{P}}(1)$ pour tout fonction $\phi \in C_b^2(\mathbb{R}^p)$
3. $\mathbb{E}(\phi(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi(X)) = o_{\mathbb{P}}(1)$ pour tout fonction $\phi \in C_c^2(\mathbb{R}^p)$.

où $C_c^2(\mathbb{R}^p)$ (resp. $C_b^2(\mathbb{R}^p)$) est l'espace des fonctions de $\mathbb{R}^p$ à support compact (resp. bornées).
(Preuve en annexe)

Théorème 4.5. (Théorème de Lévy conditionnel univarié)
Soient X_n, X variables aléatoires réelles, $(\mathcal{F}_n)$ une suite de tribus. Ces deux assertions sont équivalentes :

1. $\mathbb{E}(e^{itX_n}|\mathcal{F}_n) - \mathbb{E}(e^{itX}) = o_{\mathbb{P}}(1)$ pour tout $t \in \mathbb{R}$
2. $\mathbb{E}(\phi(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi(X)) = o_{\mathbb{P}}(1)$ pour tout fonction $\phi \in C_c^2(\mathbb{R})$

(Preuve en annexe)

Théorème 4.6 (Théorème de Lévy conditionnel multivarié)**.**
Soient X_n, X vecteurs aléatoires réels de $\mathbb{R}^p$, $(\mathcal{F}_n)$ une suite de tribus. Ces deux assertions sont équivalentes :

1. $\mathbb{E}(e^{itX_n}|\mathcal{F}_n) - \mathbb{E}(e^{itX}) = o_{\mathbb{P}}(1)$ pour tout $t \in \mathbb{R}^p$
2. $\mathbb{E}(\phi(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi(X)) = o_{\mathbb{P}}(1)$ pour tout fonction $\phi \in C_c^2(\mathbb{R}^p)$

(Preuve en annexe)

Théorème 4.7. (Théorème de Cramer Wald conditionnel)

Soit $(X_n)_{n\geq 0}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^p$ et soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^p$

$$X_n \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} X \iff \forall a \in \mathbb{R}^p \quad \langle a, X_n\rangle \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} \langle a, X\rangle$$

(Preuve en annexe)

4.4 Théorème de Lindeberg-Feller conditionnel

Théorème 4.8. (Théorème Lindeberg-Feller conditionnel univarié)

Soit $\{X_{nj}\}_{n\in\mathbb{n},j\leq r_n}$ un tableau triangulaire de variables aléatoires réelles dans $L^2(\Omega,\mathbb{P})$ et $(\mathcal{F}_n)_n$ une suite de tribus. Soit

$$\tilde{X}_{nj} = \frac{X_{nj} - \mathbb{E}(X_{nj})}{\sqrt{\sum_{j=1}^{r_n} \mathrm{Var}(X_{nj}|\mathcal{F}_n)}}$$

Si pour tout $\epsilon > 0$

$$\sum_{j=1}^{r_n} \mathbb{E}(|\tilde{X}_{nj}|^2 \mathbb{1}_{|\tilde{X}_{nj}|>\epsilon}|\mathcal{F}_n) \xrightarrow[n\to+\infty]{\mathbb{P}} 0$$

Alors

$$\sum_{j=1}^{r_n} \tilde{X}_{nj} \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} \mathcal{N}(0,1)$$

(Preuve en annexe)

Théorème 4.9. (Théorème Lindeberg Feller conditionnel multivarié)

Soit $\{X_{nj}\}_{n\in\mathbb{n},j\leq r_n}$ un tableau triangulaire de vecteurs aléatoires réelles de $\mathbb{R}^p$ dans $L^2(\Omega,\mathbb{P})$ et $\{\mathcal{F}_n\}_n$ une suite de tribu. Soit

$$\tilde{X}_{nj} = (\sum_{j=1}^{r_n} \mathrm{Cov}(X_{nj}|\mathcal{F}_n))^{-\frac{1}{2}}(X_{nj} - \mathbb{E}(X_{nj}))$$

Si pour tout $\epsilon > 0$

$$\sum_{j=1}^{r_n} \mathbb{E}(\|\tilde{X}_{nj}\|^2 \mathbb{1}_{\|\tilde{X}_{nj}\|>\epsilon}|\mathcal{F}_n) \xrightarrow[n\to+\infty]{\mathbb{P}} 0$$

Alors

$$\sum_{j=1}^{r_n} \tilde{X}_{nj} \xrightarrow[n\to+\infty]{d|\mathcal{F}_n} \mathcal{N}(0, I_p)$$

(Preuve en annexe).

5 Subsampling and Optimal subsampling

5.1 Introduction de l'estimateur par sous-échantillonage

Comme nous l'avons mentionné dans l'introduction, il n'est pas toujours possible, dans le cas de la régression logistique et dans d'autres cas d'estimation paramétrique, d'identifier un estimateur basé sur l'ensemble de l'échantillon disponible en raison de limites computationnelles. Il est donc nécessaire d'introduire des estimateurs basés sur un sous-échantillon extrait de l'échantillon original, dont les propriétés asymptotiques doivent être approfondies et ne sont pas directement déductibles des propriétés asymptotiques des estimateurs basés sur l'ensemble de l'échantillon.

Dans cette section, nous nous concentrons sur l'approfondissement d'une technique de sous-échantillonnage dans le cas de la régression logistique, telle que proposée dans l'article publié en 2018 "Optimal Subsampling for Large Sample Logistic Regression" par HaiYing Wang, Rong Zhu et Ping Ma [7].

Comme nous venons de le mentionner, nous nous intéressons à l'estimation du paramètre β de la régression logistique à l'aide d'un estimateur basé sur un sous-échantillon de taille $r = r(n)$ extrait du grand échantillon $(X_i, Y_i)_{i=1,...,n}$. Pour que la taille r de l'échantillon que nous extrayons, à la fois pour réduire effectivement le temps de calcul de l'estimateur et pour garantir des propriétés importantes de l'estimateur futur, nous demandons que $\lim_n \frac{r(n)}{n} = 0$ et que $r(n)$ tend vers l'infini quand n tend vers l'infini.

Une technique largement utilisée et proposée dans l'article que nous souhaitons approfondir consiste essentiellement à tirer $I_1, ..., I_r$ variables aléatoires indépendantes prenant des valeurs dans l'ensemble $1, ..., n$ avec des poids de probabilité $\pi_1, ..., \pi_n$ et à chercher à déduire un estimateur à partir du sous-échantillon $(X_{I_j}, Y_{I_j})_{j=1,...,r}$

À partir de cet échantillon sélectionné, on cherche à maximiser la fonction

$$l_n^*(\beta, X, Y) = \frac{1}{r} \sum_{j=1}^{r} \frac{Y_{I_j} \log(F(X_{I_j}\beta)) + (1 - Y_{I_j}) \log(1 - F(X_{I_j}\beta))}{\pi_{I_j}}$$

et on cherche le $\tilde{\beta}$ t.q.

$$\tilde{\beta}_n \in \arg\max_{\beta \in \mathbb{R}^p} l_n^*(\beta, X, Y).$$

La fonction à maximiser est choisie de cette manière car il est facile de montrer que

$$\mathbb{E}(l_n^*(\beta, X, Y)|\mathcal{F}_n) = l_n(\beta, X, Y)$$

où avec $\mathcal{F}_n$ désigne la tribu engendrée par $\{(X_i, Y_i)\}_{i=1,...,n}$. Cette propriété va jouer un rôle important dans la suite du texte et dans les preuves de théorèmes qu'on va énoncer.

Comme dans le cas de l'estimation par MLE, l'estimateur $\tilde{\beta}$ n'a de sens que s'il respecte certaines propriétés de consistance et de normalité asymptotique. Nous nous tournons donc vers l'énoncé de deux théorèmes (l'un sur la consistance et l'autre sur la normalité asymptotique) qui justifient et donnent un sens à l'utilisation d'estimateurs extraits de l'échantillon d'origine.

IMPORTANT : Toutes les hypothèses qu'on va faire dans notre théorèmes s'ajoutent aux hypothèses qu'on a déjà fait sur les théorèmes précédents d'existence unicité, consistance et normalité de l'estimateur MLE.

5.2 La consistance

Théorème 5.1.

Si les suivantes hypothèses sont respectées :

1. $n^{-1}\mathcal{H}\ l_n(\hat{\beta}, X, Y)$ tend en probabilité vers une matrice H définie positive lorsque n tend vers l'infini, c'est à dire pour tout $\epsilon > 0$

$$\mathbb{P}(\|n^{-1}\mathcal{H}\ l_n(\hat{\beta}_n, X, Y) - H\| > \epsilon) \longrightarrow 0.$$

2. $\frac{1}{n}\sum_{i=1}^{n}\|X_i\| = O_{\mathbb{P}}(1)$.
3. $\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|X_i\|^k}{\pi_i} = O_{\mathbb{P}}(1)$, pour k=2,4.
4. $\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i} = O_{\mathbb{P}}(1)$.

alors :

— $\|\tilde{\beta}_n - \hat{\beta}_n\| = o_{\mathbb{P}|\mathcal{F}_n}(1)$, c'est à dire pour tout $\epsilon, \delta > 0$

$$\mathbb{P}(\mathbb{P}(\|\tilde{\beta}_n - \hat{\beta}_n\| > \epsilon|\mathcal{F}_n) > \delta) \underset{n\to+\infty}{\longrightarrow} 0.$$

— $\|\tilde{\beta}_n - \hat{\beta}_n\| = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$, c'est à dire pour tout $\delta > 0$

$$\sup_n \mathbb{P}(\mathbb{P}(\sqrt{r}\|\tilde{\beta}_n - \hat{\beta}_n\|) > M) > \delta) \underset{M\to+\infty}{\longrightarrow} 0.$$

(Preuve en Annexe)

Remarque : Il y a un passage dans la preuve que nous n'avons pas pu prouver dans le temps qui était imparti à ce TER. Ce passage était pris comme acquis par l'auteur de l'article et il se trouve que ce n'est pas évident du tout. Nous avons mis dans les annexes notre avancée pour la preuve de ce passage et les problèmes que nous avons rencontrés.

5.3 La normalité asymptotique

On a donné un résultat de consistance de l'estimateur basé sur le sous échantillon à l'estimateur MLE et, surtout, un résultat de vitesse de convergence, vitesse qui correspond à $\sqrt{r}$. C'est exactement à cette vitesse que $\tilde{\beta} - \hat{\beta}$ converge en loi vers la loi gaussienne multivariée. Dans la suite on va noter

1. $H_n^* = n^{-1}\mathcal{H}l_n^*(\hat{\beta}, X, Y)$
2. $H_n = n^{-1}\mathcal{H}l_n(\hat{\beta}, X, Y)$
3. $V_c = \mathrm{Cov}(n^{-1}\nabla(l_n^*(\hat{\beta}, X, Y))|\mathcal{F}_n) = \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{(Y_i - F(X_i'\hat{\beta}))^2 X_i' X_i}{\pi_i}$

Théorème 5.2. Si on ajoute aux hypothèses du théorème 5.1 l'hypothèse qu'il existe un $\delta > 0$ t.q.

$$n^{-(2+\delta)}\sum_{i=1}^{n}\pi_i^{-1-\delta}\|X_i\|^{2+\delta} = O_{\mathbb{P}}(1)$$

alors on a le suivant résultat de convergence asymptotique :

$$V^{-\frac{1}{2}}(\tilde{\beta} - \hat{\beta}) \underset{n\to+\infty}{\overset{d|\mathcal{F}_n}{\longrightarrow}} \mathcal{N}(0, I_p)$$

où

$$V = H_n^{-1}V_c H_n^{-1}$$

5.4 Les poids optimaux

Nous avons finalement réussi à prouver l'existence de certains théorèmes de convergence pour notre estimateur $\tilde{\beta}$. Toutefois, pour mettre en œuvre notre estimateur, nous devons connaître les poids à associer aux éléments de l'échantillon. Jusqu'à présent, en effet, nos poids sont restés une inconnue. Nous avons demandé à nos poids de respecter certaines propriétés, à savoir les hypothèses des théorèmes de consistance et la normalité asymptotique de l'estimateur construit par sous-échantillonnage, mais nous n'avons jamais précisé s'il existe des poids optimaux et, si c'est le cas, comment les calculer. Heureusement, l'article de Wang et al.

propose deux théorèmes qui permettent d'identifier quels sont, étant donné un échantillon d'observations, les poids optimaux à choisir. Dans les suivantes lignes on va énoncer deux théorèmes qui offrent deux façons différentes de choisir les poids optimalement. Avant d'énoncer les résultats il faut définir qu'est ce que veut dire "optimal" dans ce contexte. Les résultats qu'on va énoncer montrent que avec un bon choix on est capable de réduire la norme de la matrice de covariance impliqué dans la normalité asymptotique de l'estimateur $\tilde{\beta}$. L'optimalité est donc défini au sens de la réduction de la variance asymptotique. En fait, on ne va pas parler proprement de réduction de la norme de la matrice de covariance, mais dans les deux différentes propositions on va parler respectivement de la minimisation de la trace de V et de la minimisation de la trace de V_c, où V et V_c sont les matrices définies dans le théorème 5.2.

Dans la première proposition des poids on peut voir que la minimisation de la trace de V représente approximativement une minimisation du MSE, résultat qu'on pourra apprécier graphiquement dans les simulations numériques.

Dans la deuxième proposition des poids, par contre, il y a numériquement minimisation du MSE, mais c'est plus difficile de relier théoriquement la minimisation de $tr(V_c)$ avec la minimisation de $\mathbb{E}(\|\tilde{\beta} - \hat{\beta}\|^2)$. Dans ce deuxième cas on utilisera la notation $\pi_i^{V_c}$.

Nous sommes prêts à énoncer les deux théorèmes. Pour les preuves on renvoie à [7].

Théorème 5.3. Soit

$$\pi_i^{MSE} = \frac{|Y_i - F(X_i\hat{\beta})|\|(-\mathcal{H}_n l_n(\beta, \hat{X}, Y))^{-1} X_i\|}{\sum_{j=1}^n |Y_j - F(X_j\hat{\beta})|\|(-\mathcal{H}_n l_n(\beta, \hat{X}, Y))^{-1} X_j\|}$$

alors $tr(V)$ est minimisée.

Théorème 5.4. Soit

$$\pi_i^{V_c} = \frac{|Y_i - F(X_i\hat{\beta})|\|X_i\|}{\sum_{j=1}^n |Y_j - F(X_j\hat{\beta})|\|X_j\|}$$

alors $tr(V_c)$ est minimisée.

On peut observer que nos poids dépendent fortement de $\hat{\beta}$. Calculer l'estimateur MLE pour calculer les poids pour sous-échantillonner n'est, évidemment, pas le bon choix ! Il faut donc trouver une façon de calculer les poids sans connaître l'estimateur MLE ou il faut se réorienter vers un choix de poids qui n'implique pas la connaissance du MLE.

Pour résoudre ce problème l'auteur de l'article [7] propose l'utilisation d'un algorithme en deux pas qui consiste 1) dans un calcul approximé des poids π_i et 2) dans l'implémentation effective de l'algorithme.

5.5 L'algorithme en deux pas

La description de l'algorithme en deux pas est le dernier pas de notre parcours. Une fois que l'algorithme est décrit on aura tous les outils pour procéder avec l'implémentation effective du sous-échantillonage. On explique donc le déroulement de l'algorithme :

1. Premièrement on s'intéresse à la construction des poids π_i pour sous-échantillonner. Pour savoir construire ces poids il faut savoir construire un estimateur de $\hat{\beta}$. Pour faire ça, tout simplement, on considère un sous-échantillon de taille r_0, extrait de façon uniforme, et on calcule un estimateur $\tilde{\beta}_0$ en maximisant la fonction l_n^*. On peut choisir, comme on a dit les poids du sous-échantillonage de façon uniforme, mais parfois le numéro d'éléments du grand échantillon avec $Y_i = 0$ (ou $Y_i = 1$) est trop petit. Dans ce cas on peut penser de diviser le grand échantillon en deux (d'une part les éléments avec $Y_i = 0$ et d'autre part les éléments avec $Y_i = 1$) et extraire avec poids uniformes $r_0/2$ éléments dans chaque groupe.
2. On calcule les poids en utilisant $\tilde{\beta}_0$ (π_i^{MSE} ou $\pi_i^{V_c}$ selon la préférence) et on calcule notre estimateur $\tilde{\beta}$ en maximisant l_n^* sur un échantillon de taille r_1.

L'implémentation de l'algorithme classique d'estimation de l'EMV requiert un calcul de l'ordre $O(\zeta np^2)$ où ζ est le nombre de pas de convergence de l'algorithme de Newton-Raphson, n est la dimension de l'échantillon et p est la dimension du vecteur X.

L'algorithme de sous-échantillonage avec poids π_{MSE} requiert un calcul de l'ordre de $O(np^2)$. Ce sont les poids à calculer qui sont en $O(np^2)$, la convergence de Newton-Raphson quant à elle requiert $O(\zeta rp^2)$. Comme nous supposons r négligeable devant n, l'algorithme pour ces poids est en $O(np^2)$. Donc l'algorithme de sous-échantillonnage avec poids π_{MSE} est meilleur en complexité par rapport à l'EMV, en terme de temps de calcul, seulement d'un facteur ζ.

Par contre le calcul de l'algorithme de sous-échantillonage avec poids π_{V_c} requiert seulement un calcul de poids qui est de l'ordre de $O(np)$. Newton-Raphson requiert $O(\zeta rp^2)$. La même remarque sur l'ordre de gradeur de r et n permet de conclure que l'algorithme est en $O(\zeta rp)$. L'échantillonnage avec poids π_{V_c} permet un gain d'ordre ζp par rapport à l'EMV. C'est exactement ici la puissance de l'algorithme de sous-échantillonage.

6 Les simulations numériques

6.1 Introduction

Pour tester nos théorèmes et l'optimalité effective des poids π_{V_c} et π_{MSE} nous avons suivi les simulations de l'article de Wang et al. [7] et nous avons simulé 6 échantillons différents suivant des loi différentes. Dans tous les cas nous avons pris $\beta = (0.5, ..., 0.5)$. Dans l'ordre suivant on a simulé ($p = 8$, $n = 10^5$)

1. $X = (X_1, ..., X_8)$ t.q. $(X_2, ..., X_8) \sim \mathcal{N}(0, \Sigma)$ avec Σ t.q. $\Sigma_{i,i} = 1$ et 0.5 pour les autres composantes.
2. $X = (X_1, ..., X_8)$ t.q. $(X_2, ..., X_8) \sim \mathcal{N}((1.5, ...1.5), \Sigma)$ avec Σ t.q. $\Sigma_{i,i} = 1$ et 0.5 pour les autres composantes
3. $X = (X_1, ..., X_8)$ t.q. $(X_2, ..., X_8) \sim \mathcal{N}((1.5, ...1.5), \Sigma)$ avec Σ t.q. $\Sigma_{i,i} = 1$ et $0.5\frac{1}{ij}$ pour les autres composantes
4. $X = (X_1, ..., X_8)$ t.q. $(X_1, ..., X_8) = (-1)^B * (0.5, ..., 0.5) + Z$ où $Z \sim \mathcal{N}(0, \Sigma)$ avec $\Sigma_{i,i} = 1$ et 0.5 pour les autres composantes et $B \sim Ber(1/2)$ (Mélange gaussien)
5. $X = (X_1, ..., X_8)$ t.q. $(X_2, ..., X_8)$ suit une loi de Student à 3 degrés de liberté et de covariance Σ t.q. $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$ pour $i \neq j$.
6. $X = (X_1, ..., X_8)$ t.q. $X_1, ..., X_8 \overset{iid}{\sim} \mathcal{E}(1/2)$.

Remarque : Les deux derniers échantillons ne vérifient pas les hypothèses que nous avons faits sur la loi des X pour le théorème 5.1

6.2 La consistance

Dans la figure ci-dessous nous avons tracé la norme de $\|\tilde{\beta} - \hat{\beta}\|$ pour différentes valeurs de n en prenant $r = \sqrt{n}$. On remarque que nous avons consistance de $\tilde{\beta}$ dans tous les cas que nous avons simulé. Il est cependant difficile en regardant cette figure de déterminer si notre méthode de sous-échantillonage est meilleure qu'un sous-échantillonage uniforme.

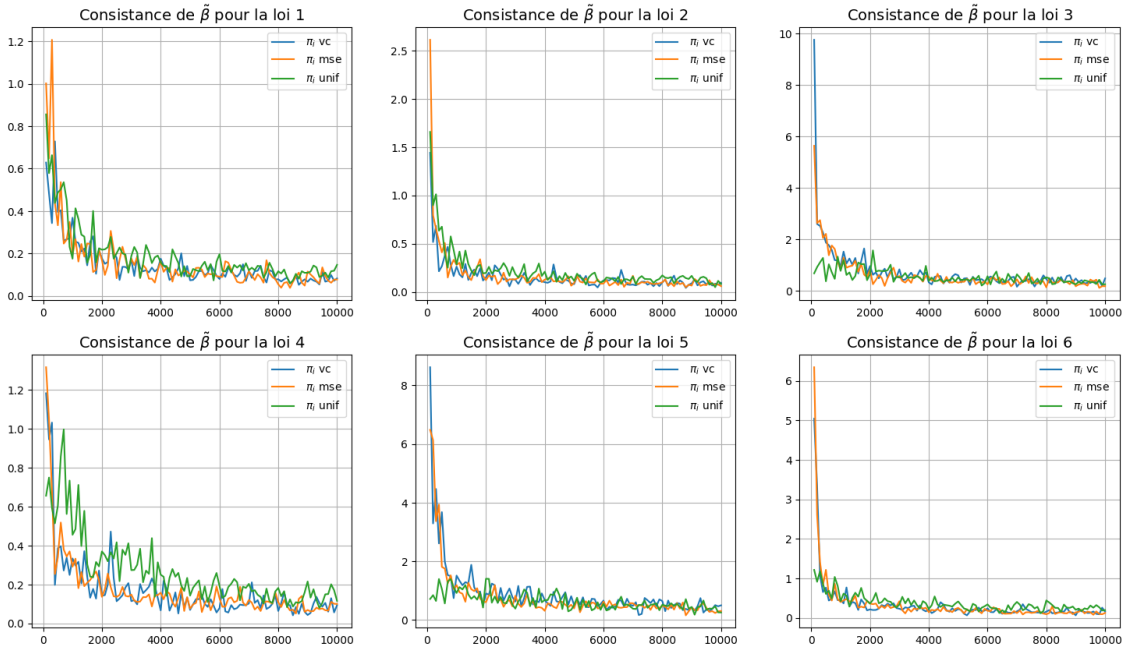

FIGURE 1 – Consistance

6.3 La normalité asymptotique

Pour vérifier le théorème 5.2, nous avons fait un histogramme de plusieurs $\tilde{\beta}$ correctement normalisé (i.e $V^{-\frac{1}{2}}(\tilde{\beta} - \hat{\beta})$). Comme $V = O_{\mathbb{P}}(\frac{1}{r})$, nous remarquons que la vitesse de convergence est bien $\sqrt{r}$ comme le

suggère le théorème 5.1. Il est à noter ici que la figure qui suit n'est pas une vérification parfaite du théorème. En effet, nous ne traçons qu'une composante pour chaque loi. Il manque aussi le fait de vérifier que nous avons bien indépendances entre les composantes (qui est un des résultats du théorème puisque la loi limite est une $\mathcal{N}(0, I)$).

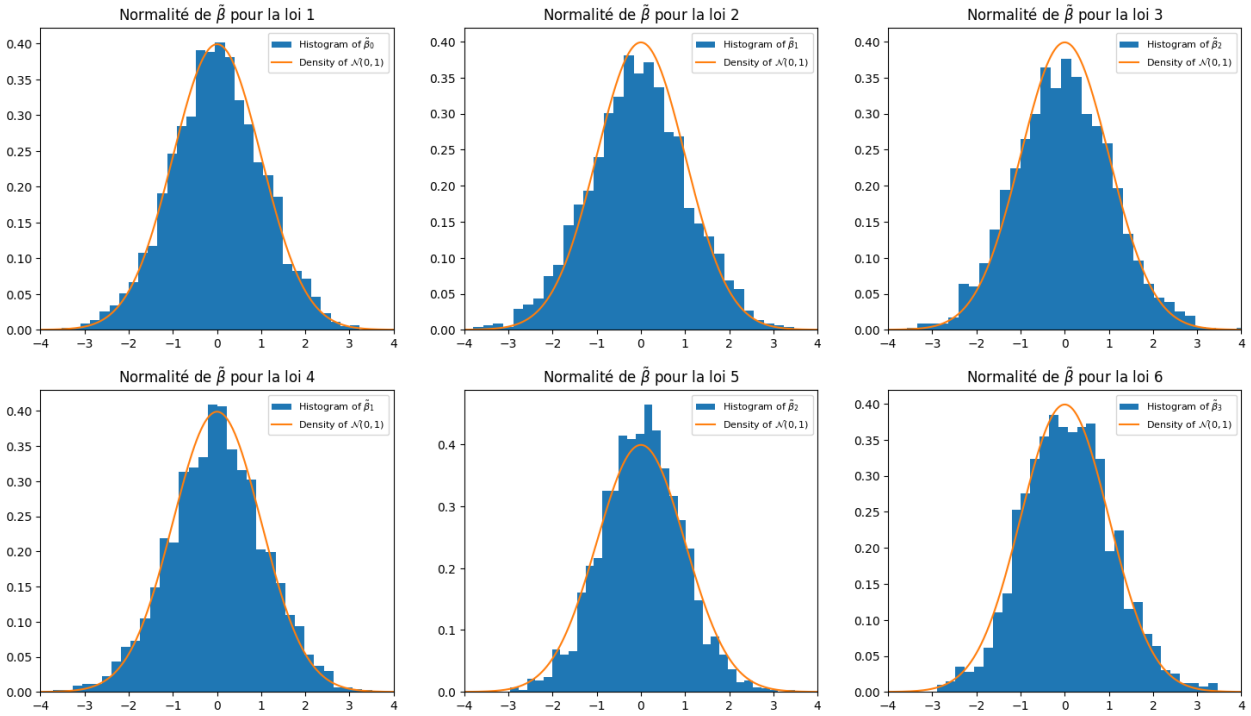

FIGURE 2 – Normalité asymptotique

6.4 L'erreur quadratique

Une des prétentions de l'article est que la méthode de sous-échantillonnage présentée est optimale au sens de l'erreur quadratique comparée à une méthode de sous-échantillonnage uniforme. C'est pour cela que nous avons calculé cette erreur quadratique par approximation Monte Carlo. Nous avons fait 1000 simulations en faisant varier la taille du sous-échantillon à chaque fois. Dans la figure ci-dessous, on peut voir que la méthode est plus efficace qu'un échantillonnage uniforme pour la majorité des lois que nous avons choisi. On voit que la méthode est moins efficace pour la loi 5 mais cela n'est pas étonnant car cette loi ne satisfait pas les hypothèses de nos théorèmes. Cependant, on voit que le sous-échantillonnage uniforme est plus efficace pour l'échantillon 3, ce qui est nécessite une étude plus pointue. Nous remarquons cependant que l'échelle de l'erreur quadratique dans ce cas est différente (elle est plus importante) des cas où notre algorithme est efficace, ce qui pourrait être une piste à explorer.

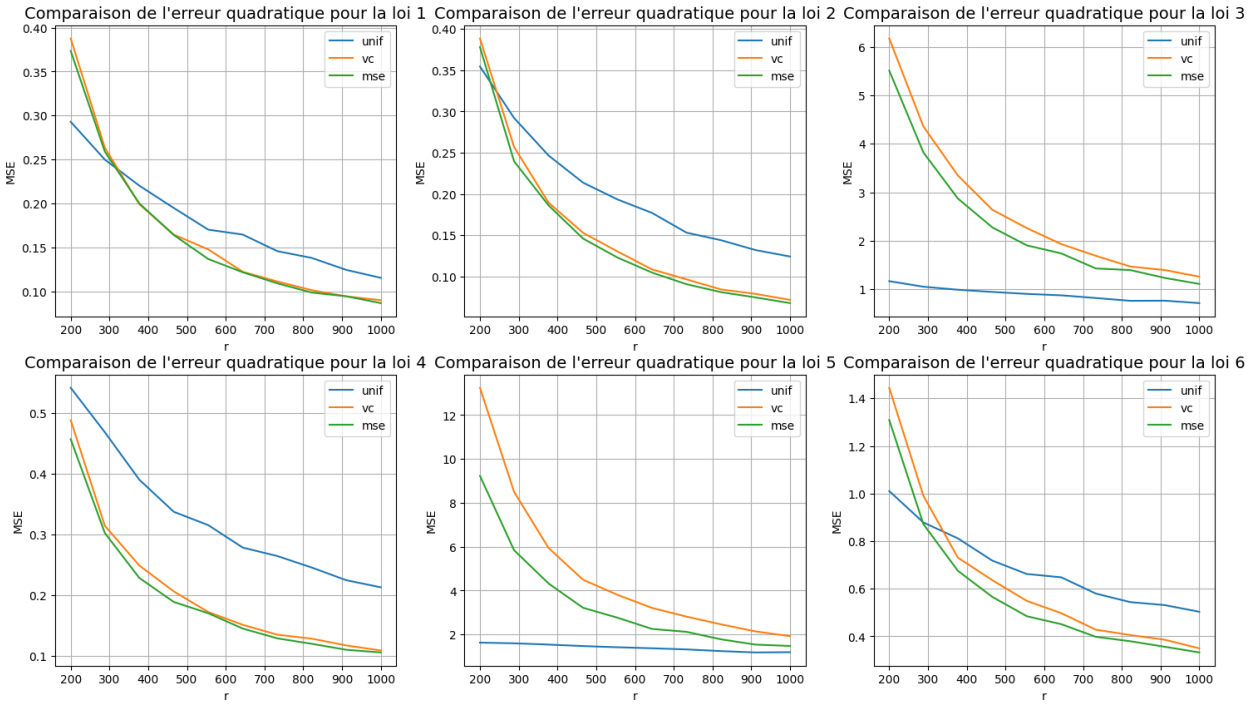

FIGURE 3 – Erreur quadratique

7 Conclusion

L'analyse de données massives représente un défi majeur dans le domaine de la statistique et de la modélisation des phénomènes complexes. Dans cette étude, nous avons abordé une problématique cruciale : le limitations computationnelles liées à l'analyse d'échantillons de grande taille dans le contexte de la régression logistique.

L'une des approches novatrices que nous avons étudiées dans ce texte est l'utilisation de sous-échantillons pour estimer les paramètres de régression. Cette méthode, analysé à travers la lentille de l'article "Optimal Subsampling for Large Sample Logistic Regression" [7] offre un compromis intéressant entre la précision des estimations et les contraintes computationnelles.

Pour comprendre et justifier cette approche, nous avons d'abord établi un cadre théorique solide. Cela a impliqué la compréhension des principaux résultats de convergence asymptotique, tel que le Théorème de Lindeberg-Feller. Ensuite, nous nous sommes penchés sur la régression linéaire et la régression logistique, en mettant l'accent sur l'estimateur du maximum de vraisemblance et ses propriétés. En parcourant divers concepts théoriques, nous avons énoncé plusieurs théorèmes de probabilité conditionnelle, résultats centraux pour nos objectifs.

L'élément clé de notre étude a été la démonstration de la consistance et de la normalité asymptotique de l'estimateur par sous-échantillonnage. Ces résultats théoriques, étayés par des simulations numériques, ont confirmé la viabilité et l'efficacité de cette approche dans des scénarios pratiques. Nous avons, en plus, énoncer les choix optimaux des poids pour obtenir une erreur quadratique minimum. Enfin, une balance entre performance d'estimation et temps de calcul nous a poussé à conclure que le meilleur choix des poids correspond aux poids π_{V_c}.

En conclusion, en combinant rigueur théorique et validation empirique, nous avons présenté une nouvelle voie pour l'estimation des paramètres de régression dans des environnements où la taille des données représente une contrainte importante. Cette approche offre des perspectives prometteuses pour des applications plus larges en matière d'analyse de données volumineuses et de modélisation statistique avancée.

8 Annexes

Dans la section suivant vous pouvez trouver les démonstrations bien énumérées des théorèmes énoncés dans le texte.

8.1 Théorème asymptotiques

Théorème 2.2.
On a

$$\begin{aligned}\Phi_{\frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}}(t) &= \Phi_{\sum_i^n X_i - n\mu}\left(\frac{t}{\sqrt{n}\sigma}\right) \\ &= \left(\Phi_{X_1-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right)^n \qquad \text{(les } X_i \text{ sont iid)}\end{aligned}$$

En utilisant le théorème de dérivation de Lebesgue et le fait que $X_1 - \mu$ est un variable centrée de variance σ^2 on peut montrer facilement que

$$\Phi_{X_1-\mu}(s) = 1 - \frac{\sigma^2 s^2}{2} + o(s^2)$$

et donc

$$\begin{aligned}\Phi_{\frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}}(t) &= 1 - \frac{\sigma^2}{2}\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + o\left(\left(\frac{t}{\sqrt{n}\sigma}\right)^2\right) \\ &= 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\end{aligned}$$

En passant à la forme exponentielle on a

$$\begin{aligned}\Phi_{\frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}}(t) &= \exp\left\{n\log\left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)\right\} \qquad \text{(pour } n \text{ assez grand)} \\ &= \exp\left\{n\left(-\frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)\right\} \\ &= \exp\left\{-\frac{t^2}{2} + o(1)\right\}\end{aligned}$$

qui, évidemment, tend pour tout t vers la fonction caractéristique de la loi centrée réduite.

$$\Phi(t) = e^{-\frac{t^2}{2}}.$$

Par le théorème de Lévy on a donc le résultat.

□

Théoreme 2.3.
On rappelle que la fonction caractéristique pour un vecteur aléatoire Y dans $\mathbb{R}^d$ de moyenne m et matrice de covariance V correspond à

$$\Phi_Y(t) = e^{i<t,m> - \frac{<t,Vt>}{2}}$$

où $< \cdot, \cdot >$ indique le produit scalaire standard.

On a donc :

$$\begin{aligned}\Phi_{\frac{\sum_i^n X_i - n\mu}{\sqrt{n}}}(t) &= \mathbb{E}\left[e^{i<t, \frac{\sum_i^n X_i - n\mu}{\sqrt{n}}>}\right] \\ &= \mathbb{E}\left[e^{i\frac{1}{\sqrt{n}}\sum_{i=1}^n <t,(X_i-\mu)>}\right] \\ &= \Phi_{\sum_{i=1}^n \frac{<t, X_i - \mu>}{\sqrt{n}}}(1)\end{aligned}$$

On a que pour tout $i \in \mathbb{N}$ la variable aléatoire $< t, X_i - \mu >$ est une variable aléatoire gaussienne centrée et de variance $< t, \Sigma t >$. Pour le TCL univarié on a

$$\sum_{i=1}^{n} \frac{< t, X_i - \mu >}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, < t, \Sigma t >\right)$$

et donc par le Théorème de Lévy

$$\Phi_{\sum_{i=1}^{n} \frac{< t, X_i - \mu >}{\sqrt{n}}}(1) \longrightarrow e^{-\frac{< t, \Sigma t >}{2}}$$

donc on déduit

$$\Phi_{\frac{\sum_i^n X_i - n\mu}{\sqrt{n}}}(t) \longrightarrow e^{-\frac{< t, \Sigma t >}{2}}$$

pour tout $t \in \mathbb{R}^d$.

Du théorème de Lévy on peut conclure le résultat. □

Lemme 1.
Pour tout $x \in \mathbb{R}$ et $n \in \mathbb{N}$, la n-ième dérivée de $x \to e^{ix}$ est $\frac{d^n}{dx^n}[e^{ix}] = i^n e^{ix}$

En appliquant la formule de Taylor avec reste intégral, on a pour tout $x \in \mathbb{R}$ et $n \in \mathbb{N}^*$

$$\begin{aligned} e^{ix} &= \sum_{k=0}^{n-1} \frac{(ix)^k}{k!} + \int_0^x \frac{(x-u)^{n-1}}{(n-1)!} i^n e^{iu} du \\ &= \sum_{k=0}^{n-1} \frac{(ix)^k}{k!} + \frac{i^n x^{n-1}}{(n-1)!} \int_0^x (1 - \frac{u}{x})^{n-1} e^{iu} du \\ &= \sum_{k=0}^{n-1} \frac{(ix)^k}{k!} + \frac{(ix)^n}{(n-1)!} \int_0^1 (1-u)^{n-1} e^{ixu} du \end{aligned}$$

Donc, pour tout $x \in \mathbb{R}$ et $n \in \mathbb{N}^*$

$$\begin{aligned} \left| e^{ix} - \sum_{k=0}^{n-1} \frac{(ix)^k}{k!} \right| &= \left| \frac{(ix)^n}{(n-1)!} \int_0^1 (1-u)^{n-1} e^{ixu} du \right| \\ &\leq \frac{|x|^n}{(n-1)!} \int_0^1 (1-u)^{n-1} du \\ &= \frac{|x|^n}{n!} \end{aligned}$$

Pour $n \geq 2$,

$$\begin{aligned} \left| e^{ix} - \sum_{k=0}^{n-1} \frac{(ix)^k}{k!} \right| &= \left| e^{ix} - \sum_{k=0}^{n-2} \frac{(ix)^k}{k!} - \frac{(ix)^{n-1}}{(n-1)!} \right| \\ &\leq \left| e^{ix} - \sum_{k=0}^{n-2} \frac{(ix)^k}{k!} \right| + \frac{|x|^{n-1}}{(n-1)!} \\ &\leq 2 \frac{|x|^{n-1}}{(n-1)!} \end{aligned}$$

Pour $n = 1$, on a bien $|e^{ix} - 1| \leq 2$

Ainsi $\left| e^{ix} - \sum_{k=0}^{n-1} \frac{(ix)^k}{k!} \right| \leq \min\left\{ \frac{|x|^n}{n!}, 2\frac{|x|^{n-1}}{(n-1)!} \right\} \qquad \forall x \in \mathbb{R} \quad \forall n \in \mathbb{N}^*$

□

Lemme 2.
Par récurrence on a :

— Pour $(n=2)$ on a $a_1a_2 - b_1b_2 = (a_1-b_1)a_2 + (a_2-b_2)b_1$. Pour l'hypothèse $|a_i|, |b_i| \leq 1$ et l'inégalité triangulaire on trouve, en passant aux modules, le résultat.

— On supposant le résultat vrai pour n, on a

$$\begin{aligned}
\left|\prod_{i=i}^{n+1} a_i - \prod_{i=i}^{n+1} b_i\right| &= \left|\prod_{i=i}^{n} a_i a_{n+1} - \prod_{i=i}^{n} b_i b_{n+1}\right| \\
&= \left|\left(\prod_{i=i}^{n} a_i - \prod_{i=i}^{n} b_i\right) a_{n+1} + (a_{n+1} - b_{n+1}) \prod_{i=i}^{n} b_i\right| \\
&\leq \left|\prod_{i=i}^{n} a_i - \prod_{i=i}^{n} b_i\right| + |a_{n+1} - b_{n+1}| \\
&\leq \sum_{i=1}^{n} |a_i - b_i| + |a_{n+1} - b_{n+1}|
\end{aligned}$$

□

Théorème 2.4.

Sans perte de généralité, on peut supposer que $s_n^2 = 1$ pour tout $n \geq 1$ (sinon il suffit de normaliser les variables adéquatement). Le but est alors de montrer que pour tout $t \in \mathbb{R}$

$$\lim_{n\to+\infty} \mathbb{E}[e^{itS_n}] = e^{-\frac{t^2}{2}}$$

On note ϕ_{nj} la fonction caractéristique de X_{nj} pour tout $n \in \mathbb{N}^*$ et $1 \leq j \leq r_n$

Montrons d'abord que $\max\limits_{1\leq j\leq r_n} \mathbb{E}[X_{nj}^2] \underset{n\to+\infty}{\longrightarrow} 0$

Soit $\epsilon > 0$,

$$\begin{aligned}
\max_{1\leq j\leq r_n} \mathbb{E}[X_{nj}^2] &= \max_{1\leq j\leq r_n} \mathbb{E}[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|>\epsilon\}}] + \mathbb{E}[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|\leq\epsilon\}}] \\
&\leq \sum_{j=1}^{r_n} \mathbb{E}[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|>\epsilon\}}] + \epsilon^2 \\
&= o(1) + \epsilon^2
\end{aligned}$$

donc $M_n := \max\limits_{1\leq j\leq r_n} \sigma_{nj}^2 \underset{n\to+\infty}{\longrightarrow} 0$

Soit $t \in \mathbb{R}$. Par ce qui précède, il existe $n_0 \in \mathbb{N}$ tel que pour tout $n \geq n_0$, $\max\{|1 - \frac{t^2}{2}\sigma_{nj}^2| : 1 \leq j \leq r_n\} \leq 1$.

On a alors

$$\begin{aligned}
\left|\mathbb{E}[e^{itS_n}] - e^{-\frac{t^2}{2}}\right| &= \left|\prod_{j=1}^{r_n} \phi_{nj}(t) - \prod_{j=1}^{r_n}(1 - \frac{t^2\sigma_{nj}^2}{2}) + \prod_{j=1}^{r_n}(1 - \frac{t^2\sigma_{nj}^2}{2}) - e^{-\frac{t^2}{2}}\right| \\
&\leq \left|\prod_{j=1}^{r_n} \phi_{nj}(t) - \prod_{j=1}^{r_n}(1 - \frac{t^2\sigma_{nj}^2}{2})\right| + \left|\prod_{j=1}^{r_n}(1 - \frac{t^2\sigma_{nj}^2}{2}) - e^{-\frac{t^2}{2}}\right| \\
&\leq \sum_{j=1}^{r_n}\left|\phi_{nj}(t) - (1 - \frac{t^2\sigma_{nj}^2}{2})\right| + \sum_{j=1}^{r_n}\left|e^{-\frac{t^2}{2}} - (1 - \frac{t^2\sigma_{nj}^2}{2})\right|. \qquad \text{(Lemme 2)}
\end{aligned}$$

On pose $A_n = \sum\limits_{j=1}^{r_n}\left|\phi_{nj}(t) - (1 - \frac{t^2\sigma_{nj}^2}{2})\right|$ et $B_n = \sum\limits_{j=1}^{r_n}\left|e^{-\frac{t^2}{2}} - (1 - \frac{t^2\sigma_{nj}^2}{2})\right|$.

• Montrons $A_n \underset{n\to+\infty}{\longrightarrow} 0$.

Soit $\epsilon > 0$,

$$\begin{aligned}
A_n &= \sum_{j=1}^{r_n} \left| \mathbb{E}[e^{itX_{nj}}] - (1 + it\mathbb{E}[X_{nj}] + \frac{(it)^2}{2}\mathbb{E}[X_{nj}^2]) \right| \\
&= \sum_{j=1}^{r_n} \left| \mathbb{E}[e^{itX_{nj}} - (1 + itX_{nj} + \frac{(itX_{nj})^2}{2})] \right| \\
&\leq \sum_{j=1}^{r_n} \left| \mathbb{E}[\min\{\frac{|tX_{nj}|^3}{3!}, |tX_{nj}|^2\}] \right| \qquad \text{(Lemme 1)} \\
&= \sum_{j=1}^{r_n} \left| \mathbb{E}[\min\{\frac{|tX_{nj}|^3}{3!}, |tX_{nj}|^2\}\mathbb{1}_{|X_{nj}|\leq\epsilon}] \right| + \sum_{j=1}^{r_n} \left| \mathbb{E}[\min\{\frac{|tX_{nj}|^3}{3!}, |tX_{nj}|^2\}\mathbb{1}_{|X_{nj}|>\epsilon}] \right| \\
&\leq \sum_{j=1}^{r_n} \mathbb{E}\left[|tX_{nj}|^3 \mathbb{1}_{|X_{nj}|\leq\epsilon} \right] + t^2 \sum_{j=1}^{r_n} \mathbb{E}\left[X_{nj}^2 \mathbb{1}_{|X_{nj}|>\epsilon} \right] \\
&\leq |t|^3\epsilon^3 + t^2 o(1).
\end{aligned}$$

• Montrons $B_n \underset{n\to+\infty}{\longrightarrow} 0$.
Remarquons d'abord que pour tout $x \in \mathbb{R}$ $|e^x - 1 - x| \leq x^2 e^{|x|}$ (Il suffit de développer l'exponentielle en série entière)

$$\begin{aligned}
B_n &= \sum_{j=1}^{r_n} \left| e^{-\frac{t^2\sigma_{nj}^2}{2}} - (1 - \frac{t^2\sigma_{nj}^2}{2}) \right| \\
&\leq \sum_{j=1}^{r_n} \left(\frac{t^2\sigma_{nj}^2}{2} \right)^2 e^{\frac{t^2\sigma_{nj}^2}{2}} \\
&\leq t^4 e^{\frac{t^2}{2}M_n} \sum_{j=1}^{r_n} \sigma_{nj}^2 M_n \\
&= t^4 e^{\frac{t^2}{2}M_n} M_n \underset{n\to+\infty}{\longrightarrow} 0.
\end{aligned}$$

Ce qui achève la preuve.

□

Théorème 2.5.
Sans perte de généralité, on suppose que $s_n^2 = 1$. On pose ϕ_{nj} la fonction caractéristique de X_{nj}.
• Montrons que pour tout $t \in \mathbb{R}$

$$\left| \prod_{j=1}^{r_n} e^{\phi_{nj}(t)-1} - \prod_{j=1}^{r_n} \phi_{nj}(t) \right| \underset{n\to+\infty}{\longrightarrow} 0.$$

Soit $t \in \mathbb{R}$,

$$\begin{aligned}
|\phi_{nj}(t) - 1| &= \left| \mathbb{E}[e^{itX_{nj}} - 1] \right| \\
&\leq \mathbb{E}[\min\{|tX_{nj}|, 2\}] \\
&\leq \mathbb{E}[2\mathbb{1}_{|X_{nj}|>\epsilon}] + |t|\mathbb{E}[|X_{nj}|\mathbb{1}_{|X_{nj}|\leq\epsilon}] \\
&\leq 2\mathbb{P}(|X_{nj}| > \epsilon) + |t|\epsilon.
\end{aligned}$$

Donc

$$\begin{aligned}
\max_{1\leq j\leq r_n} |\phi_{nj}(t) - 1| &\leq 2 \max_{1\leq j\leq r_n} \mathbb{P}(|X_{nj}| > \epsilon) + |t|\epsilon \\
&= o(1) + \epsilon.
\end{aligned}$$

De plus,

$$\begin{aligned}\sum_{j=1}^{r_n}|\phi_{nj}(t)-1| &= \sum_{j=1}^{r_n}|\mathbb{E}[e^{itX_{nj}}]-1-\mathbb{E}[itX_{nj}]| \\ &\leq \frac{t^2}{2}\sum_{j=1}^{r_n}\mathbb{E}[X_{nj}^2] \\ &= \frac{t^2}{2}.\end{aligned}$$

Pour n assez grand (pour que $\max\limits_{1\leq j\leq r_n}|\phi_{nj}(t)-1|\leq 1$), on a

$$\begin{aligned}\Big|\prod_{j=1}^{r_n}e^{\phi_{nj}(t)-1}-\prod_{j=1}^{r_n}\phi_{nj}(t)\Big| &\leq \sum_{j=1}^{r_n}|e^{\phi_{nj}(t)-1}-\phi_{nj}(t)|\prod_{j=1}^{r_n-j}\exp(|\phi_{nj}(t)-1|) \\ &\leq \sum_{j=1}^{r_n}|e^{\phi_{nj}(t)-1}-\phi_{nj}(t)|\exp(\sum_{j=1}^{r_n}|\phi_{nj}(t)-1|) \\ &\leq \sum_{j=1}^{r_n}|e^{\phi_{nj}(t)-1}-1-(\phi_{nj}(t)-1)|e^{\frac{t^2}{2}} \\ &\leq \sum_{j=1}^{r_n}|\phi_{nj}(t)-1|^2e^{|\phi_{nj}(t)-1|}e^{\frac{t^2}{2}} \\ &\leq \sum_{j=1}^{r_n}|\phi_{nj}(t)-1|^2e^{1+\frac{t^2}{2}} \\ &\leq \max_{1\leq j\leq r_n}|\phi_{nj}(t)-1|(t^2e^{1+\frac{t^2}{2}}) \\ &= o(1).\end{aligned}$$

Par la normalité asymptotique, on a $\prod\limits_{j=1}^{r_n}\phi_{nj}(t)\underset{n\to+\infty}{\longrightarrow}e^{-\frac{t^2}{2}}$. Or $\left|\prod\limits_{j=1}^{r_n}e^{\phi_{nj}(t)-1}\right|=\left|e^{\sum\limits_{j=1}^{r_n}\mathbb{E}[cos(tX_{nj})-1]}\right|$,
donc $\sum\limits_{j=1}^{r_n}\mathbb{E}[cos(tX_{nj})-1]+\frac{t^2}{2}\underset{n\to+\infty}{\longrightarrow}0$ pour tout $t\in\mathbb{R}$.
Soit $\epsilon>0$, on pose $t=\frac{4}{\epsilon}$. On a

$$\begin{aligned}o(1) &= \sum_{j=1}^{r_n}\mathbb{E}[cos(tX_{nj})-1]+\frac{t^2}{2} \\ &= \sum_{j=1}^{r_n}\mathbb{E}[\frac{t^2X_{nj}^2}{2}-1+cos(tX_{nj})] \\ &\geq \sum_{j=1}^{r_n}\mathbb{E}[(\frac{t^2X_{nj}^2}{2}-1+cos(tX_{nj})s)\mathbb{1}_{|X_{nj}|>\epsilon}] \\ &\geq \sum_{j=1}^{r_n}\mathbb{E}[(\frac{t^2X_{nj}^2}{2}-2)\mathbb{1}_{|X_{nj}|>\epsilon}] \\ &\geq \sum_{j=1}^{r_n}\mathbb{E}[(\frac{t^2}{2}-\frac{2}{X_{nj}^2})X_{nj}^2\mathbb{1}_{|X_{nj}|>\epsilon}] \\ &\geq (\frac{t^2}{2}-\frac{2}{\epsilon^2})\sum_{j=1}^{r_n}\mathbb{E}[X_{nj}^2\mathbb{1}_{|X_{nj}|>\epsilon}] \\ &= \frac{6}{\epsilon^2}\sum_{j=1}^{r_n}\mathbb{E}[X_{nj}^2\mathbb{1}_{|X_{nj}|>\epsilon}].\end{aligned}$$

donc $\sum\limits_{j=1}^{r_n}\mathbb{E}[X_{nj}^2\mathbb{1}_{|X_{nj}|>\epsilon}]\underset{n\to+\infty}{\longrightarrow}0$. □

Théorème 2.6.
Soit $\epsilon > 0$,

$$\begin{aligned}
\frac{1}{s_n^2}\sum_{j=1}^{r_n}\mathbb{E}[X_{nj}^2\mathbb{1}_{|X_n|>\epsilon s_n}] &= \sum_{j=1}^{r_n}\mathbb{E}[(\frac{X_{nj}}{s_n})^2\mathbb{1}_{|X_{nj}|>\epsilon s_n}] \\
&= \epsilon^2\sum_{j=1}^{r_n}\mathbb{E}[(\frac{X_{nj}}{\epsilon s_n})^2\mathbb{1}_{|X_{nj}|>\epsilon s_n}] \\
&\leq \epsilon^2\sum_{j=1}^{r_n}\mathbb{E}[\frac{|X_{nj}|^{2+\delta}}{\epsilon^{2+\delta}s_n^{2+\delta}}] \\
&= \frac{1}{\epsilon^\delta}\frac{1}{s_n^{2+\delta}}\sum_{j=1}^{r_n}\mathbb{E}[|X_{nj}^{2+\delta}|] \underset{n\to+\infty}{\longrightarrow} 0.
\end{aligned}$$

Par le théorème de Lindeberg-Feller, on a $\frac{S_n}{s_n} \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(0,1)$. □

Théorème 2.8.
Par le théorème de Cramer-Wold, il suffit de vérifier que pour tout $a \in \mathbb{R}^d$ tel que $\|a\| = 1$, on a

$$\langle a, \sum_{j=1}^{r_n} X_{nj}/\sqrt{r_n}\rangle \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}(0,1).$$

On montre que $\{a'X_{ni} : 1 \leq i \leq r_n\}_{n\geq 1}$ satisfait la condition de Lindeberg.
Soit $\epsilon > 0$,

$$\begin{aligned}
\frac{1}{r_n}\sum_{j=1}^{r_n}\mathbb{E}[|a'X_{nj}|^2\mathbb{1}_{|a'X_{nj}|>\epsilon\sqrt{r_n}}] &\leq \frac{1}{r_n}\sum_{j=1}^{r_n}\mathbb{E}[\|a\|^2\|X_{nj}\|^2\mathbb{1}_{\|a\|\|X_{nj}\|>\epsilon\sqrt{r_n}}] \qquad \text{(Cauchy-Schwarz)} \\
&= \frac{1}{r_n}\sum_{j=1}^{r_n}\mathbb{E}[\|X_{nj}\|^2\mathbb{1}_{\|X_{nj}\|>\epsilon\sqrt{r_n}}] \\
&= o(1).
\end{aligned}$$

□

Théorème 2.9.
On montre la premier résultat pour les vecteurs aléatoire de $\mathbb{R}^d$, $d \geq 1$.
On montre la première implication pour les vecteurs aléatoire de $\mathbb{R}^d$, $d \geq 1$.Pour la démontrer, on commence par rappeler que la convergence en loi de X_n vers X est définie comme la convergence des $\mathbb{E}[f(X_n)]$ vers $\mathbb{E}[f(X)]$ pour toute fonction continue bornée. En particulier donc, on a convergence des espérances pour les fonctions continues à support compact. En outre pour tout vecteur aléatoire dans $\mathbb{R}^d$ ou $\mathbb{R}$, il est possible de montrer que pour tout $\epsilon > 0$ il existe une boule $B(0, r_\epsilon)$ t.q. X est dans K avec probabilité $1 - \epsilon$ (Il est suffisant de considérer que $\mathbb{R}^d$ ou $\mathbb{R}$ peuvent s'obtenir comme union croissante des boules centrées en zéro)

Soit $\epsilon > 0$. Si on note $r_{\frac{\epsilon}{2}}$ le rayon t.q. $\mathbb{P}(X \in B(0, r_{\frac{\epsilon}{2}})) > 1 - \frac{\epsilon}{2}$ et on considère la fonction de $\mathbb{R}^d$ dans $\mathbb{R}$:

$$f(x) = \begin{cases} 1 & \text{si } \|x\| < r_{\frac{\epsilon}{2}} \\ 1 + r_{\frac{\epsilon}{2}} - \|x\| & \text{si } r_{\frac{\epsilon}{2}} \leq \|x\| \leq r_{\frac{\epsilon}{2}} + 1 \\ 0 & \text{si } \|x\| > r_{\frac{\epsilon}{2}} + 1 \end{cases}$$

1. En ayant

$$\mathbb{P}(X_n \in B(0, r_{\frac{\epsilon}{2}} + 1)) - \mathbb{E}[f(X)] = \mathbb{E}[\mathbb{1}_{\|X_n\| \leq r_{\frac{\epsilon}{2}}+1}] - \mathbb{E}[f(X)] \geq \mathbb{E}[f(X_n)] - \mathbb{E}[f(X)].$$

si on prend n_0 t.q. pour tout $n > n_0$,

$$\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)] \geq -\frac{\epsilon}{2}$$

on a

$$\mathbb{P}(X_n \in B(0, r_{\frac{\epsilon}{2}} + 1)) - \mathbb{E}[f(X)] \geq -\frac{\epsilon}{2}$$

donc

$$\mathbb{P}(X_n \in B(0, r_{\frac{\epsilon}{2}} + 1)) \geq \mathbb{E}[f(X)] - \frac{\epsilon}{2} > 1 - \frac{\epsilon}{2} - \frac{\epsilon}{2} = 1 - \epsilon$$

Si on choisit pour tout $i = 1, ..., n_0$ r_ϵ^i t.q.

$$P(X_i \in B(0, r_\epsilon^i)) > 1 - \epsilon$$

On a enfin

$$P(X_n \in B(0, \max\{r_\epsilon^1, ..., r_\epsilon^{n_0}, r_{\frac{\epsilon}{2}})) > 1 - \epsilon \quad \forall n \in \mathbb{N}$$

On a donc démontré que pour toute suite convergente en loi il est possible de trouver un compact qui contient les X_n avec probabilité $1 - \epsilon$, c'est à dire la définition de $O_{\mathbb{P}}(1)$.

2. On démontre cette implication seulement dans le cas des variables aléatoires. Pour démontrer cette propriété on suppose un important théorème appelé "Théorème de sélection de Helly", qui affirme que pour toute suite $\{f_n\}_n$ de fonction croissante de $\mathbb{R}$ dans $[0, 1]$ il existe une sous-suite $\{f_{n_k}\}_k$ et une fonction f t.q. $f_{n_k}(x)$ converge vers $f(x)$ pour tout $x \in [0, 1]$. Le résultat s'obtient à travers un processus d'extraction diagonale en considérant une restriction de la suite f_n sur $\mathbb{Q}$. Si on considère une suite tendue $\{X_n\}_n$ et on note $\{F_n\}_n$ la suite de fonctions de répartition correspondantes, on sait, grâce au théorème de séparation de Helly, qu'il existe une sous-suite $\{F_{n_k}\}_k$ et une fonction G croissante t.q. F_{n_k} converge simplement vers G.

   Si on définit $F(x) = \inf\{q \in [0, 1] | G(q) > x\}$ On a que :

   — F est croissante (facile à voir)

   — F est continue à droite. En effet si on prend $q \in [0, 1]$ t.q. $G(q) > x_n$ on a aussi $G(q) > x_n \geq x$ alors $q > \inf\{q | G(q) > x\} = F(x)$. Ça vaut pour tout q donc

$$F(x_n) = \inf\{q | G(q) > x_n\} > \inf\{q | G(q) > x\} = F(x)$$

   et donc

$$\lim_n F(x_n) \geq F(x).$$

   En même temps, si on prend q dans l'ensemble $\{q | \exists n \text{ t.q. } G(q) > x_n\}$, il existe n_0 t.q. $G(q) > x_{n_0}$ et donc

$$q \geq \inf_q\{q | G(q) > x_{n_0}\} \geq \inf_n \inf_q\{q | G(q) > x_n\}.$$

   Cette inégalité est vraie pour tout q donc

$$\inf_q\{q | \exists n \text{ t.q. } G(q) > x_n\} \geq \inf_n \inf_q\{q | G(q) > x_n\}.$$

   Puisque $F(x_n)$ est décroissante en n (facile à montrer)

$$lim_n F(x_n) = \inf_n F(x_n) = \inf_n \inf_q\{q | G(q) > x_n\} \leq \inf_q\{q | \exists n \text{ t.q. } G(q) > x_n\}$$

Soit maintenant q t.q. $G(q) > x$ alors $G(q) > x + \epsilon$ pour ϵ assez petit. Puisque x_n converge vers x il existe un n t.q. $x_n < x + \epsilon$ et donc pour tout q t.q. $G(q) > x$ on a aussi $G(q) > x_n$ pour un certain n. Donc

$$\{q|G(q) > x\} \subset \{q|\exists n \text{ t.q. } G(q) > x_n\}$$

qui implique

$$\lim_n F(x_n) \leq \inf_q \{q|\exists n \text{ t.q. } G(q) > x_n\} \leq \inf_q \{q|G(q) > x\} = F(x)$$

On a donc démontrer $\lim_n F(x_n) \leq F(x)$ et $\lim_n F(x_n) \geq F(x)$ pour toute suite x_n et donc on a la continuité à droite.

— On veut démontrer F_{n_k} converge vers F en tout point de continuité de F. Soit x point de continuité de F et $\epsilon > 0$. Soient $r < s < t$ t.q. $F(x) - \epsilon < F(r) \leq F(s) \leq F(x) \leq F(t) < F(x) + \epsilon$.

On a $F_{n_k}(s) \to G(s)$ et $F_{n_k}(t) \to G(t)$. On sait $G(s) > F(r)$ et $G(t) \leq F(t)$ par définition de F. Pour tout ϵ' il existe donc un K t.q pour tout $k > K_{\epsilon'}$ $F_{n_k}(t) \leq F(t) + \epsilon$ et $F_{n_k}(s) \geq F(r) - \epsilon$ et donc pour $\epsilon' = \epsilon$, pour tout $k > K_\epsilon$

$$F(x) - 2\epsilon < F(r) - \epsilon \leq F_{n_k}(s) \leq F_{n_k}(x) \leq F_{n_k}(t) \leq F(t) + \epsilon < F(x) + 2\epsilon$$

qui nous donne ce qu'on cherche.

En fin, on a donc démontrer qu'il existe une sous-suite dont les fonctions de répartition convergent en tout point de continuité vers une fonction F qui respecte toutes les propriétés d'une fonction de répartition,

— $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.

Pour le premier, on prend $\epsilon > 0$, on sait qu'il existe M_ϵ tel que $\mathbb{P}(\|X_n\| > M_\epsilon) > 1 - \epsilon$ pour tout n. Soient, $r < -M_\epsilon$ et $s > M_\epsilon$ deux points de continuité de F, qu'on peut toujours trouvé parce que une fonction bornée et croissante n'admet qu'un nombre dénombrable de points de discontinuité. On a

$$1 - F(s) + F(r) = \lim_{k \to +\infty} 1 - F_{n_k}(s) + F_{n_k}(r) \leq \lim_{k \to +\infty} 1 - F_{n_k}(M_\epsilon) + F_{n_k}(-M_\epsilon)$$

par croissance de F. En particulier

$$1 - F_{n_k}(M_\epsilon) + F_{n_k}(-M_\epsilon) = 1 - \mathbb{P}(\|X_{n_k}\| > M_\epsilon) \leq \epsilon$$

puisque X_n est tendue.

Donc pour tout ϵ on peut trouver M_ϵ t.q. pour tout s, r t.q. $|s|, |r| > M_\epsilon$ on a $1 - F(s) + F(r) \leq \epsilon$ qui veut dire :

$$\lim_{x \to +\infty} 1 - F(x) + F(-x) = 0$$

qui implique, grâce à $0 \leq F(x) \leq 1$:

$$0 = \lim_{x \to +\infty} 1 - F(x) + F(-x) \geq \lim_{x \to +\infty} 1 - F(x) \geq 0$$

$$0 = \lim_{x \to +\infty} 1 - F(x) + F(-x) = 0 + \lim_{x \to +\infty} F(-x).$$

□

8.2 La régression linéaire et logistique

Théorème 3.1.

Posons $U_i = \frac{\sqrt{n}}{\sigma}\epsilon_i (X'X)^{-\frac{1}{2}} X_i$, alors on a

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(U_i) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(\frac{\sqrt{n}}{\sigma}(X'X)^{-\frac{1}{2}}(\epsilon_i X_i)) \\
&= \frac{1}{n}\frac{n}{\sigma^2}(X'X)^{-\frac{1}{2}}(\sum_{i=1}^{n} \mathrm{Cov}(\epsilon_i X_i'))(X'X)^{-\frac{1}{2}} \\
&= \sigma^{-2}(X'X)^{-\frac{1}{2}}(\sum_{i=1}^{n} \sigma^2 X_i' X_i)(X'X)^{-\frac{1}{2}} \\
&= \sigma^{-2}(X'X)^{-\frac{1}{2}}\sigma^2 X'X(X'X)^{-\frac{1}{2}} \\
&= I_p.
\end{aligned}$$

On remarque que $h_{ii} = X_i(X'X)X_i'$ et que $\sum\limits_{i=1}^{n} h_{ii} = p$. Soit $\delta > 0$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|U_i\|_2^2 \mathbb{1}_{\|U_i\|_2 > \delta\sqrt{n}}] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\frac{\sqrt{n}}{\sigma}\epsilon_i (X'X)^{-\frac{1}{2}} X_i'\|_2^2 \mathbb{1}_{\|\frac{\sqrt{n}}{\sigma}\epsilon_i (X'X)^{-\frac{1}{2}} X_i'\|_2 > \delta\sqrt{n}}] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2 X_i (X'X)^{-1} X_i' \mathbb{1}_{\epsilon_i^2 X_i (X'X)^{-1} X_i' > \delta^2\sigma^2}] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathbb{E}[h_{ii}\epsilon_1^2 \mathbb{1}_{h_{ii}\epsilon_1^2 > \delta^2\sigma^2}] \\
&\leq \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathbb{E}[h_{ii}\epsilon_1^2 \mathbb{1}_{\max_{1\leq i\leq n} h_{ii}\epsilon_1^2 > \delta^2\sigma^2}] \\
&= \frac{p}{\sigma^2}\mathbb{E}[\epsilon_1^2 \mathbb{1}_{\max_{1\leq i\leq n} h_{ii}\epsilon_1^2 > \delta^2\sigma^2}] \\
&\underset{n\to+\infty}{\longrightarrow} 0 \qquad \text{par convergence dominée et } \mathbb{E}[\epsilon_1^2] < +\infty.
\end{aligned}$$

Donc par le théorème de Lindeberg-Feller, on a

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}(0, I_p).$$

Or

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\sqrt{n}}{\sigma}\epsilon_i (X'X)^{-\frac{1}{2}} X_i \\
&= \frac{1}{\sigma}(X'X)^{-\frac{1}{2}}\sum_{i=1}^{n} \epsilon_i X_i \\
&= \frac{1}{\sigma}(X'X)^{-\frac{1}{2}}(X'X)(\hat{\beta} - \beta) \\
&= \frac{1}{\sigma}(X'X)^{\frac{1}{2}}(\hat{\beta} - \beta).
\end{aligned}$$

D'où le résultat. □

Théorème 3.3.
On commence par montrer le fait que

$$\mathcal{H}\ l_n(\beta, X, Y) = -\sum_{i=1}^{n} X_i' X_i f(X_i\beta)$$

est définie négative. Soit z vecteur de $\mathbb{R}^p$ on a

$$z'(\mathcal{H}\ l_n(\beta, X, Y))z = -\sum_{i=1}^{n}(z'X_i')(zX_i)f(X_i\beta) = -\sum_{i=1}^{n} f(X_i\beta)(< z, X_i >)^2 \leq 0.$$

Puisque f est strictement positive et la matrice X est de rang p, si $z \neq 0$ il existe i t.q. $< z, X_i > \neq 0$ et donc $z'(\mathcal{H}\ l_n(\beta, X, Y))z < 0$.

Pour démontrer les deux points suivants, centraux pour notre étude de la régression, il faut introduire la suivante fonction $\phi_n : \mathbb{R}^p \times \Omega \longrightarrow \mathbb{R}^p$

$$\phi_n(\beta, X, Y) = \beta + (-\mathcal{H}\ l_n(\beta, X, Y))^{-1} \nabla\ l_n(\beta, X, Y)$$

et énoncer deux lemmes qu'on va utiliser sans démontrer.

Lemme 8.1. Il existe $r > 0$, $0 < c < 1$ t.q. presque sûrement et pour tout n la fonction ϕ_n est Lipschitz par rapport à la variable β sur la boule $B(\beta_0, r)$ c'est à dire pour tout ω dans un ensemble de mesure 1 pour tout $b, b' \in B(\beta_0, r)$

$$\|\phi_n(b, X, Y)(\omega) - \phi_n(b', X, Y)(\omega)\| \leq c\|b - b'\|.$$

Pour la preuve [4].

Lemme 8.2. Soit $f, g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ t.q. $f(x) = x - g(x)$ et $f(x)$ est une fonction Lipschitz sur la boule $B(x_0, r)$ avec constante de Lipschitz $0 < c < 1$. Alors la fonction $f(x)$ est telle que :

$$B(g(x_0), (1 - c)r) \subset g(B(x_0, r)).$$

Pour la preuve [3]

De ces propriétés on peut directement démontrer l'existence et consistance de $\hat{\beta}_n$.

On va parfois noter $l_n(\beta_0, X, Y)(\omega)$ comme $l_n(\beta_0, \omega)$ pour simplifier la notation et de même $\phi_n(\beta, X, Y)(\omega)$ comme $\phi_n(\beta, \omega)$

On commence par démontrer que la condition de convergence presque sure à zéro de $[-\mathcal{H}\ l_n(\beta_0, \omega)]^{-1} \nabla\ l_n(\beta_0, \omega)$ est suffisante pour l'existence et consistance.

Soit $\omega \in \Omega$ tel que $[-\mathcal{H}\ l_n(\beta_0, X, Y)]^{-1} \nabla\ l_n(\beta_0, X, Y)(\omega) \longrightarrow_n 0$, $\phi_n(\beta, X, Y)(\omega)$ est Lipschitz sur $B(\beta_0, r)$ et tel que la matrice hessienne est définie négative.

En notant

$$\psi_n(\beta, X, Y) = [\mathcal{H}\ l_n(\beta, X, Y)]^{-1} \nabla\ l_n(\beta, X, Y)$$

on a $\phi_n(\beta, X, Y) = \beta - \psi_n(\beta, X, Y)$. (On va noter aussi $\psi_n(\beta, \omega) = \psi(\beta, X, Y)(\omega)$)

Soit $\epsilon > 0$ et r_ϵ t.q. $r_\epsilon \leq \min(r, \epsilon)$. Puisque

$$B(\psi_n(\beta_0, \omega), (1 - c)r_\epsilon) \subset \psi_n(B(\beta_0, r_\epsilon), \omega).$$

et

$$\psi_n(\beta_0, \omega) \longrightarrow_n 0$$

pour tout $n > n_{\epsilon,\omega}$

$$0 \in \psi_n(B(\beta_0, r_\epsilon), \omega).$$

C'est à dire pour tout $n > n_{\epsilon,\omega}$ il existe un $\hat{\beta}_n(\omega)$ tel que

$$\psi_n(\hat{\beta}_n(\omega), \omega) = 0.$$

C'est à dire, finalement,

$$\nabla\ l_n(\hat{\beta}_n(\omega), \omega) = 0.$$

On a démontré que pour tout ω dans un ensemble de mesure 1, pour tout $\epsilon > 0$ il existe $n_{\omega,\epsilon}$ t.q. $\forall n > n_{\omega,\epsilon}$ il existe $\hat{\beta}_n(\omega)$ t.q :

$$\nabla \ell_n(\hat{\beta}_n(\omega), X, Y)(\omega) = 0$$
$$\|\hat{\beta}_n(\omega) - \beta_0\| < \epsilon$$

Puisque la matrice hessienne est définie négative sur $\mathbb{R}^p$ le point critique est unique et c'est aussi le point maximum.

On montre maintenant par l'absurde que la condition de convergence presque sure à zéro de

$$[-\mathcal{H}\ l_n(\beta_0, X, Y)]^{-1} \nabla\ l_n(\beta_0, X, Y) \longrightarrow_n 0$$

est nécessaire pour l'existence et consistance.

Soit $\hat{\beta}_n$ l'estimateur MLE défini asymptotique ment sur un ensemble Ω_0 de mesure 1 et fortement consistant sur cet ensemble. On suppose qu'il existe un ensemble Ω_1 de mesure strictement positive t.q.

$$\psi_n(\beta_0, X, Y) = [\mathcal{H}\ l_n(\beta_0, X, Y)]^{-1} \nabla\ l_n(\beta_0, X, Y)$$

ne converge pas a zéro sur cet ensemble. Soit $\omega \in \Omega_0 \bigcap \Omega_1$.

Pour l'hypothèse il existe $\epsilon > 0$ et une sous-suite $\{n_k\}_k$ t.q. pour tout k

$$\|\psi_{n_k}(\beta_0, X, Y)(\omega)\| > \epsilon.$$

Donc, pour $r = \frac{\epsilon}{2(1+c)}$, pour tout k, si on prend β dans $B(\beta_0, r)$ on a

$$\|\psi_{n_k}(\beta_0, \omega) - \psi_{n_k}(\beta, \omega)\| \leq \|\beta_0 - \beta_0\| + \|\phi_{n_k}(\beta_0, \omega) - \phi_{n_k}(\beta, \omega)\|$$

qui sont respectivement plus petit de r et de cr. Donc pour tout k , pour tout $\beta \in B(\beta_0, r)$ on a

$$\|\psi_{n_k}(\beta, X, Y)(\omega)\| > \epsilon/2.$$

En même temps il existe une suite $\hat{\beta}_{n_k}(\omega)$ qui annule $\psi_{n_k}(\beta, X, Y)(\omega)$ pour tout k et qui est pour k assez grand dans la boule $B(\beta_0, r)$. Donc pour k assez grand on ne peut pas avoir $\|\psi_{n_k}(\hat{\beta}_{n_k}(\omega), X, Y)(\omega)\| > \epsilon/2$, ce qui nous donne une absurdité et nous permet de conclure la preuve. □

Théorème 3.4.
On écrit un développement limité de Taylor-Lagrange pour $\nabla l_n(\beta, X, Y)$ et on obtient

$$\nabla l_n(\hat{\beta}, X, Y) = \nabla l_n(\beta_0, X, Y) + \mathcal{H} l_n(\beta^*, X, Y)(\hat{\beta} - \beta_0)$$

pour un certain β^* qui vérifie $\|\beta^* - \beta_0\| \leq \|\hat{\beta} - \beta_0\|$ (Cette dernière équation est valable pour n assez grand, de telle sorte que $\hat{\beta}$ existe). On obtient donc

$$\hat{\beta} - \beta_0 = -(\mathcal{H} l_n(\beta^*, X, Y))^{-1} \nabla l_n(\beta_0, X, Y).$$

Donc

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sqrt{n} \left(\underbrace{-\frac{1}{n} \mathcal{H} l_n(\beta^*, X, Y)}_{:= A_n} \right)^{-1} \frac{1}{n} \nabla l_n(\beta_0, X, Y).$$

Or $A_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i\beta_*)X_i'X_i$, donc par la loi forte des grands nombres et le théorème de continuité, on a $A_n \to A$ p.s. (L'intégrabilité vient de la première hypothèse faite sur les X). On a donc

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= \sqrt{n}(A_n^{-1}-A^{-1}+A^{-1})\frac{1}{n}\nabla l_n(\beta_0,X,Y)\\ &= A^{-1}\sqrt{n}\frac{1}{n}\nabla l_n(\beta_0,X,Y)+(A_n^{-1}-A^{-1})\sqrt{n}\frac{1}{n}\nabla l_n(\beta_0,X,Y).\end{aligned}$$

On rappelle que $\frac{1}{n}\nabla l_n(\beta_0,X,Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F(X_i\beta)X_i$. On a $\mathbb{E}[(Y_1 - F(X_1\beta)X_1] = 0$ et $Y_1 - F(X_1\beta)X_1$ est bornée donc elle admet une matrice de covariance. Donc par le théorème central limite, on a

$$\sqrt{n}\frac{1}{n}\nabla l_n(\beta_0,X,Y) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,B).$$

En particulier on a

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= A^{-1}\sqrt{n}\frac{1}{n}\nabla l_n(\beta_0,X,Y)+\underbrace{(A_n^{-1}-A^{-1})}_{o_{\mathbb{P}}(1)}\underbrace{\sqrt{n}\frac{1}{n}\nabla l_n(\beta_0,X,Y)}_{O_{\mathbb{P}}(1)}\\ &\xrightarrow[n\to\infty]{d} \mathcal{N}(0,A^{-1}BA^{-1}).\end{aligned}$$

□

8.3 Propriétés conditionnelles

Théorème 4.1 .
On démontre la première propriété. On sait que pour une suite de variables aléatoires $Z_n \in [0,1]$ la convergence en probabilité et la convergence dans $\mathcal{L}^1$ sont équivalentes. Soit $\mathcal{F}_n$, on a

$$P(|X_n| > \epsilon) = \mathbb{E}(P(|X_n| > \epsilon|\mathcal{F}_n)).$$

Donc X_n converge en probabilités vers zéro ssi $P(|X_n| > \epsilon|\mathcal{F}_n)$ converge dans $\mathcal{L}^1$ vers 0. Puisque pour une suite de variables aléatoires $Z_n \in [0,1]$ la convergence en probabilité et la convergence dans $\mathcal{L}^1$ sont équivalentes, on a X_n converge en probabilités vers zéro ssi $P(|X_n| > \epsilon|\mathcal{F}_n)$ converge en probabilités vers 0. Pour la deuxième propriété on a, grâce à l'inégalité de Markov,

$$\sup_n \mathbb{P}(\mathbb{P}(|X_n| > M|\mathcal{F}_n) > \delta) \leq \sup_n \frac{\mathbb{P}(|X_n| > M)}{\delta}$$

donc si $X_n = O_{\mathbb{P}}(1)$ alors $X_n = O_{\mathbb{P}|\mathcal{F}_n}(1)$. D'autre part

$$\mathbb{P}(|X_n| > M) = \mathbb{E}(\mathbb{P}(|X_n| > M|\mathcal{F}_n)) =$$

$$\mathbb{E}(\mathbb{P}(|X_n| > M|\mathcal{F}_n)\mathbb{1}_{\mathbb{P}(|X_n|>M|\mathcal{F}_n)>\delta}) + \mathbb{E}(\mathbb{P}(|X_n| > M|\mathcal{F}_n)\mathbb{1}_{\mathbb{P}(|X_n|>M|\mathcal{F}_n)\leq\delta}) \leq$$
$$\mathbb{P}(\mathbb{P}(|X_n| > M|\mathcal{F}_n) > \delta) + \delta \underset{M\to+\infty}{\longrightarrow} \delta$$

De la généralité de δ suit le résultat. □

Théorème 4.3.
On démontre le cas univarié. Le cas multivarié est identique.

$$\begin{aligned}\int_{\mathbb{R}} |f_n(x,.) - f(x,.)|dx &= \int_{|x|>K} |f_n(x,.) - f(x,.)|dx + \int_{|x|\leq K} |f_n(x,.) - f(x,.)|dx\\ &\leq \int_{|x|>K} 2|h(x)|dx + \int_{|x|\leq K} \sup_{x\in[-K,K]} |f_n(x,.) - f(x,.)|dx\\ &\leq \epsilon/2 + 2K_{\epsilon/2} \sup_{x\in[-K_{\epsilon/2},K_{\epsilon/2}]} |f_n(x,.) - f(x,.)|.\end{aligned}$$

Donc

$$\mathbb{P}(\int_{\mathbb{R}} |f_n(x,.) - f(x,.)|dx > \epsilon) \leq \mathbb{P}(\epsilon/2 + 2K_{\epsilon/2} \sup_{x\in[-K_{\epsilon/2},K_{\epsilon/2}} |f_n(x,.) - f(x,.)| > \epsilon)$$
$$= \mathbb{P}(\sup_{x\in[-K_{\epsilon/2},K_{\epsilon/2}]} |f_n(x,.) - f(x,.)| > \frac{\epsilon}{4K_{\epsilon/2}})$$

et le dernier terme tend vers zéro. □

Théorème 4.4.
1. implique évidemment 3. et aussi 2. implique évidemment 3. On va démontrer 3. implique 2.. La preuve de 2. implique 1. est très similaire à celle de 3. implique 2.

Pour démontrer l'implication inverse, on commence par démontrer que la suite X_n est tendue.

C'est très facile à démontrer que pour un vecteur aléatoire Y presque-sûrement fini et pour toute tribu $\mathcal{F}_n$ et pour tout δ on a $\mathbb{P}(\mathbb{P}(\|Y\|| > K|\mathcal{F}_n) > \delta)$ tend à zéro lorsque K tend vers l'infini.

Soit maintenant $\phi_k : \mathbb{R}^p \to \mathbb{R}$ une suite de fonction t.q. pour tout K $\phi_k(x) = \mathbb{1}_{\|x\|\leq k}$ pour $\|x\| \geq k+1$ et pour $\|x\| \leq k$, $\phi_k(x) \geq 0$ pour tout x et $\phi_k \in C_c^2(\mathbb{R})$, $|\phi_k(x)| \leq 1$.

Soit $\delta > 0$. On a

$$\mathbb{P}(\mathbb{P}(\|X_n\|| > K|\mathcal{F}_n) > \delta) \leq \mathbb{P}(1 - \mathbb{E}(\phi_k(X_n)|\mathcal{F}_n) > \delta)$$
$$\leq \mathbb{P}(1 - \mathbb{E}(\phi_k(X)) > \delta/2) + \mathbb{P}(|\mathbb{E}(\phi_k(X)) - \mathbb{E}(\phi_k(X_n)|\mathcal{F}_n)| > \delta/2)$$
$$\leq \mathbb{P}(P(\|X\| > k+1) > \delta/2) + \mathbb{P}(|\mathbb{E}(\phi_k(X)) - \mathbb{E}(\phi_k(X_n)|\mathcal{F}_n)| > \delta/2)$$

Soit $\epsilon > 0$ Évidemment, le premier terme est égale à zéro pour K plus grand qu'un certain K_0. Donc pour tout $K > K_0$

$$\mathbb{P}(\mathbb{P}(\|X_n\|| > K|\mathcal{F}_n) > \delta) \leq \mathbb{P}(|\mathbb{E}(\phi_k(X)) - \mathbb{E}(\phi_k(X_n)|\mathcal{F}_n)| > \delta/2).$$

Pour n plus grand d'un certain n_ϵ on a par hypothèse

$$\mathbb{P}(|\mathbb{E}(\phi_k(X)) - \mathbb{E}(\phi_k(X_n)|\mathcal{F}_n)| > \delta/2) < \epsilon.$$

On choisit maintenant pour $i \in \{1, ..., n_\epsilon\}$ un K_i t.q. $\mathbb{P}(\mathbb{P}(\|X_i\|| > K|\mathcal{F}_n) > \delta) \leq \epsilon$. On a que pour tout $K > \bar{K}$, où $\bar{K} = max\{K_0, K_1, ..., K_{n_\epsilon}\}$

$$\mathbb{P}(\mathbb{P}(\|X_n\|| > K|\mathcal{F}_n) > \delta) < \epsilon$$

donc

$$sup_n\mathbb{P}(\mathbb{P}(\|X_n\|| > K|\mathcal{F}_n) > \delta) < \epsilon$$

Pour tout ϵ on a trouvé $\bar{K}$ t.q. $sup_n\mathbb{P}(\mathbb{P}(\|X_n\|| > K|\mathcal{F}_n) > \delta) < \epsilon$ qui implique que $X_n = o_{\mathbb{P}}(1)$. La suite X_n est donc tendue.

Soit maintenant la $\psi \in C_b^2(\mathbb{R}^p)$. Soit une nouvelle suite ϕ_k t.q. pour tout K $\phi_k(x) = \psi(x)$ pour tout x dans la boule de rayon K, $\phi_k(x) = 0$ pour tout x au dehors de la boule de rayon k+1, $\phi_k \in C_c^2(\mathbb{R}^p)$ et $|\phi_k(x)| \leq |\psi_k(x)|$ pour tout x.

Soit $\epsilon > 0$. On veut montrer $\mathbb{E}(\psi(X_n)|\mathcal{F}_n) - \mathbb{E}(\psi(X)) = o_{\mathbb{P}}(1)$. On a

$$\mathbb{P}(\mathbb{E}(\psi(X_n)|\mathcal{F}_n) - \mathbb{E}(\psi(X)) > \delta)$$

$$\leq \mathbb{P}(\mathbb{E}(|\psi(X_n) - \phi_k(X_n)||\mathcal{F}_n) > \delta/3) + \mathbb{P}(\mathbb{E}(\phi_k(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi_k(X)) > \delta/3) + \mathbb{P}(\mathbb{E}(\phi_k(X) - \psi(X)) > \delta/3)$$

$$\mathbb{P}(\mathbb{E}(|\psi(X_n)\mathbb{1}_{\|X_n\|>K}||\mathcal{F}_n) > \delta/3) + \mathbb{P}(\mathbb{E}(\phi_k(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi_k(X)) > \delta/3) + \mathbb{P}(\mathbb{E}(|\psi(X)|\mathbb{1}_{\|X\|} > K) > \delta/3)$$

$$\mathbb{P}(\mathbb{E}(C\mathbb{1}_{\|X_n\|>K}|\mathcal{F}_n) > \delta/3) + \mathbb{P}(\mathbb{E}(\phi_k(X_n)|\mathcal{F}_n) - \mathbb{E}(\phi_k(X)) > \delta/3) + \mathbb{P}(\mathbb{E}(C\mathbb{1}_{\|X\|} > K) > \delta/3)$$

Pour n assez grand le deuxième terme est plus petit de ϵ et pour K assez grand le premier terme e le dernier sont plus petit de ϵ. On peut donc déduire que $\mathbb{E}(\psi(X_n)|\mathcal{F}_n) - \mathbb{E}(\psi(X)) = o_{\mathbb{P}}(1)$. □

Théorème 4.5.

L'implication de 2. à 1. est évidente.

On démontre seulement 1. implique 2.

Soit g est une fonction de classe C^2 à support compact, comme g est à support compact, alors g, g', g'' sont dans $L^1(\mathbb{R}, \lambda_{\mathbb{R}})$.

On peut donc considérer leur transformée de Fourier, qu'on note avec un chapeau. Puisque pour tout $t \in \mathbb{R}$, $\hat{g''}(t) = -t^2\hat{g}(t)$, on a $|\hat{g}(t)| = O(1/t^2)$ puisque $\hat{g''}(t)$ est bornée. Donc $\hat{g}(t)$ est dans $L^1(\mathbb{R}, \lambda_{\mathbb{R}})$ et elle est bornée.

Pour la formule d'inversion de Fourier on peut écrire :

$$g(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{g}(s)e^{isx}ds$$

et donc en passant aux espérances pour X_n

$$\mathbb{E}(g(X_n)|\mathcal{F}_n) = \frac{1}{2\pi}\mathbb{E}(\int_{\mathbb{R}} \hat{g}(s)e^{isX_n}ds|\mathcal{F}_n)$$

$$= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{g}(s)\mathbb{E}(e^{isX_n}|\mathcal{F}_n)ds.$$

On peut appliquer Fubini presque sûrement parce que $|\hat{g}(s)\mathbb{E}(e^{isX_n}|\mathcal{F}_n)| \leq \hat{g}(s)$ p.s. . En plus on a

$$|\hat{g}(s)\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \hat{g}(s)\mathbb{E}(e^{isX})| = o_{\mathbb{P}}(1).$$

Pour conclure il faut démontrer que $|\hat{g}(s)\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \hat{g}(s)\mathbb{E}(e^{isX})|$ converge uniformément en probabilité à zéro sur tout intervalle fermé.

Dans les prochaine lignes on va donc démontrer la convergence uniforme en probabilité à zéro sur un intervalle fermé $[-M, M]$. de $|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{isX})|$. Si on démontre ça, puisque $\hat{g}$ est bornée on démontre aussi la convergence uniforme à zéro de $|\hat{g}(s)\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \hat{g}(s)\mathbb{E}(e^{isX})|$.

Pour arriver à démontrer cette convergence uniforme il faut démontrer que

— X_n est tendue

— $\mathbb{E}(e^{isX_n}|\mathcal{F}_n)$ est uniformément equicontinue en probabilité

Le fait que la suite X_n est tendu est facile à voir. En effet puisque,

$$\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{isX}) = o_{\mathbb{P}}(1).$$

et

$$|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{isX}) = o_{\mathbb{P}}(1)| \leq 2$$

on a convergence dans L^1, c'est à dire

$$\mathbb{E}(e^{isX_n}) \to_n \mathbb{E}(e^{isX}).$$

Cette dernière propriété implique la convergence en loi et donc la suite est tendue.
Pour l'equicontinuité, premièrement on peut noter que

$$|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{itX_n}|\mathcal{F}_n)| \leq |\mathbb{E}(e^{i(s-t)X_n} - 1|\mathcal{F}_n)| \leq \mathbb{E}(|e^{i(s-t)X_n} - 1||\mathcal{F}_n)$$

$$\leq \mathbb{E}(|e^{i(s-t)X_n} - 1|\mathbb{1}_{|X_n|>K}|\mathcal{F}_n)) + \mathbb{E}(|(s-t)X_n|\mathbb{1}_{|X_n|\leq K}|\mathcal{F}_n)) \leq 2\mathbb{E}(\mathbb{1}_{|X_n|>K}|\mathcal{F}_n)) + K|s-t|$$

.

On a donc

$$\mathbb{P}(|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{itX_n}|\mathcal{F}_n)| > \delta) \leq \mathbb{P}(2\mathbb{E}(\mathbb{1}_{|X_n|>K}|\mathcal{F}_n)) > \delta/2) + \mathbb{1}_{|s-t|K>\delta/2}$$

$$\leq \mathbb{P}(|X_n| > K)/(\delta/2) + \mathbb{1}_{|s-t|K>\delta/2}$$

(la dernière inégalité vient de Markov).

En passant au sup on a

$$\sup_n \mathbb{P}(|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{itX_n}|\mathcal{F}_n)| > \delta) \leq \sup_n \mathbb{P}(|X_n| > K)/(\delta/2) + \mathbb{1}_{|s-t|K>\delta/2}$$

Puisque la suite est tendue, on arrive à démontrer l'uniforme equicontinuité.

Soit maintenant un intervalle $[-M, M]$. Soit ϵ, δ arbitraires et δ_ϵ pour avoir l'equicontinuité.
Puisque $[-M, M]$ est compact, il existent $t_1, ..., t_d$ t.q. $[-M, M] \subset U_{i=1,...,d} B(t_i, \delta_\epsilon)$.
Pour tout s dans $[-M, M]$, s est dans une boule (on appelle le centre de cette boule t_k)
On a

$$|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{isX})| \leq |\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{it_kX_n}|\mathcal{F}_n)| + \max_{1,...,d} |\mathbb{E}(e^{it_iX_n}|\mathcal{F}_n) - \mathbb{E}(e^{it_iX})| + |\mathbb{E}(e^{it_kX}) - \mathbb{E}(e^{isX})|$$

En passant au probabilité

$$\begin{aligned}\mathbb{P}(|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{isX})| > \delta) \leq{}& \mathbb{P}(|\mathbb{E}(e^{isX_n}|\mathcal{F}_n) - \mathbb{E}(e^{it_kX_n}|\mathcal{F}_n)| > \delta/3) \\ &+ \mathbb{P}(\max_{1,...,d} |\mathbb{E}(e^{it_iX_n}|\mathcal{F}_n) - \mathbb{E}(e^{it_iX})| > \delta/3) \\ &+ \mathbb{1}_{|\mathbb{E}(e^{i(t_k-s)X})-1|>\delta/3}\end{aligned}$$

— Le premier terme est plus petit de ϵ
— Le deuxième terme est plus petit de ϵ pour n assez grand
— Le troisième terme est zéro pour δ_ϵ assez petit.

On a donc la convergence uniforme en probabilité sur tout compact. Grâce au théorème de convergence des intégrales en probabilité on peut conclure

$$\mathbb{E}(g(X_n)|mathcalF_n) - \mathbb{E}(g(X)) = o_{\mathbb{P}}(1)$$

□

Théorème 4.6.
Comme dans le cas univarié, on peut montrer que la convergence ponctuel en probabilités de $|\mathbb{E}(e^{itX_n}|\mathcal{F}_n) - \mathbb{E}(e^{itX})|$ implique la convergence en probabilités pour le sup sur tout compact $K \subset \mathbb{R}^p$, c'est à dire

$$\sup_{t\in K} |\mathbb{E}(e^{itX_n}|\mathcal{F}_n) - \mathbb{E}(e^{itX})| = o_{\mathbb{P}(1)}.$$

Soit maintenant $g \in C_c^2(\mathbb{R}^p)$. On g est uniformément continue et borné et on peut écrire

$$\begin{aligned}|\mathbb{E}(g(X_n)|\mathcal{F}_n) - \mathbb{E}(g(X))| \leq{}& |\mathbb{E}(g(X_n)|\mathcal{F}_n) - \mathbb{E}(g(X_n + Z)|\mathcal{F}_n)| \\ &+ |\mathbb{E}(g(X)) - \mathbb{E}(g(X + Z))| \\ &+ |\mathbb{E}(g(X_n + Z)|\mathcal{F}_n) - \mathbb{E}(g(X + Z))|\end{aligned}$$

où Z est de loi $\mathcal{N}(0, \sigma^2 I_p)$.
Soit ϵ et le δ_ϵ de la continuité uniforme. Pour le premier terme, on a

$$|\mathbb{E}(g(X_n)|\mathcal{F}_n) - \mathbb{E}(g(X_n + Z)|\mathcal{F}_n)| \leq 2C\mathbb{P}(\|Z\| > \delta_\epsilon|\mathcal{F}_n) + |\mathbb{E}(\epsilon\mathbb{1}_{\|Z\|\leq\delta_\epsilon}|\mathcal{F}_n)|$$

où C est la constante qui borne g. On peut écrire la même chose pour le deuxième terme et arriver à démontrer que, pour σ assez petit

$$|\mathbb{E}(g(X_n)|\mathcal{F}_n) - \mathbb{E}(g(X))| \leq 4\epsilon + |\mathbb{E}(g(X_n + Z)|\mathcal{F}_n) - \mathbb{E}(g(X + Z))|.$$

On peut écrire, en considérant Z complètement indépendant de $\{\mathcal{F}_n\}_n$,

$$\mathbb{E}(g(X_n+Z)|\mathcal{F}_n) =^{p.s.} \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(x+z)\phi(z)dzd\mu_{X_n|\mathcal{F}_n}(x)$$

(Φ densité de Z) et avec la transformation $u = x+z, x = x$

$$=^{p.s.} \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\Phi(u-x)dud\mu_{X_n|\mathcal{F}_n}(x)$$

De façon analogue on peut écrire

$$\mathbb{E}(g(X+Z)) = \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\Phi(u-x)dud\mu_X(x)$$

On a

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\phi(u-x)dud\mu_{X_n|\mathcal{F}_n}(x) =^{p.s.} \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\prod_{j=1}^{p}\phi(u_j-x_j)dud\mu_{X_n|\mathcal{F}_n}(x)$$

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\phi(u-x)dud\mu_X(x) = \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\prod_{j=1}^{p}\phi(u_j-x_j)dud\mu_X(x)$$

avec ϕ densité de Z_1.
Par calcul on obtient

$$\phi(x_j-u_j) = \int_{\mathbb{R}} exp(it_j(u_j-x_j)-\sigma^2t_j^2/2)dt_j$$

et donc on peut écrire

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\phi(u-x)dud\mu_{X_n|\mathcal{F}_n}(x) =^{p.s.} \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\prod_{j=1}^{p}\int_{\mathbb{R}} exp(it_j(u_j-x_j)-\sigma^2t_j^2/2)dt_jdud\mu_{X_n|\mathcal{F}_n}(x)$$

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\phi(u-x)dud\mu_X(x) = \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\prod_{j=1}^{p}\int_{\mathbb{R}} exp(it_j(u_j-x_j)-\sigma^2t_j^2/2)dt_jdud\mu_X(x)$$

Par Fubini,

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\prod_{j=1}^{p}\int_{\mathbb{R}} exp(it_j(u_j-x_j)-\sigma^2t_j^2/2)dt_jdud\mu_{X_n|\mathcal{F}_n}(x) =^{p.s.} \int_{\mathbb{R}^p}\int_{\mathbb{R}} g(u)e^{it'u-\sigma^2\frac{<t,t>}{2}}\mathbb{E}(e^{i<-t,X_n>}|\mathcal{F}_n)dtdu$$

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)\prod_{j=1}^{p}\int_{\mathbb{R}} exp(it_j(u_j-x_j)-\sigma^2t_j^2/2)dt_jdud\mu_X(x) = \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} g(u)e^{it'u-\sigma^2\frac{<t,t>}{2}}\mathbb{E}(e^{i<-t,X>})dtdu$$

Si on appelle K la constante qui borne $f(u,t) = g(u)e^{it'u-\sigma^2\frac{<t,t>}{2}}$

$$|\mathbb{E}(g(X_n+Z)|\mathcal{F}_n) - \mathbb{E}(g(X+Z))| \leq \int_{\mathbb{R}^p} K|\mathbb{E}(e^{i<-t,X_n>}|\mathcal{F}_n) - \mathbb{E}(e^{i<-t,X>})|dt$$

qui, pour le théorème de convergence dominée en probabilité est un $o_{\mathbb{P}}(1)$.
On a donc demontré que pour tout ϵ

$$|\mathbb{E}(g(X_n)|\mathcal{F}_n) - \mathbb{E}(g(X))| \leq 4\epsilon + o_{\mathbb{P}}(1).$$

Ce qui nous permet de conclure la convergence en loi conditionnelle aussi dans le cas multivarié.

□

Théorème 4.7.
En regardant la preuve du théorème de Cramer Wald classique et en notant que

$$\mathbb{E}(e^{i1<t,X_n>}|\mathcal{F}_n) - \mathbb{E}(e^{i1<t,X>}) = \mathbb{E}(e^{i<t,X_n>}|\mathcal{F}_n) - \mathbb{E}(e^{i<t,X>}) = o_{\mathbb{P}}(1).$$

□

Théorème 4.8.
Soit $\epsilon > 0$. En suivant exactement les mêmes passages du théorème de Lindeberg - Feller univarié classique on arrive a pouvoir écrire

$$|\mathbb{E}(exp(it\sum_{j=1}^{r_n}\tilde{X}_{nj}) - exp(-\frac{t^2}{2})|\mathcal{F}_n)|$$

$$\leq \sum_{j=1}^{r_n}|\mathbb{E}(e^{it\tilde{X}_{nj}}|\mathcal{F}_n) - (1 - \frac{t^2\mathrm{Var}(\tilde{X}_{nj}|\mathcal{F}_n)}{2})| + \sum_{j=1}^{r_n}|(1 - \frac{t^2\mathrm{Var}(\tilde{X}_{nj}|\mathcal{F}_n)}{2} - exp(\frac{t^2\mathrm{Var}(\tilde{X}_{nj}|\mathcal{F}_n)}{2})|$$

en analogie, on appelle le premier terme A_n et le deuxième B_n. On a

$$A_n \leq t^3\epsilon^3\sum_{j=1}^{r_n}\mathrm{Var}(\tilde{X}_{nj}|\mathcal{F}_n) + t^2 o_{\mathbb{P}}(1)$$

où la somme des variances est p.s. 1.

$$B_n \leq t^4 \max_{1\leq j\leq r_n}\{\mathrm{Var}(\tilde{X}_{nj}|\mathcal{F}_n)\}exp(t^2\max_{1\leq j\leq r_n}\{\mathrm{Var}(\tilde{X}_{nj}|\mathcal{F}_n)\})$$

$$\leq t^4\epsilon^2 + o_{\mathbb{P}}(1)exp(t^2\epsilon^2 + o_{\mathbb{P}}(1)).$$

Ces deux inégalités permettent de conclure. □

Théorème 4.9.
Par Cramer Wald, soit $a \in \mathbb{R}^{\mathbb{P}}$ et $\epsilon > 0$. On a

$$0 \leq \sum_{j=1}^{r_n}\mathbb{E}(|<a,\tilde{X}_{nj}>|^2\mathbb{1}_{|<a,\tilde{X}_{nj}>|>\epsilon}|\mathcal{F}_n) \leq \|a\|^2\sum_{j=1}^{r_n}\mathbb{E}(\|\tilde{X}_{nj}\|^2\mathbb{1}_{\|\tilde{X}_{nj}\|>\frac{\epsilon}{\|a\|}}|\mathcal{F}_n) \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} 0$$

Donc pour tout $a \in \mathbb{R}$

$$<a,\sum_{j=1}^{r_n}\tilde{X}_{nj}> = \sum_{j=1}^{r_n}<a,\tilde{X}_{nj}> \underset{n\to+\infty}{\overset{d|\mathcal{F}_n}{\longrightarrow}} \mathcal{N}(0,\|a\|^2) = <a,\mathcal{N}(0,I_p)>.$$

Par Cramer Wald on en deduit

$$\sum_{j=1}^{r_n}\tilde{X}_{nj} \underset{n\to+\infty}{\overset{d|\mathcal{F}_n}{\longrightarrow}} \mathcal{N}(0,I_p)$$

□

8.4 Subsampling et Optimal Subsampling

Théoreme 5.1.

Pour démontrer le théorème il faut d'abord démontrer la propriété suivante Avec les hypothèses que nous avons fait on a :

— $n^{-1}(\mathcal{H}\ l_n^*(\hat{\beta}_n, X, Y) - n^{-1}\mathcal{H}\ l_n(\hat{\beta}_n, X, Y)) = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$
— $n^{-1}\nabla\ l_n^*(\hat{\beta}, X, Y) = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$

(La constante c qu'on va utiliser dans la preuve indique l'équivalence entre la norme matricielle et la somme des ses éléments en valeur absolue.)

Un calcul direct et assez facile (on se souvenant que les I_j sont iid) nous permet de montrer que :

$$n^{-1}\mathcal{H}\ l_n^*(\hat{\beta}_n, X, Y) = -(nr)^{-1}\sum_{j=1}^{r}\frac{X'_{I_j}X_{I_j}f(X_{I_j}\hat{\beta}_n)}{\pi_{I_j}}$$

et que

$$\mathbb{E}(n^{-1}\mathcal{H}\ l_n^*(\hat{\beta}_n, X, Y)|\mathcal{F}_n) = n^{-1}\mathcal{H}\ l_n(\hat{\beta}_n, X, Y)$$

Dans la suite on va noter $H_n^* = n^{-1}\mathcal{H}\ l_n^*(\hat{\beta}_n, X, Y)$ et $H_n = n^{-1}\mathcal{H}\ l_n(\hat{\beta}_n, X, Y)$. On a pour tout élément j_1, j_2 que :

$$\begin{aligned}
\mathrm{Var}(H_n^{*j_1,j_2}|\mathcal{F}_n) &= \mathbb{E}\bigg(\bigg(-(nr)^{-1}\sum_{j=1}^{r}\frac{X_{I_j}^{j_1\prime}X_{I_j}^{j_2}f(X_{I_j}\hat{\beta}_n)}{\pi_{I_j}} - H_n^{j_1,j_2}\bigg)^2\bigg|\mathcal{F}_n\bigg)\\
&= \frac{1}{n^2r}\mathbb{E}\bigg(\bigg(\frac{X_{I_1}^{j_1\prime}X_{I_1}^{j_2}f(X_{I_1}\hat{\beta}_n)}{\pi_{I_1}}^2\bigg)\bigg|\mathcal{F}_n\bigg) - \frac{(H_n^{j_1,j_2})^2}{r}\\
&\leq \frac{1}{n^2r}\sum_{i=1}^{n}\frac{(X_i^{j_1\prime}X_i^{j_2}f(X_i\hat{\beta}_n))^2}{\pi_i}\\
&\leq \frac{1}{16n^2r}\sum_{i=1}^{n}\frac{\|X_i\|^4}{\pi_i}
\end{aligned}$$

Par hypothèse, ce terme multiplié par r est un $O_{\mathbb{P}}(1)$ donc c'est un $O_{\mathbb{P}}(r^-1)$. Au final

$$0 \leq \mathrm{Var}(H_{nj_1,j_2}^*|\mathcal{F}_n) \leq \frac{1}{16n^2r}\sum_{i=1}^{n}\frac{\|X_i\|^4}{\pi_i} = O_{\mathbb{P}}(r^{-1}).$$

Comme $\mathrm{Var}(H_{nj_1,j_2}^*|\mathcal{F}_n) = O_{\mathbb{P}}(r^{-1})$, on peut voir que

$$\sup_n \mathbb{P}(\mathbb{P}(\sqrt{r}\|H_n^* - H_n\| > M|\mathcal{F}_n) > \delta) = \sup_n \mathbb{P}(\mathbb{P}(c\sqrt{r}\sum_{i,j=1}^{p}|H_n^{*i,j} - H_n^{i,j}| > M|\mathcal{F}_n) > \delta)$$

Pour tout terme on a

$$\begin{aligned}
\sup_n \mathbb{P}(\mathbb{P}(c\sqrt{r}|H_n^{*i,j} - H_n^{i,j}| > M|\mathcal{F}_n) > \delta) &\leq \sup_n \mathbb{P}(c^2r\frac{\mathrm{Var}(H_n^{*i,j}|\mathcal{F}_n)}{M^2} > \delta)\\
&= \sup_n \mathbb{P}(c^2r\mathrm{Var}(H_n^{*i,j}|\mathcal{F}_n) > M^2\delta).
\end{aligned}$$

qui tend à zéro lorsque M tend vers l'infini pour ce qu'on vient de voir. On a donc

$$\|H_n^* - H_n\| = c\sum_{i,j=1}^{p}|H_n^{*i,j} - H_n^{i,j}| = \sum_{i,j=1}^{p}O_{\mathbb{P}|\mathcal{F}_n}(r^{-1/2})$$

donc

$$\|H_n^* - H_n\| = O_{\mathbb{P}|\mathcal{F}_n}(r^{-1/2})$$

parce que la somme de $O_{\mathbb{P}|\mathcal{F}_n}$ est un $O_{\mathbb{P}|\mathcal{F}_n}$.On a donc la première propriété.
Pour la deuxième on peut d'abord noter que

$$\mathbb{E}(n^{-1}\nabla\ l_n^*(\hat{\beta}, X, Y)|\mathcal{F}_n) = \frac{1}{n}\mathbb{E}(\nabla\ l_n(\hat{\beta}, X, Y)) = 0$$

Donc on a que la matrice de covariance est telle que

$$\begin{aligned}
\|\mathrm{Cov}(n^{-1}\nabla\ l_n^*(\hat{\beta}, X, Y)|\mathcal{F}_n)\| &= \|\mathbb{E}((n^{-1}\nabla\ l_n^*(\hat{\beta}, X, Y))^2|\mathcal{F}_n)\|\\
&= \left\|\frac{1}{n^2r}\sum_{i=1}^{n}\frac{(Y_i - F(X_i'\hat{\beta}))^2X_i'X_i}{\pi_i}\right\|\\
&\leq \frac{1}{n^2r}\sum_{i=1}^{n}\left\|\frac{(Y_i - F(X_i'\hat{\beta}))^2X_i'X_i}{\pi_i}\right\|\\
&\leq \frac{1}{n^2r}\sum_{i=1}^{n}\left\|\frac{X_i'X_i}{\pi_i}\right\|\\
&= \frac{1}{n^2r}\sum_{i=1}^{n}\frac{\|X_i\|^2}{\pi_i}\\
&= r^{-1}O_{\mathbb{P}}(1) = O_{\mathbb{P}}(r^{-1}).
\end{aligned}$$

Toujours par Markov (en appelant $D = (D_1, ..., D_p)$ la matrice $n^{-1}\nabla\ l_n^*(\hat{\beta}, X, Y)$)

$$\begin{aligned}
\sup_n \mathbb{P}(\mathbb{P}(r^{1/2}\|D\| > M|\mathcal{F}_n) > \delta) &= \sup_n \mathbb{P}(\mathbb{P}(r^{1/2}\|D\|^2 > M|\mathcal{F}_n) > \delta) \\
&= \sup_n \mathbb{P}(\mathbb{P}(r^1\|D\|^2 > M^2|\mathcal{F}_n) > \delta) \\
&= \sup_n \mathbb{P}(\mathbb{P}(r\sum_{i=1}^p D_i^2 > M^2|\mathcal{F}_n) > \delta) \\
&\leq \sup_n \mathbb{P}\left(r\sum_{i=1}^p \frac{\mathbb{E}(D_i^2|\mathcal{F}_n)}{M^2} > \delta\right) \\
&= \sup_n \mathbb{P}\left(r\sum_{i=1}^p \mathbb{E}(D_i^2|\mathcal{F}_n) > \delta M^2\right) \\
&\leq \sup_n \mathbb{P}\left(cr\|\mathrm{Cov}(D)\||\mathcal{F}_n) > \delta M^2\right).
\end{aligned}$$

Puisque on a montré $\mathrm{Cov}(n^{-1}\nabla\ l_n^*(\hat{\beta}, X, Y) = O_{\mathbb{P}}(r^{-1})$, alors le dernier terme tend vers 0 lorsque M tend vers l'infini et on a démontré que

$$n^{-1}\nabla\ l_n^*(\hat{\beta}, X, Y) = O_{\mathbb{P}|\mathcal{F}_n}(r^{-1/2}).$$

Ces résultats en poche, attaquons la preuve du théorème. On note $t_i(\beta) = y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$. On a

$$\begin{aligned}
\mathbb{E}[l_n^*(\beta, X, Y)^2|\mathcal{F}_n] &= \frac{1}{r^2}\mathbb{E}[\sum_{j,k=1}^r \frac{t_{I_j}(\beta)}{\pi_{I_j}}\frac{t_{I_k}(\beta)}{\pi_{I_k}}|\mathcal{F}_n] \\
&= \frac{1}{r^2}\sum_{j,k=1}^r \mathbb{E}[\frac{t_{I_j}(\beta)}{\pi_{I_j}}\frac{t_{I_k}(\beta)}{\pi_{I_k}}|\mathcal{F}_n] \\
&= \frac{1}{r^2}\sum_{i=1}^r \mathbb{E}[(\frac{t_{I_j}(\beta)}{\pi_{I_j}})^2|\mathcal{F}_n] + \sum_{j\neq k}\mathbb{E}[\frac{t_{I_j}(\beta)}{\pi_{I_j}}|\mathcal{F}_n]\mathbb{E}[\frac{t_{I_k}(\beta)}{\pi_{I_k}}|\mathcal{F}_n] \\
&= \frac{1}{r}\mathbb{E}[(\frac{t_{I_1}(\beta)}{\pi_{I_1}})^2] + \frac{r^2 - r}{r^2}(\mathbb{E}[(\frac{t_{I_1}(\beta)}{\pi_{I_1}})])^2 \\
&= \frac{1}{r}\sum_{i=1}^n \frac{t_i(\beta)^2}{\pi_i} + (\sum_{i=1}^n \frac{t_i(\beta)}{\pi_i}\pi_i)^2 - \frac{1}{r}(\sum_{i=1}^n \frac{t_i(\beta)}{\pi_i}\pi_i)^2 \\
&= \frac{1}{r}\sum_{i=1}^n \frac{t_i(\beta)^2}{\pi_i} + l_n(\beta, X, Y)^2 - \frac{1}{r}(\sum_{i=1}^n t_i(\beta))^2
\end{aligned}$$

Comme $\mathbb{E}[l_n^*(\beta, X, Y)|\mathcal{F}_n] = l_n(\beta, X, Y)$, on a

$$\mathbb{E}[(\frac{l_n^*(\beta, X, Y)}{n} - \frac{l_n(\beta, X, Y)}{n})^2|\mathcal{F}_n] = \frac{1}{r}[\frac{1}{n^2}\sum_{i=1}^n \frac{t_i(\beta)^2}{\pi_i} - (\frac{1}{n}\sum_{i=1}^n t_i(\beta))^2].$$

En remarquant que

$$\begin{aligned}
|t_i(\beta)| &= \begin{cases} |\log(\frac{e^{X_i\beta}}{1+e^{X_i\beta}})| & \text{si } Y_i = 1 \\ |\log(\frac{1}{1+e^{X_i\beta}})| & \text{si } Y_i = 0 \end{cases} \\
&= \begin{cases} \log(1 + e^{-X_i\beta}) & \text{si } Y_i = 1 \\ \log(1 + e^{X_i\beta}) & \text{si } Y_i = 0 \end{cases} \\
&\leq \log(1 + e^{|X_i\beta|}) \\
&\leq \log(2e^{|X_i\beta|}) \\
&\leq \log(2) + \|X_i\|\|\beta\|,
\end{aligned}$$

on a

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}\frac{t_i(\beta)^2}{\pi_i}-(\frac{1}{n}\sum_{i=1}^{n}t_i(\beta))^2 &\leq \frac{1}{n^2}\sum_{i=1}^{n}\frac{t_i(\beta)^2}{\pi_i}\\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\frac{(\log(2)+\|X_i\|\|\beta\|)^2}{\pi_i}\\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\frac{2\log(2)^2+2\|X_i\|^2\|\beta\|^2}{\pi_i}\\ &= 2\frac{\log(2)^2}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}+2\frac{\|\beta\|^2}{n^2}\sum_{i=1}^{n}\frac{\|X_i\|^2}{\pi_i}\end{aligned}$$

Ces deux termes sont des $O_{\mathbb{P}}(1)$ par hypothèse. Donc $\mathbb{E}[(\frac{l_n^*(\beta,X,Y)}{n}-\frac{l_n(\beta,X,Y)}{n})^2|\mathcal{F}_n]=O_{\mathbb{P}}(r^{-1})$. En suivant la procédure qu'on a utilisé précédemment, on a pour tout $\delta>0$

$$\sup_n \mathbb{P}(\mathbb{P}(\sqrt{r}(\frac{l_n^*(\beta,X,Y)}{n}-\frac{l_n(\beta,X,Y)}{n})>M|\mathcal{F}_n)>\delta)\leq \mathbb{P}(r\mathbb{E}[(\frac{l_n^*(\beta,X,Y)}{n}-\frac{l_n(\beta,X,Y)}{n})^2|\mathcal{F}_n]>M^2\delta)$$
$$\underset{M\to+\infty}{\longrightarrow} 0$$

Donc $\frac{l_n^*(\beta,X,Y)}{n}-\frac{l_n(\beta,X,Y)}{n}=O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$ et c'est donc un $o_{\mathbb{P}|\mathcal{F}_n}(1)$ et, donc, aussi un $o_{\mathbb{P}}(1)$.

L'auteur de l'article conclut la consistance juste en citant le théorème 5.7 du livre Asymptotic Statistics de Van der Vaart [6].

Attention!!! En fait l'application du théorème de Van Der Waart dans ce cas ne marche pas et dans l'article étudié il n'y a pas les outils nécessaires pour démontrer ce résultat.
On donne dans les prochaines lignes une idée de comment on aurait appliqué le théorème de Van der Vaart. Au lecteur le défi de comprendre si et comment est il possible d'appliquer ce théorème.

Pour pouvoir montrer la consistance de l'estimateur de $\tilde{\beta}$ à $\hat{\beta}$ il faut avoir

1. $\sup_{\beta\in B}\|n^{-1}l_n(\beta)-n^{-1}l_n^*(\beta)\|=o_{\mathbb{P}}(1)$.
2. Il existe $\eta>0$ t.q. pour tout n : $\sup_{\beta\in\mathbb{R}^p:\|\beta-\hat{\beta}\|>\epsilon} n^{-1}l_n(\beta,x,y)<n^{-1}l_n(\hat{\beta}(x,y),x,y)-\eta$.

En effet, en regardant la preuve du théorème 5.7 de Van der Vaart [6], on peut remarquer que

$$\begin{aligned}0&<n^{-1}l_n(\hat{\beta})-n^{-1}l_n(\tilde{\beta})\\ &=n^{-1}l_n(\hat{\beta})-n^{-1}l_n^*(\tilde{\beta})+n^{-1}l_n^*(\tilde{\beta})-n^{-1}l^*(\hat{\beta})+n^{-1}l_n^*(\hat{\beta})-n^{-1}l_n(\tilde{\beta})\\ &\leq 2\sup_{\beta\in B}\|n^{-1}l_n(\beta)-n^{-1}l_n^*(\beta)\|+n^{-1}l_n^*(\hat{\beta})-n^{-1}l_n(\tilde{\beta})\leq 2\sup_{\beta\in B}\|n^{-1}l_n(\beta)-n^{-1}l_n^*(\beta)\|.\end{aligned}$$

Donc

$$\begin{aligned}0&\leq \mathbb{P}(\|\tilde{\beta}-\hat{\beta}\|>\epsilon|\mathcal{F}_n)\\ &\leq \mathbb{P}(n^{-1}l(\hat{\beta},X,Y)-n^{-1}l(\tilde{\beta},X,Y)>\eta|\mathcal{F}_n)\\ &\leq \mathbb{P}(2\sup_{\beta\in B}\|n^{-1}l_n(\beta)-n^{-1}l_n^*(\beta)\|>\eta|\mathcal{F}_n)\\ &=o_{\mathbb{P}}(1).\end{aligned}$$

Dont on peut déduire la consistance

$$\|\tilde{\beta}-\hat{\beta}\|=o_{\mathbb{P}|\mathcal{F}_n}(1).$$

Malheureusement on n'a pas été capable de démontrer les hypothèses. Par rapport à la première hypothèse on ne sait pas comment passer de la convergence ponctuelle de $n^{-1}l_n(\beta)-n^{-1}l_n^*(\beta)$ à la convergence uniforme

sur un compact.
Pour la deuxième hypothèse aussi on ne sait pas si et comment démontrer l'existence d'un tel η. On est capable de démontrer l'existence d'un η qui fait un travail semblable, mais notre η dépend de l'échantillon et de n. En effet $\hat{\beta}$ est unique (grâce au théorème 3.2) et la matrice $\mathcal{H}_n l_n(\beta, X, Y)$ est définie négative pour tout $\beta \in \mathbb{R}^p$ (hypothèse du théorème 3.2). Ces deux propriété de l_n permettent d'écrire

$$\sup_{\beta \in \mathbb{R}^p : \|\beta - \hat{\beta}\| > \epsilon} n^{-1} l_n(\beta, x, y) < n^{-1} l_n(\hat{\beta}(x,y), x, y)$$

c'est à dire il existe $\eta(x, y, n)$

$$\sup_{\beta \in \mathbb{R}^p : \|\beta - \hat{\beta}\| > \epsilon} n^{-1} l_n(\beta, x, y) \leq n^{-1} l_n(\hat{\beta}(x,y), x, y) - \eta(x, y, n).$$

D'ici on reprend la preuve où on l'avait laissé et la parenthèse sur le théorème de Van der Vaart peut être considérée conclue. Á partir d'ici on va considérer la consistance comme acquise.
Le développement de Taylor de $\partial_j l_n^*(\beta, X, Y)$ au voisinage de $\hat{\beta}$ donne :

$$0 = \frac{\partial_j l_n^*(\hat{\beta}, X, Y)}{n} + \frac{1}{n} \nabla \partial_j l_n^*(\hat{\beta}, X, Y)(\tilde{\beta} - \hat{\beta}) + \frac{1}{2n}(\tilde{\beta} - \hat{\beta})' \int_0^1 (1-t) \mathcal{H} \partial_j l_n^*(\hat{\beta} + t(\tilde{\beta} - \hat{\beta}), X, Y) dt (\tilde{\beta} - \hat{\beta})$$

Des calculs de gradients donnent que

$$\begin{aligned} \|\mathcal{H} \partial_j l_n^*(\beta, X, Y)\| &= \|\frac{1}{r} \sum_{i=1}^{r} \frac{p_{I_i}(1-p_{I_i})(1-2p_{I_i})}{\pi_{I_i}} X_{I_i}^j X_{I_i} X_{I_i}'\| \\ &\leq \frac{1}{r} \sum_{i=1}^{r} \frac{\|X_{I_i}\|^3}{\pi_{I_i}} \end{aligned}$$

Or pour tout $\delta > 0$ et $M > 0$

$$\begin{aligned} \sup_n \mathbb{P}(\mathbb{P}(\frac{1}{nr} \sum_{i=1}^{r} \frac{\|X_{I_i}\|^3}{\pi_{I_i}} > M | \mathcal{F}_n) \geq \delta) &\leq \sup_n \mathbb{P}(\mathbb{E}[\sum_{i=1}^{r} \frac{\|X_{I_i}\|^3}{\pi_{I_i}} | \mathcal{F}_n] \geq M\delta) \\ &= \sup_n \mathbb{P}(\frac{1}{n} \sum_{i=1}^{n} \|X_i\|^3 \geq M\delta) \\ &\underset{M \to \infty}{\longrightarrow} 0 \end{aligned}$$

Donc $\frac{1}{nr} \sum_{i=1}^{r} \frac{\|X_{I_i}\|^3}{\pi_{I_i}} = O_{\mathbb{P}|\mathcal{F}_n}(1)$.Donc

$$\begin{aligned} &\|\frac{1}{2n}(\tilde{\beta} - \hat{\beta})' \int_0^1 (1-t) \mathcal{H} \partial_j l_n^*(\hat{\beta} + t(\tilde{\beta} - \hat{\beta}), X, Y) dt (\tilde{\beta} - \hat{\beta})\| \\ &\leq \|\tilde{\beta} - \hat{\beta}\|^2 \frac{1}{2nr} \sum_{i=1}^{r} \frac{\|X_{I_i}\|^3}{\pi_{I_i}} \\ &= O_{\mathbb{P}|\mathcal{F}_n}(\|\tilde{\beta} - \hat{\beta}\|^2) \end{aligned}$$

L'écriture matricielle du développement de Taylor donne que

$$\tilde{\beta} - \hat{\beta} = H_n^{*-1}(\frac{\nabla l_n^*(\hat{\beta}, X, Y)}{n} + O_{\mathbb{P}|\mathcal{F}_n}(\|\tilde{\beta} - \hat{\beta}\|^2)$$

car un vecteur composé de $O_{\mathbb{P}|\mathcal{F}_n}(\|\tilde{\beta} - \hat{\beta}\|^2)$ est lui même un $O_{\mathbb{P}|\mathcal{F}_n}(\|\tilde{\beta} - \hat{\beta}\|^2)$
De plus, on a par le lemme 1 que $H_n^* - H_n = O_{\mathbb{P}|\mathcal{F}_n}(r^{-1/2})$ donc $H_n^* - H_n = o_{\mathbb{P}}(1)$. De plus $H_n \xrightarrow{\mathbb{P}} H$ par hypothèse donc on a $H_n^* \xrightarrow{\mathbb{P}} H$. Comme H est définie positive (donc inversible) et que l'inversion matricielle est continue, on a $H_n^{*-1} \xrightarrow{\mathbb{P}} H^{-1}$ et finalement $H_n^{*-1} = O_{\mathbb{P}|\mathcal{F}_n}(1)$.
De plus $\frac{\nabla l^*(\hat{\beta}, X, Y)}{n} = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$ et $\|\tilde{\beta} - \hat{\beta}\| = o_{\mathbb{P}|\mathcal{F}_n}(1)$. Au final on a

$$\tilde{\beta} - \hat{\beta} = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}}) + o_{\mathbb{P}|\mathcal{F}_n}(\|\tilde{\beta} - \hat{\beta}\|)$$

Le deuxième terme vient du fait que

$$O_{\mathbb{P}}(\|\tilde{\beta} - \hat{\beta}\|^2) = \|\tilde{\beta} - \hat{\beta}\|\|\tilde{\beta} - \hat{\beta}\|O_{\mathbb{P}}(1) = \|\tilde{\beta} - \hat{\beta}\|o_{\mathbb{P}}(1)O_{\mathbb{P}}(1) = \|\tilde{\beta} - \hat{\beta}\|o_{\mathbb{P}}(1) = o_{\mathbb{P}}(\|\tilde{\beta} - \hat{\beta}\|)$$

Ce qui permet de conclure que (? ? ?)

$$\tilde{\beta} - \hat{\beta} = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$$

□

Théorème 5.2.
On a

$$\frac{\nabla l_n^*(\hat{\beta}, X, Y)}{n} = \frac{1}{r}\sum_{i=1}^{r} \frac{(Y_{I_i} - p_{I_i}(\hat{\beta}))X_{I_i}}{n\pi_{I_i}}$$

On note $\eta_i := \frac{(Y_{I_i} - p_{I_i}(\hat{\beta}))X_{I_i}}{n\pi_{I_i}}$. Sachant $\mathcal{F}_n$, les $(\eta_i)_{i=1}^n$ sont i.i.d de moyenne et de variance

$$\begin{aligned}
\mathbb{E}[\eta_1|\mathcal{F}_n] &= \mathbb{E}[\frac{(Y_{I_1} - p_{I_1}(\hat{\beta}))X_{I_1}}{n\pi_{I_1}}|\mathcal{F}_n] \\
&= \sum_{i=1}^{n} \frac{(Y_i - p_i(\hat{\beta}))X_i}{n\pi_i}\pi_i \\
&= \frac{\nabla l_n(\hat{\beta}, X, Y)}{n} \\
&= 0 \\
\mathrm{Cov}[\eta_1|\mathcal{F}_n] &= \mathbb{E}[\eta_1'\eta_1|\mathcal{F}_n] \\
&= \mathbb{E}[\frac{(Y_{I_1} - p_{I_1}(\hat{\beta}))^2 X_{I_1}' X_{I_1}}{n^2\pi_{I_1}^2}|\mathcal{F}_n] \\
&= \sum_{i=1}^{n} \pi_i \frac{(Y_i - p_i(\hat{\beta}))^2 X_i' X_i}{n^2\pi_i^2} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \frac{(Y_i - p_i(\hat{\beta}))^2 X_i' X_i}{\pi_i} \\
&= O_{\mathbb{P}}(1) \qquad \text{(Par hypothèse)}
\end{aligned}$$

De plus, pour tout $\epsilon > 0$, on a

$$\begin{aligned}
&\sum_{i=1}^{r} \mathbb{E}[\|r^{-\frac{1}{2}}\eta_i\|^2 \mathbb{1}_{\|\eta_i\| > r^{\frac{1}{2}}\epsilon}|\mathcal{F}_n] \\
\leq &\sum_{i=1}^{r} \frac{1}{r^{1+\frac{\delta}{2}}\epsilon^\delta} E[\|\eta_i\|^{2+\delta} \mathbb{1}_{\|\eta_i\| > r^{\frac{1}{2}}\epsilon}|\mathcal{F}_n] \\
\leq &\sum_{i=1}^{r} \frac{1}{r^{1+\frac{\delta}{2}}\epsilon^\delta} \mathbb{E}[\frac{|Y_{I_i} - F(X_{I_i}\hat{\beta})|^{2+\delta}}{n^{2+\delta}\pi_{I_i}^{2+\delta}}\|X_{I_i}\|^{2+\delta}|\mathcal{F}_n] \\
\leq &\sum_{i=1}^{r} \frac{1}{r^{1+\frac{\delta}{2}}\epsilon^\delta} \sum_{j=1}^{n} \frac{|Y_j - F(X_j\hat{\beta})|^{2+\delta}}{n^{2+\delta}\pi_j^{2+\delta}}\|X_j\|^{2+\delta}\pi_j \\
= &\frac{1}{r^{\frac{\delta}{2}}\epsilon^\delta} \sum_{j=1}^{n} \frac{|Y_j - F(X_j\hat{\beta})|^{2+\delta}}{n^{2+\delta}\pi_j^{1+\delta}}\|X_j\|^{2+\delta} \\
\leq &\frac{1}{r^{\frac{\delta}{2}}}\frac{1}{\epsilon^\delta}(\frac{1}{n^{2+\delta}}\sum_{j=1}^{n} \pi_j^{-(1+\delta)}\|X_j\|^{2+\delta}) \\
= &\frac{1}{r^{\frac{\delta}{2}}}\frac{1}{\epsilon^\delta} O_{\mathbb{P}}(1) \\
= &o_{\mathbb{P}}(1)
\end{aligned}$$

De plus, nous avons fait l'hypothèse que $\sum_{i=1}^{r} \text{Cov}[r^{-\frac{1}{2}}\eta_1|\mathcal{F}_n] = \text{Cov}[\eta_1|\mathcal{F}_n] \xrightarrow[n\to\infty]{\mathbb{P}} B$, on a par le théorème de Lindeberg-Feller conditionnel que

$$\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\eta_i \xrightarrow[n\to\infty]{d|\mathcal{F}_n} \mathcal{N}(0,B)$$

ou par Slutsky conditionnel que

$$\frac{1}{\sqrt{r}}(\text{Cov}[\eta_1|\mathcal{F}_n])^{-\frac{1}{2}}\sum_{i=1}^{r}\eta_i \xrightarrow[n\to\infty]{d|\mathcal{F}_n} \mathcal{N}(0,I)$$

Dans la preuve du lemme 1, nous avons prouvé que

$$\tilde{\beta}-\hat{\beta} = \frac{1}{n}H_n^{*-1}\nabla l_n^*(\hat{\beta},X,Y) + O_{\mathbb{P}|\mathcal{F}_n}(r^{-1})$$

De plus, en utilisant le Lemme 1, on a

$$H_n^{-1} - H_n^{*-1} = H_n^{-1}(H_n^* - H_n)H_n^{*-1} = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$$

car $H_n^{-1} \xrightarrow{\mathbb{P}} H^{-1}$ et $H_n^{*-1} \xrightarrow{\mathbb{P}} H^{-1}$
On rappelle que $V_c = \frac{1}{r}\text{Cov}[\eta_1|\mathcal{F}_n]$, on a donc

$$rV = rH_n^{-1}V_cH_n^{-1} = H_n^{-1}(rV_c)H_n^{-1} \xrightarrow[n\to\infty]{\mathbb{P}} H^{-1}BH^{-1}$$

Par le theorème de continuité on a donc

$$r^{-\frac{1}{2}}V^{-\frac{1}{2}} \xrightarrow[n\to\infty]{\mathbb{P}} H^{\frac{1}{2}}B^{-\frac{1}{2}}H^{\frac{1}{2}}$$

et donc $r^{-\frac{1}{2}}V^{-\frac{1}{2}} = O_{\mathbb{P}|\mathcal{F}_n}(1)$ et ainsi $V^{-\frac{1}{2}} = O_{\mathbb{P}|\mathcal{F}_n}(r^{\frac{1}{2}})$. On a donc

$$\begin{aligned}
V^{-\frac{1}{2}}(\tilde{\beta}-\hat{\beta}) &= V^{-\frac{1}{2}}(\frac{1}{n}H_n^{*-1}\nabla l_n^*(\hat{\beta},X,Y) + O_{\mathbb{P}|\mathcal{F}_n}(r^{-1})) \\
&= V^{-\frac{1}{2}}\frac{1}{n}H_n^{*-1}\nabla l_n^*(\hat{\beta},X,Y) + O_{\mathbb{P}|\mathcal{F}_n}(r^{\frac{1}{2}})O_{\mathbb{P}|\mathcal{F}_n}(r^{-1}) \\
&= V^{-\frac{1}{2}}{H_n}^{-1}\frac{1}{n}\nabla l_n^*(\hat{\beta},X,Y) + V^{-\frac{1}{2}}(H_n^{*-1} - {H_n}^{-1})\frac{1}{n}\nabla l_n^*(\hat{\beta},X,Y) + O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})
\end{aligned}$$

Comme $H_n^{*-1} - {H_n}^{-1} = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$ et $\frac{1}{n}\nabla l_n^*(\hat{\beta},X,Y) = O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$

$$V^{-\frac{1}{2}}(\tilde{\beta}-\hat{\beta}) = V^{-\frac{1}{2}}{H_n}^{-1}V_c^{\frac{1}{2}}V_c^{-\frac{1}{2}}\frac{1}{n}\nabla l_n^*(\hat{\beta},X,Y) + O_{\mathbb{P}|\mathcal{F}_n}(r^{-\frac{1}{2}})$$

Or

- $V^{-\frac{1}{2}}{H_n}^{-1}V_c^{\frac{1}{2}}V_c^{-\frac{1}{2}} = (rV)^{-\frac{1}{2}}{H_n}^{-1}(rV_c)^{\frac{1}{2}} \xrightarrow[n\to\infty]{\mathbb{P}} H^{\frac{1}{2}}B^{-\frac{1}{2}}H^{\frac{1}{2}}H^{-1}B^{\frac{1}{2}}$
- $V^{-\frac{1}{2}}{H_n}^{-1}V_c^{\frac{1}{2}}(V^{-\frac{1}{2}}{H_n}^{-1}V_c^{\frac{1}{2}})' = V^{-\frac{1}{2}}{H_n}^{-1}V_c^{\frac{1}{2}}V_c^{\frac{1}{2}}{H_n}^{-1}V^{-\frac{1}{2}} = V^{-\frac{1}{2}}{H_n}^{-1}V_c{H_n}^{-1}V^{-\frac{1}{2}} = V^{-\frac{1}{2}}VV^{-\frac{1}{2}} = I$
- $\frac{1}{n}V_c^{-\frac{1}{2}}\nabla l_n^*(\hat{\beta},X,Y) \xrightarrow[n\to\infty]{d|\mathcal{F}_n} \mathcal{N}(0,I)$

Par Slutsky conditionnel on a

$$V^{-\frac{1}{2}}(\tilde{\beta}-\hat{\beta}) \xrightarrow[n\to\infty]{d|\mathcal{F}_n} \mathcal{N}(0,I)$$

□

Références

[1] A. Albert et J. A. Anderson. « On the existence of maximum likelihood estimates in logistic regression models ». en. In : *Biometrika* 71.1 (1984), p. 1-10. issn : 0006-3444, 1464-3510. doi : 10.1093/biomet/71.1.1. url : https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/71.1.1 (visité le 10/02/2024).

[2] Joseph Berkson. « Application of the Logistic Function to Bio-Assay ». In : *Journal of the American Statistical Association* 39.227 (1944), p. 357-365. issn : 01621459. url : http://www.jstor.org/stable/2280041 (visité le 24/05/2024).

[3] H. Cartan. *Calcul différentiel*. Calcul différentiel vol. 2. Hermann, 1967. url : https://books.google.fr/books?id=YmEZAQAAIAAJ.

[4] Christian Gourieroux et Alain Monfort. « Asymptotic properties of the maximum likelihood estimator in dichotomous logit models ». en. In : *Journal of Econometrics* 17.1 (sept. 1981), p. 83-97. issn : 03044076. doi : 10.1016/0304-4076(81)90060-9. url : https://linkinghub.elsevier.com/retrieve/pii/0304407681900609 (visité le 22/02/2024).

[5] C. Radhakrishna Rao et C. Radhakrishna Rao, éd. *Linear models and generalizations : least squares and alternatives*. 3rd extended ed. Springer series in statistics. OCLC : ocn173807301. Berlin ; New York : Springer, 2008. isbn : 9783540742265.

[6] A. W. van der Vaart. *Asymptotic Statistics*. Anglais. Cambridge : Cambridge University Press, 2000.

[7] H.Y. Wang, R. Zhu et P. Ma. « Optimal Subsampling for Large Sample Logistic Regression ». In : *Journal of the American Statistical Association* 113.522 (2018), p. 829-844.

[8] Shifeng Xiong et Guoying Li. « Some results on the convergence of conditional distributions ». In : *Statistics & Probability Letters* 78.18 (déc. 2008), p. 3249-3253. issn : 0167-7152. doi : 10.1016/j.spl.2008.06.026. url : https://www.sciencedirect.com/science/article/pii/S0167715208003088 (visité le 24/03/2024).