Statistique et Apprentissage

Catherine Matias

2025-2026

Ce document contient les notes du cours Statistique et Apprentissage donné dans le cadre de la spécialité Probabilités et Modèles Aléatoires du master Mathématiques et Applications de Sorbonne Université. Son contenu utilise tout ou partie des documents suivants :

- G. Biau (2014). *Statistique et Apprentissage*; première version de ce cours, reprise et complétée par Lorenzo Zambotti puis Irina Kourkova.
- F. Bach (2024). *Learning Theory from First Principles*, https://www.di.ens.fr/~fbach/ltfp_book.pdf.
- L. Devroye, L. Györfi et G. Lugosi (1996). *A probabilistic theory of pattern recognition*, Springer, New York.
- A. van der Vaart (1998). *Asymptotic statistics*, Cambridge University Press.

Ce cours est constitué de plusieurs parties essentiellement indépendantes : la première partie contient les Chapitres 1, 2, 3 et porte sur la classification supervisée et la régression non paramétrique. Le Chapitre 4 est une introduction au clustering (ou classification non supervisée). Le Chapitre 5 contient des rappels de statistique paramétrique et étend ces notions par des considérations de statistique asymptotique (i.e. quand la taille d'échantillon grandit). Enfin le Chapitre 6 porte sur l'estimation (paramétrique) par moindres carrés.

Table des matières

I Apprentissage supervisé **5**

1 Introduction à l'apprentissage supervisé **6**
- 1.1 Objectifs 6
- 1.2 Fonction de perte et risque 7
- 1.3 L'apprentissage et la minimisation du risque empirique . . . 11
- 1.4 Cas de la classification binaire et d'une classe de cardinal fini 15

2 Théorie de Vapnik-Chervonenkis pour la classification **18**
- 2.1 Passage du $\sup_{g \in \mathscr{G}}$ au $\sup_{A \in \mathscr{A}}$ 18
- 2.2 Théorème de Vapnik-Chervonenkis 20
- 2.3 Aspects combinatoires 29
- 2.4 Application à la minimisation du risque empirique 31

3 Théorème de Stone et plus proches voisins **36**
- 3.1 Liens entre classification et régression 37
- 3.2 Le théorème de Stone 39
- 3.3 Estimateur de Nadaraya-Watson pour la régression 46
- 3.4 k-plus proches voisins pour la classification 50

II Introduction au clustering **57**

4 Quantification et clustering **58**
- 4.1 Principe de la quantification 58
- 4.2 Quantification empirique et clustering 62
- 4.3 Consistance et vitesse 66

III Statistique paramétrique **78**

5 Statistique paramétrique asymptotique **79**
5.1 Rappels sur les estimateurs 79
5.2 M- et Z-estimateurs 88
5.3 Théorème de Wald 89
5.4 Vitesse de convergence et loi limite 93
5.5 Tests asymptotiques 98

6 Échantillons gaussiens et modèle linéaire **105**
6.1 Notations 105
6.2 Rappels sur les vecteurs gaussiens 106
6.3 Théorème de Cochran 108
6.4 Échantillons gaussiens 110
6.5 Régression linéaire des moindres carrés 115

Première partie

Apprentissage supervisé

Chapitre 1

Introduction à l'apprentissage supervisé

1.1 Objectifs

On considère un couple (X, Y) de variables aléatoires à valeurs dans $\mathbb{R}^d \times \mathcal{Y}$, avec $\mathcal{Y} = \{0, 1\}$ (classification binaire) ou $\mathcal{Y} = \{1, 2, \ldots, M\}$ (classification multi-classes) ou encore $\mathcal{Y} = \mathbb{R}$ ou un sous-ensemble borné de $\mathbb{R}$ (régression). La variable X est appelée variable explicative et la variable Y est appelée label, classe ou étiquette (dans la cas de la classification) ou encore réponse ou variable à prédire (dans le cas de la régression). L'apprentissage supervisé, qu'il s'agisse de classification supervisée (binaire ou multi-classes) ou de régression, consiste à prédire au mieux Y à partir de X, c'est-à-dire à construire une fonction borélienne $g : \mathbb{R}^d \to \mathcal{Y}$ qui, à un x donné (réalisation de X) associe une valeur $y \in \mathcal{Y}$ qui correspond à son label supposé (cas $\mathcal{Y}$ discret) ou à sa réponse (cas continu réél). Pour prendre un exemple en classification binaire, on peut penser à X comme un vecteur de variables aléatoires représentant les fréquences d'un certain nombre de mots-clés dans un email, et à Y comme la variable associée exprimant le fait que l'email est sain (label 0) ou bien spam (label 1). Dans le contexte de la régression, X peut-être la dose d'insuline injectée à un patient diabétique et Y le taux de glucose dans le sang de ce patient après 30 minutes. La fonction g s'appelle un prédicteur ou règle de décision.

Dans la suite, la loi du couple (X, Y) sera notée ν, tandis que la marginale en X est notée μ (i.e. $A \in \mathscr{B}(\mathbb{R}^d), \mu(A) = \mathbb{P}(X \in A)$) et r est la fonction de

régression de Y sur X, définie par

$$r(x) = \mathbb{E}(Y|X = x) = \int_{\mathcal{Y}} y \mathrm{d}\nu(x, y).$$

C'est donc l'espérance conditionnelle de la variable à prédire, sachant la variable explicative. On peut noter que dans le cas de la classification binaire, la loi du couple (X, Y) est entièrement caractérisée par le couple (μ, r) et $r(x) = \mathbb{P}(Y = 1|X = x)$ puisque dans ce cas Y prend ses valeurs dans $\{0, 1\}$.

Attention ! Dans ce modèle, Y n'est pas nécessairement lié à X de manière fonctionnelle, i.e. rien ne dit qu'il existe une fonction φ telle que $Y = \varphi(X)$. Pour s'en convaincre, il suffit de penser à l'exemple des emails, au sein duquel le mot « livraison » peut être associé à un label spam ou non. De même, le taux du glucose d'un patient n'est pas exactement déterminé par la dose d'insuline absorbée. La modélisation nous dit seulement que nous envisageons Y comme une fonction bruitée (aléatoire) de X.

Bien entendu, n'importe quelle fonction borélienne $g : \mathbb{R}^d \to \mathcal{Y}$ fournit un prédicteur et il est donc nécessaire d'adjoindre un critère de qualité à chaque décision.

1.2 Fonction de perte et risque

On se donne donc une **fonction de perte** (ou de coût) $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ (ou le plus souvent $\mathbb{R}_+$), où $\ell(y, z)$ mesure l'erreur (la perte) lorsque l'on prédit z tandis que la vraie valeur est y.

Exemples.

1. Dans le cas de la classification binaire, $\mathcal{Y} = \{0, 1\}$, une erreur de classification se produit lorsque $z \neq y$. On définit naturellement la fonction de coût $\ell(y, z) = \mathbb{1}_{y \neq z}$. Cette fonction s'appelle perte 0-1.
2. En classification multi-classes, on peut également utiliser la perte 0-1, i.e. $\ell(y, z) = \mathbb{1}_{y \neq z}$. De façon parfaitement équivalente, en utilisant la notation $\mathcal{Y} = \{0, 1\}^M$ (plutôt que $\{1, \ldots, M\}$) on définit le coût de Hamming $\ell(y, z) = \sum_{j=1}^{M} \mathbb{1}_{y_j \neq z_j}$.
3. En régression, $\mathcal{Y} = \mathbb{R}$ et $\ell(y, z) = (y - z)^2$ est le coût quadratique tandis que $\ell(y, z) = |y - z|$ est le coût absolu.

À partir d'une fonction de coût ℓ, on introduit le **risque attendu** (ou erreur de généralisation) d'une fonction $g : \mathbb{R}^d \to \mathcal{Y}$ défini par

$$\mathcal{R}(g) = \mathbb{E}\ell(Y, g(X)) = \int_{\mathbb{R}^d \times \mathcal{Y}} \ell(y, g(x)) \mathrm{d}\nu(x, y). \tag{1.1}$$

Cette quantité dépend de la loi inconnue ν du couple (X, Y).

Exemples

1. En classification binaire, avec la perte 0-1, on obtient $\mathcal{R}(g) = \mathbb{P}(Y \neq g(X))$. En classification multi-classes avec le coût de Hamming, on a de même $\mathcal{R}(g) = \mathbb{P}(Y \neq g(X))$.
2. En régression avec le coût quadratique, $\ell(y, z) = (y - z)^2$, on obtient $\mathcal{R}(g) = \mathbb{E}(Y - g(X))^2$ qui est l'erreur quadratique moyenne (MSE pour mean-squared error en anglais).

Dans la suite, la fonction de perte (et donc le risque attendu) sont fixés. Pour la classification, il s'agit de la perte 0-1 et pour la régression, de la perte quadratique.

La quantité $\mathcal{R}(g)$, qui mesure la pertinence de la règle g, permet donc de hiérarchiser les fonctions de décision agissant sur le couple (X, Y). Il est alors légitime de se poser la question de l'existence éventuelle d'une règle meilleure que les autres. Ce champion existe et s'appelle le **prédicteur de Bayes**. Pour l'introduire, il nous faut la notion de **risque conditionnel**. Pour tous $(x, z) \in \mathbb{R}^d \times \mathcal{Y}$ on note

$$r(z|x) = \mathbb{E}\big(\ell(Y, z)|X = x\big)$$

et on remarque que

$$\mathcal{R}(g) = \mathbb{E}\mathbb{E}\big(\ell(Y, g(X))|X\big) = \int_{\mathbb{R}^d} r(g(x)|x) \mathrm{d}\mu(x).$$

Proposition 1 (Risque et prédicteur de Bayes)**.** *Le risque attendu $\mathcal{R}$ est minimum pour un prédicteur de Bayes $g^\star : \mathbb{R}^d \to \mathcal{Y}$ qui satisfait*

$$g^\star(x) \in \underset{z \in \mathcal{Y}}{Argmin}\, \mathbb{E}\big(\ell(Y, z)|X = x\big) = \underset{z \in \mathcal{Y}}{Argmin}\, r(z|x), \quad \forall x \in R^d. \tag{1.2}$$

Le risque de Bayes $\mathcal{R}^\star$ est le risque de n'importe quel prédicteur de Bayes et vaut

$$\mathcal{R}^\star = \mathcal{R}(g^\star) = \mathbb{E} \inf_{z \in \mathcal{Y}} \mathbb{E}\big(\ell(Y, z)|X = x\big) = \int_{\mathbb{R}^d} \inf_{z \in \mathcal{Y}} \mathbb{E}\big(\ell(Y, z)|X = x\big) \mathrm{d}\mu(x).$$

Démonstration. Pour toute fonction borélienne $g : \mathbb{R}^d \to \mathcal{Y}$, on a

$$\mathcal{R}(g) - \mathcal{R}^\star = \mathcal{R}(g) - \mathcal{R}(g^\star) = \int_{\mathbb{R}^d} \left[r(g(x)|x) - \inf_{z \in \mathcal{Y}} r(z|x) \right] \mathrm{d}\mu(x),$$

ce qui prouve le résultat annoncé. □

Dans la foulée, on définit l'**excès de risque** d'un prédicteur g comme la différence $\mathcal{R}(g) - \mathcal{R}^\star \geq 0$.

Le prédicteur de Bayes n'est pas nécessairement unique, comme le montre l'exemple suivant en classification binaire, mais tous les choix possibles induisent le même risque minimum.

Exemple de la classification binaire. En classification binaire (et pour la perte 0-1), le prédicteur de Bayes est défini par

$$g^\star(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x) \\ 0 & \text{sinon.} \end{cases}$$

(Ici, les égalités sont rompues en faveur de 0 par convention, mais l'autre choix conduirait également à un prédicteur de Bayes.) De façon équivalente,

$$g^\star(x) = \begin{cases} 1 & \text{si } r(x) > 1/2 \\ 0 & \text{sinon.} \end{cases}$$

On vérifie aisément que quelle que soit la règle de décision $g : \mathbb{R}^d \to \{0, 1\}$, on a

$$\mathcal{R}(g^\star) \leq \mathcal{R}(g).$$

En effet, puisque $\mathbb{P}(g(X) \neq Y) = 1 - \mathbb{P}\left(g(X) = Y\right)$, on a

$$\begin{aligned} \mathbb{P}(g(X) \neq Y) - \mathbb{P}(g^\star(X) \neq Y) &= \mathbb{P}\left(g^\star(X) = Y\right) - \mathbb{P}\left(g(X) = Y\right) \\ &= \mathbb{E}\left(\mathbb{P}(g^\star(X) = Y|X) - \mathbb{P}(g(X) = Y|X)\right) \\ &\geq 0. \end{aligned}$$

L'inégalité ci-dessus provient du fait que

$$\begin{aligned} \mathbb{P}(g(X) = Y|X) &= \mathbb{P}(g(X) = 1, Y = 1|X) + \mathbb{P}(g(X) = 0, Y = 0|X) \\ &= \mathbb{1}_{g(X)=1}\mathbb{P}(Y = 1|X) + \mathbb{1}_{g(X)=0}\mathbb{P}(Y = 0|X) \end{aligned}$$

et, de façon similaire,

$$\begin{aligned}\mathbb{P}(g^\star(X) = Y|X) &= \mathbb{1}_{g^\star(X)=1}\mathbb{P}(Y=1|X) + \mathbb{1}_{g^\star(X)=0}\mathbb{P}(Y=0|X) \\ &= \max\left(\mathbb{P}(Y=0|X), \mathbb{P}(Y=1|X)\right),\end{aligned}$$

par définition de $g^\star$. Il est clair que

$$\begin{aligned}&\mathbb{1}_{g(X)=1}\mathbb{P}(Y=1|X) + \mathbb{1}_{g(X)=0}\mathbb{P}(Y=0|X) - \max\left(\mathbb{P}(Y=0|X), \mathbb{P}(Y=1|X)\right) \\ &\quad = \mathbb{1}_{g(X)=1}\Big(\mathbb{P}(Y=1|X) - \max(\mathbb{P}(Y=0|X), \mathbb{P}(Y=1|X))\Big) \\ &\quad +\mathbb{1}_{g(X)=0}\Big(\mathbb{P}(Y=0|X) - \max(\mathbb{P}(Y=0|X), \mathbb{P}(Y=1|X))\Big) \leq 0.\end{aligned}$$

Le résultat est donc démontré.

Par ailleurs, toujours en classification binaire, on note en particulier que

$$\mathcal{R}^\star = \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}(g(X) \neq Y),$$

où l'infimum est évalué sur toutes les fonctions de décision. Il est également instructif de remarquer que $\mathcal{R}^\star = 0$ si et seulement si $Y = g^\star(X)$ $\mathbb{P}$-p.s., i.e. si et seulement si Y est une fonction borélienne de X. Dans le jargon de la classification supervisée, les probabilités $\mathbb{P}(Y=0|X=x)$ et $\mathbb{P}(Y=1|X=x)$ sont dites probabilités a posteriori.

Observons enfin que

$$\begin{aligned}\mathcal{R}(g) &= 1 - \mathbb{P}(g(X) = Y) \\ &= 1 - \mathbb{E}\left(\mathbb{P}(g(X) = Y|X)\right) \\ &= 1 - \mathbb{E}\left[\mathbb{1}_{g(X)=1} r(X) + \mathbb{1}_{g(X)=0}\left(1 - r(X)\right)\right].\end{aligned}$$

En conséquence,

$$\mathcal{R}^\star = 1 - \mathbb{E}\left[\mathbb{1}_{r(X)>1/2} r(X) + \mathbb{1}_{r(X)\leq 1/2}\left(1 - r(X)\right)\right] = 1 - \mathbb{E}[\max(r(X), 1 - r(X))].$$

Ceci montre qu'en classification binaire

$$\mathcal{R}^\star = \mathbb{E}\left[\min\left(r(X), 1 - r(X)\right)\right] = \frac{1}{2} - \frac{1}{2}\mathbb{E}|2r(X) - 1|,$$

et nous fournit donc des écritures alternatives pour $\mathcal{R}^\star$.

L'exemple de la classification binaire s'étend facilement au cas multi-classe, le prédicteur de Bayes (pour la fonction de perte 0-1) vérifiant dans ce cas

$$g^\star(x) \in \operatorname*{Argmax}_{m \in \{1,\ldots,M\}} \mathbb{P}(Y = m|X = x), \quad \forall x \in \mathbb{R}^d.$$

Exemple de la régression. Avec la perte quadratique, le prédicteur de Bayes satisfait

$$g^\star(x) \in \operatorname*{Argmin}_{z\in\mathbb{R}} \mathbb{E}\big[(Y-z)^2|X=x\big].$$

Or on a

$$\mathbb{E}\big[(Y-z)^2|X=x\big] = \mathbb{E}\big[(Y-\mathbb{E}(Y|X=x))^2|X=x\big] + (z-\mathbb{E}(Y|X=x))^2 + 0.$$

Le premier terme ne dépend plus de z tandis que le second est toujours positif et vaut 0 pour $z = \mathbb{E}(Y|X=x)$. On en déduit que

$$g^\star(x) = \mathbb{E}(Y|X=x),$$

i.e. le prédicteur de Bayes est donné par l'espérance conditionnelle de la variable à prédire Y, c'est-à-dire la fonction de régression r elle-même. Par ailleurs, le risque de Bayes est l'espérance de la variance conditionnelle $\mathcal{R}^\star = \mathbb{E}(Y-\mathbb{E}(Y|X))^2$.

Problème. Le prédicteur optimal $g^\star$ dépend de la loi ν du couple (X,Y). Puisque cette loi est (en général) inconnue, $g^\star$ et $\mathcal{R}^\star$ sont inaccessibles et il faut alors faire appel à un échantillon i.i.d. $(X_1,Y_1),\dots,(X_n,Y_n)$, de même loi que (X,Y), pour espérer récupérer de l'information sur ces deux quantités.

1.3 L'apprentissage et la minimisation du risque empirique

On suppose donc à partir de maintenant que l'on a accès à un n-échantillon i.i.d. (également appelé dans ce contexte base de données ou base d'apprentissage) formé de n couples $(X_1,Y_1),\dots,(X_n,Y_n)$ de variables aléatoires indépendantes entre elles, de même loi que (X,Y) et indépendantes de ce dernier couple. Pour abréger, on note $\mathscr{D}_n = (X_1,Y_1),\dots,(X_n,Y_n)$. C'est à partir de cet échantillon que l'on va s'attacher à construire un prédicteur $g_n(x) = g_n(x;\mathscr{D}_n)$ à valeurs dans $\mathcal{Y}$ dont les performances se rapprochent de celles de la règle de Bayes $g^\star$. C'est le mécanisme d'apprentissage. Puisque les observations sont en nombre fini (n), il s'agira « d'interpoler », voire « d'extrapoler » ce qui est observé pour construire notre prédicteur.

La qualité d'un prédicteur g_n est mesurée par le risque (conditionnel) attendu

$$\mathcal{R}(g_n) = \mathbb{E}(\ell(Y, g_n(X))|\mathscr{D}_n) = \int_{\mathbb{R}^d \times \mathcal{Y}} \ell(y, g_n(x)) \mathrm{d}\mu(x, y).$$

Il convient de remarquer que, tout comme g_n, le risque $\mathcal{R}(g_n)$ est aléatoire par l'intermédiaire de $\mathscr{D}_n$. Le conditionnement par $\mathscr{D}_n$ permet de distinguer l'aléatoire provenant de l'échantillon de celui issu du couple générique (X, Y). On notera au passage que $\mathbb{E}\mathcal{R}(g_n) = \mathbb{E}(\ell(Y, g_n(X)))$.

A partir de là, il est raisonnable de s'interroger sur le comportement du risque attendu lorsque la taille de l'échantillon tend vers l'infini. On est en particulier en droit d'attendre d'une « bonne règle » que son risque attendu se rapproche de $\mathcal{R}^\star$ lorsque n croît. Comme $\mathcal{R}(g_n)$ est aléatoire (contrairement à $\mathcal{R}^\star$), il convient de bien préciser le sens des convergences. C'est l'objet de la définition qui suit.

Définition 1. *Un prédicteur g_n est convergent si $\mathbb{E}\mathcal{R}(g_n) \to \mathcal{R}^\star$. Il est fortement convergent si $\mathcal{R}(g_n) \to \mathcal{R}^\star$, $\mathbb{P}$-p.s.*

Comme $\mathcal{R}(g_n) \geq \mathcal{R}^\star$, on notera que la propriété $\mathbb{E}\mathcal{R}(g_n) \to \mathcal{R}^\star$ est équivalente à $\mathcal{R}(g_n) \to \mathcal{R}^\star$ dans $\mathbb{L}^1(\nu)$. On pourra aussi montrer (exercice), que la convergence dans $\mathbb{L}^1(\nu)$ équivaut dans ce cas à la convergence en probabilité de $\mathcal{R}(g_n)$ vers $\mathcal{R}^\star$. On en déduit en particulier que si g_n est fortement convergent, il est aussi convergent.

La minimisation du risque empirique fait partie des grands paradigmes de l'apprentissage statistique. Le principe général est le suivant. Donnons-nous un n-échantillon i.i.d. $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (et indépendant de) (X, Y) et une famille $\mathscr{G}$ de prédicteurs candidats. On se pose alors le problème de choisir dans $\mathscr{G}$, en utilisant $\mathscr{D}_n$, une règle particulière $g_n^\star$ telle que $\mathcal{R}(g_n^\star) = \mathbb{E}(\ell(Y, g_n^\star(X))|\mathscr{D}_n)$ soit proche de $\inf_{g \in \mathscr{G}} \mathcal{R}(g) = \inf_{g \in \mathscr{G}} \mathbb{E}(\ell(Y, g(X)))$. (Cette dernière quantité n'est pas $\mathcal{R}^\star$, qui lui est un infimum sur tous les prédicteurs possibles). En d'autres termes, on cherche à utiliser au mieux la base de données afin de sélectionner la meilleure technique de prévision possible au sein d'une collection $\mathscr{G}$ de règles fixée a priori. Il peut par exemple s'agir de prédicteurs linéaires (i.e. de la forme $x \mapsto \theta^\intercal x$ en régression, ou qui décident 0 ou 1 selon que l'on tombe d'un

côté ou de l'autre d'un hyperplan en classification binaire), de règles polynomiales (des fonctions polynomiales de la variable explicative en régression, ou qui décident 0 ou 1 en fonction du signe d'un polynôme pour la classification binaire), mais bien d'autres exemples sont possibles.

Afin d'atteindre cet objectif, une façon naturelle de procéder consiste à sélectionner dans $\mathscr{G}$ une règle $g_n^\star$ qui minimise le **risque empirique**

$$\hat{\mathcal{R}}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, g(X_i))$$

parmi tous les éléments de $\mathscr{G}$, soit donc

$$g_n^\star \in \operatorname*{Argmin}_{g\in\mathscr{G}} \hat{\mathcal{R}}_n(g).$$

Exemples.

1. En classification binaire, le risque empirique $\hat{\mathcal{R}}_n(g) = n^{-1}\sum_{i=1}^n \mathbb{1}_{g(X_i)\neq Y_i}$ est le nombre d'erreurs moyen sur l'échantillon d'apprentissage $\mathcal{D}_n$. Un minimiseur $g_n^\star$ est une règle qui, parmi la famille de règles considérée $\mathscr{G}$, commet le moins d'erreurs possibles sur $\mathcal{D}_n$.
2. En régression, le risque empirique $\hat{\mathcal{R}}_n(g) = n^{-1}\sum_{i=1}^n (Y_i - g(X_i))^2$ est l'erreur des moindres carrés empirique. Par exemple pour la classe $\mathscr{G}$ qui est l'ensemble des applications linéaires sur $\mathbb{R}^d$, i.e. $g(x) = g_\theta(x) = \theta^\intercal x$ pour $\theta \in \mathbb{R}^d$, on est ramenés au problème de la régression linéaire (voir chapitre 6).

En rappelant que $\mathcal{R}(g_n^\star) = \mathbb{E}(\ell(Y, g_n^\star(X))|\mathscr{D}_n)$, on espère donc naturellement que $\mathcal{R}(g_n^\star) \approx \inf_{g\in\mathscr{G}} \mathcal{R}(g)$. Remarquons d'emblée que

$$\mathcal{R}(g_n^\star) - \mathcal{R}^\star = \underbrace{\left[\mathcal{R}(g_n^\star) - \inf_{g\in\mathscr{G}} \mathcal{R}(g)\right]}_{\text{erreur d'estimation}} + \underbrace{\left[\inf_{g\in\mathscr{G}} \mathcal{R}(g) - \mathcal{R}^\star\right]}_{\text{erreur d'approximation}}.$$

Cette égalité, simple mais fondamentale, montre que l'erreur commise par $\mathcal{R}(g_n^\star)$ en tant qu'estimateur de $\mathcal{R}^\star$ se décompose en deux termes, respectivement appelés erreur d'estimation et erreur d'approximation. L'erreur d'estimation est aléatoire et reflète l'écart entre la règle sélectionnée et le champion local dans $\mathscr{G}$. L'erreur d'approximation est déterministe et mesure la proximité entre la famille $\mathscr{G}$ et la règle optimale de Bayes.

Il est facile de voir que les deux termes d'erreur varient en sens inverse avec la taille de la classe $\mathscr{G}$, qui doit donc être suffisamment grande pour que l'erreur d'approximation soit petite, mais aussi suffisamment petite pour que l'erreur d'estimation soit contrôlée ! Pour s'en convaincre, il suffit d'envisager la situation extrême où $\mathscr{G}$ est constituée de toutes les fonctions mesurables de $\mathbb{R}^d$ dans $\mathcal{Y}$. Dans ce cas, l'erreur d'approximation est nulle, mais l'erreur d'estimation peut être importante, comme le montre le choix de la règle

$$g_n^\star(x) = \begin{cases} Y_i & \text{si } x = X_i,\ 1 \le i \le n \\ 0 & \text{sinon,} \end{cases}$$

dont le risque empirique est nul ! (En effet, cette règle impose $g_n^\star(x) = 0$ pour tout x différent des données $X_1, \dots, X_n$). Ce phénomène indésirable, qui traduit une accroche trop importante aux données, est appelé sur-apprentissage (« overfitting » en anglais) et nous donnerons dans la suite des conditions précises sur $\mathscr{G}$ permettant de l'éviter. À partir de maintenant, nous supposons donc la classe $\mathscr{G}$ fixée une fois pour toutes et cherchons à contrôler le terme d'estimation.

Lemme 1. *Un prédicteur $g_n^\star$ minimisant le risque empirique sur la classe $\mathscr{G}$ vérifie*

(**i**) $\mathcal{R}(g_n^\star) - \inf_{g\in\mathscr{G}} \mathcal{R}(g) \le 2 \sup_{g\in\mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$

(**ii**) $|\hat{\mathcal{R}}_n(g_n^\star) - \mathcal{R}(g_n^\star)| \le \sup_{g\in\mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$.

Démonstration. Le point (i) est une borne de l'erreur d'estimation du prédicteur. En introduisant le risque empirique de ce prédicteur, on écrit

$$\mathcal{R}(g_n^\star) - \inf_{g\in\mathscr{G}} \mathcal{R}(g) \le \big|\mathcal{R}(g_n^\star) - \hat{\mathcal{R}}_n(g_n^\star)\big| + \big|\hat{\mathcal{R}}_n(g_n^\star) - \inf_{g\in\mathscr{G}} \mathcal{R}(g)\big|.$$

Clairement,

$$\big|\mathcal{R}(g_n^\star) - \hat{\mathcal{R}}_n(g_n^\star)\big| \le \sup_{g\in\mathscr{G}} \big|\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)\big|,$$

et par définition de $g_n^\star$,

$$\big|\hat{\mathcal{R}}_n(g_n^\star) - \inf_{g\in\mathscr{G}} \mathcal{R}(g)\big| = \big|\inf_{g\in\mathscr{G}} \hat{\mathcal{R}}_n(g) - \inf_{g\in\mathscr{G}} \mathcal{R}(g)\big| \le \sup_{g\in\mathscr{G}} \big|\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)\big|.$$

(La dernière inégalité provient de la définition de inf et de sup.) Cela prouve la première assertion. La preuve de la seconde est immédiate. □

Le Lemme 1 montre qu'en contrôlant la quantité $\sup_{g\in\mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$, on fait coup double, puisque l'on maîtrise non seulement la sous-optimalité de $g_n^\star$ dans $\mathscr{G}$ vis-à-vis du vrai risque $\mathcal{R}$, mais aussi l'erreur $|\hat{\mathcal{R}}_n(g_n^\star) - \mathcal{R}(g_n^\star)|$ commise lorsque $\hat{\mathcal{R}}_n(g_n^\star)$ est utilisée pour estimer $\mathcal{R}(g_n^\star)$, le véritable risque du prédicteur sélectionné. Il est donc désormais légitime de faire porter nos efforts sur l'analyse du terme $\sup_{g\in\mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$. Historiquement, dans le problème de classification supervisée, la théorie de Vapnik-Chervonenkis a eu une influence considérable et nous allons la présenter dans le chapitre suivant. Pour la motiver, nous commençons simplement, en examinant le cas où la classe de fonctions $\mathscr{G}$ a un cardinal fini.

1.4 Cas de la classification binaire et d'une classe de cardinal fini

Dorénavant, nous considérons le problème de la classification binaire, avec $\mathcal{Y} = \{0,1\}$, la fonction de perte 0-1 donnée par $\ell(y,z) = \mathbb{1}_{y\neq z}$ et le risque empirique $\hat{\mathcal{R}}_n(g) = n^{-1}\sum_{i=1}^n \mathbb{1}_{g(X_i)\neq Y_i}$. (Le label est supposé binaire pour simplifier mais la théorie s'étend sans trop de difficultés au cas multi-labels).

Commençons par rappeler l'inégalité de Hoeffding.

Théorème 1 (Inégalité de Hoeffding)**.** *Soit $Z_1,\dots,Z_n$ des variables aléatoires réelles indépendantes telles que $a_i \leq Z_i \leq b_i$, $\mathbb{P}$-p.s. ($a_i < b_i$). Alors, pour tout $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n (Z_i - \mathbb{E}Z_i)\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

En particulier, si Z désigne une variable aléatoire de loi binomiale $\mathscr{B}(n,p)$, alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{Z}{n} - p\right| \geq \varepsilon\right) = \mathbb{P}\left(|Z - np| > n\varepsilon\right)\big) \leq 2\exp\left(-2n^2\varepsilon^2 / \sum_{i=1}^n 1\right) = 2e^{-2n\varepsilon^2}.$$

En remarquant que pour un prédicteur fixé g (non aléatoire), la quantité

$$n\hat{\mathcal{R}}_n(g) = \sum_{i=1}^{n} \ell(Y_i, g(X_i)) = \sum_{i=1}^{n} \mathbb{1}_{g(X_i)\neq Y_i}$$

suit une loi $\mathscr{B}(n, \mathcal{R}(g))$, on en conclut que

$$\mathbb{P}\left(\left|\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2},$$

ce qui conduit au premier résultat fondamental suivant :

Théorème 2. *Supposons que la classe $\mathcal{G}$ soit de cardinal fini majoré par N. Alors, pour tout $\varepsilon > 0$,*

$$\mathbb{P}\Big(\sup_{g\in\mathcal{G}} \left|\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)\right| \geq \varepsilon\Big) \leq 2Ne^{-2n\varepsilon^2}. \tag{1.3}$$

Il faut noter que cette inégalité est déjà remarquable car la majoration de la probabilité est universelle, au sens où elle ne dépend pas de la loi du couple (X, Y). On en déduit en particulier, en utilisant le lemme de Borel-Cantelli, que

$$\sup_{g\in\mathcal{G}} \left|\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)\right| \to 0, \quad \mathbb{P}\text{-p.s.}$$

et donc, d'après le Lemme 1, que

$$\mathcal{R}(g_n^\star) - \inf_{g\in\mathcal{G}} \mathcal{R}(g) \to 0, \quad \mathbb{P}\text{-p.s.}$$

Ce résultat signifie que pourvu que la classe $\mathcal{G}$ soit de cardinal fini, l'erreur d'estimation pour la classification tend p.s. vers 0 lorsque n tend vers l'infini ; en d'autres termes, l'apprentissage est asymptotiquement optimal. Tout ceci s'étend sans difficulté au contrôle de l'espérance $\mathbb{E}(\sup_{g\in\mathcal{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|)$, via le lemme technique suivant.

Lemme 2. *Soit Z une variable aléatoire à valeurs dans $\mathbb{R}_+$. Supposons qu'il existe une constante $C \geq 1$ telle que, pour tout $\varepsilon > 0$,*

$$\mathbb{P}(Z \geq \varepsilon) \leq Ce^{-2n\varepsilon^2}.$$

Alors

$$\mathbb{E}Z \leq \sqrt{\frac{\log(Ce)}{2n}}.$$

Démonstration. En partant de l'identité

$$\mathbb{E}Z^2 = \int_0^{+\infty} \mathbb{P}(Z^2 > \varepsilon)\mathrm{d}\varepsilon,$$

on a, pour tout $u \geq 0$,

$$\begin{aligned}
\mathbb{E}Z^2 &= \int_0^u \mathbb{P}(Z^2 > \varepsilon)\mathrm{d}\varepsilon + \int_u^{+\infty} \mathbb{P}(Z^2 > \varepsilon)\mathrm{d}\varepsilon \\
&= \int_0^u \mathbb{P}(Z^2 > \varepsilon)\mathrm{d}\varepsilon + \int_u^{+\infty} \mathbb{P}(Z > \sqrt{\varepsilon})\mathrm{d}\varepsilon \\
&\leq u + C\int_u^{+\infty} e^{-2n\varepsilon}\mathrm{d}\varepsilon \\
&= u + \frac{C}{2n}e^{-2nu}.
\end{aligned}$$

Avec le choix $u^\star = \frac{\log C}{2n}$ (qui minimise la borne de droite), on en déduit que $\mathbb{E}Z^2 \leq \frac{\log C}{2n} + \frac{1}{2n} = \frac{\log(Ce)}{2n}$, d'où le résultat par l'inégalité de Cauchy-Schwarz. □

Le lemme précédent, couplé à l'inégalité (1.3), montre que

$$\mathbb{E}\Big(\sup_{g\in\mathscr{G}} \big|\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)\big|\Big) \leq \sqrt{\frac{\log(2eN)}{2n}}.$$

Le Lemme 1 nous permet alors de conclure que

$$\mathbb{E}\mathcal{R}(g_n^\star) - \inf_{g\in\mathscr{G}} \mathcal{R}(g) \leq 2\sqrt{\frac{\log(2eN)}{2n}},$$

ce qui montre que, pour une classe $\mathscr{G}$ de cardinal fini, l'espérance de l'erreur d'estimation reste sous contrôle (avec une borne plus ou moins grande selon sa taille N) et tend vers 0 à la vitesse $1/\sqrt{n}$ lorsque n tend vers l'infini.

Néanmoins, lorsque $\mathscr{G}$ n'est pas de cardinal fini (comme c'est le cas dans la plupart des problèmes intéressants), l'approche que nous venons de présenter ne fonctionne plus et il faut trouver de nouveaux outils pour appréhender la « taille » de $\mathscr{G}$. C'est l'objet du chapitre suivant, qui présente la théorie de Vapnik-Chervonenkis.

Chapitre 2

Théorie de Vapnik-Chervonenkis pour la classification

Dans tout ce chapitre, on considère le problème de la classification supervisée : le couple (X, Y) est à valeurs dans $\mathbb{R}^d \times \mathcal{Y}$ où $\mathcal{Y}$ est fini. Par souci de simplification, on choisit de présenter uniquement le cas $\mathcal{Y} = \{0, 1\}$. La fonction de perte est le coût 0-1.

2.1 Passage du $\sup_{g \in \mathcal{G}}$ au $\sup_{A \in \mathcal{A}}$

Etant donné un n-échantillon i.i.d. $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (et indépendant de) $(X, Y) \in \mathbb{R}^d \times \mathcal{Y}$ et une famille $\mathcal{G}$ de règles de décision candidates, le chapitre précédent a montré le rôle essentiel joué par le terme $\sup_{g \in \mathcal{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$, qu'il faut donc apprendre à contrôler avec la plus grande généralité possible.

On rappelle que ν désigne la loi du couple (X, Y) et on note ν_n la mesure empirique associée à $\mathcal{D}_n$, i.e., pour tout $A \in \mathcal{B}(\mathbb{R}^d \times \{0, 1\})$,

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i, Y_i) \in A}.$$

À une règle de décision quelconque $g \in \mathcal{G}$, nous pouvons associer le borélien

$$A_g = \left\{ (x, y) \in \mathbb{R}^d \times \{0, 1\} : g(x) \neq y \right\}.$$

En utilisant cette notation, il est alors facile de voir que, d'une part,

$$\mathcal{R}(g) = \mathbb{P}(g(X) \neq Y) = \nu(A_g)$$

et, d'autre part, que

$$\hat{\mathcal{R}}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(X_i)\neq Y_i} = \nu_n(A_g).$$

On constate ainsi que

$$\left\{ \sup_{g\in\mathscr{G}} \left|\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)\right| \right\} = \left\{ \sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \right\},$$

où, par définition, $\mathscr{A} = \{A_g : g \in \mathscr{G}\}$. Ce jeu d'écriture nous montre donc que pour analyser le comportement probabiliste du terme $\sup_{g\in\mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$, il faut avant tout comprendre comment se comporte la déviation maximale de la mesure empirique ν_n par rapport à la vraie mesure ν, sur une classe d'ensembles mesurables $\mathscr{A}$ donnée. On peut d'ores et déjà observer que, pour un ensemble A fixé,

$$|\nu_n(A) - \nu(A)| \to 0, \quad \mathbb{P}\text{-p.s.}$$

d'après la loi des grands nombres. D'autre part, si le cardinal de $\mathscr{A}$ est fini et majoré par N, un raisonnement similaire à celui du Théorème 2 nous apprend que, pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \geq \varepsilon \Big) \leq 2Ne^{-2n\varepsilon^2}, \tag{2.1}$$

d'où l'on déduit (lemme de Borel-Cantelli) que, pour toute loi ν,

$$\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \to 0, \quad \mathbb{P}\text{-p.s.}$$

En revanche, si la classe $\mathscr{A}$ est trop grande, ce comportement n'est plus assuré. On s'en convaincra facilement en remarquant que si $\mathscr{A}$ désigne l'ensemble de tous les boréliens de $\mathbb{R}^d \times \{0,1\}$, alors on peut trouver des lois ν telles que

$$\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| = 1, \quad \mathbb{P}\text{-p.s.}$$

Il suffit de prendre $\nu = \rho \otimes (1/2\delta_0 + 1/2\delta_1)$ où ρ est une loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ (par exemple la loi d'un vecteur gaussien à densité). Prenons ensuite pour ω fixé l'ensemble

$A(\omega) = (\mathbb{R}^d \times \{0,1\}) \setminus \Big\{(X_1(\omega), Y_1(\omega)), \ldots, (X_n(\omega), Y_n(\omega))\Big\}$. Alors $\nu_n(A(\omega)) = 0$ mais $\nu(A(\omega)) = 1$ donc $\sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| = 1$.

La conclusion de tout ceci est qu'il faut parvenir, d'une manière ou d'une autre, à contrôler la « taille » de la classe d'ensembles $\mathscr{A}$. Pour atteindre cet objectif, il convient au préalable d'introduire quelques outils combinatoires nouveaux.

2.2 Théorème de Vapnik-Chervonenkis

Soit $\mathscr{A}$ une famille de sous-ensembles de $\mathbb{R}^p$, de cardinal (pas nécessairement fini) strictement supérieur à 1 (cette hypothèse sera implicite dans la suite). Etant donné n points $z_1, \ldots, z_n$ de $\mathbb{R}^p$, on définit la quantité $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n)$ par

$$\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n) = \big|\big\{\{z_1, \ldots, z_n\} \cap A : A \in \mathscr{A}\big\}\big|.$$

En d'autres termes, $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n)$ représente le nombre de sous-ensembles de $\{z_1, \ldots, z_n\}$ que l'on peut obtenir en intersectant ces n points par les ensembles de $\mathscr{A}$. Bien entendu, on a toujours $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n) \leq 2^n$, et lorsque $\mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n) = 2^n$, on dit que la classe $\mathscr{A}$ *pulvérise* l'ensemble $\{z_1, \ldots, z_n\}$. Afin de ne pas être gêné par le choix arbitraire de $z_1, \ldots, z_n$, on pose

$$\mathbf{S}_{\mathscr{A}}(n) = \max_{(z_1,\ldots,z_n) \in \mathbb{R}^{pn}} \mathcal{N}_{\mathscr{A}}(z_1, \ldots, z_n)$$

et on appelle cet indice le coefficient de pulvérisation de n points par la classe $\mathscr{A}$.

Clairement, $\mathbf{S}_{\mathscr{A}}(n) \leq 2^n$. D'autre part, $\mathbf{S}_{\mathscr{A}}(1) = 2$ (pourquoi ?) et si l'on a $\mathbf{S}_{\mathscr{A}}(k) < 2^k$ pour un certain entier $k > 1$ alors $\mathbf{S}_{\mathscr{A}}(n) < 2^n$ pour tout $n \geq k$ (pourquoi ?). Il est donc naturel de s'interroger sur l'existence d'un plus grand entier n tel que $\mathbf{S}_{\mathscr{A}}(n) = 2^n$. C'est l'objet de la définition suivante.

Définition 2. *Soit $\mathscr{A}$ une famille de sous-ensembles de $\mathbb{R}^p$. On appelle dimension de Vapnik-Chervonenkis de $\mathscr{A}$, notée $V_{\mathscr{A}}$, le plus grand entier $n_0 \geq 1$ tel que $\mathbf{S}_{\mathscr{A}}(n_0) = 2^{n_0}$. Si $\mathbf{S}_{\mathscr{A}}(n) = 2^n$ pour tout $n \geq 1$, on pose $V_{\mathscr{A}} = +\infty$.*

La dimension de Vapnik-Chervonenkis mesure, en un certain sens, la « taille » (la « dimension ») de la famille $\mathscr{A}$ et généralise ainsi la notion de cardinal. Il s'agit d'un concept combinatoire important qui, comme nous le verrons dans la suite, joue un rôle clé dans la théorie de l'apprentissage statistique. Examinons auparavant quelques exemples (les preuves sont de difficultés variées et laissées au lecteur).

Exemples.

1. Supposons $|\mathscr{A}| < \infty$. Dans ce cas, $\mathbf{S}_{\mathscr{A}}(n) \leq |\mathscr{A}|$. D'autre part, par définition de $V_{\mathscr{A}}$, on a $\mathbf{S}_{\mathscr{A}}(V_{\mathscr{A}}) = 2^{V_{\mathscr{A}}}$, d'où l'on déduit que
$$V_{\mathscr{A}} \leq \log_2 |\mathscr{A}|.$$

2. En dimension $p = 1$, si $\mathscr{A} = \{(-\infty, a] : a \in \mathbb{R}\}$, alors $\mathbf{S}_{\mathscr{A}}(n) = n + 1$. En effet, si on a n points, $x_1 < x_2 < \ldots < x_n$, on peut obtenir tous les sous-ensembles de points consécutifs à partir du premier point : $\emptyset, \{x_1\}, \{x_1, x_2\}, \{x_1, x_2, x_3, ..\}, \ldots, \{x_1, x_2, \ldots, x_n\}$ en intersectant avec $\mathscr{A}$ et seulement ceux-ci. Comme $\mathbf{S}_{\mathscr{A}}(1) = 2$ et $\mathbf{S}_{\mathscr{A}}(2) = 3 < 4$, on obtient $V_{\mathscr{A}} = 1$.

   Si $\mathscr{A} = \{[a, b] : (a, b) \in \mathbb{R}^2\}$, alors $\mathbf{S}_{\mathscr{A}}(n) = \frac{n(n+1)}{2} + 1$. En effet, en intersectant n points avec des segments, on pourra obtenir n ensembles de un seul point ; $n - 1$ ensembles de 2 points qui doivent être consécutifs ($\{x_1, x_2\}, \{x_2, x_3\}, \ldots, \{x_{n-1}, x_n\}$) ; $n - 2$ ensembles de 3 points qui doivent aussi être consécutifs ($\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}$, $\ldots, \{x_{n-2}, x_{n-1}, x_n\}$); $\cdots$; 2 ensembles de $n - 1$ points ($\{x_1, \ldots, x_{n-1}\}$, $\{x_2, \ldots, x_n\}$) et un ensemble de tous les n points. N'oublions pas l'ensemble vide qu'il est aussi possible d'obtenir en intersectant avec un segment qui ne contient aucun de ces n points. Il est facile de voir que nous ne pourrons obtenir aucun ensemble de points avec des « trous » , c'est-à-dire qui contient par exemple x_i et x_{i+2} mais pas x_{i+1} pour un $i = 2, 3, \ldots, n - 1$. On compte donc $\mathbf{S}_{\mathscr{A}}(n) = (1 + 2 + \cdots + n) + 1 = n(n + 1)/2 + 1$. Comme $\mathbf{S}_{\mathscr{A}}(2) = 4 = 2^2$ mais $\mathbf{S}_{\mathscr{A}}(3) = 7 < 2^3$, on obtient $V_{\mathscr{A}} = 2$.
3. Soit $p = 2$. Si
$$\mathscr{A} = \left\{(-\infty, a_1] \times (-\infty, a_2] : (a_1, a_2) \in \mathbb{R}^2\right\},$$
   alors $V_{\mathscr{A}} = 2$. En effet pour $n = 2$, il est facile d'obtenir les 4 sous-ensembles de points de $\{z_1 = (1, 0), z_2 = (0, 1)\}$ en prenant $(a_1, a_2) =$

$(0,0),(1,0),(0,1)$ et $(1,1)$ dans la définition de A. Donc $V_{\mathscr{A}} \geq 2$. Lorsque $n = 3$, soient $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ trois points quelconques de $\mathbb{R}^2$. On peut choisir un sous-ensemble de deux points parmi ceux-ci tels que le maximum de coordonnées x_1, x_2, x_3 est atteint sur une de leurs deux abssisses et le maximum de coordonnées y_1, y_2, y_3 est atteint sur une de leurs deux ordonnées. Alors tout ensemble A qui contient ce sous-ensemble de 2 points doit contenir aussi le 3ème point car ses deux coordonnées sont inférieures ou égales à $\max(x_1, x_2, x_3)$ et $\min(y_1, y_2, y_3)$ respectivement. Alors $\mathbf{S}_{\mathscr{A}}(3) < 2^3$ et donc $V_{\mathscr{A}} = 2$.

Si $\mathscr{A} = \{\text{rectangles de } \mathbb{R}^2\}$, alors $V_{\mathscr{A}} = 4$. En effet, lorsque $n = 2$, nous pouvons obtenir tous les sous-ensembles de $\{z_1 = (1,0), z_2 = (0,1), z_3 = (0,-1), z_4 = (-1,0)\}$ par intersections avec des rectangles. Par contre si on prend 5 points quelconques, on ne peut pas obtenir tous les sous-ensembles : un rectangle A qui contient 4 points avec les premières et deuxièmes coordonnées maximales et minimales, doit contenir obligatoirement le 5eme point qui reste (faire un dessin !).

4. En dimension p quelconque, si

$$\mathscr{A} = \left\{(-\infty, a_1] \times \cdots \times (-\infty, a_p] : (a_1, \ldots, a_p) \in \mathbb{R}^p\right\},$$

alors $V_{\mathscr{A}} = p$. Si $\mathscr{A} = \{\text{rectangles de } \mathbb{R}^p\}$, alors $V_{\mathscr{A}} = 2p$. La preuve qui généralise le cas $p = 2$ est laissée en exercice.

5. En revanche, pour $\mathscr{A} = \{\text{polygones convexes de } \mathbb{R}^2\}$, on a $V_{\mathscr{A}} = +\infty$. En effet, pour tout $n \geq 1$, si on considère n points sur un cercle, on pourra obtenir n'importe lequel des 2^n sous-ensembles de ces points en intersectant avec des polygones convexes.

6. (**Important.**) Soit $\mathscr{F}$ un espace vectoriel de fonctions de $\mathbb{R}^p \to \mathbb{R}$, de dimension finie $\dim(\mathscr{F})$. Alors, si

$$\mathscr{A} = \left\{\{x \in \mathbb{R}^p : f(x) \geq 0\} : f \in \mathscr{F}\right\},$$

on a $V_{\mathscr{A}} \leq \dim(\mathscr{F})$.

Soit $\dim(\mathscr{F}) = m$. Prenons $x_1, \ldots, x_{m+1}$, des points quelconques de $\mathbb{R}^p$. Considérons l'application linéaire $L : \mathscr{F} \to \mathbb{R}^{m+1}$ définie par $L(f) = (f(x_1), f(x_2), \ldots, f(x_{m+1}))$. Comme $\dim(\mathscr{F}) = m$, alors $\dim L(\mathscr{F}) \leq m$ et donc il existe un vecteur non-nul $\gamma = (\gamma_1, \ldots, \gamma_{m+1}) \in \mathbb{R}^{m+1}$ orthogonal à $L(\mathscr{F})$. Donc pour tout $f \in \mathscr{F}$:

$$\gamma_1 f(x_1) + \gamma_2 f(x_2) + \cdots + \gamma_{m+1} f(x_{m+1}) = 0.$$

Comme γ est non nul, quitte à le multiplier par (-1), on peut dire que $\{i : \gamma_i < 0\} \neq \emptyset$. Alors

$$\sum_{i:\gamma_i \geq 0} \gamma_i f(x_i) = - \sum_{i:\gamma_i < 0} \gamma_i f(x_i). \tag{2.2}$$

(La somme à gauche peut être vide de termes, mais la somme à droite est non-vide de termes). Montrons qu'il est impossible d'obtenir un sous-ensemble de points $\{x_i : \gamma_i \geq 0\}$. En effet si pour une certaine fonction $f : f(x_i) \geq 0$ pour tout i tel que $\gamma_i \geq 0$ et $f(x_i) < 0$ pour tout i tel que $\gamma_i < 0$, alors la partie droite dans (2.2) est supérieure ou égale à zéro alors que la partie gauche est strictement négative, ce qui est impossible. Donc $S_{\mathscr{A}}(m+1) < 2^{m+1}$ et donc $V_{\mathscr{A}} \leq m$.

7. En particulier, si $\mathscr{A}$ désigne la famille des $1/2$-espaces linéaires, i.e. les sous-ensembles de $\mathbb{R}^p$ de la forme $\{x \in \mathbb{R}^p : a^\intercal x + b \geq 0\}$ pour $a \in \mathbb{R}^p, b \in \mathbb{R}$, il vient $V_{\mathscr{A}} \leq p+1$. En effet l'espace $\mathscr{F} = \{x \mapsto a^\intercal x + b : a \in \mathbb{R}^p, b \in \mathbb{R}\}$ est de dimension $p+1$.

Nous sommes désormais équipés pour énoncer le théorème fondamental suivant, appelé théorème de Vapnik-Chervonenkis.

Théorème 3 (Vapnik-Chervonenkis)**.** *Soit $Z_1, \ldots, Z_n$ des variables aléatoires indépendantes, de même loi ν sur $\mathbb{R}^p$, et soit ν_n la mesure empirique correspondante. Alors, pour toute famille borélienne $\mathscr{A} \subset \mathbb{R}^p$ et pour tout $\varepsilon > 0$, on a*

$$\mathbb{P}\left(\sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \right) \leq 8 \mathbf{S}_{\mathscr{A}}(n) e^{-n\varepsilon^2/32}.$$

Avant de prouver ce théorème, il convient de souligner quelques points essentiels.

1. La borne est universelle, dans le sens où elle ne dépend pas de la loi particulière ν.
2. Ce résultat généralise l'inégalité (2.1) qui n'était valable que pour une classe $\mathscr{A}$ de cardinal fini. Grosso modo, le cardinal de $\mathscr{A}$ est remplacé par le coefficient de pulvérisation.
3. D'après le lemme de Borel-Cantelli, il s'ensuit que

$$\sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| \to 0, \quad \mathbb{P}\text{-p.s.}$$

dès que la série de terme général $\mathbf{S}_{\mathscr{A}}(n)e^{-n\varepsilon^2/32}$ est sommable. C'est par exemple le cas si $|\mathscr{A}| < \infty$ ou si $\mathbf{S}_{\mathscr{A}}(n)$ est un polynôme en n. En revanche, il est impossible de conclure si $\mathbf{S}_{\mathscr{A}}(n) = 2^n$ pour tout n (ou, c'est équivalent, si $V_{\mathscr{A}} = +\infty$).

4. La preuve du Théorème 3 n'est pas compliquée et repose sur quelques arguments clés que l'on rencontre fréquemment en théorie de l'apprentissage. En un mot, le principe consiste à faire sortir le supremum de la parenthèse pour le placer devant la probabilité. Ce grand saut est effectué en jouant sur les propriétés combinatoires de la classe $\mathscr{A}$ telles que décrites par $\mathbf{S}_{\mathscr{A}}(n)$.

Démonstration du Théorème 3. Dans toute la preuve, on suppose $\varepsilon > 0$ fixé et on choisit n assez grand de sorte que $n\varepsilon^2 \geq 2$. Dans le cas contraire, il est facile de voir que le résultat annoncé est correct car la borne du théorème est alors plus grande que 1. La preuve s'organise en 4 étapes.

Etape 1 : Symétrisation. En sus du n-échantillon i.i.d. original $Z_1, \ldots, Z_n$, on considère un second échantillon i.i.d. $Z'_1, \ldots, Z'_n$ de la loi ν, indépendant du premier. On note ν_n la mesure empirique relative à $Z_1, \ldots, Z_n$ et ν'_n celle relative à $Z'_1, \ldots, Z'_n$. La première étape consiste à montrer que

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon \Big) \leq 2\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu'_n(A)| > \frac{\varepsilon}{2} \Big).$$

Pour tout $\omega \in \Omega$ tel que $\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \epsilon$ choisissons un ensemble $A^\star = A^\star(\omega)$ (dépendant de l'échantillon initial $Z_1, \ldots, Z_n$) tel que $|\nu_n(A^\star) - \nu(A^\star)| > \epsilon$. Pour tout $\omega \in \Omega$ tel que $\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \leq \epsilon$, posons $A^\star = \mathbb{R}^p$. Dans ce cas $|\nu_n(A^\star) - \nu(A^\star)| = |1 - 1| = 0$. Autrement dit,

$$\Big\{ \omega \in \Omega : \sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \epsilon \Big\} = \Big\{ \omega \in \Omega : |\nu_n(A^\star) - \nu(A^\star)| > \epsilon \Big\}.$$

Par ailleurs,

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \frac{\varepsilon}{2}\Big) \\
&\quad = \mathbb{E}\bigg(\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \frac{\varepsilon}{2} \,\Big|\, Z_1,\dots,Z_n\Big)\bigg) \\
&\quad \geq \mathbb{E}\bigg(\mathbb{P}\Big(\big|\nu_n(A^\star) - \nu'_n(A^\star)\big| > \frac{\varepsilon}{2} \,\Big|\, Z_1,\dots,Z_n\Big)\bigg) \\
&\quad = \mathbb{P}\Big(\big|\nu_n(A^\star) - \nu'_n(A^\star)\big| > \frac{\varepsilon}{2}\Big).
\end{aligned}$$

On en déduit, en utilisant l'inégalité triangulaire et la définition de $A^\star$, que

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu'_n(A)\big| > \frac{\varepsilon}{2}\Big) \\
&\quad \geq \mathbb{P}\bigg(\Big\{ \big|\nu_n(A^\star) - \nu(A^\star)\big| > \varepsilon\Big\} \cap \Big\{ \big|\nu'_n(A^\star) - \nu(A^\star)\big| < \frac{\varepsilon}{2}\Big\}\bigg) \\
&\quad = \mathbb{E}\bigg(\mathbb{1}_{|\nu_n(A^\star)-\nu(A^\star)|>\varepsilon} \mathbb{P}\Big(\big|\nu'_n(A^\star) - \nu(A^\star)\big| < \varepsilon/2 \,\Big|\, Z_1,\dots,Z_n\Big)\bigg).
\end{aligned}$$

L'inégalité de Bienaymé-Tchebytchev montre que

$$\begin{aligned}
&\mathbb{P}\Big(\big|\nu'_n(A^\star) - \nu(A^\star)\big| < \frac{\varepsilon}{2} \,\Big|\, Z_1,\dots,Z_n\Big) \\
&\qquad \geq 1 - \frac{\mathbb{E}\Big(\big(\nu'_n(A^\star) - \nu(A^\star)\big)^2 \,\Big|\, Z_1,\dots,Z_n\Big)}{\varepsilon^2/4}.
\end{aligned}$$

En observant que, conditionnellement à $Z_1,\dots,Z_n$, la variable $n\nu'_n(A^\star)$ suit une loi $\mathscr{B}(n,\nu(A^\star))$, on en déduit en particulier que

$$\begin{aligned}
\mathbb{P}\Big(\big|\nu'_n(A^\star) - \nu(A^\star)\big| < \varepsilon/2 \,\Big|\, Z_1,\dots,Z_n\Big) &\geq 1 - \frac{\mathbb{V}\big(\nu'_n(A^\star)\big|Z_1,\dots,Z_n\big)}{\varepsilon^2/4} \\
&= 1 - \frac{\nu(A^\star)(1-\nu(A^\star))}{n\varepsilon^2/4} \\
&\geq 1 - \frac{1}{n\varepsilon^2},
\end{aligned}$$

car $\sup_{u\in[0,1]} u(1-u) = 1/4$. Ainsi, puisque $n\varepsilon^2 \geq 2$,

$$\begin{aligned}\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu_n'(A)\big| > \varepsilon/2\Big) &\geq \mathbb{E}\left(\mathbb{1}_{|\nu_n(A^\star)-\nu(A^\star)|>\varepsilon}\Big(1-\frac{1}{n\varepsilon^2}\Big)\right)\\ &\geq \frac{1}{2}\,\mathbb{P}\Big(\,|\nu_n(A^\star) - \nu(A^\star)| > \varepsilon\Big)\\ &= \frac{1}{2}\,\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon\Big).\end{aligned}$$

On en conclut bien que

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon\Big) \leq 2\,\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu_n'(A)\big| > \frac{\varepsilon}{2}\Big).$$

Etape 2 : Signes aléatoires. On se donne maintenant n variables aléatoires $\sigma_1,\ldots,\sigma_n$, indépendantes et chacune de loi de Rademacher, i.e. telles que $\mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = +1) = 1/2$. On suppose en outre que les variables $\sigma_1,\ldots,\sigma_n$ sont indépendantes de $Z_1,\ldots,Z_n,Z_1',\ldots,Z_n'$. Il est alors facile de voir que

$$n\sup_{A\in\mathscr{A}} \big|\nu_n(A) - \nu_n'(A)\big| = \sup_{A\in\mathscr{A}} \left|\sum_{i=1}^n \big(\mathbb{1}_A(Z_i) - \mathbb{1}_A(Z_i')\big)\right|$$

a même loi que

$$\sup_{A\in\mathscr{A}} \left|\sum_{i=1}^n \sigma_i\big(\mathbb{1}_A(Z_i) - \mathbb{1}_A(Z_i')\big)\right|.$$

En effet, si une variable aléatoire U est de loi symétrique et une variable σ est indépendante de U à valeurs ±1 avec probabilités $1/2$, alors σU est de même loi que U : pour tout B borélien $\mathbb{P}(\sigma U \in B) = 1/2\mathbb{P}(U \in B) + 1/2\mathbb{P}(-U \in B) = 1/2\mathbb{P}(U \in B) + 1/2\mathbb{P}(U \in B) = \mathbb{P}(U \in B)$. Si des variables aléatoires $U_1,\ldots,U_n,\sigma_1,\ldots,\sigma_n$ sont indépendantes, les U_i étant toutes de même loi symétrique, et les σ_i étant de même loi que σ pour $i = 1,\ldots,n$, alors $\sigma_1U_1,\ldots,\sigma_nU_n$ sont indépendantes et de même loi que U. Il reste à remarquer que $\mathbb{1}_A(Z_i) - \mathbb{1}_A(Z_i')$ sont de loi symétrique car Z_i et Z_i' sont indépendantes et de même loi.

Dès lors, en utilisant le résultat de la première étape, nous pouvons écrire que

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > \varepsilon\Big) \leq 2\mathbb{P}\Big(\sup_{A\in\mathscr{A}} \frac{1}{n}\Big|\sum_{i=1}^n \sigma_i\big(\mathbb{1}_A(Z_i) - \mathbb{1}_A(Z_i')\big)\Big| > \frac{\varepsilon}{2}\Big),$$

et donc (par symétrie de la loi)

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}}|\nu_n(A)-\nu(A)|>\varepsilon\Big)\leq 4\mathbb{P}\Big(\sup_{A\in\mathscr{A}}\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big).$$

Etape 3 : Le saut du sup. En poursuivant le calcul précédent, on a

$$\begin{aligned}&\mathbb{P}\Big(\sup_{A\in\mathscr{A}}|\nu_n(A)-\nu(A)|>\varepsilon\Big)\\&\quad\leq 4\mathbb{E}\bigg(\mathbb{P}\Big(\sup_{A\in\mathscr{A}}\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)\bigg).\end{aligned}$$

Majorons alors le terme

$$\begin{aligned}&\mathbb{P}\Big(\sup_{A\in\mathscr{A}}\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)\\&\quad\leq\mathbb{P}\Big(\exists A\in\mathscr{A}:\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big).\end{aligned}$$

Une fois fixés les points $z_1,\dots,z_n$, le vecteur $(\mathbb{1}_A(z_1),\dots,\mathbb{1}_A(z_n))$ prend $\mathcal{N}_{\mathscr{A}}(z_1,\dots,z_n)$ valeurs distinctes lorsque A varie dans $\mathscr{A}$, soit donc un maximum de $\mathbf{S}_{\mathscr{A}}(n)$ valeurs. Du coup,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{A\in\mathscr{A}}\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)\\&\quad\leq\mathbb{P}\Big(\exists A\in\mathscr{A}_0:\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big),\end{aligned}$$

où $\mathscr{A}_0$ est un ensemble fini (dépendant de $Z_1,\dots,Z_n$) de cardinal au plus $\mathbf{S}_{\mathscr{A}}(n)$. Il s'ensuit que

$$\begin{aligned}&\mathbb{P}\Big(\sup_{A\in\mathscr{A}}\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)\\&\quad\leq\sum_{A\in\mathscr{A}_0}\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)\\&\quad\leq\mathbf{S}_{\mathscr{A}}(n)\sup_{A\in\mathscr{A}}\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big).\end{aligned}$$

On notera ici le saut du sup de l'intérieur vers l'extérieur de la probabilité, accompli grâce à l'introduction du coefficient de pulvérisation. Ainsi,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{A\in\mathscr{A}}|\nu_n(A)-\nu(A)|>\varepsilon\Big)\\&\quad\le 4\mathbf{S}_{\mathscr{A}}(n)\mathbb{E}\Big(\sup_{A\in\mathscr{A}}\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)\Big).\end{aligned}\tag{2.3}$$

Etape 4 : Inégalité de Hoeffding et conclusion. Conditionnellement à $Z_1,\dots,Z_n$, la variable aléatoire $\sum_{i=1}^n\sigma_i\mathbb{1}_A(Z_i)$ est la somme de n variables aléatoires indépendantes, **centrées** (c'est là que les signes aléatoires jouent un rôle primordial !) et comprises entre -1 et 1. Ainsi, d'après l'inégalité de Hoeffding (Théorème 1)

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)&=\mathbb{P}\Big(\Big|\sum_{i=1}^n\sigma_i\mathbb{1}_A(Z_i)\Big|>\frac{n\varepsilon}{4}\Big|Z_1,\dots,Z_n\Big)\\&\le 2e^{-2n^2\varepsilon^2/(4^2(1-(-1))^2n)}=2e^{-n\varepsilon^2/32}.\end{aligned}$$

On conclut alors en utilisant (2.3) que

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}}|\nu_n(A)-\nu(A)|>\varepsilon\Big)\le 8\mathbf{S}_{\mathscr{A}}(n)e^{-n\varepsilon^2/32},$$

ce qui est bien le résultat annoncé. □

Application : théorème de Glivenko-Cantelli. Plaçons-nous sur la droite réelle et considérons un n-échantillon $Z_1,\dots,Z_n$ de variables aléatoires i.i.d., de loi commune ν. En prenant $\mathscr{A}=\{(-\infty,z]:z\in\mathbb{R}\}$, il est facile de voir que, pour tout $A=(-\infty,z]\in\mathscr{A}$, on a $\nu(A)=F(z)$ et $\nu_n(A)=F_n(z)$, où F (respectivement F_n) est la fonction de répartition associée à la loi ν (respectivement, la fonction de répartition empirique associée à $Z_1,\dots,Z_n$). D'autre part, nous avons vu ci-dessus (exemple 2) que $\mathbf{S}_{\mathscr{A}}(n)=n+1$. Ainsi, en utilisant le théorème de Vapnik-Chervonenkis, on montre que

$$\begin{aligned}\mathbb{P}\Big(\sup_{z\in\mathbb{R}}|F_n(z)-F(z)|>\varepsilon\Big)&=\mathbb{P}\Big(\sup_{A\in\mathscr{A}}|\nu_n(A)-\nu(A)|>\varepsilon\Big)\\&\le 8(n+1)e^{-n\varepsilon^2/32}.\end{aligned}$$

Le lemme de Borel-Cantelli implique alors que

$$\sup_{z\in\mathbb{R}} |F_n(z) - F(z)| \to 0, \quad \mathbb{P}\text{-p.s.},$$

c'est-à-dire que la fonction de répartition empirique converge presque sûrement vers la fonction de répartition, au sens de la convergence uniforme des fonctions. Ce résultat remarquable porte le nom de théorème de Glivenko-Cantelli. Il permet d'approximer la fonction de répartition inconnue F par la fonction de repartition empirique F_n.

Avant de tirer les conséquences du Théorème 3 pour la théorie de l'apprentissage, il convient de préciser quelques propriétés élémentaires de la dimension de Vapnik-Chervonenkis.

2.3 Aspects combinatoires

Nous admettrons le résultat combinatoire suivant (la preuve se fait par récurrence), connu sous le nom de lemme de Sauer :

Théorème 4 (Lemme de Sauer)**.** *Soit $\mathscr{A}$ une famille d'ensembles admettant une dimension de Vapnik-Chervonenkis finie $V_{\mathscr{A}}$. Alors, pour tout $n \geq 1$,*

$$\mathbf{S}_{\mathscr{A}}(n) \leq \sum_{i=0}^{V_{\mathscr{A}}} \binom{n}{i}.$$

Dans la suite, c'est surtout le corollaire suivant qui nous sera utile :

Corollaire 1. *Soit $\mathscr{A}$ une famille d'ensembles admettant une dimension de Vapnik-Chervonenkis finie $V_{\mathscr{A}}$. Alors, pour tout $n \geq 1$,*

$$\mathbf{S}_{\mathscr{A}}(n) \leq (n+1)^{V_{\mathscr{A}}}.$$

Démonstration. On a

$$\begin{aligned}(n+1)^{V_{\mathscr{A}}} = \sum_{i=0}^{V_{\mathscr{A}}} \binom{V_{\mathscr{A}}}{i} n^i &= \sum_{i=0}^{V_{\mathscr{A}}} \frac{n^i V_{\mathscr{A}}!}{i!(V_{\mathscr{A}}-i)!} \\ &\geq \sum_{i=0}^{V_{\mathscr{A}}} \frac{n^i}{i!} \\ &\geq \sum_{i=0}^{V_{\mathscr{A}}} \binom{n}{i} \geq \mathbf{S}_{\mathscr{A}}(n),\end{aligned}$$

où la dernière minoration provient du Lemme de Sauer. □

On déduit en particulier du corollaire précédent qu'un coefficient de pulvérisation tombe forcément dans l'une des deux catégories suivantes :

- ▷ Ou bien $V_{\mathscr{A}} = +\infty$ et dans ce cas $\mathbf{S}_{\mathscr{A}}(n) = 2^n$ pour tout $n \geq 1$.
- ▷ Ou bien $V_{\mathscr{A}} < \infty$ et dans ce cas $\mathbf{S}_{\mathscr{A}}(n) \leq (n+1)^{V_{\mathscr{A}}}$.

On ne peut donc **jamais** avoir des situations intermédiaires, comme par exemple $\mathbf{S}_{\mathscr{A}}(n) \sim 2^{\sqrt{n}}$.

Enfin, en combinant le théorème de Vapnik-Chervonenkis, le Lemme technique 2 et le Corollaire 1, on conclut que pour toute famille d'ensembles mesurables $\mathscr{A}$ de $\mathbb{R}^p$ admettant une dimension de Vapnik-Chervonenkis finie $V_{\mathscr{A}}$,

$$\mathbb{P}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| > 8\varepsilon \Big) \leq 8\mathbf{S}_{\mathscr{A}}(n) e^{-2n\epsilon^2}$$

et donc (par le Lemme 2)

$$\begin{aligned}\mathbb{E}\Big(\sup_{A\in\mathscr{A}} |\nu_n(A) - \nu(A)| \Big) &\leq 8\sqrt{\frac{\log\left(8e\mathbf{S}_{\mathscr{A}}(n)\right)}{2n}} \\ &\leq 8\sqrt{\frac{V_{\mathscr{A}}\log(n+1)+4}{2n}} \\ &= O\left(\sqrt{\frac{V_{\mathscr{A}}\log n}{n}}\right).\end{aligned}$$

Il est à noter qu'il est possible de se débarrasser du terme logarithmique en utilisant des techniques dites de chaînage, dont la présentation dépasse le cadre de ce cours.

2.4 Application à la minimisation du risque empirique

Nous pouvons à présent peaufiner les bornes sur l'erreur d'estimation dans le problème de classification supervisée. Rappelons, pour mémoire, que l'on considère un n-échantillon i.i.d. $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (et indépendant de) $(X, Y) \in \mathbb{R}^d \times \mathcal{Y}$ et une famille $\mathscr{G}$ de règles de décision candidates. En désignant par $g_n^\star$ un minimiseur du risque empirique dans $\mathscr{G}$, nous savons que, d'une part,

$$\mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g) \leq 2 \sup_{g \in \mathscr{G}} \left| \hat{\mathcal{R}}_n(g) - \mathcal{R}(g) \right|$$

et d'autre part,

$$\left\{ \sup_{g \in \mathscr{G}} \left| \hat{\mathcal{R}}_n(g) - \mathcal{R}(g) \right| \right\} = \left\{ \sup_{A \in \mathscr{A}} |\nu_n(A) - \nu(A)| \right\},$$

où, par définition, $\mathscr{A} = \{A_g : g \in \mathscr{G}\}$, avec

$$A_g = \Big\{ (x, y) \in \mathbb{R}^d \times \{0, 1\} : g(x) \neq y \Big\}.$$

Il est alors clair, de par le théorème de Vapnik-Chervonenkis, que le coefficient de pulvérisation $\mathbf{S}_{\mathscr{A}}(n)$ va jouer un rôle fondamental dans le contrôle du terme $\sup_{g \in \mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$. Néanmoins, la classe $\mathscr{A}$, composée de sous-ensembles de $\mathbb{R}^d \times \{0, 1\}$, revêt une structure un peu complexe qui ne se prête pas bien à l'analyse combinatoire. Fort heureusement, les choses se simplifient grâce à la proposition suivante.

Proposition 2. *Soit $\mathscr{A} = \{A_g : g \in \mathscr{G}\}$ et $\bar{\mathscr{A}} = \{\{x \in \mathbb{R}^d : g(x) = 1\} : g \in \mathscr{G}\}$. Alors, pour tout $n \geq 1$, $\mathbf{S}_{\bar{\mathscr{A}}}(n) = \mathbf{S}_{\mathscr{A}}(n)$. En particulier, $V_{\bar{\mathscr{A}}} = V_{\mathscr{A}}$.*

Démonstration. Nous allons montrer que pour tous $z_1, \ldots, z_n \in \mathbb{R}^d$ et tous $e_1, \ldots, e_n \in \{0, 1\}$ fixés,

$$|\mathcal{N}_{\mathscr{A}}((z_1, e_1), (z_2, e_2), \ldots, (z_n, e_n))| = |\mathcal{N}_{\bar{\mathscr{A}}}(z_1, z_2, \ldots z_n)|. \tag{2.4}$$

Sans perte de généralité, on va supposer pour simplifier les notations que $e_1 = 0, e_2 = 0, \ldots, e_k = 0, e_{k+1} = 1, \ldots, e_n = 1$. On note les premières

coordonnées par $V_k = (z_1, \ldots, z_k)$ et $W_{n-k} = (z_{k+1}, \ldots, z_n)$. Soit $g \in \mathscr{G}$. Alors l'intersection de A_g avec $\{(z_1, e_1), (z_2, e_2) \ldots (z_n, e_n)\}$ donne un premier sous-ensemble de points parmi $\{(z_1, e_1), (z_2, e_2), \ldots, (z_k, e_k)\}$ tels que $g(z_i) = 1$ (notons le sous-ensemble de leurs premières coordonnées par $Q_g \subset V_k$) et un second sous-ensemble de points parmi $\{(z_{k+1}, e_{k+1}), \ldots, (z_n, e_n)\}$ tels que $g(z_i) = 0$ (notons le sous-ensemble de leurs premières coordonnées par $R_g \subset W_{n-k}$). Alors l'intersection de $\bar{A}_g = \{x \in \mathbb{R}^d : g(x) = 1\}$ avec $\{z_1, \ldots, z_n\}$ vaut $Q_g \cup (W_{n-k} \setminus R_g)$.

Remarquons que $|\mathcal{N}_{\mathscr{A}}((z_1, e_1), (z_2, e_2), \ldots, (z_n, e_n))| = |\{Q_g \cup R_g : g \in \mathscr{G}\}|$ et que $|\mathcal{N}_{\mathscr{A}}(z_1, z_2, \ldots z_n)| = |\{Q_g \cup (W_{n-k} \setminus Q_g) : g \in \mathscr{G}\}|$. Mais le nombre d'ensembles différents $Q_g \cup R_g$ associés à toutes les fonctions $g \in \mathscr{G}$ est le même que le nombre d'ensembles différents $Q_g \cup (W_{n-k} \setminus R_g)$ associés à toutes les fonctions $g \in \mathscr{G}$. Autrement dit $|\{Q_g \cup R_g : g \in \mathscr{G}\}| = |\{Q_g \cup (W_{n-k} \setminus Q_g) : g \in \mathscr{G}\}|$. En effet : $Q_{g_1} \cup R_{g_1} = Q_{g_2} \cup R_{g_2}$ ssi $Q_{g_1} \cup (W_{n-k} \setminus Q_{g_1}) = Q_{g_2} \cup (W_{n-k} \setminus Q_{g_2})$. L'égalité (2.4) est démontrée. □

Nous sommes désormais en mesure d'énoncer le principal résultat de ce chapitre, dont la preuve découle du Lemme 1, du théorème de Vapnik-Chervonenkis (et du Lemme technique 2 pour la seconde assertion).

Théorème 5. *Soit $V_{\mathscr{A}} < \infty$. On a, pour tout $n \geq 1$,*

$$\mathbb{P}\Big(\big|\mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g)\big| > \varepsilon\Big) \leq 8\mathbf{S}_{\mathscr{A}}(n) e^{-n\varepsilon^2/128}.$$

En outre,

$$\mathbb{E}\mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g) \leq 16\sqrt{\frac{\log\left(8e\mathbf{S}_{\mathscr{A}}(n)\right)}{2n}}.$$

Démonstration. Par le Lemme 1,

$$0 \geq \mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g) \leq 2 \sup_{g \in \mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)|$$

donc par le Théorème 3 et la Proposition 2,

$$\begin{aligned}\mathbb{P}\Big(\big|\mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g)\big| > \varepsilon\Big) &\leq \mathbb{P}\Big(\sup_{g \in \mathscr{G}} |\hat{\mathcal{R}}_n(g) - \mathcal{R}(g)| > \frac{\varepsilon}{2}\Big) \\ &\leq 8\mathbf{S}_{\mathscr{A}}(n) e^{-n\varepsilon^2/(2^2 \times 32)}.\end{aligned}$$

D'après le lemme de Borel-Cantelli, il suit de ce résultat que

$$\mathcal{R}(g_n^\star) - \inf_{g\in\mathcal{G}} \mathcal{R}(g) \to 0, \quad \mathbb{P}\text{-p.s.}$$

dès que la série de terme général $\mathbf{S}_{\mathscr{A}}(n)e^{-n\varepsilon^2/128}$ est sommable. Or, d'après le Corollaire 1, c'est précisément le cas dès que $V_{\mathscr{A}}$ (ou $V_{\mathscr{A}}$) est finie puisqu'alors $\mathbf{S}_{\mathscr{A}}(n)$ a une croissance au plus polynomiale en n. On retiendra donc de tout ceci que la condition $V_{\mathscr{A}} < \infty$ est suffisante pour assurer la convergence presque sûre du terme d'estimation vers 0. Dans ce cas,

$$\mathbb{P}\left(\left|\mathcal{R}(g_n^\star) - \inf_{g\in\mathcal{G}} \mathcal{R}(g)\right| > 2^4\varepsilon\right) \leq 8\mathbf{S}_{\mathscr{A}}(n)\exp(-2n\varepsilon^2)$$

d'où par le Lemme 2,

$$\mathbb{E}\mathcal{R}(g_n^\star) - \inf_{g\in\mathcal{G}} \mathcal{R}(g) \leq 16\sqrt{\frac{\log(8e\mathbf{S}_{\mathscr{A}}(n))}{2n}} \leq 16\sqrt{\frac{\log(8e) + V_{\mathscr{A}}\log(n+1)}{2n}},$$

autrement dit

$$\mathbb{E}\mathcal{R}(g_n^\star) - \inf_{g\in\mathcal{G}} \mathcal{R}(g) = O\left(\sqrt{\frac{V_{\mathscr{A}}\log n}{n}}\right).$$

□

Exemples.

1. **Classification linéaire.** En notant $x = (x^{(1)}, \ldots, x^{(d)})$, on considère des règles de classification très simples, de la forme

$$g(x) = \begin{cases} 1 & \text{si } \sum_{j=1}^d a_j x^{(j)} + a_0 > 0 \\ 0 & \text{sinon,} \end{cases}$$

où $(a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1}$ est un paramètre vectoriel. Chaque fonction g de ce type subdivise l'espace $\mathbb{R}^d$ en deux demi-espaces par la droite $\sum_{j=1}^d a_j x^{(j)} + a_0 = 0$. Pour la variable X on attribue $Y = 1$ si X tombe dans le demi-plan « positif », i.e. $\sum_{j=1}^d a_j X^{(j)} + a_0 \geq 0$, et $Y = 0$ si X tombe dans le demi-plan « négatif » i.e. $\sum_{j=1}^d a_j X^{(j)} + a_0 < 0$. L'ensemble $\mathcal{G}$ est bien sur infini mais de dimension finie.

Dans ce cas,

$$\mathscr{A} \subset \left\{ \{x \in \mathbb{R}^d : a^\intercal x + a_0 \geq 0\} : a \in \mathbb{R}^d, a_0 \in \mathbb{R} \right\}$$

et, d'après les propriétés de la dimension de Vapnik-Chervonenkis vues plus haut, on a $V_{\mathscr{A}} \leq d+1$ (voir exemple 6 ci-dessus). Ainsi le Théorème de Vapnik-Chervonenkis s'applique,

$$\mathbb{P}\left(\left| \mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g) \right| > \varepsilon \right) \leq 8(n+1)^{d+1} e^{-n\varepsilon^2/128}$$

et

$$\mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g) \to 0, \quad \mathbb{P}\text{-p.s.}$$

En pratique, ce théorème nous recommande de chercher $g_n^\star$ qui minimise le risque empirique et nous montre que son erreur d'estimation est très proche de l'erreur d'estimation optimale sur $\mathscr{G}$.

Pour obtenir $g_n^\star$, on peut visualiser dans l'espace $\mathbb{R}^d$ les données $X_1, \ldots, X_n$ en les coloriant avec deux couleurs : X_i en rouge si $Y_i = 1$ et X_i en bleu si $Y_i = 0$. On cherche un hyperplan $\sum_{j=1}^d a_j^\star x^{(j)} + a_0^\star = 0$ qui divise $\mathbb{R}^d$ en deux demi-espaces : celui $\sum_{j=1}^d a_j^\star x^{(j)} + a_0^\star > 0$ contenant le maximum de points rouges, l'autre contenant le maximum d points bleus. La fonction $g_n^\star(X) = 1_{\sum_{j=1}^d a_j^\star x^{(j)} + a_0^\star > 0}(X)$ sera alors la fonction minimisant le risque empirique.

2. **Classification par des boules fermées.** La classe $\mathscr{G}$ est composée de toutes les indicatrices des boules fermées de $\mathbb{R}^d$. Ainsi, la fonction de décision $g(x)$ est l'indicatrice de n'importe quelle boule B_g, i.e. pour la donnée X on attribue $Y = 1$ si X est à l'intérieur de B_g et $Y = 0$ sinon. Dans ce cas,

$$\mathscr{A} = \left\{ \left\{ x \in \mathbb{R}^d : \sum_{j=1}^d |x^{(j)} - a_j|^2 \leq a_0 \right\} : (a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1} \right\}.$$

En remarquant que

$$a_0 - \sum_{j=1}^d |x^{(j)} - a_j|^2 = a_0 - \sum_{j=1}^d \left(x^{(j)}\right)^2 + 2\sum_{j=1}^d x^{(j)} a_j - \sum_{j=1}^d a_j^2,$$

on voit que $\mathscr{A}$ est inclus dans une famille d'ensembles de la forme $\{\{x \in \mathbb{R}^d : f(x) \geq 0\} : f \in \mathscr{F}\}$, où $\mathscr{F}$ est un espace vectoriel de dimension $d+2$. En effet, c'est un espace avec les fonctions $1, x^{(1)}, x^{(2)}$, $\ldots, x^{(d)}, \sum_{j=1}^{d}(x^{(j)})^2$ formant une base. On conclut comme précédemment que

$$\mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g) \to 0, \quad \mathbb{P}\text{-p.s.}$$

En pratique, pour minimiser la probabilité d'erreur, on cherche donc $g_n^\star$ qui minimise le risque empirique parmi toutes les indicatrices de boules. On visualise à nouveau dans l'espace $\mathbb{R}^d$ les données $X_1, \ldots, X_n$ en les coloriant avec deux couleurs : X_i en rouge si $Y_i = 1$ et X_i en bleu si $Y_i = 0$. On essaie de trouver une boule $\mathscr{B}^\star$ de centre $(a_1^\star, \ldots, a_d^\star)$ et de rayon $\sqrt{a_0^\star}$ qui inclut le maximum de points rouges et laisse à l'extérieur le maximum de points bleus ; soit l'inverse en échangeant rouge et bleu. La fonction $g_n^\star(X) = 1_{\mathscr{B}^\star}(X)$ sera la fonction minimisant le risque empirique.

3. **Classification par des convexes.** On prend pour $\mathscr{A}$ l'ensemble de tous les polygones convexes de $\mathbb{R}^2$, famille pour laquelle nous avons déjà vu que $V_{\mathscr{A}} = +\infty$. Cette classe d'ensembles est trop massive pour que l'erreur d'estimation puisse être raisonnablement contrôlée par la théorie de Vapnik-Chervonenkis.
4. **Classification linéaire généralisée.** On se place dans $\mathbb{R}^d$ et on se donne $\psi_1, \ldots, \psi_{d^\star}$ un nombre fixe de fonctions mesurables de $\mathbb{R}^d \to \mathbb{R}$. Les règles de classification considérées sont alors de la forme

$$g(x) = \begin{cases} 1 & \text{si } \sum_{j=1}^{d^\star} a_j \psi_j(x) + a_0 > 0 \\ 0 & \text{sinon,} \end{cases}$$

où $(a_0, a_1, \ldots, a_{d^\star}) \in \mathbb{R}^{d^\star}$ est un paramètre vectoriel. Lorsque $\psi_j(x) = x^{(j)}$, on retrouve la famille des règles linéaires. Néanmoins, bien d'autres choix sont possibles. En prenant par exemple pour les ψ_j les applications coordonnées et produits de ces coordonnées, on voit que $\mathscr{A}$ est contenu dans une famille d'ensembles du type

$$\left\{ a_0 + \sum_{j=1}^{d} a_j x^{(j)} + \sum_{j=1}^{d} b_j \left(x^{(j)}\right)^2 + \sum_{1 \leq j_1 < j_2 \leq d} c_{j_1} c_{j_2} x^{(j_1)} x^{(j_2)} \geq 0 \right\}.$$

Dans ce cas, $d^\star = 1 + 2d + \frac{d(d-1)}{2}$ et par ailleurs $V_{\mathscr{A}} \leq d^\star + 1$, et donc

$$\mathcal{R}(g_n^\star) - \inf_{g \in \mathscr{G}} \mathcal{R}(g) \to 0, \quad \mathbb{P}\text{-p.s.}$$

Chapitre 3

Théorème de Stone et plus proches voisins

Les chapitres précédents ont mis en lumière le rôle essentiel joué par le principe de minimisation du risque empirique pour l'apprentissage supervisé. Dans le cas de la classification, on a également vu le rôle du théorème de Vapnik-Chervonenkis dans le contrôle de l'erreur d'estimation. Il s'avère cependant que les familles de règles de décision admettant une dimension de Vapnik-Chervonenkis finie sont presque toujours trop petites et ne permettent pas d'approcher correctement le risque de Bayes $\mathcal{R}^\star$. On peut par exemple montrer que pour n'importe quelle famille de règles $\mathcal{G}$ dont la classe de boréliens associée $\bar{\mathscr{A}_\mathcal{G}} = \{\{x : g(x) = 1\} : g \in \mathscr{G}\}$ admet une dimension de Vapnik-Chervonenkis finie, et pour tout $\varepsilon \in (0, 1/2)$, il existe un couple de variables aléatoires (X, Y) tel que

$$\inf_{g \in \mathcal{G}} \mathcal{R}(g) - \mathcal{R}^\star > 1/2 - \varepsilon.$$

Il existe cependant d'autres façons de procéder. Une stratégie concurrente de la minimisation du risque empirique consiste à utiliser l'échantillon $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ pour estimer la fonction de régression $r(x) = \mathbb{E}(Y|X = x)$, et la remplacer par son estimateur $r_n(x) = r_n(x, \mathscr{D}_n)$ dans la règle de classification.

Dans ce chapitre, nous revenons au cadre général de l'apprentissage supervisé, avec un couple de variables aléatoires (X, Y) dans $\mathbb{R}^d \times \mathcal{Y}$. Le cas $\mathcal{Y} = \{0, 1\}$ (ou plus généralement un ensemble fini) correspond au problème de classification tandis que $\mathcal{Y} = \mathbb{R}$ correspond à la régression. Nous nous plaçons toujours dans le cadre où on veut prédire Y à partir de X, en

utilisant pour cela un n-échantillon de variables $(X_1, Y_1), \ldots, (X_n; Y_n)$ i.i.d. de même loi que (X, Y) (et indépendantes de celles-ci).

3.1 Liens entre classification et régression

Considérons tout d'abord le cas $\mathcal{Y} = \{0, 1\}$ de la classification binaire. Pour faire le lien avec le chapitre précédent, partons de la caractérisation du classifieur de Bayes

$$g^\star(x) = \begin{cases} 1 & \text{si } r(x) > 1/2 \\ 0 & \text{sinon,} \end{cases} \tag{3.1}$$

où $r(x) = \mathbb{P}(Y = 1 \mid X = x) = \mathbb{E}(Y \mid X = x)$. Nous allons donc utiliser l'échantillon $\mathscr{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ pour estimer la fonction de régression, et la remplacer par son estimateur $r_n(x, \mathscr{D}_n)$ dans (3.1). La règle de classification résultante, dite règle plug-in, s'écrit donc naturellement

$$g_n(x) = \begin{cases} 1 & \text{si } r_n(x) > 1/2 \\ 0 & \text{sinon.} \end{cases}$$

Le théorème qui suit précise le lien entre g_n et r_n, en termes d'erreurs

$$\mathcal{R}(g_n) = \mathbb{P}(g_n(X) \neq Y \mid \mathscr{D}_n), \quad \mathcal{R}^\star = \mathbb{P}(g^\star(X) \neq Y).$$

On rappelle que μ désigne la loi de la variable aléatoire X.

Théorème 6. *Soit r_n un estimateur de la fonction de régression et g_n la règle de décision plug-in associée. Alors*

$$0 \leq \mathcal{R}(g_n) - \mathcal{R}^\star \leq 2 \int_{\mathbb{R}^d} |r_n(x) - r(x)| \, \mu(\mathrm{d}x).$$

En particulier, pour tout $p \geq 1$,

$$0 \leq \mathcal{R}(g_n) - \mathcal{R}^\star \leq 2 \left(\int_{\mathbb{R}^d} |r_n(x) - r(x)|^p \, \mu(\mathrm{d}x) \right)^{1/p},$$

et

$$0 \leq \mathbb{E}\mathcal{R}(g_n) - \mathcal{R}^\star \leq 2 \big(\mathbb{E} \, |r_n(X) - r(X)|^p \big)^{1/p}.$$

Démonstration. Remarquons que

$$\mathbb{1}_{g_n(X)\neq Y} = \mathbb{1}_{g_n(X)=1}\mathbb{1}_{Y=0} + \mathbb{1}_{g_n(X)=0}\mathbb{1}_{Y=1},$$

d'où il vient

$$\begin{aligned}\mathbb{P}(g_n(X) \neq Y|X, \mathcal{D}_n) &= \mathbb{1}_{g_n(X)=1}\mathbb{P}(Y=0|X,\mathcal{D}_n) + \mathbb{1}_{g_n(X)=0}\mathbb{P}(Y=1|X,\mathcal{D}_n)\\ &= \mathbb{1}_{g_n(X)=1}(1-r(X)) + \mathbb{1}_{g_n(X)=0}r(X)\end{aligned}$$

où, dans la dernière égalité, nous avons utilisé l'indépendance entre le couple (X, Y) et $\mathcal{D}_n$. De façon similaire, (on rappelle que $g^\star$ est déterministe)

$$\mathbb{P}(g^\star(X) \neq Y|X) = \mathbb{1}_{g^\star(X)=1}(1-r(X)) + \mathbb{1}_{g^\star(X)=0}r(X).$$

Ainsi,

$$\begin{aligned}&\mathbb{P}(g_n(X) \neq Y|X, \mathcal{D}_n) - \mathbb{P}(g^\star(X) \neq Y|X)\\ &\quad= r(X)(\mathbb{1}_{g_n(X)=0} - \mathbb{1}_{g^\star(X)=0}) + \big(1-r(X)\big)(\mathbb{1}_{g_n(X)=1} - \mathbb{1}_{g^\star(X)=1})\\ &\quad= (2r(X)-1)(\mathbb{1}_{g_n(X)=0} - \mathbb{1}_{g^\star(X)=0})\\ &\quad= |2r(X)-1|\mathbb{1}_{g_n(X)\neq g^\star(X)}.\end{aligned}$$

En effet, remarquons que $\mathbb{1}_{g_n(X)=0} - \mathbb{1}_{g^\star(X)=0} = -(\mathbb{1}_{g_n(X)=1} - \mathbb{1}_{g^\star(X)=1})$. Par ailleurs, si $\mathbb{1}_{g_n(X)=0} \neq \mathbb{1}_{g^\star(X)=0}$ alors soit $\mathbb{1}_{g_n(X)=0} = 1$ et $\mathbb{1}_{g^\star(X)=0} = 0$, auquel cas $g^\star(X) = 1$ et par définition on a $r(X) > 1/2$ et donc $2r(X) - 1 > 0$; soit $\mathbb{1}_{g_n(X)=0} = 0$ et $\mathbb{1}_{g^\star(X)=0} = 1$ et auquel cas par définition de $g^\star$ on a $r(X) \leq 1/2$ et donc $2r(X) - 1 \leq 0$.

Finalement,

$$\begin{aligned}\mathbb{P}(g_n(X) \neq Y|\mathcal{D}_n) - \mathcal{R}^\star &= \mathbb{E}\Big[\mathbb{P}(g_n(X) \neq Y|X,\mathcal{D}_n) - \mathbb{P}(g^\star(X) \neq Y|X)\Big]\\ &= 2\int_{\mathbb{R}^d} |r(x) - 1/2|\mathbb{1}_{g_n(x)\neq g^\star(x)}\mu(\mathrm{d}x)\\ &\leq 2\int_{\mathbb{R}^d} |r_n(x) - r(x)|\,\mu(\mathrm{d}x),\end{aligned}$$

puisque $g_n(x) \neq g^\star(x)$ implique $|r_n(x) - r(x)| \geq |r(x) - 1/2|$. En effet si $g_n(x) \neq g^\star(x)$ soit $r_n(x) > 1/2$ et $r(x) \leq 1/2$, soit $r_n(x) \leq 1/2$ et $r(x) > 1/2$. Dans les deux cas $|r_n(x) - r(x)| \geq |r(x) - 1/2|$.

La 2ème assertion découle de la première par l'inégalité de Hölder.

Pour déduire la 3ème assertion du théorème, on remarque que

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)| \mu(dx) = \mathbb{E}(|r_n(X) - r(X)| \mid \mathscr{D}_n).$$

On prend l'espérance des deux parties de la première assertion du théorème et on obtient :

$$0 \leq \mathbb{E}\mathcal{R}(g_n) - L^\star \leq 2\mathbb{E}|r_n(X) - r(X)|.$$

Il reste à appliquer ensuite l'inégalité de Jensen qui donne

$$\mathbb{E}|r_n(X) - r(X)| \leq (\mathbb{E}|r_n(X) - r(X)|^p)^{1/p}.$$

□

Remarque : On retiendra du Théorème 6 que si l'on dispose d'un estimateur r_n de la fonction de régression qui soit tel que

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \, \mu(\mathrm{d}x) \to 0 \tag{3.2}$$

dans $\mathbb{L}^1(\mathbb{P})$ (ou $\mathbb{P}$-presque sûrement), alors la règle de classification associée g_n est automatiquement convergente par l'assertion 3 pour $p = 2$ (ou fortement convergente par l'assertion 2 pour $p = 2$). (Rappel : voir la définition 1 pour la convergence et la convergence forte.) Ce sera le point de départ de la preuve du théorème de Stone.

Il nous reste donc à savoir comment construire des estimateurs de la fonction de régression qui possèdent la propriété de convergence (3.2) ; c'est l'objet du théorème de Stone.

3.2 Le théorème de Stone

Dans cette section, $\mathcal{Y} = \{0, 1\}$ ou $\mathbb{R}$ et nous allons construire des estimateurs de la fonction de régression $r(x) = \mathbb{E}(Y|X = x)$ à partir du n-échantillon $\mathscr{D}_n$. Une façon canonique de procéder consiste à écrire

$$r_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \quad x \in \mathbb{R}^d, \tag{3.3}$$

où chaque

$$W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$$

est une fonction borélienne réelle de x et $X_1, \ldots, X_n$ (pas $Y_1, \ldots, Y_n$).

Il est intuitivement clair que les couples (X_i, Y_i) pour lesquels X_i est « proche » de x (en un sens qui reste à préciser) devraient apporter davantage d'information sur $r(x)$ que leurs homologues plus éloignés. En conséquence, les poids W_{ni} devront en règle générale être plus grands autour de x, de telle sorte que $r_n(x)$ ainsi défini se présente comme une moyenne pondérée des Y_i correspondants aux X_i situés dans un voisinage de x. Voilà pourquoi un estimateur r_n de la forme (3.3) est appelé estimateur **de type moyenne locale**. Bien souvent (mais pas toujours), les $W_{ni}(x)$ sont positifs et normalisés à 1, de telle sorte que $(W_{n1}(x), \ldots, W_{nn}(x))$ est en fait un vecteur de probabilités.

Un exemple typique d'estimateur de type moyenne locale est l'estimateur à noyau, qui est obtenu en prenant

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)},$$

où K est une fonction positive mesurable sur $\mathbb{R}^d$ avec le maximum en 0 (appelée noyau) et h est un paramètre strictement positif (appelé fenêtre), en pratique fonction de n. (Si le dénominateur est nul, on pose $W_{ni}(x) = 1/n$.) En d'autres termes, pour $x \in \mathbb{R}^d$, l'estimateur à noyau de la fonction de régression r, appelé estimateur de **Nadaraya-Watson**, est donné par

$$r_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}. \tag{3.4}$$

(Si le dénominateur est nul, on pose $r_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i$.) En particulier, pour le choix de noyau dit naïf $K(z) = \mathbb{1}_{\|z\| \leq 1}$, on obtient

$$r_n(x) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\|x-X_i\| \leq h} Y_i}{\sum_{j=1}^{n} \mathbb{1}_{\|x-X_j\| \leq h}},$$

ce qui montre que $r(x)$ est estimé par la moyenne des Y_i tels que la distance euclidienne entre x et X_i ne dépasse pas h. Pour un noyau plus général K,

le poids de Y_i dépend de la distance entre x et X_i par l'intermédiaire de la forme du noyau. Les noyaux les plus classiques sont le noyau d'Epanechnikov $K(z) = (1 - \|z\|^2)\mathbb{1}_{\|z\|\leq 1}$ et le noyau gaussien $K(z) = e^{-\|z\|^2}$.

Un second exemple important d'estimateur de type moyenne locale nous est fourni par l'estimateur des **plus proches voisins** :

$$r_n(x) = \sum_{i=1}^{n} v_{ni} Y_{(i)}(x), \quad x \in \mathbb{R}^d,$$

pour lequel $(v_{n1}, \ldots, v_{nn})$ est un vecteur de poids déterministes normalisés à 1, et la suite $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ est la permutation de $(X_1, Y_1), \ldots, (X_n, Y_n)$ correspondante aux distances croissantes des $\|X_i - x\|$ (en cas d'égalité $\|X_i - x\| = \|X_j - x\|$ avec $i < j$, X_i sera arbitrairement déclaré plus proche de x que X_j). En d'autres termes,

$$\|X_{(1)}(x) - x\| \leq \|X_{(2)}(x) - x\| \leq \cdots \leq \|X_{(n)}(x) - x\|.$$

Pour s'assurer que cet estimateur est bien de la forme (3.3), il suffit de poser

$$W_{ni}(x) = v_{n\sigma_i(x, X_1, \ldots, X_n)},$$

où $(\sigma_1(x, X_1, \ldots, X_n), \sigma_2(x, X_1, \ldots, X_n) \ldots, \sigma_n(x, X_1, \ldots, X_n))$ est la permutation de $(1, \ldots, n)$ telle que X_i est le σ_i-ème plus proche voisin de x.

Parmi tous les choix possibles de vecteurs de poids $(v_{n1}, \ldots, v_{nn})$, un cas particulier important est obtenu en posant $v_{ni} = 1/k$ pour $1 \leq i \leq k$ et $v_{ni} = 0$ autrement, avec $\{k\} = \{k_n\}$ une suite d'entiers strictement positifs ne dépassant pas n. L'estimateur résultant s'appelle estimateur des k**-plus proches voisins** et s'écrit donc

$$r_n(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(x), \quad x \in \mathbb{R}^d.$$

Le principe de cet estimateur est naturel : pour estimer la fonction de régression autour de x, on regarde les k observations X_i les plus proches de x et on fait la moyenne des Y_i correspondants.

Le théorème ci-après, connu sous le nom de théorème de Stone, donne des conditions suffisantes sur les poids $W_{ni}(x)$ garantissant que la fonction de régression de type moyenne locale vérifie la convergence (3.2) dans $\mathbb{L}^1(\mathbb{P})$

dès que la loi de (X, Y) vérifie $\mathbb{E}Y^2 < +\infty$. Pour simplifier, nous supposerons désormais que les poids $W_{ni}(x)$ sont positifs et normalisés à 1 (i.e., $\sum_{i=1}^n W_{ni}(x) = 1$), ce qui fait de $(W_{n1}(x), \ldots, W_{nn}(x))$ un vecteur de probabilités.

Théorème 7 (Stone). *Supposons que, quelle que soit la loi de X, les poids W_{ni} satisfont les 3 conditions suivantes :*

1. *Il existe une constante c telle que, pour toute fonction borélienne $f : \mathbb{R}^d \to \mathbb{R}$ telle que $\mathbb{E}|f(X)| < \infty$,*

$$\mathbb{E}\left(\sum_{i=1}^n W_{ni}(X) \, |f(X_i)| \right) \leq c\mathbb{E} \, |f(X)| \,, \quad \textit{pour tout } n \geq 1.$$

2. *Pour tout $a > 0$,*

$$\mathbb{E}\left(\sum_{i=1}^n W_{ni}(X) \mathbb{1}_{\|X_i - X\| > a} \right) \to 0.$$

3. *On a*

$$\mathbb{E}\left(\max_{1 \leq i \leq n} W_{ni}(X) \right) \to 0.$$

Alors, l'estimateur r_n de la régression défini en (3.3) satisfait

$$\mathbb{E} \, (r_n(X) - r(X))^2 = \mathbb{E} \int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(\mathrm{d}x) \to 0,$$

quelle que soit la loi du couple (X, Y), dès que $\mathbb{E}Y^2 < +\infty$.

La condition 2 exprime le fait que la contribution des poids à l'extérieur de n'importe quelle boule fermée centrée en X doit être asymptotiquement négligeable. En d'autres termes, seuls les points situés dans un voisinage local de la cible sont importants pour l'évaluation de la moyenne. La condition 3 interdit à un seul point d'avoir une influence disproportionnée sur le calcul de l'estimateur. Enfin, l'hypothèse 1, parfois appelée condition de Stone, est essentiellement de nature technique. Insistons bien sur le fait que le résultat du théorème est *universel*, au sens où la convergence est valable *quelle que soit la loi* du couple (X, Y), dès que $\mathbb{E}Y^2 < +\infty$. En particulier en classification $\mathcal{Y} = \{0, 1\}$, noter que cette dernière condition d'ordre est toujours satisfaite.

Démonstration du Théorème 7. On introduit

$$\widetilde{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) r(X_i).$$

(Notez que ce n'est pas un estimateur, c'est une quantité qui dépend de la fonction de régression inconnue r.) En utilisant l'inégalité $(a+b)^2 \leq 2(a^2+b^2)$, on a

$$\begin{aligned}\mathbb{E}\left(r_n(X) - r(X)\right)^2 &= \mathbb{E}\left(r_n(X) - \widetilde{r}_n(X) + \widetilde{r}_n(X) - r(X)\right)^2 \\ &\leq 2\Big(\mathbb{E}\left(r_n(X) - \widetilde{r}_n(X)\right)^2 + \mathbb{E}\left(\widetilde{r}_n(X) - r(X)\right)^2\Big). \quad (3.5)\end{aligned}$$

Il suffit donc de montrer que chacun des deux termes de la borne ci-dessus tend vers 0 lorsque n tend vers l'infini. Comme les poids $W_{ni}(x)$ sont positifs et normalisés ($\sum_{i=1}^{n} W_{n,i}(x) = 1$), l'inégalité de Jensen permet d'écrire que

$$\begin{aligned}\mathbb{E}\left(\widetilde{r}_n(X) - r(X)\right)^2 &= \mathbb{E}\Big(\sum_{i=1}^{n} W_{ni}(X)\left(r(X_i) - r(X)\right)\Big)^2 \\ &\leq \mathbb{E}\sum_{i=1}^{n} W_{ni}(X)\left(r(X_i) - r(X)\right)^2.\end{aligned}$$

Par densité dans $\mathbb{L}^2(\mu)$ des fonctions continues à support compact, pour tout $\varepsilon > 0$, on peut trouver r' continue à support compact telle que

$$\mathbb{E}(r(X) - r'(X))^2 = \int |r(x) - r'(x)|^2 \mu(\mathrm{d}x) \leq \varepsilon.$$

Alors, avec $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, on a

$$\begin{aligned}\mathbb{E}\sum_{i=1}^{n} W_{ni}(X)(r(X_i) - r(X))^2 \leq\ & 3\mathbb{E}\sum_{i=1}^{n} W_{ni}(X)(r(X_i) - r'(X_i))^2 \\ &+ 3\mathbb{E}\sum_{i=1}^{n} W_{ni}(X)(r'(X_i) - r'(X))^2 \\ &+ 3\mathbb{E}\sum_{i=1}^{n} W_{ni}(X)(r'(X) - r(X))^2.\end{aligned}$$

En utilisant la première condition et $\sum_{i=1}^n W_{ni}(X) = 1$, ceci implique

$$\begin{aligned}
&\mathbb{E}\sum_{i=1}^n W_{ni}(X)(r(X_i) - r(X))^2 \\
&\quad \leq 3c\mathbb{E}(r(X) - r'(X))^2 + 3\mathbb{E}\sum_{i=1}^n W_{ni}(X)(r'(X_i) - r'(X))^2 + 3\mathbb{E}(r'(X) - r(X))^2 \\
&\quad \leq 3(c+1)\varepsilon + 3\mathbb{E}\sum_{i=1}^n W_{ni}(X)(r'(X_i) - r'(X))^2.
\end{aligned}$$

Considérons le second terme de droite. Puisque r' est continue à support compact, elle est aussi uniformément continue. Ainsi, il existe $\rho > 0$ tel que $\|x - x'\| \leq \rho$ implique $|r'(x) - r'(x')|^2 \leq \varepsilon$. Par ailleurs, r' est aussi bornée. Ainsi,

$$\begin{aligned}
&\mathbb{E}\sum_{i=1}^n W_{ni}(X)(r'(X_i) - r'(X))^2 \\
&\quad \leq 4\|r'\|_\infty^2 \mathbb{E}\left(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{\|X_i - X\| > \rho}\right) + \mathbb{E}\left(\sum_{i=1}^n W_{ni}(X)\varepsilon\right) \\
&\quad = 4\|r'\|_\infty^2 \mathbb{E}\left(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{\|X_i - X\| > \rho}\right) + \varepsilon.
\end{aligned}$$

Ainsi, d'après la condition 2, puisque ε est arbitraire, le terme ci-dessus peut-être rendu arbitrairement petit et on obtient

$$\mathbb{E}\left(\sum_{i=1}^n W_{ni}(X)\,(r(X_i) - r(X))^2\right) \to 0$$

ce qui implique $\mathbb{E}\,(\widetilde{r}_n(X) - r(X))^2 \to 0$.

Il nous reste à contrôler le premier terme du membre de droite de l'inégalité

(3.5). Observons pour cela que, pour $i \neq j$,

$$\begin{aligned}
&\mathbb{E}\left(W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\big(Y_j - r(X_j)\big)\right) \\
&\quad = \mathbb{E}\Big[\mathbb{E}\left(W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\big(Y_j - r(X_j)\big)|X, X_1, \ldots, X_n, Y_i\right)\Big] \\
&\quad = \mathbb{E}\Big[W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\mathbb{E}(Y_j - r(X_j)|X, X_1, \ldots, X_n, Y_i)\Big] \\
&\quad = \mathbb{E}\Big[W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\mathbb{E}(Y_j - r(X_j)|X_j)\Big] \\
&\qquad \text{(par indépendence entre } (X_j, Y_j) \text{ et } X, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n, Y_i) \\
&\quad = \mathbb{E}\Big[W_{ni}(X)\big(Y_i - r(X_i)\big)W_{nj}(X)\big(r(X_j) - r(X_j)\big)\Big] \\
&\quad = 0.
\end{aligned}$$

Du coup,

$$\begin{aligned}
\mathbb{E}\left(r_n(X) - \widetilde{r}_n(X)\right)^2 &= \mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X)\left(Y_i - r(X_i)\right)\right)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left(W_{ni}(X)\left(Y_i - r(X_i)\right)W_{nj}(X)\left(Y_j - r(X_j)\right)\right) \\
&= \sum_{i=1}^{n}\mathbb{E}\left(W_{ni}^2(X)\left(Y_i - r(X_i)\right)^2\right).
\end{aligned}$$

On note

$$\sigma^2(x) = \mathbb{E}((Y - r(X))^2|X = x)$$

et on remarque que puisque $\mathbb{E}Y^2 < +\infty$, on a $\mathbb{E}\sigma^2(X) = \mathbb{E}(Y - r(X))^2 = \mathbb{E}Y^2 - \mathbb{E}(r(X))^2 < +\infty$. Ainsi,

$$\begin{aligned}
\mathbb{E}\left(r_n(X) - \widetilde{r}_n(X)\right)^2 &= \sum_{i=1}^{n}\mathbb{E}\left(W_{ni}^2(X)\mathbb{E}\Big[\left(Y_i - r(X_i)\right)^2|X_i\Big]\right) \\
&= \sum_{i=1}^{n}\mathbb{E}\left(W_{ni}^2(X)\sigma^2(X_i)\right).
\end{aligned}$$

Si $\sigma^2(\cdot)$ est une fonction bornée, on conclut en utilisant la condition 3 via :

$$\begin{aligned}\mathbb{E}\left(r_n(X)-\widetilde{r}_n(X)\right)^2 &\leq \|\sigma^2\|_\infty \mathbb{E}\left(\sum_{i=1}^n W_{ni}^2(X)\right)\\ &\leq \|\sigma^2\|_\infty \mathbb{E}\left(\max_{1\leq i\leq n} W_{ni}(X)\sum_{j=1}^n W_{nj}(X)\right)\\ &= \|\sigma^2\|_\infty \mathbb{E}\left(\max_{1\leq i\leq n} W_{ni}(X)\right)\to 0.\end{aligned}$$

Sinon, puisque $\sigma^2 \in \mathbb{L}^1(\mu)$, en utilisant à nouveau un argument de densité, pour tout $\varepsilon > 0$, il existe une fonction continue bornée $\tilde{\sigma}^2$ telle que

$$\mathbb{E}|\tilde{\sigma}^2(X)-\sigma^2(X)|\leq \varepsilon.$$

Par ailleurs,

$$\begin{aligned}\mathbb{E}\left(r_n(X)-\widetilde{r}_n(X)\right)^2 &= \sum_{i=1}^n \mathbb{E}\left(W_{ni}^2(X)\sigma^2(X_i)\right)\\ &\leq \sum_{i=1}^n \mathbb{E}\left(W_{ni}^2(X)\tilde{\sigma}^2(X_i)\right)+\sum_{i=1}^n \mathbb{E}\left(W_{ni}^2(X)|\sigma^2(X_i)-\tilde{\sigma}^2(X_i)|\right)\end{aligned}$$

et on utilise la première condition pour traiter le second terme. ceci termine la preuve. □

3.3 Estimateur de Nadaraya-Watson pour la régression

Dans cette section, on se place dans le cadre de la régression, avec $\mathcal{Y} = \mathbb{R}$. Nous allons nous intéresser à l'estimateur à noyau de la fonction de régression (3.4) et utiliser le théorème de Stone (théorème 7) pour prouver la convergence universelle de cet estimateur sous des conditions générales sur le noyau K et la fenêtre h. On se place dans la cas d'un noyau à support compact, c'est le cas par exemple du noyau d'Epanechnikov. La condition que K est non nul sur une boule au voisinage de 0 est faible.

Théorème 8. *On suppose que $K : \mathbb{R}^d \to \mathbb{R}$ est à support compact et qu'il existe $\eta > 0, b > 0$ tels que $K(x) > b\mathbb{1}_{x \in B(0,\eta)}$. Alors, si $h_n \to 0$ avec $nh_n^d \to +\infty$, l'estimateur à noyau (3.4) est universellement consistant pour le risque quadratique intégré, i.e.*

$$\mathbb{E}(r_n(X) - r(X))^2 = \mathbb{E}\int_{\mathbb{R}^d}(r_n(x) - r(x))^2\mu(\mathrm{d}x) \to 0,$$

quelle que soit la loi du couple (X, Y), dès que $\mathbb{E}Y^2 < +\infty$.

Démonstration. Pour prouver le résultat, il suffit de vérifier que les conditions $1-3$ du théorème de Stone (Théorème 7) sont satisfaites sous les conditions du théorème. On note $K_h(\cdot) = K(\cdot/h)$. On rappelle que les poids W_{ni} sont définis par

$$W_{ni}(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^n K_h(x - X_j)}.$$

Commençons par vérifier la première condition du théorème de Stone. Soit f une fonction borélienne telle que $\mathbb{E}|f(X)| < +\infty$. Alors

$$\begin{aligned}
\mathbb{E}\sum_{i=1}^n W_{ni}(X)|f(X_i)| &= \mathbb{E}\sum_{i=1}^n \frac{K_h(X - X_i)}{\sum_{j=1}^n K_h(X - X_j)}|f(X_i)| \\
&= n\mathbb{E}\frac{K_h(X - X_1)}{\sum_{j=1}^n K_h(X - X_j)}|f(X_1)| \\
&= n\int_{\mathbb{R}^d}|f(u)|\mathbb{E}\frac{K_h(X - u)}{K_h(X - u) + \sum_{j=2}^n K_h(X - X_j)}\mu(\mathrm{d}u) \\
&= n\int_{\mathbb{R}^d}|f(u)|\mathbb{E}\int_{\mathbb{R}^d}\frac{K_h(x - u)}{K_h(x - u) + \sum_{j=2}^n K_h(x - X_j)}\mu(\mathrm{d}x)\mu(\mathrm{d}u).
\end{aligned}$$

Il suffit donc de montrer qu'il existe $c > 0$ tel que pour tout $u \in \mathbb{R}^d$,

$$\mathbb{E}\int_{\mathbb{R}^d}\frac{K_h(x - u)}{K_h(x - u) + \sum_{j=2}^n K_h(x - X_j)}\mu(\mathrm{d}x) \leq \frac{c}{n}.$$

Puisque le support de K est compact, il peut être recouvert par une union finie de boules $(B(c_i, \eta/2))_{1 \leq i \leq L}$. Alors, pour tout $x \in \mathbb{R}^d, u \in \mathbb{R}^d$,

$$\begin{aligned}
K_h(x - u) &= \sum_{i=1}^L K_h(x - u)\mathbb{1}\Big\{\frac{x - u}{h} \in B(c_i, \eta/2)\Big\} \\
&= \sum_{i=1}^L K_h(x - u)\mathbb{1}\{x \in u + hc_i + B(0, h\eta/2)\}.
\end{aligned}$$

On obtient

$$\begin{aligned}
&\mathbb{E}\int_{\mathbb{R}^d} \frac{K_h(x-u)}{K_h(x-u)+\sum_{j=2}^n K_h(x-X_j)}\mu(\mathrm{d}x)\\
&= \sum_{i=1}^{L}\mathbb{E}\int_{u+hc_i+B(0,h\eta/2)} \frac{K_h(x-u)}{K_h(x-u)+\sum_{j=2}^n K_h(x-X_j)}\mu(\mathrm{d}x)\\
&\leq \sum_{i=1}^{L}\mathbb{E}\int_{u+hc_i+B(0,h\eta/2)} \frac{1}{1+\sum_{j=2}^n K_h(x-X_j)/\|K\|_\infty}\mu(\mathrm{d}x).
\end{aligned}$$

Par ailleurs, si $x \in u+hc_i+B(0,h\eta/2)$ alors $u+hc_i+B(0,h\eta/2) \subset x+B(0,h\eta)$ (faire un dessin !) et par hypothèse $K(x) > b\mathbb{1}_{x\in B(0,\eta)}$ d'où il vient

$$\begin{aligned}
&\mathbb{E}\int_{\mathbb{R}^d} \frac{K_h(x-u)}{K_h(x-u)+\sum_{j=2}^n K_h(x-X_j)}\mu(\mathrm{d}x)\\
&\leq \sum_{i=1}^{L}\mathbb{E}\int_{u+hc_i+B(0,h\eta/2)} \frac{1}{1+b/\|K\|_\infty\sum_{j=2}^n \mathbb{1}\{X_j\in x+B(0,h\eta)\}}\mu(\mathrm{d}x)\\
&\leq \sum_{i=1}^{L}\mathbb{E}\int_{u+hc_i+B(0,h\eta/2)} \frac{1}{1+b/\|K\|_\infty\sum_{j=2}^n \mathbb{1}\{X_j\in u+hc_i+B(0,h\eta/2)\}}\mu(\mathrm{d}x)\\
&\leq \frac{\|K\|_\infty}{b}\sum_{i=1}^{L}\mathbb{E}\frac{\mu(u+hc_i+B(0,h\eta/2))}{1+\sum_{j=2}^n \mathbb{1}\{X_j\in u+hc_i+B(0,h\eta/2)\}},
\end{aligned}$$

(puisque $b/\|K\|_\infty \leq 1$). On utilise alors le lemme technique suivant : Si $U \sim \mathcal{B}(n,p)$ alors

$$\mathbb{E}\frac{1}{1+U} \leq \frac{1}{(n+1)p}.$$

En effet,

$$\begin{aligned}
\mathbb{E}\frac{1}{1+U} &= \sum_{k=0}^{n}\frac{1}{k+1}\binom{n}{k}p^k(1-p)^{n-k}\\
&= \frac{1}{(n+1)p}\sum_{k=0}^{n}\binom{n+1}{k+1}p^{k+1}(1-p)^{n-k}\\
&\leq \frac{1}{(n+1)p}\sum_{j=0}^{n+1}\binom{n+1}{j}p^{j}(1-p)^{n-j+1}\\
&= \frac{1}{(n+1)p}(p+(1-p))^{n+1} = \frac{1}{(n+1)p}.
\end{aligned}$$

Ainsi, la variable aléatoire $\sum_{j=2}^{n} \mathbb{1}\{X_j \in u + hc_i + B(0, h\eta/2)\}$ suit une loi $\mathcal{B}(n-1, p)$ avec $p = \mu(u + hc_i + B(0, h\eta/2))$ donc

$$\begin{aligned}\mathbb{E}\int_{\mathbb{R}^d} \frac{K_h(x-u)}{K_h(x-u)+\sum_{j=2}^{n} K_h(x-X_j)}\mu(\mathrm{d}x) \\ \leq \quad \frac{\|K\|_\infty}{b}\sum_{i=1}^{L}\frac{\mu(u+hc_i+B(0,h\eta/2))}{n\mu(u+hc_i+B(0,h\eta/2))} = \frac{L\|K\|_\infty}{bn},\end{aligned}$$

ce qui termine la vérification de la première condition du théorème de Stone.

La seconde condition de ce théorème est vraie puisque K est à support compact donc pour tout $a > 0$, dès que $h = h_n$ est assez petit,

$$\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\|X_i - X\| > a} = \frac{\sum_{i=1}^{n} K_h(X-X_i)\mathbb{1}_{\|X_i-X\|>a}}{\sum_{j=1}^{n} K_h(X-X_j)} = 0.$$

Vérifions enfin la troisième condition du théorème. Si $\sum_{j=1}^{n} K_h(X-X_j) = 0$ p.s, alors on rappelle que $W_{ni}(X) = 1/n$ pour tout $1 \leq i \leq n$, et la condition est triviale. Sinon, on choisit $M > 0$ et

$$\begin{aligned}\mathbb{E}\max_{1\leq i\leq n} W_{ni}(X) &= \mathbb{E}\max_{1\leq i\leq n}\frac{K_h(X-X_i)}{\sum_{j=1}^{n} K_h(X-X_j)} \\ &\leq \frac{\|K\|_\infty}{b}\mathbb{E}\frac{1}{\sum_{j=1}^{n}\mathbb{1}_{\|X-X_j\|\in B(0,h\eta)}}\mathbb{1}_{\sum_{j=1}^{n}\mathbb{1}_{\|X-X_j\|\in B(0,h\eta)}>0} \\ &\leq \frac{\|K\|_\infty}{b}\left[\mathbb{E}\frac{\mathbb{1}_{X\in B(0,M)}\mathbb{1}_{\sum_{j=1}^{n}\mathbb{1}_{\|X-X_j\|\in B(0,h\eta)}>0}}{\sum_{j=1}^{n}\mathbb{1}_{\|X-X_j\|\in B(0,h\eta)}} + \mu(B(0,M)^c)\right].\end{aligned}$$

Le premier élément du terme entre crochets s'écrit

$$\mathbb{E}\Big[\mathbb{1}_{X\in B(0,M)}\mathbb{E}\Big(\frac{\mathbb{1}_{\sum_{j=1}^{n}\mathbb{1}_{X_j\in B(X,h\eta)}}}{\sum_{j=1}^{n}\mathbb{1}_{X_j\in B(X,h\eta)}}|X\Big)\Big] = \mathbb{E}\Big[\mathbb{1}_{X\in B(0,M)}\mathbb{E}\Big(\frac{\mathbb{1}_{U_X>0}}{U_X}|X\Big)\Big].$$

Or $U_X = \sum_{j=1}^{n}\mathbb{1}_{X_j\in B(X,h\eta)}$ est une variable aléatoire de loi conditionnelle à X, binomiale $\mathcal{B}(n,p)$ avec $p = \mu(B(X,h\eta))$. On peut montrer comme précédemment (exercice) que

$$\mathbb{E}\Big(\frac{\mathbb{1}_{U_X>0}}{U_X}|X\Big) \leq \frac{2}{(n+1)\mu(B(X,h\eta))}.$$

On en déduit

$$\mathbb{E} \max_{1\leq i\leq n} W_{ni}(X) \leq \frac{\|K\|_\infty}{b}\left[\mathbb{E}\frac{2\mathbb{1}_{X\in B(0,M)}}{(n+1)\mu(B(X,h\eta))} + \mu(B(0,M)^c)\right].$$

Pour tout $\varepsilon > 0$, on peut choisir $M > 0$ assez grand pour avoir $\mu(B(0,M)^c) \leq \varepsilon$. Pour finir, il faut donc vérifier que

$$\mathbb{E}\frac{\mathbb{1}_{X\in B(0,M)}}{(n+1)\mu(B(X,h\eta))}$$

tend vers 0. Pour tout $n \geq 1$, la boule $B(0,M)$ peut être recouverte par un nombre $k_n \leq C/h_n^d$ de boules de rayon $h\eta/2 = h_n\eta/2$, i.e.

$$B(0,M) \subset \cup_{i=1}^{k_n} B(c_i, h_n\eta/2), \quad k_n \leq \frac{C}{h_n^d}.$$

(Pour le voir, on peut raisonner sur la norme uniforme, faire une grille de pas $1/h_n$ dans chacune des dimensions, et utiliser les équivalences de norme). Alors,

$$\begin{aligned}
\mathbb{E}\frac{\mathbb{1}_{X\in B(0,M)}}{(n+1)\mu(B(X,h\eta))} &= \int_{B(0,M)} \frac{1}{(n+1)\mu(B(x,h_n\eta))}\mathrm{d}\mu(x) \\
&\leq \sum_{i=1}^{k_n} \int \frac{\mathbb{1}_{x\in B(c_i,h_n\eta/2)}}{n\mu(B(x,h_n\eta))}\mathrm{d}\mu(x) \\
&\leq \sum_{i=1}^{k_n} \int \frac{\mathbb{1}_{x\in B(c_i,h_n\eta/2)}}{n\mu(B(c_i,h_n\eta/2))}\mathrm{d}\mu(x) = \frac{k_n}{n} \leq \frac{C}{nh_n^d},
\end{aligned}$$

puisque si $x \in B(c_i, h_n\eta/2)$ alors $B(c_i, h_n\eta/2) \subset B(x, h_n\eta)$. On obtient que sous la condition $nh_n^d \to 0$, la 3ème condition du théorème de Stone est satisfaite. Ceci achève la preuve de notre théorème. □

3.4 k-plus proches voisins pour la classification

Dans cette section, on se place dans le cadre $\mathcal{Y} = \{0,1\}$ et le problème de classification supervisée. Conformément à ce que nous avons dit en introduction de ce chapitre, on associe naturellement à un estimateur de type moyenne locale la règle de classification plug-in

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n W_{ni}(x)Y_i > 1/2 \\ 0 & \text{sinon,} \end{cases}$$

ou, de façon équivalente lorsque les poids sont normalisés (i.e. lorsque $\sum_{i=1}^n W_{ni}(x) = 1$),

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n W_{ni}(x)\mathbb{1}_{Y_i=1} > \sum_{i=1}^n W_{ni}(x)\mathbb{1}_{Y_i=0} \\ 0 & \text{sinon.} \end{cases}$$

Dans la suite, $k = k_n$ est un entier strictement positif compris entre 1 et n (et fonction de n). On rappelle que $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$ désigne le réordonnement de l'échantillon original $(X_1, Y_1), \ldots, (X_n, Y_n)$ suivant les distances euclidiennes croissantes des X_i à x.

Nous avons vu dans la section 3.2 que la règle de classification des k-plus proches voisins a pour expression

$$g_n(x) = \begin{cases} 1 & \text{si } \frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{1}_{Y_{(i)}(x)=1} > \frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{1}_{Y_{(i)}(x)=0} \\ 0 & \text{sinon} \end{cases}$$

ou, de façon équivalente,

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^{k_n} \mathbb{1}_{Y_{(i)}(x)=1} > \sum_{i=1}^{k_n} \mathbb{1}_{Y_{(i)}(x)=0} \\ 0 & \text{sinon.} \end{cases}$$

Le prochain théorème, dont la preuve utilise le théorème de Stone, établit la convergence universelle de la règle g_n, pourvu que k croisse avec n mais pas trop vite.

Théorème 9. *Supposons que $k_n \to +\infty$ et $k_n/n \to 0$. Alors la règle de classification des k-plus proches voisins est universellement convergente, i.e.*

$$\mathbb{E}\mathcal{R}(g_n) \to \mathcal{R}^\star$$

quelle que soit la loi du couple (X, Y).

Pour prouver le résultat, il suffit simplement de s'assurer que les conditions $1-3$ du théorème de Stone (Théorème 7) sont effectivement vérifiées par la règle des k-plus proches voisins. Pour ce faire, nous aurons au préalable besoin de quelques lemmes techniques. Pour simplifier un peu, nous supposerons dans la suite que les égalités entre distances $\|X_i - x\| = \|X_j - x\|$ se produisent avec probabilité zéro (c'est par exemple le cas lorsque $\|X - x\|$

admet une densité par rapport à la mesure de Lebesgue). La preuve du Théorème 9 s'étend au cas général, au prix de quelques petits aménagements techniques pour gérer les distances ex-aequo. On rappelle que le support de la loi μ est défini comme l'ensemble des $x \in \mathbb{R}^d$ tels que, pour tout $\varepsilon > 0$, $\mu(B(x,\varepsilon)) > 0$, avec $B(x,\varepsilon)$ la boule fermée de centre x et de rayon ε. Alternativement, il s'agit du plus petit ensemble fermé de μ-mesure 1.

Lemme 3. *Soit x dans le support de μ. Alors, si $k_n/n \to 0$, on a*

$$\|X_{(k_n)}(x) - x\| \to 0, \quad \mathbb{P}\text{-}p.s.$$

Démonstration. Fixons $\varepsilon > 0$ et observons, puisque x appartient au support de μ, que $\mu(B(x,\varepsilon)) > 0$. Notons également l'égalité suivante entre événements :

$$\left\{ \|X_{(k_n)}(x) - x\| > \varepsilon \right\} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in B(x,\varepsilon)} < \frac{k_n}{n} \right\}.$$

Or, d'après la loi forte des grands nombres,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in B(x,\varepsilon)} \to \mu\left(B(x,\varepsilon)\right) > 0, \quad \mathbb{P}\text{-p.s.}$$

Comme $k_n/n \to 0$, on en conclut immédiatement que $\|X_{(k_n)}(x) - x\| \to 0, \mathbb{P}$-p.s. □

Lemme 4. *Soit ν une mesure de probabilité sur $\mathbb{R}^d$. Fixons $x' \in \mathbb{R}^d$ et définissons, pour $a \geq 0$,*

$$\mathcal{E}_a(x') = \left\{ x \in \mathbb{R}^d : \nu\left(B(x, \|x' - x\|)\right) \leq a \right\}.$$

Alors

$$\nu\left(\mathcal{E}_a(x')\right) \leq \gamma_d a,$$

où γ_d est une constante strictement positive ne dépendant que de d.

Démonstration. Fixons $x' \in \mathbb{R}^d$ et considérons une famille $\mathscr{C}_1, \ldots, \mathscr{C}_{\gamma_d}$ de demi-cônes d'angle $\pi/6$ centrés en x', suffisamment nombreux pour que leur union recouvre $\mathbb{R}^d$. En d'autres termes,

$$\bigcup_{j=1}^{\gamma_d} \mathscr{C}_j = \mathbb{R}^d.$$

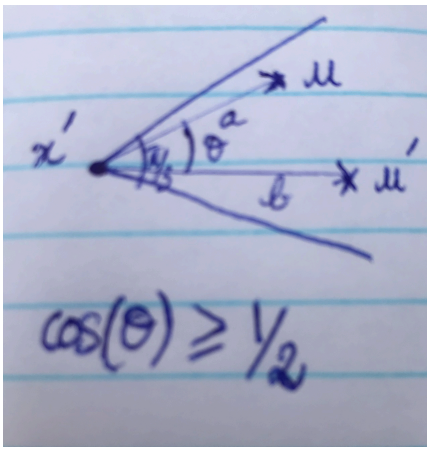

FIGURE 3.1 – Illustration de la propriété : si $u, u' \in \mathscr{C}_j$ et $\|u - x'\| \leq \|u' - x'\|$, alors $\|u - u'\| \leq \|u' - x'\|$

Commençons par montrer que si $u, u' \in \mathscr{C}_j$ et $\|u - x'\| \leq \|u' - x'\|$, alors $\|u - u'\| \leq \|u' - x'\|$ (voir Figure 3.4). refaire ce dessin En effet, notons $u - x' = a$ et $u' - x' = b$, alors $u - u' = a - b$. Par hypothèse $\|a\| \leq \|b\|$ et $\langle a, b\rangle / (\|a\| \times \|b\|) \geq \cos(\pi/3) = 1/2$. Ainsi, $\|a - b\|^2 = \|a\|^2 + \|b\|^2 - 2\langle a, b\rangle \leq \|a\|^2 + \|b\|^2 - \|a\| \times \|b\| \leq \|b\|^2 + \|a\|(\|a\| - \|b\|) \leq \|b\|^2$, d'où $\|a - b\| \leq \|b\|$.

On a

$$\nu\left(\mathscr{C}_j \cap \mathcal{E}_a(x')\right) = \lim_{R \to R_0^-} \nu\left(\mathscr{C}_j \cap B(x', R) \cap \mathcal{E}_a(x')\right).$$

où $R_0 = \sup\{R : \exists x^* \in \mathscr{C}_j \cap \mathcal{E}_a(x') : \|x' - x^*\| = R\}$ (il se peut que $R_0 = \infty$). Or pour tout $R < R_0$

$$\nu\left(\mathscr{C}_j \cap B(x', R) \cap \mathcal{E}_a(x')\right) \leq \nu\left(\mathscr{C}_j \cap B(x', R)\right) = \nu\left(\mathscr{C}_j \cap B(x', \|x^* - x'\|)\right)$$

avec $x^* \in \mathscr{C}^j \cap \mathcal{E}_a(x')$. Par la propriété de cônes ci-dessus

$$\mathscr{C}_j \cap B(x', \|x^* - x'\|) \subset B(x^*, \|x^* - x'\|).$$

En effet, si $x \in \mathscr{C}_j \cap B(x', \|x^* - x'\|)$, alors $\|x - x'\| \leq \|x^* - x'\|$, et comme $x, x^* \in \mathscr{C}_j$ on a $\|x - x^*\| \leq |x^* - x'\|$, donc $x \in B(x^*, \|x^* - x'\|)$.

Or, comme $x^* \in \mathcal{E}_a(x')$, on déduit par la définition de $\mathcal{E}_a(x')$:

$$\nu\left(B(x^*, \|x^* - x'\|)\right) \leq a.$$

Finalement pour tout $R > 0$ tel que $\exists x^* \in \mathscr{C}_j \cap \mathcal{E}_a(x') : \|x' - x^*\| = R$ on a

$$\nu\left(\mathscr{C}_j \cap B(x', R) \cap \mathcal{E}_a(x')\right) \leq a.$$

Alors

$$\nu(\mathscr{C}_j \cap \mathcal{E}_a(x')) \leq a$$

ce qu'il fallait démontrer. □

Le corollaire suivant énonce une conséquence fondamentale du lemme précédent : le nombre de points dans $\{X_1, \ldots, X_n\}$ pour lesquels X est l'un des k plus proches voisins, ne dépasse pas une constante fois k. Dans la suite, l'abréviation k-ppv signifie « k-plus proches voisins » .

Corollaire 2. *Si les égalités entre distances se produisent avec probabilité zéro, alors $\mathbb{P}$-p.s., le nombre de X_i tels que X soit parmi ses k-ppv est borné par $k_n\gamma_d$, i.e.*

$$\sum_{i=1}^{n} \mathbb{1}\{X \text{ est parmi les } k\text{-ppv de } X_i \text{ dans } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}\} \leq k_n\gamma_d,$$

$\mathbb{P}$-*p.s.*

Démonstration. On applique le Lemme 4 avec $a = k_n/n$ et ν la mesure empirique μ_n associée à $X_1, \ldots, X_n$. Avec ce choix, on a

$$\mathcal{E}_{k_n/n}(X) = \left\{x \in \mathbb{R}^d : \mu_n\left(B(x, \|X - x\|)\right) \leq k_n/n\right\}$$

et, $\mathbb{P}$-p.s.,

$$\begin{aligned} &X_i \in \mathcal{E}_{k_n/n}(X) \\ &\Leftrightarrow \mu_n\left(B(X_i, \|X - X_i\|)\right) \leq k_n/n \\ &\Leftrightarrow X \text{ est parmi les } k_n\text{-ppv de } X_i \text{ dans } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}. \end{aligned}$$

(Noter que la seconde équivalence utilise le fait que les égalités entre distances se produisent avec probabilité zéro !). Ainsi, en appliquant le Lemme

4, il vient, $\mathbb{P}$-p.s.,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{1}\{X \text{ est parmi les } k_n\text{-ppv de } X_i \text{ dans } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}\} \\
&\quad = \sum_{i=1}^{n} \mathbb{1}\{X_i \in \mathcal{E}_{k_n/n}(X)\} \\
&\quad = n \times \mu_n\left(\mathcal{E}_{k_n/n}(X)\right) \\
&\quad \leq k_n \gamma_d.
\end{aligned}$$

□

Lemme 5. *Supposons que les égalités entre distances se produisent avec probabilité zéro. Alors, pour toute fonction borélienne $f : \mathbb{R}^d \to \mathbb{R}$ telle que $\mathbb{E}|f(X)| < \infty$, on a*

$$\sum_{i=1}^{k_n} \mathbb{E}\left|f\left(X_{(i)}(X)\right)\right| \leq k_n \gamma_d \mathbb{E}\left|f(X)\right|,$$

où γ_d est une constante strictement positive ne dépendant que de d.

Remarque : Avant de montrer ce lemme, remarquons que la relation « être parmi les k-ppv » n'est pas une relation symétrique. Ainsi, si on se donne un ensemble de points $\mathscr{U} = \{u_1, \ldots, u_n\} \in \mathbb{R}^{dn}$, on peut avoir u_i est un k-ppv de u_j dans $\mathscr{U} \setminus \{u_j\}$ sans que u_j soit un k-ppv de u_i dans $\mathscr{U} \setminus \{u_i\}$.

Démonstration. Prenons f une fonction comme dans l'énoncé. Alors

$$\begin{aligned}
&\sum_{i=1}^{k_n} \mathbb{E}\left|f\left(X_{(i)}(X)\right)\right| \\
&= \mathbb{E}\left(\sum_{i=1}^{n} |f(X_i)| \, \mathbb{1}_{X_i \text{ est parmi les } k_n\text{-ppv de } X \text{ dans } \{X_1, \ldots, X_n\}} \right) \\
&= \mathbb{E}\Bigg(|f(X)| \\
&\times \sum_{i=1}^{n} \mathbb{1}_{X \text{ est parmi les } k_n\text{-ppv de } X_i \text{ dans } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}} \Bigg) \\
&\quad \text{(en échangeant } X \text{ et } X_i \text{ qui sont de même loi !)} \\
&\leq \mathbb{E}\left(|f(X)| \, k_n \gamma_d\right),
\end{aligned}$$

d'après le Corollaire 2. Ceci conclut la preuve du lemme. □

Nous sommes désormais en mesure de démontrer le Théorème 9. Il suffit pour cela de vérifier les conditions du théorème de Stone (théorème 7), avec $W_{ni}(x) = 1/k_n$ si X_i est parmi les k_n plus proches voisins de X et $W_{ni}(x) = 0$ sinon.

Démonstration du Théorème 9. La condition 3 est évidente dans la mesure où $k_n \to +\infty$. Pour la condition 2, on note que

$$\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\|X_i - X\|]>a}\right) = \mathbb{E}\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\mathbb{1}_{\|X_{(i)}(X)-X\|>a}\right),$$

de sorte que par le lemme de Cesàro,

$$\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\|X_i - X\|>a}\right) \to 0$$

dès que, pour tout $a > 0$, le terme général

$$\mathbb{P}\big(\|X_{(k_n)}(X) - X\| > a\big) \to 0.$$

Or,

$$\mathbb{P}\big(\|X_{(k_n)}(X) - X\| > a\big) = \int_{\mathbb{R}^d} \mathbb{P}\big(\|X_{(k_n)}(x) - x\| > a\big)\mu(\mathrm{d}x).$$

Pour x fixé dans le support de μ, le Lemme 3 indique que la convergence

$$\mathbb{P}\big(\|X_{(k_n)}(x) - x\| > a\big) \to 0$$

a lieu lorsque $k_n/n \to 0$. Le résultat s'en déduit par convergence dominée, en notant que le support de μ est de μ-mesure 1.

Examinons pour terminer la condition 1. Il s'agit de voir que, pour toute fonction f telle que $\mathbb{E}|f(X)| < \infty$, on a

$$\mathbb{E}\left(\frac{1}{k_n}\sum_{i=1}^{n}|f(X_i)|\,\mathbb{1}_{X_i \text{ est parmi les } k_n\text{-ppv de } X}\right) \leq c\mathbb{E}\,|f(X)|\,,$$

pour une certaine constante c. Comme

$$\mathbb{E}\left(\frac{1}{k_n}\sum_{i=1}^{n}|f(X_i)|\,\mathbb{1}_{X_i \text{ est parmi les } k_n\text{-ppv de } X}\right) = \mathbb{E}\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\left|f\big(X_{(i)}(X)\big)\right|\right),$$

c'est précisément l'énoncé du Lemme 5. □

Deuxième partie

Introduction au clustering

Chapitre 4

Quantification et clustering

Le clustering est une technique de classification non supervisée : contrairement à la partie précédente, on ne dispose pas d'un échantillon dont les étiquettes sont observées pour faire notre apprentissage. Nous présentons ici le clustering à travers ses liens avec la quantification.

4.1 Principe de la quantification

La quantification est un principe probabiliste dont l'objectif est de compresser l'information contenue dans une variable aléatoire X à valeurs dans $(\mathbb{R}^d, \|\cdot\|)$, où $\|\cdot\|$ désigne la norme euclidienne. On se donne dorénavant une telle variable X, en notant μ sa loi et en supposant que $\mathbb{E}\|X\|^2 < \infty$ ou, ce qui est équivalent, que

$$\int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x) < \infty.$$

Définition 3. *Soit k un entier ≥ 1. Un quantifieur q d'ordre k est une fonction mesurable $q : \mathbb{R}^d \to \mathscr{C} \subset \mathbb{R}^d$ avec $|\mathscr{C}| \leq k$.*

Un quantifieur q d'ordre k est donc caractérisé par :

- ▷ Un alphabet $\mathscr{C} = \{c_1, \ldots, c_k\}$.
- ▷ Une partition $\mathscr{P} = \{A_1, \ldots, A_k\}$ de $\mathbb{R}^d$, avec la numérotation imposée par
$$q(x) = c_j \Leftrightarrow x \in A_j.$$

On écrira dans la suite $q = (\mathscr{C}, \mathscr{P})$. Un quantifieur apparaît ainsi comme un outil de compression de l'information. L'étape suivante consiste alors à se doter d'un critère mesurant la pertinence de la compression de la variable aléatoire X (ou de sa loi μ) au travers de q.

Définition 4. *La distorsion (pour X ou μ) d'un quantifieur $q = (\mathscr{C}, \mathscr{P})$ d'ordre k est définie par*

$$D(\mu, q) = \mathbb{E}\|X - q(X)\|^2 = \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu(\mathrm{d}x).$$

La distorsion minimale à l'ordre k est

$$D_k^\star(\mu) = \inf_q D(\mu, q),$$

où l'infimum est évalué sur tous les quantifieurs d'ordre k.

Bien entendu, plus la distorsion est faible, meilleure est la compression. Par ailleurs, comme on s'en doute, la qualité d'une quantification s'améliore lorsque k grandit. Ce phénomène est précisé dans le lemme ci-dessous.

Lemme 6. *La suite des distorsions minimales à l'ordre k décroît vers 0 lorsque k grandit, i.e. $D_k^\star(\mu) \searrow 0$ si $k \to +\infty$.*

Démonstration. Tout d'abord, il est clair que la distorsion minimale décroît à mesure que son ordre augmente. Puis, comme $\mathbb{R}^d$ est un espace métrique complet, la mesure bornée ν définie pour tout borélien A de $\mathbb{R}^d$ par

$$\nu(A) = \int_A \|x\|^2 \mu(\mathrm{d}x)$$

est tendue, i.e. pour tout $\varepsilon \in (0, 1]$, il existe un compact K tel que $\nu(K) \geq \nu(\mathbb{R}^d) - \varepsilon$. On note $\{c_1, c_2, \ldots\}$ un sous-ensemble dénombrable dense dans $\mathbb{R}^d$, alors pour tout $\epsilon > 0$, on a $K \subset \bigcup_{j=1}^{\infty} B(c_j, \sqrt{\varepsilon})$. Comme K est compact, il existe un $k > 0$ (assez grand) tel que

$$K \subset B := \bigcup_{j=1}^{k} B(c_j, \sqrt{\varepsilon}).$$

On a donc $\nu(B) \geq \nu(\mathbb{R}^d) - \varepsilon$. Notons maintenant q_{k+1} le quantifieur d'ordre $k+1$ d'alphabet $\{c_1, \ldots, c_k, 0\}$ (en supposant, sans perte de généralité, que

$0 \notin \{c_1, c_2, \ldots\}$) et de partition $\{A_1, \ldots, A_k, B^c\}$, avec $A_1 = B(c_1, \sqrt{\varepsilon})$ et, pour $j = 2, \ldots, k$, $A_j = B(c_j, \sqrt{\varepsilon}) \setminus (\cup_{l=1}^{j-1} A_l)$. Comme $\|x - c_j\| \leq \sqrt{\varepsilon}$ si $x \in A_j$, on a

$$\begin{aligned} D_{k+1}^{\star}(\mu) \leq D_{k+1}(\mu, q_{k+1}) &= \int_{\mathbb{R}^d} \|x - q_{k+1}(x)\|^2 \mu(\mathrm{d}x) \\ &= \sum_{j=1}^{k} \int_{A_j} \|x - c_j\|^2 \mu(\mathrm{d}x) + \int_{B^c} \|x\|^2 \mu(\mathrm{d}x) \\ &\leq \varepsilon \mu\left(\bigcup_{j=1}^{k} A_j \right) + \nu(B^c) \leq 2\varepsilon, \end{aligned}$$

ce qui achève la preuve. □

Parmi toutes les façons possibles de compresser l'information, la classe des quantifieurs de type plus proches voisins, que nous abrégerons désormais en quantifieurs PPV, joue un rôle bien particulier. Dans la suite, on suppose que les quantifieurs sont d'ordre k et on note, pour un alphabet $\mathscr{C} = \{c_1, \ldots, c_k\} \subset \mathbb{R}^d$ de taille k, $\mathscr{P}_{\mathrm{V}}(\mathscr{C})$ la partition de Voronoï associée à $\mathscr{C}$, définie par

$$\begin{aligned} A_1 &= \left\{ x \in \mathbb{R}^d : \|x - c_1\| \leq \|x - c_\ell\|, \forall \ell = 1, \ldots, k \right\}, \text{ et} \\ A_j &= \left\{ x \in \mathbb{R}^d : \|x - c_j\| \leq \|x - c_\ell\|, \forall \ell = 1, \ldots, k \right\} \setminus \bigcup_{t=1}^{j-1} A_t, \end{aligned}$$

pour $j = 2, \ldots, k$.

Définition 5. *Un quantifieur d'ordre k est un quantifieur PPV si sa partition est une partition de Voronoï associée à son alphabet. En d'autres termes, un quantifieur PPV s'écrit $q = (\mathscr{C}, \mathscr{P}_{\mathrm{V}}(\mathscr{C}))$, avec $\mathscr{C} \subset \mathbb{R}^d$ de cardinal inférieur ou égal à k.*

Un quantifieur PPV noté q est donc entièrement caractérisé par son alphabet $\mathscr{C}$ (dont les éléments sont appelés centres ou centroïdes), via la règle

$$\|x - q(x)\| = \min_{c_j \in \mathscr{C}} \|x - c_j\|,$$

les égalités entre distances sur le bord des cellules étant brisées en faveur des plus petits indices. On notera les propriétés élémentaires suivantes.

Proposition 3. *Soit q_{ppv} un quantifieur PPV d'alphabet $\mathscr{C} = \{c_1, \ldots, c_k\}$. Alors*

$$D(\mu, q_{\mathrm{ppv}}) = \mathbb{E} \min_{1 \leq j \leq k} \|X - c_j\|^2 = \int_{\mathbb{R}^d} \min_{1 \leq j \leq k} \|x - c_j\|^2 \mu(\mathrm{d}x).$$

En outre, pour tout quantifieur avec le même alphabet et une autre partition $q = (\mathscr{C}, \mathscr{P})$, on a $D(\mu, q_{\mathrm{ppv}}) \leq D(\mu, q)$.

Démonstration. Pour la première propriété, en désignant par $\mathscr{P}_{\mathrm{V}}(\mathscr{C}) = \{A_{\mathrm{V},1}, \ldots, A_{\mathrm{V},k}\}$ la partition de Voronoï associée à $\mathscr{C}$:

$$\begin{aligned} D(\mu, q_{\mathrm{ppv}}) &= \int_{\mathbb{R}^d} \|x - q_{\mathrm{ppv}}(x)\|^2 \mu(\mathrm{d}x) = \sum_{j=1}^{k} \int_{A_{\mathrm{V},j}} \|x - c_j\|^2 \mu(\mathrm{d}x) \\ &= \sum_{j=1}^{k} \int_{A_{\mathrm{V},j}} \min_{1 \leq i \leq k} \|x - c_i\|^2 \mu(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d} \min_{1 \leq i \leq k} \|x - c_i\|^2 \mu(\mathrm{d}x). \end{aligned}$$

Puis, pour la seconde propriété, si $\mathscr{P} = \{A_1, \ldots, A_k\}$ est la partition d'un autre quantificateur q, on a :

$$\begin{aligned} D(\mu, q_{\mathrm{ppv}}) &= \int_{\mathbb{R}^d} \min_{1 \leq i \leq k} \|x - c_i\|^2 \mu(\mathrm{d}x) \\ &= \sum_{j=1}^{k} \int_{A_j} \min_{1 \leq i \leq k} \|x - c_i\|^2 \mu(\mathrm{d}x) \\ &\leq \sum_{j=1}^{k} \int_{A_j} \|x - c_j\|^2 \mu(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu(\mathrm{d}x) = D(\mu, q), \end{aligned}$$

par définition de la distorsion. □

La conséquence fondamentale de cette dernière proposition est que les quantifieurs de distorsion minimale, s'ils existent, sont à rechercher parmi les quantifieurs du type $q_{\mathrm{ppv}} = (\mathbf{c}, \mathscr{P}_{\mathrm{V}}(\mathbf{c}))$ avec $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$, (noter l'abus de notation), de distorsion

$$W(\mu, \mathbf{c}) := \int_{\mathbb{R}^d} \min_{1 \leq j \leq k} \|x - c_j\|^2 \mu(\mathrm{d}x) = D(\mu, q_{\mathrm{ppv}}).$$

Théorème 10. *Parmi les quantificateurs d'ordre k, il existe un quantifieur de distorsion minimale.*

Esquisse de démonstration. D'après la Proposition 3, on peut restreindre l'étude aux quantifieurs PPV. Il s'agit donc de montrer qu'il existe $\mathbf{c}^\star \in \mathbb{R}^{dk}$ tel que

$$W(\mu, \mathbf{c}^\star) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, \mathbf{c}).$$

On prouve d'abord (admis ici) qu'il existe $R > 0$ tel que

$$\inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, \mathbf{c}) = \inf_{\|\mathbf{c}\| \leq R} W(\mu, \mathbf{c}).$$

On établit ensuite que la fonction $\mathbb{R}^{dk} \ni \mathbf{c} \mapsto W(\mu, \mathbf{c})$ est continue. Observons pour cela que la fonction $x \mapsto \min_{1 \leq j \leq k} \|x - c_j\|$ est continue. Dès lors, pour $\mathbf{c}_0 = (c_{1,0}, \ldots, c_{k,0}) \in \mathbb{R}^{dk}$ fixé, on a

$$\begin{aligned}
\lim_{\mathbf{c} \to \mathbf{c}_0} W(\mu, \mathbf{c}) &= \int_{\mathbb{R}^d} \lim_{\mathbf{c} \to \mathbf{c}_0} \min_{1 \leq j \leq k} \|x - c_j\|^2 \mu(\mathrm{d}x) \\
&\quad \text{(d'après le théorème de Lebesgue)} \\
&= \int_{\mathbb{R}^d} \min_{1 \leq j \leq k} \|x - c_{j,0}\|^2 \mu(\mathrm{d}x) \\
&\quad \text{(par continuité)} \\
&= W(\mu, \mathbf{c}_0),
\end{aligned}$$

ce qui montre bien que $W(\mu, \cdot)$ est continue.

On déduit de cette dernière propriété et de la compacité de la boule $B(0, R)$ de $\mathbb{R}^{dk}$ qu'il existe $\mathbf{c}^\star \in \mathbb{R}^{dk}$ minimum de $W(\mu, \cdot)$. Le quantifieur $q^\star = (\mathbf{c}^\star, \mathscr{P}_{\mathrm{ppv}}(\mathbf{c}^\star))$ est alors de distorsion minimale car

$$W(\mu, \mathbf{c}^\star) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu, c) = \inf_q D(\mu, q) = D^\star(\mu).$$

□

4.2 Quantification empirique et clustering

En pratique, la loi μ de la variable aléatoire X est inconnue et il est donc, par voie de conséquence, impossible de procéder à sa quantification optimale.

On dispose cependant bien souvent d'un n-échantillon i.i.d. $X_1, \ldots, X_n$ formé de variables aléatoires indépendantes entre elles, de même loi que X et indépendante de cette dernière. C'est à partir de cet échantillon que l'on va s'attacher à construire un quantifieur empirique $q_n(\cdot) = q_n(\cdot; X_1, \ldots, X_n)$ dont les performances se rapprochent si possible de celles du quantifieur optimal.

Dans ce contexte, la distorsion pour μ du quantifieur empirique q_n (d'ordre k) est naturellement définie par

$$D(\mu, q_n) = \int_{\mathbb{R}^d} \|x - q_n(x)\|^2 \mu(\mathrm{d}x).$$

Noter qu'il s'agit d'une variable aléatoire qui dépend de $X_1, \ldots, X_n$ à travers l'estimateur q_n et que par ailleurs

$$D(\mu, q_n) = \mathbb{E}\big(\|X - q_n(X)\|^2 | X_1, \ldots, X_n\big).$$

Comme cette quantité est inconnue (elle dépend de la mesure μ inconnue), on l'approche par son équivalent empirique. Ainsi, on suppose toujours que $\mathbb{E}\|X\|^2 < \infty$ et on désigne par μ_n la mesure empirique associée à $X_1, \ldots, X_n$, i.e.

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

On introduit *la distorsion empirique* d'un quantifieur quelconque q, elle prend la forme

$$D(\mu_n, q) = \int_{\mathbb{R}^d} \|x - q(x)\|^2 \mu_n(\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - q(X_i)\|^2.$$

Dans le cas particulier d'un quantifieur de type PPV, $q_{\mathrm{ppv}} = (\mathbf{c}, \mathscr{P}_{\mathrm{V}}(\mathbf{c}))$ avec $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$, on obtient

$$D(\mu_n, q_{\mathrm{ppv}}) = W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le j \le k} \|X_i - c_j\|^2.$$

Pour se doter d'outils qui assurent que la méthode de quantification empirique est performante, on introduit la définition qui suit.

Définition 6. *Soit q_n un quantifieur empirique. On dit qu'il est consistant si $\mathbb{E}D(\mu, q_n) \to D^\star(\mu)$. On dit qu'il est de vitesse $(v_n)_n$ si $\mathbb{E}D(\mu, q_n) - D^\star(\mu) = O(1/v_n)$, avec $v_n \to +\infty$.*

On aura noté au passage que, puisque $D(\mu, q_n) \geq D^\star(\mu)$, la propriété $\mathbb{E}D(\mu, q_n) \to D^\star(\mu)$ est équivalente à $D(\mu, q_n) \to D^\star(\mu)$ dans $\mathbb{L}^1(\mu)$.

Le quantifieur empirique $q_n^\star$ le plus naturel est obtenu en minimisant la distorsion empirique sur tous les quantifieurs PPV. En d'autres termes, on cherche les centres optimaux $\mathbf{c}_n^\star = (c_{n,1}^\star, \ldots, c_{n,k}^\star)$ tels que

$$W(\mu_n, \mathbf{c}_n^\star) = \inf_{\mathbf{c} \in \mathbb{R}^{dk}} W(\mu_n, \mathbf{c}). \tag{4.1}$$

On a donc

$$q_n^\star = \left(\mathbf{c}_n^\star, \mathscr{P}_{\mathrm{V}}(\mathbf{c}_n^\star)\right).$$

(Observons que $\mathbf{c}_n^\star$ et donc $q_n^\star$ existent en vertu du Théorème 10.)

Un quantifieur empirique q_n (d'ordre k) est naturellement associé à une méthode de regroupement (ou clustering) des données $X_1, \ldots, X_n$ en k classes, en décidant que l'observation X_i est rangée dans la classe j ($1 \leq j \leq k$) si $q_n(X_i) = j$.

Pour le quantifieur empirique PPV optimal $q_n^\star$, le j-ème cluster est constitué des observations X_i telles que $\|X_i - c_{n,j}^\star\| \leq \|X_i - c_{n,\ell}^\star\|$, $\forall \ell = 1, \ldots, k$.

On parle parfois, en lieu et place de clustering, de classification (ou apprentissage) non supervisé, l'adjectif « non supervisé » renvoyant au fait qu'il n'y a pas d'information annexe apportée par des variables réponses Y_i. Le problème consiste ici à regrouper les données $X_1, \ldots, X_n$ « à l'aveugle » , de la façon la plus pertinente possible et sans information supplémentaire.

Algorithme des k-means. En pratique, l'approche (4.1) par minimisation de la distorsion empirique est difficile à mettre en œuvre, surtout en grande dimension (problème NP-complet [1]). On a alors recours à une technique approchée, appelée algorithme des k-means. Pour une partition quelconque $\mathscr{P} = \{A_1^0, \ldots, A_k^0\}$ et un alphabet quelconque $\mathscr{C} = \{c_1^0, \ldots, c_k^0\} \subset \mathbb{R}^d$, on définit $q^0 = (\mathscr{C}, \mathscr{P})$. On définit ensuite $q^1 = (\mathscr{C}^1, \mathscr{P}^1)$ de la manière suivante. On calcule d'abord $\mathscr{C}^1 = \{c_1^1, \ldots, c_k^1\}$ tel que, pour tout $j = 1, \ldots, k$,

$$c_j^1 = \operatorname*{Argmin}_{y \in \mathbb{R}^d} \sum_{i=1}^n \|X_i - y\|^2 \mathbb{1}_{X_i \in A_j^0}. \tag{4.2}$$

1. En théorie de la complexité, un problème de décision est NP-complet lorsqu'il est possible de vérifier une solution en temps polynomial (NP) et tous les problèmes de la classe NP se ramènent à lui via une réduction polynomiale.

Il est facile de déduire par un calcul élémentaire que ce minimum est atteint pour

$$y_j = \frac{\sum_{i=1}^n X_i \mathbb{1}_{X_i \in A_j^0}}{\sum_{i=1}^n \mathbb{1}_{X_i \in A_j^0}} = c_j^1, \quad j = 1, \dots, k. \tag{4.3}$$

Noter que c_j^1 est une espérance conditionnelle pour la mesure empirique, $\mathbb{E}_{\mu_n}(X \mid X \in A_j^0)$. On construit ensuite la partition $\mathscr{P}^1 = (A_1^1, \dots, A_k^1)$ qui est la partition de Voronoï associé à l'alphabet $\mathscr{C}^1$. Par ailleurs, il est facile maintenant d'attribuer à chaque donnée sa cellule : $X_i \in A_j^1$ si

$$\min_{1 \le l \le k} \|X_i - c_l^1\| = \|X_i - c_j^1\|, \quad \text{pour } i = 1, \dots, k.$$

On continue cette procedure par récurrence. Soient la partition $\mathscr{P}^{m-1} = \{A_1^{m-1}, \dots, A_k^{m-1}\}$, l'alphabet $\mathscr{C} = \{c_1^{m-1}, \dots, c_k^{m-1}\} \subset \mathbb{R}^d$ et $q^{m-1} = (\mathscr{C}, \mathscr{P})$ déjà définis. On construit $q^m = (\mathscr{C}^m, \mathscr{P}^m)$, avec $\mathscr{C}^m = \{c_1^m, \dots, c_k^m\}$ tel que, pour tout $j = 1, \dots, k$,

$$c_j^m = \operatorname*{Argmin}_{y \in \mathbb{R}^d} \sum_{i=1}^n \|X_i - y\|^2 \mathbb{1}_{X_i \in A_j^{m-1}}. \tag{4.4}$$

Ce minimum est atteint pour

$$y_j = \frac{\sum_{i=1}^n X_i \mathbb{1}_{X_i \in A_j^{m-1}}}{\sum_{i=1}^n \mathbb{1}_{X_i \in A_j^{m-1}}} = c_j^m, \quad j = 1, \dots, k. \tag{4.5}$$

Autrement dit $c_j^m = \mathbb{E}_{\mu_n}(X \mid X \in A_j^{m-1})$. On construit ensuite la partition $\mathscr{P}^m = (A_1^m, \dots, A_k^m)$ qui est la partition de Voronoï associé à l'alphabet $\mathscr{C}^m$. Par ailleurs $X_i \in A_j^m$ si

$$\min_{1 \le l \le k} \|X_i - c_l^m\| = \|X_i - c_j^m\|, \quad \text{pour } i = 1, \dots, k.$$

L'algorithme s'arrête lorsque plus rien ne bouge, i.e. $\mathbf{c}^{(m+1)} = \mathbf{c}^{(m)}$.

Remarquons que la distorsion empirique décroit à chaque étape

$$\begin{aligned} D(\mu_n, q^{m-1}) &= \frac{1}{n} \sum_{i=1}^n \|X_i - q^{m-1}(X_i)\|^2 = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^n \|X_i - c_j^{m-1}\|^2 \mathbb{1}_{X_i \in A_j^{m-1}} \\ &\ge \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^n \|X_i - c_j^m\|^2 \mathbb{1}_{X_i \in A_j^{m-1}} \\ &= D(\mu_n, \tilde{q}^m). \end{aligned} \tag{4.6}$$

où $\widetilde{q}^m$ est le quantificateur avec l'alphabet $\mathscr{C}^m$ mais avec la partition encore $\mathscr{P}^{m-1}$. Ensuite on passe de $\mathscr{P}^{m-1}$ à la partition de Voronoï $\mathscr{P}^m$ associée à $\mathscr{C}^m$ et on remarque

$$D(\mu_n, \widetilde{q}^m) \geq D(\mu_n, q^m)$$

en utilisant simplement la Proposition 3 valide pour n'importe quelle mesure de probabilité μ et en particulier pour μ_n. Ainsi l'algorithme de k-means fait décroitre la distorsion empirique :

$$D(\mu_n, q^{m-1}) \geq D(\mu_n, q^m), \quad m = 1, 2, \ldots$$

Néanmoins, même s'il est assuré que la distorsion empirique décroît entre deux itérations et que l'algorithme s'arrête au bout d'un nombre d'itérations fini, rien ne garantit cependant que les centres ainsi définis soient proches des centres optimaux $\mathbf{c}_n^\star$. Il s'agit d'une méthode approchée que l'on manipulera donc avec prudence. En pratique, il peut arriver que pour un point de départ $\mathbf{c}^{(0)}$ particulier, l'algorithme s'arrête dans un minimum local et pas global. C'est pourquoi on utilise en général l'algorithme k-means avec 10 points de départ (aléatoires) et on sélectionne le résultat donnant la plus faible distorsion.

4.3 Consistance et vitesse

L'outil indispensable pour établir la consistance du quantifieur empirique PPV optimal défini via (4.1) est la distance de Wasserstein.

Définition 7. *Soit ν_1 et ν_2 des probabilités d'ordre 2 sur $\mathbb{R}^d$. La distance de Wasserstein ρ_W entre ν_1 et ν_2 est définie par*

$$\rho_W(\nu_1, \nu_2) = \inf_{X \sim \nu_1, Y \sim \nu_2} \sqrt{\mathbb{E}\|X - Y\|^2}.$$

Il s'agit d'une distance usuelle sur les mesures de probabilité. Mentionnons sans preuve deux de ses propriétés fondamentales :

Propriétés.

1. Pour ν_1, ν_2 des probabilités d'ordre 2 sur $\mathbb{R}^d$, il existe un couple de variables aléatoires (X_0, Y_0) telles que $X_0 \sim \nu_1$, $Y_0 \sim \nu_2$ et
$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$

2. Soit $(\nu_n)_n$ et ν des probabilités d'ordre 2 sur $\mathbb{R}^d$. On a $\rho_W(\nu_n, \nu) \to 0$ si
$$\nu_n \Rightarrow \nu \quad \text{et} \quad \int_{\mathbb{R}^d} \|x\|^2 \nu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \|x\|^2 \nu(\mathrm{d}x).$$
Ici $\Rightarrow$ signifie la convergence faible (i.e. étroite) de mesures (pour toute fonction f continue bornée $\int f d\nu_n \to \int f d\nu$).

Proposition 4.
$$\rho_W(\mu_n, \mu) \to 0, \quad \mathbb{P} - p.s.$$

Démonstration. On applique la propriété 2 ci-dessus. Remarquons tout d'abord que $\int_{\mathbb{R}^d} \|x\|^2 \mu_n(\mathrm{d}x) = \frac{1}{n}\sum_{i=1}^n \|X_i\|^2 \to \mathbb{E}X^2 = \int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x)$, $\mathbb{P}$- p.s. par la loi des grands nombres. Quant à la convergence faible $\mu_n \Rightarrow \mu, \mathbb{P}-$p.s., on pourrait penser qu'elle découle également de la loi des grands nombres : pour toute $f : \mathbb{R}^d \to \mathbb{R}$ continue et bornée, $\mathbb{E}|f(X)| < \infty$, alors $\int_{\mathbb{R}^d} f d\mu_n = \frac{1}{n}\sum_{i=1}^n f(X_i(\omega)) \to \mathbb{E}f(X) = \int_{\mathbb{R}^d} f d\mu$, p.s.. En effet, l'ensemble $A_f = \{\omega : \int_{\mathbb{R}^d} f(x) d\mu_n \not\to \int_{\mathbb{R}^d} f d\mu\}$ est de probabilité $\mathbb{P}(A_f) = 0$ mais cet ensemble dépend de f. Et il n'est pas garanti que $\mathbb{P}(\bigcup_{f \text{ continues bornées}} A_f) = 0$. La preuve de $\mu_n \Rightarrow \mu$ p.s. est beaucoup plus délicate et fait l'objet du théorème qui suit : le Théorème de Varadarajan. L'application de ce théorème termine la preuve de $\rho_W(\mu_n, \mu) \to 0$ p.s. □

Théorème 11 (Varadarajan). $\mu_n \Rightarrow \mu, \quad \mathbb{P}- p.s.$

Démonstration. Pour prouver cette convergence en loi, on utilise le théorème de Portemanteau, qui dit (entre autres) qu'une suite de mesures de probabilités ν_n sur $\mathbb{R}^d$ converge en loi vers ν si et seulement si pour pour tout ensemble O ouvert de $\mathbb{R}^d$, on a $\nu(O) \leq \liminf_{n\to\infty} \nu_n(O)$.

Nous devons donc prouver que pour tout O ouvert dans $\mathbb{R}^d$ et tout $\omega \in \Omega \setminus \mathcal{N}$, on a $\mu(O) \leq \liminf_{n\to\infty} \mu_n(O, \omega)$ avec $\mathbb{P}(\mathcal{N}) = 0$ et $\mathcal{N}$ est universel pour tous les ouverts de $\mathbb{R}^d$ (i.e. ne dépend pas de l'ouvert O considéré).

Soit $\mathcal{D}$ un ensemble dénombrable dense dans $\mathbb{R}^d$, soit $\mathcal{B}$ l'ensemble des boules de centres dans $\mathcal{D}$ et de rayons rationnels. L'ensemble $\mathcal{B}$ est dénombrable, on note B_i pour $i = 1, 2, \ldots$, les boules de cette ensemble. Par la loi des grands nombres, pour tout ensemble borélien $B \subset \mathbb{R}^d$, il existe $\mathcal{N}_B$ avec $\mathbb{P}(\mathcal{N}_B) = 0$ tel que

$$\lim_{n\to\infty} \mu_n(B, \omega) \to \mu(B), \quad \forall \omega \in \Omega \setminus \mathcal{N}_B.$$

(Ici $\mu_n(B, \omega)$ indique que $X_1(\omega), \ldots, X_n(\omega)$ sont fixés dans la mesure empirique μ_n). Prenons

$$\mathcal{N} = \cup_{k\geq 1} \cup_{1\leq i_1<\cdots<i_k} \mathcal{N}_{B_{i_1}\cap\cdots\cap B_{i_k}},$$

où B_i sont les boules de $\mathcal{B}$. Alors

$$\mathbb{P}(\mathcal{N}) = 0$$

et pour tout $k \geq 1$ et tous $1 \leq i_1 < i_2 < \cdots < i_k$,

$$\lim_{n\to\infty} \mu_n(B_{i_1} \cap \cdots \cap B_{i_k}, \omega) \to \mu(B_{i_1} \cap \cdots \cap B_{i_k}), \quad \forall \omega \in \Omega \setminus \mathcal{N}.$$

Pour tout ouvert O, il existe un ensemble dénombrable de boules dans $\mathcal{B}$ tel que $O = \cup_{i=1}^{\infty} B(d_i, r_i)$. En effet, prenons tous les points $d_i \in \mathcal{D} \cap O$ et tous les rayons r_i rationnels tels que $B(d_i, r_i) \subset O$, alors $\cup_{i=1}^{\infty} B(d_i, r_i) \subset O$. Réciproquement, supposons que $x \in O$. Alors comme O est ouvert, il existe r rationnel tel que $B(x, r) \subset O$. De plus comme l'ensemble $\mathcal{D}$ est dense, il existe un point $d \in \mathcal{D} \cap B(x, r/5)$. Alors $B(d, 3r/5) \subset B(x, r) \subset O$ et $x \in B(d, 3r/5)$. Donc pour tout point $x \in O$, on obtient $x \in \cup_{i=1}^{\infty} B(d_i, r_i)$.

Par le théorème de convergence monotone,

$$\mu(O) = \mu(\cup_{i=1}^{\infty} B(d_i, r_i)) = \lim_{\ell\to\infty} \mu(\cup_{i=1}^{\ell} B(d_i, r_i)),$$

donc pour tout $\epsilon > 0$, il existe $L \geq 1$ tel que $\forall \ell \geq L$,

$$\mu(\cup_{i=1}^{\ell} B(d_i, r_i)) \geq \mu(O) - \epsilon.$$

Notons par la suite $B(d_i, r_i) = B_i$. La formule de Poincaré (ou formule du

crible) donne

$$\begin{aligned}\mu(\cup_{i=1}^{\ell} B(d_i, r_i)) &= \sum_{k=1}^{\ell} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le \ell} \mu(B_{i_1} \cap \cdots \cap B_{i_k}) \\ &= \sum_{k=1}^{\ell} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le \ell} \lim_{n\to\infty} \mu_n(B_{i_1} \cap \cdots \cap B_{i_k}, \omega) \\ &= \lim_{n\to\infty} \mu_n(\cup_{i=1}^{\ell} B_i, \omega),\end{aligned}$$

pour tout $\omega \in \Omega \setminus \mathcal{N}$. On en déduit puisque $\cup_{i=1}^{\ell} B_i \subset O$,

$$\mu(O) - \epsilon \le \lim_{n\to\infty} \mu_n(\cup_{i=1}^{\ell} B_i, \omega) \le \liminf_{n\to\infty} \mu_n(O, \omega), \quad \forall \omega \in \Omega \setminus \mathcal{N}.$$

En faisant tendre $\epsilon > 0$ vers 0, on obtient

$$\mu(O) \le \liminf_{n\to\infty} \mu_n(O, \omega), \quad \forall \omega \in \Omega \setminus \mathcal{N}.$$

□

Remarque 1. Le théorème de Varadarajan porte parfois le titre de principe fondamental de la statistique. En effet, il justifie la méthode d'approcher et même de « remplacer » la mesure μ inconnue par la mesure empirique μ_n calculée à partir de n expériences indépendantes.

Remarque 2. La distance de Wasserstein n'est bien sûr pas la seule distance qu'on pourrait définir sur l'ensemble des mesures. On pourrait considérer par exemple la distance en variation totale

$$d_{TV}(\nu_1, \nu_2) = \sup_{A \subset \mathbb{R}^d} |\nu_1(A) - \nu_2(A)|.$$

Mais la convergence de μ_n vers μ n'est pas assurée dans cette distance. Par exemple, si la loi μ est sans atomes,

$$\begin{aligned}&d_{TV}(\mu_n(\omega), \mu) \\ &\quad \ge |\mu_n(\{X_1(\omega)\} \cup \cdots \cup \{X_n(\omega)\}, \omega) - \mu(\{X_1(\omega)\} \cup \cdots \cup \{X_n(\omega)\})| \\ &\quad = |1 - 0| = 1.\end{aligned}$$

Le lien entre distorsion de quantifieurs PPV et distance de Wasserstein est établi dans la proposition qui suit.

Proposition 5. *Soient ν_1 et ν_2 des probabilités d'ordre 2 sur $\mathbb{R}^d$. Si q est un quantifieur PPV, alors*

$$\left| D(\nu_1, q)^{1/2} - D(\nu_2, q)^{1/2} \right| \leq \rho_W(\nu_1, \nu_2).$$

Démonstration. Soit (X_0, Y_0) tel que $X_0 \sim \nu_1$, $Y_0 \sim \nu_2$ et

$$\rho_W(\nu_1, \nu_2) = \sqrt{\mathbb{E}\|X_0 - Y_0\|^2}.$$

Si $q = (\mathbf{c}, \mathscr{P}_V(\mathbf{c}))$, alors sa distorsion s'écrit (voir Proposition 3) :

$$\begin{aligned}
D(\nu_1, q)^{1/2} &= \sqrt{\mathbb{E} \min_{1 \leq j \leq k} \|X_0 - c_j\|^2} \\
&= \sqrt{\mathbb{E}\big(\min_{1 \leq j \leq k} \|X_0 - c_j\| \big)^2} \\
&\leq \sqrt{\mathbb{E}\Big(\min_{1 \leq j \leq k} \big(\|X_0 - Y_0\| + \|Y_0 - c_j\| \big) \Big)^2} \\
&= \sqrt{\mathbb{E}\Big(\|X_0 - Y_0\| + \min_{1 \leq j \leq k} \|Y_0 - c_j\| \Big)^2} \\
&\leq \sqrt{\mathbb{E}\|X_0 - Y_0\|^2} + \sqrt{\mathbb{E} \min_{1 \leq j \leq k} \|Y_0 - c_j\|^2} \\
&\quad \text{(en utilisant l'inégalité de Cauchy-Schwarz)} \\
&= \rho_W(\nu_1, \nu_2) + D(\nu_2, q)^{1/2}.
\end{aligned}$$

(En effet pour $a, b \geq 0$, on a $\mathbb{E}(ab) \leq (\mathbb{E}a^2 \mathbb{E}b^2)^{1/2}$ par l'inégalité de Cauchy-Schwartz, alors $\mathbb{E}(a+b)^2 \leq \mathbb{E}a^2 + \mathbb{E}b^2 + 2(\mathbb{E}a^2)^{1/2}(\mathbb{E}b^2)^{1/2} = (\sqrt{\mathbb{E}a^2} + \sqrt{\mathbb{E}b^2})^2$, d'où $\sqrt{\mathbb{E}(a+b)^2} \leq \sqrt{\mathbb{E}a^2} + \sqrt{\mathbb{E}b^2}$).

Par symétrie des rôles de ν_1, ν_2, on a pareillement $D(\nu_2, q)^{1/2} \leq \rho_W(\nu_1, \nu_2) + D(\nu_1, q)^{1/2}$, d'où la proposition. □

On considère à partir de maintenant le quantifieur empirique optimal PPV $q_n^\star$, défini via (4.1) par ses centres $\mathbf{c}_n^\star = (c_{n,1}^\star, \ldots, c_{n,k}^\star)$:

$$q_n^\star = (\mathbf{c}_n^\star, \mathscr{P}_V(\mathbf{c}_n^\star)).$$

Théorème 12. *La distorsion $D(\mu, q_n^\star) \to D^\star(\mu)$, $\mathbb{P}$-p.s. et $\mathbb{E}D(\mu, q_n^\star) \to D^\star(\mu)$. Le quantifieur $q_n^\star$ est donc consistant.*

Démonstration. (Rappel : l'ordre k des quantifieurs est fixé et omis dans les notations). Si $q^\star$ est un quantifieur optimal PPV pour la loi μ (i.e. $D(\mu, q^\star) = \inf_q D(\mu, q) = D^\star(\mu)$), la Proposition 5 nous donne en considérant le quantifieur empirique $q_n^\star$

$$\begin{aligned} 0 &\leq D(\mu, q_n^\star)^{1/2} - D^\star(\mu)^{1/2} \\ &= \left[D(\mu, q_n^\star)^{1/2} - D(\mu_n, q_n^\star)^{1/2}\right] + \left[D(\mu_n, q_n^\star)^{1/2} - D(\mu, q^\star)^{1/2}\right] \\ &\leq \left[D(\mu, q_n^\star)^{1/2} - D(\mu_n, q_n^\star)^{1/2}\right] + \left[D(\mu_n, q^\star)^{1/2} - D(\mu, q^\star)^{1/2}\right] \\ &\leq 2\rho_W(\mu, \mu_n). \end{aligned} \tag{4.7}$$

En effet, $D(\mu_n, q_n^\star) \leq D(\mu_n, q^\star)$ car $q^\star$ est un quantifieur PPV et $q_n^\star$ est le quantifieur PPV optimal par rapport à μ_n. Par ailleurs, la Proposition 5 peut être appliquée deux fois car $q_n^\star$ et $q^\star$ sont tous les deux des quantifieurs PPV.

Or, $\rho_W(\mu_n, \mu) \to 0$, $\mathbb{P}$-p.s. par le Théorème 4.

Pour prouver la seconde assertion, introduisons $\mathscr{M}(\mu, \mu_n)$ l'ensemble (aléatoire) des probabilités sur $\mathbb{R}^d \times \mathbb{R}^d$ admettant μ et μ_n comme marginales. Par définition, le carré de la distance de Wasserstein entre μ et μ_n s'écrit

$$\rho_W^2(\mu, \mu_n) = \inf_{\nu \in \mathscr{M}(\mu, \mu_n)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \nu(\mathrm{d}x, \mathrm{d}y).$$

C'est une variable aléatoire car μ_n est une mesure aléatoire.

Soit C une constante arbitraire strictement positive et soit $\mathscr{A}$ le sous-ensemble de $\mathbb{R}^d \times \mathbb{R}^d$ défini par

$$\mathscr{A} = \left\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \max(\|x\|, \|y\|) \leq C\right\}.$$

On a, pour tout $\nu \in \mathscr{M}(\mu, \mu_n)$,

$$\begin{aligned}
&\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\quad = \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + \int_{\mathscr{A}^c} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\quad \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 2\int_{\mathscr{A}^c} \|x\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 2\int_{\mathscr{A}^c} \|y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\qquad (\text{car } \|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2) \\
&\quad \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\qquad + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{\|x\|>C} \mu(\mathrm{d}x) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{\|x\|\leq C, \|y\|>C} \nu(\mathrm{d}x, \mathrm{d}y) \\
&\qquad + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{\|y\|>C} \mu_n(\mathrm{d}y) + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{\|x\|>C, \|y\|\leq C} \nu(\mathrm{d}x, \mathrm{d}y) \\
&\quad \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\qquad + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{\|x\|>C} \mu(\mathrm{d}x) + 2C^2 \mu_n\left(\|y\| > C\right) \\
&\qquad + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{\|y\|>C} \mu_n(\mathrm{d}y) + 2C^2 \mu\left(\|x\| > C\right).
\end{aligned}$$

Ainsi, en appliquant l'inégalité de Markov, il vient

$$\begin{aligned}
&\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\quad \leq \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) \\
&\qquad + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{\|x\|>C} \mu(\mathrm{d}x) + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{\|y\|>C} \mu_n(\mathrm{d}y) \\
&\qquad + 2\int_{\mathbb{R}^d} \|y\|^2 \mathbb{1}_{\|y\|>C} \mu_n(\mathrm{d}y) + 2\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{\|x\|>C} \mu(\mathrm{d}x).
\end{aligned}$$

En prenant l'infimum à droite sur $\mathscr{M}(\mu, \mu_n)$ puis l'espérance des deux côtés, on en conclut que

$$\mathbb{E}\rho_W^2(\mu, \mu_n) \leq \mathbb{E} \inf_{\nu \in \mathscr{M}(\mu, \mu_n)} \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x, \mathrm{d}y) + 8\int_{\mathbb{R}^d} \|x\|^2 \mathbb{1}_{\|x\|>C} \mu(\mathrm{d}x).$$

(En effet $\mathbb{E}\int_{\mathbb{R}^d} f d\mu_n = \int f d\mu$ pour f qui est μ-intégrable).

Pour $C > 0$ fixé, le premier terme du membre de droite tend vers 0 lorsque n tend vers l'infini d'après le Théorème 4 et le théorème de convergence

dominée. En effet

$$\begin{aligned}\eta_n &= \inf_{\nu \in \mathscr{M}(\mu,\mu_n)} \int_{\mathscr{A}} \|x-y\|^2 \nu(\mathrm{d}x,\mathrm{d}y) \\ &\leq \inf_{\nu \in \mathscr{M}(\mu,\mu_n)} \int_{\mathbb{R}^d} \|x-y\|^2 \nu(\mathrm{d}x,\mathrm{d}y) = \rho_W^2(\mu_n,\mu) \to 0, \quad \mathbb{P}-p.s.\end{aligned}$$

Par ailleurs $\eta_n \leq 4C^2$ à cause de la définition de $\mathscr{A}$ et du fait que $\|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$. Donc $\mathbb{E}\eta_n \to 0$ par convergence dominée.

Puisque $\int_{\mathbb{R}^d} \|x\|^2 \mu(\mathrm{d}x) < \infty$, le second terme peut être rendu arbitrairement petit en prenant C suffisamment grand. Au final, $\mathbb{E}\rho_W^2(\mu,\mu_n)$ tend vers 0. Il reste à voir que le résultat de convergence souhaité, $\mathbb{E}D(\mu,q_n^\star) \to D^\star(\mu)$, est alors une conséquence facile de $\mathbb{E}\rho_W^2(\mu,\mu_n) \to 0$ et de l'inégalité (4.7). En effet, par (4.7)

$$D(\mu,q_n^\star) \leq D^\star(\mu) + 4\rho_W^2(\mu,\mu_n) + 4(D^\star(\mu))^{1/2}\rho_W(\mu,\mu_n).$$

Alors

$$0 \leq D(\mu,q_n^\star) - D^\star(\mu) \leq 4\rho_W^2(\mu,\mu_n) + 4(D^\star(\mu))^{1/2}\rho_W(\mu,\mu_n)$$

et en prenant l'espérance, puisque

$$\mathbb{E}((D^\star(\mu))^{1/2}\rho_W(\mu,\mu_n)) = (D^\star(\mu))^{1/2}\mathbb{E}\rho_W(\mu,\mu_n) \leq D^\star(\mu)(\mathbb{E}\rho_W(\mu,\mu_n)^2)^{1/2}$$

tend aussi vers 0 quand $n \to \infty$, on obtient le résultat souhaité. □

Analysons maintenant la vitesse de convergence de $q_n^\star$. Pour ce faire, nous supposerons qu'il existe une constante $R \geq 0$ telle que $\|X\| \leq R$, $\mathbb{P}$-p.s. Cette hypothèse est parfois appelée **contrainte de pic** dans le vocabulaire de la quantification.

Théorème 13. *S'il existe une constante $R \geq 0$ telle que $\|X\| \leq R, \mathbb{P}$-p.s., alors pour tout ordre $k \geq 1$,*

$$0 \leq \mathbb{E}D(\mu,q_n^\star) - D_k^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

Énonçons tout d'abord, sans preuve, un outil fondamental dans l'étude de la mesure empirique :

Lemme 7 (Principe de contraction). *Soit $\sigma_1, \dots, \sigma_n$ des variables aléatoires i.i.d. de loi de Rademacher, indépendantes de $X_1, \dots, X_n$, et soit $\mathscr{F}$ un ensemble borné de fonctions réelles, définies sur $\mathbb{R}^d$. On a*

$$\mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \, |f(X_i)| \leq \mathbb{E} \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i).$$

Remarques préliminaires.

1. Si $\|X\| \leq R$, $\mathbb{P}$-p.s, alors les centres optimaux $\mathbf{c}^\star$ sont dans $B_R := B(0, R)$. En effet, si $c \in \mathbb{R}^d$ avec $\|c\| > R$ et p est la projection sur B_R, alors, par définition de la projection, on a, $\forall x \in B_R$,

$$\begin{aligned} \|x - c\|^2 &= \|x - p(c)\|^2 + \|p(c) - c\|^2 - 2\langle x - p(c), c - p(c)\rangle \\ &\geq \|x - p(c)\|^2. \end{aligned}$$

   On a donc une distorsion plus petite pour des centres dans B_R.
2. Si $X \sim \mu$, on a

$$\begin{aligned} W(\mu, \mathbf{c}) &= \mathbb{E} \min_{1 \leq j \leq k} \|X - c_j\|^2 \\ &= \mathbb{E}\|X\|^2 + \mathbb{E} \min_{1 \leq j \leq k} \left(-2\langle X, c_j \rangle + \|c_j\|^2 \right). \end{aligned}$$

Ces deux observations nous conduisent à la conclusion suivante : plutôt que de minimiser $W(\mu, \cdot)$ sur $\mathbb{R}^{dk}$, il suffit donc de minimiser, sur B_R^k,

$$\bar{W}(\mu, \mathbf{c}) := \mathbb{E} \min_{1 \leq j \leq k} f_{c_j}(X), \quad \text{avec } f_c(x) = -2\langle x, c \rangle + \|c\|^2.$$

La même observation est valable en remplaçant μ par μ_n et en se rappelant que la distorsion empirique est une espérance par rapport à la mesure empirique μ_n, i.e.

$$\bar{W}(\mu_n, \mathbf{c}) := \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} f_{c_j}(X_i).$$

Démonstration du Théorème 13. On a

$$\begin{aligned}
&D(\mu, q_n^\star) - D^\star(\mu) \\
&\quad = W(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c}\in B_R^k} W(\mu, \mathbf{c}) \\
&\quad = \bar{W}(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c}\in B_R^k} \bar{W}(\mu, \mathbf{c}) \\
&\quad = \left[\bar{W}(\mu, \mathbf{c}_n^\star) - \bar{W}(\mu_n, \mathbf{c}_n^\star)\right] + \left[\inf_{\mathbf{c}\in B_R^k} \bar{W}(\mu_n, \mathbf{c}) - \inf_{\mathbf{c}\in B_R^k} \bar{W}(\mu, \mathbf{c})\right] \\
&\qquad \text{(par définition de } \mathbf{c}_n^\star\text{)} \\
&\quad \le \sup_{\mathbf{c}\in B_R^k} \left(\bar{W}(\mu, \mathbf{c}) - \bar{W}(\mu_n, \mathbf{c})\right) + \sup_{\mathbf{c}\in B_R^k} \left(\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})\right).
\end{aligned}$$

Dans la suite, nous cherchons à majorer le terme

$$\sup_{\mathbf{c}\in B_R^k} \left(\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})\right),$$

l'autre terme se bornant de façon similaire. En utilisant un n-échantillon indépendant annexe $X_1', \ldots, X_n'$ et un argument de symétrisation similaire à celui employé dans la preuve du théorème de Vapnik-Chervonenkis (Théorème 3), il vient

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{c}\in B_R^k} \left(\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})\right) \\
&= \mathbb{E} \sup_{\mathbf{c}\in B_R^k} \frac{1}{n} \sum_{i=1}^n \left(\min_{1\le j\le k} f_{c_j}(X_i) - \mathbb{E} \min_{1\le j\le k} f_{c_j}(X)\right) \\
&= \mathbb{E} \sup_{\mathbf{c}\in B_R^k} \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n \left(\min_{1\le j\le k} f_{c_j}(X_i) - \min_{1\le j\le k} f_{c_j}(X_i')\right) \middle| X_1, \ldots, X_n\right].
\end{aligned}$$

Ainsi, en observant que $\sup \mathbb{E}(\cdot) \le \mathbb{E} \sup(\cdot)$,

$$\begin{aligned}
\mathbb{E} \sup_{\mathbf{c}\in B_R^k} \left(\bar{W}(\mu_n, \mathbf{c}) - \bar{W}(\mu, \mathbf{c})\right) &\le \mathbb{E} \sup_{\mathbf{c}\in B_R^k} \frac{1}{n} \sum_{i=1}^n \left(\min_{1\le j\le k} f_{c_j}(X_i) - \min_{1\le j\le k} f_{c_j}(X_i')\right) \\
&\le 2\mathbb{E} \sup_{\mathbf{c}\in B_R^k} \frac{1}{n} \sum_{i=1}^n \sigma_i \min_{1\le j\le k} f_{c_j}(X_i).
\end{aligned}$$

Pour le traitement de ce dernier terme, nous allons procéder par itération sur l'ordre k des quantifieurs, en nous appuyant sur le principe de contraction. On note

$$S_k = \mathbb{E} \sup_{(c_1,\dots,c_k)\in B_R^k} \frac{1}{n}\sum_{i=1}^n \sigma_i \min_{1\le j\le k} f_{c_j}(X_i).$$

Cas $k = 1$. Comme $\|X\| \le R$:

$$\begin{aligned}
S_1 &= \mathbb{E}\sup_{c\in B_R} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(-2\langle X_i, c\rangle + \|c\|^2\big) \\
&\le 2\mathbb{E}\sup_{c\in B_R} \frac{1}{n}\sum_{i=1}^n \sigma_i\langle X_i, c\rangle + \mathbb{E}\sup_{c\in B_R}\frac{\|c\|^2}{n}\sum_{i=1}^n \sigma_i \\
&\le 2\mathbb{E}\sup_{c\in B_R} \frac{1}{n}\sum_{i=1}^n \sigma_i\langle X_i, c\rangle + \frac{R^2}{n}\mathbb{E}\left|\sum_{i=1}^n \sigma_i\right| \\
&\le 2\mathbb{E}\sup_{c\in B_R} \frac{1}{n}\sum_{i=1}^n \sigma_i\langle X_i, c\rangle + \frac{R^2}{\sqrt{n}} \\
&\qquad \text{(par l'inégalité de Cauchy-Schwarz).}
\end{aligned}$$

Ainsi, en utilisant que pour tout $u \in B_R$ on a $\sup_{c\in B_R}\langle u, c\rangle = R\|u\|$,

$$\begin{aligned}
S_1 &\le 2\mathbb{E}\sup_{c\in B_R}\frac{1}{n}\left|\left\langle \sum_{i=1}^n \sigma_i X_i, c\right\rangle\right| + \frac{R^2}{\sqrt{n}} \\
&= \frac{2R}{n}\mathbb{E}\left\|\sum_{i=1}^n \sigma_i X_i\right\| + \frac{R^2}{\sqrt{n}} \\
&\le 2R\sqrt{\frac{\mathbb{E}\|X\|^2}{n}} + \frac{R^2}{\sqrt{n}} \\
&\qquad \text{(par l'inégalité de Cauchy-Schwarz)} \\
&\le \frac{3R^2}{\sqrt{n}}.
\end{aligned}$$

Cas $k = 2$. Comme $\min(a, b) = \frac{a+b}{2} - \frac{|a-b|}{2}$, on a

$$\begin{aligned}
S_2 &= \mathbb{E}\sup_{(c_1,c_2)\in B_R^2}\frac{1}{2n}\sum_{i=1}^n \sigma_i\big(f_{c_1}(X_i) + f_{c_2}(X_i) - |f_{c_1}(X_i) - f_{c_2}(X_i)|\,\big) \\
&\le S_1 + \mathbb{E}\sup_{(c_1,c_2)\in B_R^2}\frac{1}{2n}\sum_{i=1}^n \sigma_i\big|f_{c_1}(X_i) - f_{c_2}(X_i)\big|.
\end{aligned}$$

En appliquant le principe de contraction, on obtient

$$S_2 \leq S_1 + \mathbb{E} \sup_{(c_1,c_2)\in B_R^2} \frac{1}{2n} \sum_{i=1}^{n} \sigma_i \big(f_{c_1}(X_i) - f_{c_2}(X_i)\big) \leq 2S_1.$$

Cas $k = 3$**.** Comme $S_2 \leq 2S_1$,

$$S_3 \leq \frac{S_1 + S_2}{2} + \frac{S_1 + S_2}{2} \leq 3S_1.$$

En itérant le procédé, on trouve

$$S_k \leq kS_1 \leq \frac{3kR^2}{\sqrt{n}}.$$

Finalement,

$$\mathbb{E}D(\mu, q_n^\star) - D^\star(\mu) \leq 4S_k \leq \frac{12kR^2}{\sqrt{n}},$$

d'où le théorème. □

Troisième partie

Statistique paramétrique

Chapitre 5

Statistique paramétrique asymptotique

Dans tout le chapitre, $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ désigne un modèle statistique paramétrique avec $\mathscr{H} \subset \mathbb{R}^d$ et $\Theta \subset \mathbb{R}^k$. Le paramètre d'intérêt est $g(\theta)$ avec $g : \Theta \to \mathbb{R}^p$ une fonction connue. L'objectif consiste dans un premier temps à estimer $g(\theta)$ à partir de l'observation $\mathbb{X} = (X_1, \dots, X_n)$ issue du modèle ; puis dans un second temps à faire des tests d'hypothèse sur le paramètre inconnu $g(\theta)$.

5.1 Rappels sur les estimateurs

Définition 8. *Une statistique est une fonction borélienne de l'observation* $\mathbb{X} = (X_1, \dots, X_n)$. *Un estimateur de* $g(\theta)$ *est une statistique qui prend ses valeurs dans un sur-ensemble de* $g(\Theta)$.

Dans la suite, $\mathbb{E}_\theta$ désigne l'espérance sous une loi paramétrée par θ et $\mathbb{V}_\theta Z$ désigne la matrice de variance-covariance (ou la variance si $s = 1$) de $Z \in \mathbb{L}^2(\mu_\theta)$ sous la loi μ_θ, i.e. pour une variable aléatoire intégrable Z à valeurs dans $\mathbb{R}^s$ et de loi μ_θ,

$$\mathbb{E}_\theta Z = \int_{\mathbb{R}^s} Z(x)\mu_\theta(\mathrm{d}x) \quad \text{et} \quad \mathbb{V}_\theta Z = \mathbb{E}_\theta \left(Z - \mathbb{E}_\theta Z\right)\left(Z - \mathbb{E}_\theta Z\right)^\intercal = \mathbb{E}_\theta ZZ^\intercal - \left(\mathbb{E}_\theta Z\right)\left(\mathbb{E}_\theta Z\right)^\intercal .$$

Une statistique $S(\mathbb{X})$ est d'ordre $q \in \mathbb{N}$ si $S(\mathbb{X}) \in \mathbb{L}^q(\mu_\theta)$ pour chaque $\theta \in \Theta$, i.e.

$$\mathbb{E}_\theta \left\|S(\mathbb{X})\right\|^q = \int_{\mathscr{H}^n} \left\|S(x)\right\|^q \mu_\theta(\mathrm{d}x) < \infty, \qquad \forall \theta \in \Theta.$$

Définition 9 (Biais)**.** *Soit $\hat{g}$ un estimateur d'ordre 1. On appelle biais la fonction $\theta \mapsto \mathbb{E}_\theta \hat{g} - g(\theta)$. L'estimateur $\hat{g}$ est dit sans biais lorsque cette fonction est nulle, i.e. $\mathbb{E}_\theta \hat{g} = g(\theta), \forall \theta \in \Theta$. Il est asymptotiquement sans biais lorsque $\lim_{n\to+\infty} \mathbb{E}_\theta \hat{g} = g(\theta), \forall \theta \in \Theta$.*

Exemples. Dans les exemples qui suivent, on se place dans le cadre d'un n-échantillon $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi $P_\theta = Q_\theta^{\otimes n}$.

1. Supposons que $\mathscr{H} \subset \mathbb{R}$ et que la probabilité Q_θ admette un moment d'ordre 2. La variance empirique S_n^2

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2,$$

   est alors un estimateur biaisé de $g(\theta) = \mathbb{V}_\theta X_1$. En effet, on a (exercice facile)

$$\mathbb{E}_\theta S_n^2 = \frac{n-1}{n} g(\theta).$$

   Voilà pourquoi, lorsque $n > 1$, on considère plutôt l'estimateur

$$\widetilde{S}_n^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

   appelé variance empirique corrigée qui, lui, estime sans biais $g(\theta) = \mathbb{V}_\theta X_1$. Notons également que la variance empirique S_n^2 est asymptotiquement sans biais.
2. Supposons que chaque X_i suive la loi $\mathscr{U}([0,\theta])$, $\theta > 0$. Dans ce modèle, l'estimateur $\hat{\theta} = 2\bar{X}_n$ obtenu par la méthode des moments (cf ci-dessous) est sans biais. Il n'en est pas de même pour l'EMV $\hat{\theta} = X_{(n)}$ (cf ci-dessous), car $\mathbb{E}_\theta \hat{\theta} < \theta$. On montre d'ailleurs facilement que $\mathbb{E}_\theta \hat{\theta} = \frac{n}{n+1}\theta$.

Maximum de vraisemblance. On suppose dans ce paragraphe que le modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est dominé par une mesure σ-finie ν, avec $\mathscr{H} \subset \mathbb{R}^d$ et $\Theta \subset \mathbb{R}^k$.

Définition 10. *La vraisemblance du modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est l'application $L_n : \mathscr{H}^n \times \Theta \to \mathbb{R}_+$ telle que, pour chaque $\theta \in \Theta$, $L_n(\cdot;\theta) : \mathscr{H}^n \to \mathbb{R}_+$ est un élément de la classe d'équivalence de la densité de P_θ par rapport à ν.*

Dans un modèle à échantillonnage i.i.d., l'expression de la vraisemblance se simplifie.

Proposition 6. *Soit L la vraisemblance du modèle $(\mathscr{H}, \{Q_\theta\}_{\theta\in\Theta})$ dominé par la mesure μ. Si, pour chaque $\theta \in \Theta$, $P_\theta = Q_\theta^{\otimes n}$, alors la fonction*

$$\begin{array}{rccc} L_n & : & \mathscr{H}^n \times \Theta & \to \mathbb{R}_+ \\ & & (x_1, \ldots, x_n, \theta) & \mapsto \prod_{i=1}^n L(x_i; \theta) \end{array}$$

est la vraisemblance du modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ pour $\nu = \mu^{\otimes n}$.

Démonstration. Il suffit de remarquer que, pour chaque $\theta \in \Theta$, l'application

$$(x_1, \ldots, x_n) \mapsto \prod_{i=1}^n L(x_i; \theta),$$

définie sur $\mathscr{H}^n$, est une version de la densité de $Q_\theta^{\otimes n}$ par rapport à $\nu = \mu^{\otimes n}$. □

Les deux cas les plus classiques en échantillonnage i.i.d. sont ceux où μ est la mesure de comptage sur $\mathscr{H}$ (cas discret) ou la mesure de Lebesgue (cas continu). On utilise alors souvent la notation $\mathbb{P}_\theta(X_1 = x)$ (cas discret) ou $f_\theta(x)$ (cas continu) en lieu et place de $L(x; \theta)$, de sorte que

$$L_n(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i)$$

pour le cas discret et

$$L_n(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n f_\theta(x_i)$$

pour le cas continu.

Définition 11. *Un estimateur du maximum de vraisemblance (EMV) est un estimateur $\hat{\theta}$ qui vérifie*

$$\hat{\theta} \in \underset{\theta\in\Theta}{Argmax}\, L_n(X_1, \ldots, X_n; \theta).$$

Remarques.

1. En pratique, la vraisemblance se maximise en θ à $X_1, \ldots, X_n$ « fixés » et l'éventuel EMV s'écrit comme une fonction de $X_1, \ldots, X_n$.
2. Ni l'existence, ni l'unicité des EMV ne sont en général acquises, mais on parle souvent par abus de « l'EMV » au lieu de dire « un EMV ». De plus, sous réserve d'existence, l'EMV peut ne pas avoir de représentation explicite ; dans ce cas, le recours à une méthode d'optimisation numérique est nécessaire afin de déterminer sa valeur en l'observation.
3. Un EMV, noté $\hat{\theta}$, est donc un estimateur du paramètre θ du modèle. Si le paramètre d'intérêt est $g(\theta)$, avec g une fonction borélienne connue définie sur Θ, on considère l'estimateur $g(\hat{\theta})$ dit *estimateur plug-in*. Par abus de langage, $g(\hat{\theta})$ est parfois qualifié d'EMV de $g(\theta)$.
4. Lorsque $\mathbb{X} = (X_1, \ldots, X_n)$ est un n-échantillon i.i.d., on préfère parfois calculer l'EMV en maximisant la log-vraisemblance [1]
$$\log L_n(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log L(x_i; \theta),$$
pour $(x_1, \ldots, x_n) \in \mathscr{H}^n$ et $\theta \in \Theta$. L'intérêt pratique est clair, l'étape de maximisation étant en principe plus facile à mener.
5. Sous certaines conditions de régularité du modèle, l'EMV possède de bonnes propriétés (existence, unicité, convergence, etc.).

Méthode des moments. Dans le cas particulier où le paramètre d'intérêt $g(\theta)$ est un moment de la loi Q_θ ou, par extension, une fonction de plusieurs moments de cette loi, la méthode des moments permet de construire des estimateurs naturels, en substituant à Q_θ la mesure empirique issue de l'échantillon. Ainsi, si $g(\theta) = \Psi(g_1(\theta), \ldots, g_q(\theta))$, où $g_j(\theta) = \mathbb{E}_\theta \Phi_j(X_1)$ et $\mathbb{E}_\theta \|\Phi_j(X_1)\| < \infty$, la méthode des moments consiste à utiliser l'estimateur
$$\hat{g} = \Psi\left(\frac{1}{n}\sum_{i=1}^{n} \Phi_1(X_i), \ldots, \frac{1}{n}\sum_{i=1}^{n} \Phi_q(X_i)\right).$$

Plus généralement, la méthode des moments propose d'estimer un paramètre θ comme la solution (si elle existe) d'un système d'équations
$$\frac{1}{n}\sum_{i=1}^{n} \Phi_j(X_i) = \mathbb{E}_\theta \Phi_j(X), \quad j = 1, \ldots, q, \tag{5.1}$$

1. Tous les logarithmes sont des logarithmes népériens.

pour un choix fixé de fonctions Φ_j. Le choix de $\Phi_j(x) = x^j$ correspond à la méthode des moments la plus simple. Lorsque le paramètre θ est k-dimensionnel, on cherche usuellement à résoudre le système induit par les k premiers moments de la loi Q_θ.

Remarque. Dans le cas particulier où la loi Q_θ est à support fini de taille k, les k premiers moments (théoriques) de cette loi caractérisent entièrement la distribution.

Avantage : L'estimateur a souvent de bonnes propriétés, obtenues via la loi des grands nombres ou le théorème central limite. Par ailleurs, pour les modèles issus de la famille exponentielle de rang plein [2], EMV et estimateur des moments coïncident.

Inconvénient : Pour utiliser cette méthode, il faut soit pouvoir exprimer $g(\theta)$ comme une fonction des moments de la loi Q_θ, ce qui n'est pas toujours possible (ou facile) ; soit être capable de résoudre le système des équations de moment. De plus, cette approche est en général moins efficace (en termes de variance asymptotique) que la méthode du maximum de vraisemblance.

Exemples.

1. $\mathbb{X} = (X_1, \dots, X_n)$ i.i.d. de loi commune Poisson $\mathscr{P}(\theta)$, $\theta > 0$. Ici encore, $\theta = \mathbb{E}_\theta X_1$, et l'on choisit donc $\hat{\theta} = \bar{X}_n$. Mais comme $\theta = \mathbb{V}_\theta X_1$, on peut aussi prendre $\hat{\theta} = S_n^2$.
2. $\mathbb{X} = (X_1, \dots, X_n)$ i.i.d., de loi commune $\mathscr{U}([0, \theta])$, $\theta > 0$. Il est facile de voir que $\theta = 2\mathbb{E}_\theta X_1$, d'où l'estimateur $\hat{\theta} = 2\bar{X}_n$. Noter que, pour ce modèle, l'estimateur obtenu par la méthode des moments est différent de l'EMV.

Risque quadratique et décomposition biais-variance. La proximité entre l'estimateur et le paramètre d'intérêt peut être évaluée par leur distance dans $\mathbb{L}^2(P_\theta)$.

Définition 12. *Soit $\hat{g}$ un estimateur d'ordre 2.*

1. *Pour $\theta \in \Theta$, le risque quadratique de $\hat{g}$ sous P_θ est*

$$\mathcal{R}(\hat{g}; \theta) = \mathbb{E}_\theta \left\| \hat{g} - g(\theta) \right\|^2 .$$

2. Le modèle de la famille exponentielle n'est pas détaillé dans ce document

2. *$\hat{g}$ est dit préférable à l'estimateur $\hat{g}'$ d'ordre 2 lorsque*

$$\mathcal{R}(\hat{g};\theta) \leq \mathcal{R}(\hat{g}';\theta), \quad \forall\theta \in \Theta.$$

On a la relation fondamentale suivante (dite décomposition « biais-variance ») :

$$\mathcal{R}(\hat{g};\theta) = \|\mathbb{E}_\theta\hat{g} - g(\theta)\|^2 + \mathbb{E}_\theta\|\hat{g} - \mathbb{E}_\theta\hat{g}\|^2, \forall\theta \in \Theta. \tag{5.2}$$

En effet,

$$\mathcal{R}(\hat{g};\theta) = \mathbb{E}_\theta \left\|\hat{g} - \mathbb{E}_\theta\hat{g} + \mathbb{E}_\theta\hat{g} - g(\theta)\right\|^2$$

$$= \mathbb{E}_\theta \left\|\hat{g} - \mathbb{E}_\theta\hat{g}\right\|^2 + \mathbb{E}_\theta \left\|\mathbb{E}_\theta\hat{g} - g(\theta)\right\|^2 + 2\mathbb{E}_\theta < \hat{g} - \mathbb{E}_\theta\hat{g}, \mathbb{E}_\theta\hat{g} - g(\theta) >$$

Or, comme dans le produit scalaire le 2ème terme est constant, on obtient :

$$\mathbb{E}_\theta < \hat{g} - \mathbb{E}_\theta\hat{g}, \mathbb{E}_\theta\hat{g} - g(\theta) >=< \mathbb{E}_\theta(\hat{g} - \mathbb{E}_\theta\hat{g}), \mathbb{E}_\theta\hat{g} - g(\theta) >=< 0, \mathbb{E}_\theta\hat{g} - g(\theta) >= 0,$$

ce qui prouve la décomposition biais-variance.

En particulier, pour $p = 1$,

$$\mathcal{R}(\hat{g};\theta) = \text{biais}^2(\theta) + \mathbb{V}_\theta\hat{g}.$$

L'intérêt de la décomposition (5.2) est qu'elle montre que, pour un risque quadratique donné, abaisser le biais revient à augmenter le terme de variance $\mathbb{E}_\theta\|\hat{g} - \mathbb{E}_\theta\hat{g}\|^2$, et réciproquement. Il est alors naturel de s'intéresser aux estimateurs qui minimisent uniformément la variance parmi les estimateurs sans biais de $g(\theta)$.

Définition 13 (VUMSB)**.** *Un estimateur $\hat{g}$ d'ordre 2 est de variance uniformément minimum parmi les estimateurs sans biais (VUMSB) s'il est sans biais et préférable à tout autre estimateur sans biais d'ordre 2.*

L'existence d'un estimateur VUMSB n'est en général pas acquise. Pour l'instant nous allons néanmoins noter que la variance de tous les estimateurs sans biais admet une borne inférieure, la borne de Cramer-Rao (5.5), que nous allons expliquer dans le paragraphe ci-dessous brièvement, sans préciser de bonnes hypothèses et seulement dans le cas $\Theta \subset \mathbb{R}$ ($k = 1$).

Information de Fisher et borne de Cramer-Rao. On suppose à nouveau que le modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est un modèle statistique dominé par une mesure σ-finie ν, avec $\mathscr{H} \subset \mathbb{R}^d$ et $\Theta \subset \mathbb{R}^k$. Notons

$$I_n(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big)^2\Big].$$

Cette quantité s'appelle *l'information de Fisher*.

Notons que sous de « bonnes » hypothèses d'interversion entre dérivées et intégrales et en supposant que le support S de la fonction $(x_1,\dots,x_n) \mapsto L_n(x_1,\dots,x_n,\theta)$ est le même pour tout $\theta \in \Theta$, on obtient :

$$\begin{aligned}
\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big] &= \int \frac{\frac{\partial}{\partial\theta}L_n(x_1,\dots,x_n,\theta)}{L_n(x_1,\dots,x_n,\theta)} L_n(x_1,\dots,x_n,\theta)d\nu \\
&= \int \frac{\partial}{\partial\theta}L_n(x_1,\dots,x_n,\theta)1_{L(x_1,\dots,x_n,\theta)>0}d\nu \\
&= \int \frac{\partial}{\partial\theta}L_n(x_1,\dots,x_n,\theta)1_S(x_1,\dots,x_n)d\nu \\
&= \frac{\partial}{\partial\theta}\int_S L_n(x_1,\dots,x_n,\theta)d\nu \\
&= \frac{\partial}{\partial\theta}1 = 0. \qquad (5.3)
\end{aligned}$$

Par conséquent, sous ces hypothèses, l'information de Fisher prend la forme équivalente

$$I_n(\theta) = \mathbb{V}_\theta\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big). \qquad (5.4)$$

Soit $\hat{g} = h(\mathbb{X})$ un estimateur *sans biais* de $g(\theta)$. Toujours sous de bonnes hypothèses d'interversion entre dérivées et intégrales, l'inégalité suivante a lieu :

$$\mathbb{V}_\theta(\hat{g}) \geq \frac{(g'(\theta))^2}{I_n(\theta)}. \qquad (5.5)$$

C'est la borne de Cramer-Rao.

Pour la prouver, on remarque que $h(X_1,\dots,X_n)$ étant sans biais, on a

$$g(\theta) = \int_S h(x_1,\dots,x_n)L_n(x_1,\dots,x_n,\theta)d\nu \text{ pour tout } \theta \in \Theta.$$

On calcule :

$$\begin{aligned} g'(\theta) &= \frac{\partial}{\partial\theta}\int_S h(x_1,\dots,x_n)L_n(x_1,\dots,x_n,\theta)d\nu \\ &= \int_S h(x_1,\dots,x_n)\Big(\frac{\partial}{\partial\theta}L_n(x_1,\dots,x_n,\theta)\Big)d\nu \\ &= \int_S h(x_1,\dots,x_n)\Big(\frac{\partial}{\partial\theta}\log L_n(x_1,\dots,x_n,\theta)\Big)L_n(x_1,\dots,x_n,\theta)d\nu \\ &= \mathbb{E}_\theta\Big(h(\mathbb{X})\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big)\Big). \end{aligned}$$

En tenant compte de la remarque (5.3)

$$g'(\theta) = \mathbb{E}_\theta\Big(h(\mathbb{X})\Big[\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big) - \mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big)\Big]\Big)$$

Et comme $\mathbb{E}_\theta h(\mathbb{X})$ est constante, on déduit finalement

$$g'(\theta) = \mathbb{E}_\theta\Big(\Big[h(\mathbb{X}) - \mathbb{E}_\theta h(\mathbb{X})\Big]\Big[\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big) - \mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big)\Big]\Big).$$

L'inégalité de Chauchy-Swartz conduit alors à

$$|g'(\theta)| \le \sqrt{\mathbb{V}_\theta\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big)} \times \sqrt{\mathbb{V}_\theta h(\mathbb{X})}$$

ce qui prouve la borne (5.5).

Exemple. $\mathbb{X} = (X_1,\dots,X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in (0,1)$. L'estimateur $\bar{X}_n$ de θ est sans biais, et sa variance vaut

$$\mathbb{V}_\theta \bar{X}_n = \frac{\mathbb{V}_\theta X_1}{n} = \frac{\theta(1-\theta)}{n},$$

car les variables aléatoires $X_1,\dots,X_n$ sont indépendantes et de même loi $\mathcal{B}(\theta)$. Par suite, d'après la décomposition biais-variance (5.2) :

$$\mathcal{R}(\bar{X}_n;\theta) = \frac{\theta(1-\theta)}{n}.$$

En augmentant n, l'estimateur $\bar{X}_n$ gagne donc en précision. Ce n'est pas le cas pour l'estimateur X_1, de risque quadratique $\mathcal{R}(X_1;\theta) = \theta(1-\theta)$. Comme on pouvait s'y attendre, $\bar{X}_n$ est donc préférable à X_1.

Par ailleurs, calculons l'information de Fisher. On peut commencer par remarquer que dans un modèle i.i.d., partant de (5.4), on obtient

$$I_n(\theta) = \mathbb{V}_\theta\Big(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n L(X_i,\theta)\Big) = \mathbb{V}_\theta\Big(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log L(X_i,\theta)\Big)$$
$$= n\mathbb{V}_\theta\Big(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\Big),$$

que l'on notera également $nI(\theta)$. Dans le cas de variables de loi de Bernoulli, on obtient

$$\log L(X_1,\theta) = X_1\log\theta + (1-X_1)\log(1-\theta)$$

et

$$I_n(\theta) = \mathbb{V}_\theta\Big(\frac{\partial}{\partial\theta}\log L_n(\mathbb{X},\theta)\Big) = n\mathbb{V}_\theta\Big(\frac{X_1}{\theta} - \frac{1-X_1}{1-\theta}\Big) = n\mathbb{V}_\theta\frac{X_1}{\theta(1-\theta)}$$
$$= n\frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Comme $g'(\theta) = 1$, nous voyons que

$$\mathbb{V}_\theta\bar{X}_n = \frac{g'(\theta)}{I_n(\theta)}.$$

La borne de Rao-Cramer est atteinte, donc $\bar{X}_n$ est VUMSB.

Exemple. $\mathbb{X} = (X_1,\dots,X_n)$ de loi commune de Poisson $\mathcal{P}(\theta)$. Nous avons déjà présenté deux estimateurs sans biais pour $g(\theta) = \theta$, à savoir $\bar{X}_n$ et $\widetilde{S}_n^2$. Calculons l'information de Fisher $I_n(\theta) = nI(\theta)$. En se servant de la remarque (5.3) pour $n = 1$:

$$\mathbb{V}_\theta\Big(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\Big) = \mathbb{E}_\theta\Big(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\Big)^2$$
$$= \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}(-\theta + X_1\log\theta - \log X_1!\Big)^2\Big]$$
$$= \mathbb{E}_\theta\Big[\Big(\frac{X_1}{\theta} - 1\Big)^2\Big] = \frac{1}{\theta^2}\mathbb{V}_\theta X_1 = \frac{1}{\theta}.$$

Alors l'information de Fisher $I_n(\theta) = \frac{n}{\theta}$. Notons que $\mathbb{V}\bar{X}_n = \frac{1}{n}\mathbb{V}_\theta X_1 = \frac{\theta}{n}$. Comme $g(\theta) = \theta$, nous pouvons conclure que

$$\mathbb{V}_\theta\bar{X}_n = \frac{g'(\theta)}{I_n(\theta)}.$$

La borne de Rao-Cramer est atteinte, donc $\bar{X}_n$ est un estimateur VUMSB.

5.2 M- et Z-estimateurs

De façon générale, les estimateurs sont souvent construits comme maximas d'un critère empirique, ou de façon souvent équivalente, comme « zeros » (solutions d'une équation) d'un critère empirique (la dérivée du précédent).

Définition 14. *Dans un modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, un M-estimateur a la forme générale*[3]

$$\hat{\theta}^M \in \underset{\theta\in\Theta}{Argmax}\ \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i), \tag{5.6}$$

où pour tout $\theta \in \Theta$, $x \mapsto m_\theta(x)$ est une fonction réelle connue (on peut autoriser à valeurs dans $\bar{\mathbb{R}} = [-\infty, +\infty]$).

Lorsque le critère est différentiable, on peut également chercher les points critiques, d'où la définition suivante d'un Z-estimateur.

Définition 15. *Dans un modèle $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$, un Z-estimateur noté $\hat{\theta}^Z$ est obtenu comme solution de l'équation*

$$\frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = 0,$$

où pour tout $\theta \in \Theta$, $x \mapsto \psi_\theta(x)$ est une fonction vectorielle connue.

L'étude des M- et Z- estimateurs se fait dans un cadre asymptotique, lorsque la taille d'échantillon grandit. Il est donc naturel d'élargir les définitions au cas où les conditions ci-dessus sont réalisées de façon approchée, i.e. à $o(1)$ près qui tend vers 0 lorsque la taille d'échantillon augmente. Noter que les restes sont (en général) des variables aléatoires ; il faut donc préciser le type de convergence (généralement, en probabilités, ce que l'on note parfois $o_P(1)$).

Les estimateurs de moment, sont en fait des Z-estimateurs. En effet, la caractérisation donnée dans (5.1) indique que estimateurs de moment sont des Z-estimateurs pour $\psi_\theta(x) = \Phi(x) - \mathbb{E}_\theta\Phi(X)$ avec $\Phi = (\Phi_1, \ldots, \Phi_q)$.

3. La notation Argmax désigne l'ensemble des maximas (supremums) d'une fonction.

Dans un modèle à échantillonnage i.i.d., un estimateur du maximum de vraisemblance est en fait un M-estimateur pour la fonction $m_\theta = \log L(\cdot;\theta)$ où L est la vraisemblance du modèle $(\mathscr{H}, Q_\theta)$. Par ailleurs, si $\theta \mapsto \log L(\cdot;\theta)$ est différentiable, alors l'EMV est souvent défini comme Z-estimateur avec $\psi_\theta = \nabla_\theta \log L(\cdot;\theta)$. En fait, les EMV ne sont des M-estimateurs que dans les modèles à échantillonnage i.i.d.. Cependant, dans les modèles qui ne sont pas à échantillonnage i.i.d., on peut construire des estimateurs du maximum de *pseudo-vraisemblance*, qui sont des M-estimateurs. Même si le modèle n'est pas à échantillonnage i.i.d., ces estimateurs ont parfois de bonnes propriétés.

Un exemple plus exotique est donné par la médiane empirique.

Exemple. Supposons que $\mathscr{H} \subset \mathbb{R}$ (i.e. $d = 1$) et que le paramètre θ est la médiane de la loi Q_θ, i.e. $Q_\theta(X \leq \theta) \geq 1/2$ et $Q_\theta(X \geq \theta) \geq 1/2$. On se place dans le cas d'une loi Q_θ sans atomes. On constate que la médiane empirique, définie par

$$\inf\left\{t \in \mathbb{R}; \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} \geq \frac{1}{2}\right\}$$

est un Z-estimateur (approché) de θ avec $\psi_\theta(x) = \operatorname{sign}(x-\theta) = \mathbb{1}_{x>\theta} - \mathbb{1}_{x<\theta}$. En effet, le critère vaut alors

$$\frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = \frac{2}{n}\left|\left\{i \in \{1,\dots,n\}; X_i > \theta\right\}\right| - 1 - \frac{1}{n}\left|\left\{i \in \{1,\dots,n\}; X_i = \theta\right\}\right|.$$

Comme la loi est sans atomes, le dernier terme est d'espérance nulle et par la loi des grands nombres, il tend P_θ-ps vers 0. La médiane empirique $\hat{\theta}^{\text{med}}$ vérifie

$$\left|\left\{i \in \{1,\dots,n\}; X_i > \hat{\theta}^{\text{med}}\right\}\right| \geq \frac{n}{2} \text{ et } \left|\left\{i \in \{1,\dots,n\}; X_i < \hat{\theta}^{\text{med}}\right\}\right| \geq \frac{n}{2}$$

c'est donc un zéro approché du critère.

On peut généraliser cet exemple au p-ième quantile empirique, $p \in (0,1)$, qui est un Z-estimateur approché pour la fonction $\psi_\theta(x) = \mathbb{1}_{x \leq \theta} - p$.

5.3 Théorème de Wald

L'observation $\mathbb{X} = (X_1, \dots, X_n)$ contient de plus en plus d'information sur la vraie valeur du paramètre à mesure que sa taille n croît. De ce fait, on est

amené à s'intéresser aux propriétés asymptotiques des estimateurs. Dans la suite, sauf mention explicite du contraire, toute propriété de convergence sera entendue pour une taille d'échantillon n qui tend vers l'infini.

Définition 16. *L'estimateur $\hat{g}$ est dit consistant lorsque*

$$\hat{g} \xrightarrow{P_\theta} g(\theta), \quad \forall \theta \in \Theta.$$

Exemple. L'estimateur $\bar{X}_n$ de $g(\theta) = \mathbb{E}_\theta X_1$ construit avec un n-échantillon i.i.d. $\mathbb{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{dn}$ satisfaisant $\mathbb{E}_\theta \|X_1\| < \infty$ est consistant car, d'après la loi faible des grands nombres :

$$\bar{X}_n \xrightarrow{P_\theta} \mathbb{E}_\theta X_1, \quad \forall \theta \in \Theta.$$

Pour $d = 1$, il en est de même de la variance empirique dès que $\mathbb{E}_\theta X_1^2 < \infty$ puisque, toujours par la loi faible des grands nombres,

$$S_n^2 \xrightarrow{P_\theta} \mathbb{V}_\theta X_1, \quad \forall \theta \in \Theta.$$

Remarque. Consistance et absence de biais asymptotique ne sont pas les mêmes notions. Par exemple, pour se convaincre qu'un estimateur consistant n'est pas nécessairement asymptotiquement sans biais, considérons le modèle statistique $(\mathbb{R}^n, \{\mathcal{N}(\theta, 1)^{\otimes n}\}_{\theta \in (0,1)})$ et l'estimateur $\hat{\theta}$ de θ issu de $\mathbb{X} = (X_1, \ldots, X_n) \sim P_\theta = \mathcal{N}(\theta, 1)^{\otimes n}$ défini par

$$\hat{\theta} = \bar{X}_n + \frac{1}{\Phi(-\sqrt{n})} \mathbb{1}_{\{\bar{X}_n \leq 0\}},$$

où Φ désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$. L'estimateur $\bar{X}_n$ est consistant d'après la loi faible des grands nombres. En effet, comme $\theta > 0$, pour chaque $\varepsilon > 0$:

$$\lim_{n \to +\infty} \mathbb{P}\left(\frac{1}{\Phi(-\sqrt{n})} \mathbb{1}_{\{\bar{X}_n \leq 0\}} > \varepsilon \right) = \lim_{n \to +\infty} \mathbb{P}(\bar{X}_n \leq 0) = 0.$$

On en déduit la consistance de $\hat{\theta}$. Or, comme $\bar{X}_n$ suit la loi $\mathcal{N}(\theta, 1/n)$ et $\theta \leq 1$:

$$\begin{aligned} \mathbb{P}(\bar{X}_n \leq 0) &= \frac{1}{\sqrt{2\pi \times \frac{1}{n}}} \int_{-\infty}^{0} \exp\Big(- \frac{(x-\theta)^2}{\frac{2}{n}} \Big) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\theta\sqrt{n}} e^{-t^2/2} \mathrm{d}t = \Phi(-\theta\sqrt{n}) \\ &\geq \Phi(-\sqrt{n}). \end{aligned}$$

En conséquence,

$$\mathbb{E}_\theta \hat{\theta} = \mathbb{E}_\theta \bar{X}_n + \frac{1}{\Phi(-\sqrt{n})} P_\theta(\bar{X}_n \leq 0) \geq \theta + 1,$$

donc $\hat{\theta}$ est biaisé, et même asymptotiquement biaisé.

On se place à présent dans le cas d'un n-échantillon $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d. Le théorème de Wald est le résultat classique qui donne la consistance des M-estimateurs (5.6). Historiquement, Wald a prouvé en 1949 la consistance du maximum de vraisemblance par cette méthode. Notez qu'il en découle aussi la consistance des Z-estimateurs en remarquant que ce dernier est un M-estimateur pour $m_\theta = -\|\psi_\theta\|$. Nous l'énonçons ici dans une forme simplifiée et renvoyons au livre de van der Vaart pour des hypothèses plus générales.

Théorème 14 (Wald). *On suppose que*
a) pour tout x, $\theta \mapsto m_\theta(x)$ est continue ;
b) pour toute boule ouverte $U \subset \Theta$ assez petite, $\mathbb{E} \sup_{\theta' \in U} m_{\theta'}(X_1) < +\infty$;
c) l'espace Θ est compact.
Alors l'estimateur $\hat{\theta}^M$ défini par (5.6) converge en probabilité vers l'ensemble des maxima de $\theta \mapsto \mathbb{E} m_\theta(X_1)$.

Démonstration. On note $\Theta_0 = \mathrm{Argmax}_{\theta \in \Theta} \mathbb{E} m_\theta(X_1)$ l'ensemble des maxima et on suppose que cet ensemble est non vide (sinon il n'y a rien à prouver).

Fixons $\theta \in \Theta$ et $(U_j)_{j \geq 1}$ une suite de boules ouvertes autour de θ qui décroît vers $\{\theta\}$. La suite $(\sup_{\theta' \in U_j} m_{\theta'}(x))_{j \geq 1}$ est une suite décroissante, minorée par $m_\theta(x)$ et d'après a) elle converge vers $m_\theta(x)$. D'après b), le théorème de convergence monotone s'applique et donne la convergence de $\mathbb{E} \sup_{\theta' \in U_j} m_{\theta'}(X_1)$ vers $\mathbb{E} m_\theta(X_1)$.

Supposons que $\theta \notin \Theta_0$, i.e. $\mathbb{E} m_\theta(X_1) < \sup_{\theta' \in \Theta} \mathbb{E} m_{\theta'}(X_1)$. D'après le résultat de convergence qui précède, il existe une boule ouverte U_θ autour de θ telle que $\mathbb{E} \sup_{\theta' \in U_\theta} m_{\theta'}(X_1) < \sup_{\theta' \in \Theta} \mathbb{E} m_{\theta'}(X_1)$. Fixons $\varepsilon > 0$. Comme l'ensemble $B_\varepsilon = \{\theta \in \Theta ; d(\theta, \Theta_0) \geq \varepsilon\}$ est compact et recouvert par l'union des boules U_θ pour $\theta \in B_\varepsilon$, on peut extraire un sous-recouvrement fini $U_{\theta_1}, \ldots, U_{\theta_r}$, d'où il vient

$$\sup_{\theta \in B_\varepsilon} \frac{1}{n} \sum_{i=1}^n m_\theta(X_i) \leq \sup_{\theta \in B_\varepsilon} \frac{1}{n} \sum_{i=1}^n \sup_{\theta' \in U_\theta} m_{\theta'}(X_i) \leq \max_{1 \leq j \leq r} \frac{1}{n} \sum_{i=1}^n \sup_{\theta' \in U_{\theta_j}} m_{\theta'}(X_i).$$

Par la loi des grands nombres (qui s'applique sous l'hypothèse b)), le terme de droite converge $\mathbb{P}$-presque sûrement vers

$$\max_{1\le j\le r} \mathbb{E} \sup_{\theta'\in U_{\theta_j}} m_{\theta'}(X_1) < \sup_{\theta'\in\Theta} \mathbb{E} m_{\theta'}(X_1) = \mathbb{E} m_{\theta_0}(X_1),$$

pour tout $\theta_0 \in \Theta_0$. Ainsi, si $\hat{\theta}^M \in B_\varepsilon$, alors

$$\frac{1}{n}\sum_{i=1}^n m_{\hat{\theta}^M}(X_i) \le \sup_{\theta\in B_\varepsilon} \frac{1}{n}\sum_{i=1}^n m_\theta(X_i) < \mathbb{E} m_{\theta_0}(X_1) = \frac{1}{n}\sum_{i=1}^n m_{\theta_0}(X_i) + R_n,$$

où

$$R_n = \mathbb{E} m_{\theta_0}(X_1) - \frac{1}{n}\sum_{i=1}^n m_{\theta_0}(X_i)$$

tend $\mathbb{P}$-ps vers 0 en utilisant à nouveau la loi des grands nombres (cette fois sur m_{θ_0} au lieu du sup). Ainsi,

$$\{\hat{\theta}^M \in B_\varepsilon\} \subset \Big\{ \frac{1}{n}\sum_{i=1}^n m_{\hat{\theta}^M}(X_i) \le R_n + \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n m_\theta(X_i) \Big\} = \{R_n \ge 0\},$$

par définition de $\hat{\theta}^M$. La probabilité de cet évènement tend vers 0, ce qui termine la preuve. □

Remarque : Dans le cas du maximum de vraisemblance d'un n-échantillon $\mathbb{X} = (X_1, \dots, X_n)$ i.i.d., on a $m_\theta = \log L(\cdot;\theta)$ et lorsque le modèle est identifiable, l'ensemble des maxima Θ_0 de la fonction $\theta \mapsto \mathbb{E}\log L(X_1;\theta)$ est réduit au « vrai » paramètre. En effet, notons $\theta^\star$ le paramètre de la loi de $\mathbb{X}$, i.e. on travaille sous $\mathbb{E} = \mathbb{E}_{\theta^\star}$. Par ailleurs, supposons pour simplifier les notations que la loi commune des X_i a la densité $f_{\theta^\star}$ (par rapport à une mesure dominante μ). Alors

$$\ell(\theta) := \mathbb{E}\log L(X_1;\theta) = \int \log f_\theta(x) f_{\theta^\star}(x)\mu(dx)$$

et par ailleurs

$$\begin{aligned}\ell(\theta) - \ell(\theta^\star) &= \int \log \frac{f_\theta(x)}{f_{\theta^\star}(x)} f_{\theta^\star}(x)\mu(dx) \\ &\le \log \int \frac{f_\theta(x)}{f_{\theta^\star}(x)} f_{\theta^\star}(x)\mu(dx) = 0,\end{aligned}$$

d'après l'inégalité de Jensen et car f_θ est une densité. De plus, l'inégalité ci-dessus est stricte, sauf si $f_\theta = f_{\theta^\star}$ presque sûrement. Dans un modèle identifiable, on obtient donc $\ell(\theta) \le \ell(\theta^\star)$ avec égalité seulement en $\theta = \theta^\star$. La consistance de l'EMV est donc une conséquence du théorème de Wald.

5.4 Vitesse de convergence et loi limite

La consistance ne doit être vue que comme une propriété minimale que doit satisfaire un estimateur. Elle ne permet cependant pas de préciser l'erreur commise, d'où la définition qui suit.

Définition 17. *Soit $(v_n)_{n\geq 1}$ une suite de réels positifs telle que $v_n \to +\infty$. L'estimateur $\hat{g}$ est dit de vitesse v_n si, pour chaque $\theta \in \Theta \subset \mathbb{R}^k$, il existe une loi $\ell(\theta)$ sur $\mathbb{R}^p$ différente de la loi de Dirac en 0, appelée loi limite de $\hat{g}$, telle que*

$$v_n \left(\hat{g} - g(\theta)\right) \xrightarrow{\mathscr{L}(P_\theta)} \ell(\theta).$$

Si toutes les lois $\ell(\theta)$ sont gaussiennes, $\hat{g}$ est dit asymptotiquement normal.

La qualité d'un estimateur est ainsi évaluée sur sa vitesse car il est alors d'autant plus proche de $g(\theta)$ qu'elle est rapide, mais aussi sur la variance de la loi limite, qui doit idéalement être faible afin que l'estimateur se concentre sur le paramètre d'intérêt.

Exemples.

1. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in (0,1)$. L'estimateur $\hat{\theta} = \bar{X}_n$ de θ est consistant. Il est aussi asymptotiquement normal de vitesse $\sqrt{n}$ car, pour chaque $\theta \in (0,1)$:

   $$\sqrt{n}\left(\bar{X}_n - \theta\right) \xrightarrow{\mathscr{L}(P_\theta)} \mathcal{N}\left(0, \theta(1-\theta)\right),$$

   d'après le théorème central limite. Noter que la variance de la loi limite prend ses valeurs les plus faibles lorsque θ est proche de 0 ou de 1 et ses valeurs les plus grandes lorsque θ est proche de $1/2$. De ce fait, l'estimation de θ par $\bar{X}_n$ est d'autant meilleure que θ est proche de 0 ou de 1 car la loi limite de l'estimateur $\bar{X}_n$ est alors très peu dispersée.
2. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{P}(\theta)$, $\theta > 0$. Ici encore, l'estimateur $\hat{\theta} = \bar{X}_n$ est consistant et asymptotiquement normal de vitesse $\sqrt{n}$, car

   $$\sqrt{n}\left(\bar{X}_n - \theta\right) \xrightarrow{\mathscr{L}(P_\theta)} \mathcal{N}(0, \theta),$$

   toujours d'après le théorème central limite.

3. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{U}([0,\theta])$, $\theta > 0$. Il est facile de voir (exercice) que l'EMV $\hat{\theta} = X_{(n)}$ est consistant et de vitesse n, car
$$n(X_{(n)} - \theta) \xrightarrow{\mathscr{L}(P_\theta)} Z,$$
où $-Z \sim \mathscr{E}(1/\theta)$. (Suggestion : calculer la fonction de répartition de $n(X_{(n)} - \theta)$.) De ce point de vue, il est plus performant (malgré son caractère biaisé) que l'estimateur $2\bar{X}_n$ obtenu par la méthode des moments, qui ne converge qu'à la vitesse $\sqrt{n}$.

Pour fixer les idées, on suppose dans la suite que l'estimateur $\hat{\theta}$ de θ est de vitesse v_n, i.e.
$$v_n(\hat{\theta} - \theta) \xrightarrow{\mathscr{L}(P_\theta)} \ell(\theta), \quad \forall \theta \in \Theta, \tag{5.7}$$
avec $\ell(\theta)$ une loi sur $\mathbb{R}^k$ différente de la mesure de Dirac en 0 et $v_n \to +\infty$.

La loi de l'erreur renormalisée $v_n(\hat{\theta} - \theta)$ est proche de la loi $\ell(\theta)$ pour les grandes valeurs de n. Or, $\ell(\theta)$ est inconnu car θ est inconnu, donc comment peut-on préciser cette erreur d'approximation ? De plus, comment exploiter cette propriété asymptotique lorsque le paramètre d'intérêt est $g(\theta)$? Sous réserve d'hypothèses supplémentaires, nous allons examiner de quelle manière il est possible d'apporter des réponses à ces questions. Commençons au préalable par énoncer le lemme très utile suivant dont la preuve, facile, (passant par les fonctions caractéristiques) est laissée à la lectrice.

Lemme 8 (Lemme de Slutsky)**.** *Soit $(Z_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ des suites de variables aléatoires à valeurs dans $\mathbb{R}^k$ et $\mathbb{R}^q$ telles que $(Z_n)_{n\geq 1}$ converge en loi vers une variable aléatoire Z et $(Y_n)_{n\geq 1}$ converge en loi vers une constante $y \in \mathbb{R}^q$. Alors, la suite des couples $((Z_n, Y_n))_{n\geq 1}$ converge en loi vers le couple (Z, y).*

Le plus souvent, on applique à cette convergence jointe $(Z_n, Y_n) \longrightarrow^{\mathscr{L}(P_\theta)} (Z, y)$ une fonction continue h (somme, multiplication, etc.) et l'on en tire que $h(Z_n, Y_n) \longrightarrow^{\mathscr{L}(P_\theta)} h(Z, y)$.

On notera en particulier que la convergence (5.7) implique que $\hat{\theta}$ tend vers θ en probabilité. En effet, comme $v_n^{-1} \to 0$ lorsque $n \to \infty$, alors par le lemme de Slutsky : $(v_n^{-1}, v_n(\hat{\theta} - \theta)) \longrightarrow^{\mathscr{L}(P_\theta)} (0, \ell(\theta))$, donc : $v_n^{-1} \times v_n(\hat{\theta} - \theta) = (\hat{\theta} - \theta) \xrightarrow{\mathscr{L}} 0 \times \ell(\theta) = 0$. Donc $\hat{\theta} - \theta \to 0$ en loi et par conséquent aussi en probabilité.

Pour éviter tout malentendu, un rappel élémentaire : la convergence en probabilité implique toujours la convergence en loi. La convergence en loi de manière générale n'implique pas la convergence en probabilité. Cependant, la convergence en loi vers une constante implique la convergence en probabilité.

Estimation de la variance limite - cas de θ. Supposons qu'il existe une fonction connue $\sigma : \Theta \to \mathbb{R}^\star$ et une loi connue τ sur $\mathbb{R}^k$ telles que pour chaque $\theta \in \Theta$, $\ell(\theta) = \sigma(\theta)\tau$. Pourvu que l'on dispose d'un estimateur consistant $\hat{\sigma}$ de $\sigma(\theta)$, on déduit du lemme de Slutsky que

$$\left(v_n(\hat{\theta} - \theta), \hat{\sigma}\right) \xrightarrow{\mathscr{L}(P_\theta)} \left(\sigma(\theta)W, \sigma(\theta)\right), \quad \forall \theta \in \Theta,$$

où W est une variable aléatoire de loi τ. Comme la convergence en loi est préservée par la composition des fonctions continues,

$$\frac{v_n}{\hat{\sigma}}(\hat{\theta} - \theta) \xrightarrow{\mathscr{L}(P_\theta)} W, \quad \forall \theta \in \Theta.$$

Ainsi, la loi de l'erreur renormalisée $(v_n/\hat{\sigma})(\hat{\theta} - \theta)$ est proche de celle de τ pour les grandes valeurs de n. Cette dernière ne dépend plus de θ inconnu.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta), \theta \in (0,1)$. Le théorème central limite donne

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathscr{L}(P_\theta)} \mathcal{N}\left(0, \theta(1-\theta)\right), \quad \forall \theta \in (0,1).$$

De plus, $\sqrt{\bar{X}_n(1-\bar{X}_n)}$ est un estimateur consistant de $\sqrt{\theta(1-\theta)}$ d'après la loi des grands nombres. La loi asymptotique de l'erreur renormalisée est donc $\mathcal{N}(0,1)$, car

$$\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}(\bar{X}_n - \theta) \xrightarrow{\mathscr{L}(P_\theta)} \mathcal{N}(0,1), \quad \forall \theta \in (0,1).$$

Ce résultat peut alors être exploité pour encadrer le paramètre inconnu θ.

Vitesse de convergence de $g(\theta)$. Revenons au problème plus général de l'estimation du paramètre $g(\theta) \in \mathbb{R}^p$. Comme l'indique le résultat qui suit, le calcul de la vitesse de l'estimateur $g(\hat{\theta})$ est immédiat, sous réserve que g possède les propriétés analytiques adéquates.

Théorème 15 (δ-méthode). *Soit $(v_n)_{n\geq 1}$ une suite de réels qui tend vers $+\infty$, $z \in \mathbb{R}^k$ et $(Z_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^k$ telle que $v_n(Z_n - z)$ converge en loi vers une loi ℓ sur $\mathbb{R}^k$. Si $g : \mathbb{R}^k \to \mathbb{R}^p$ est de classe $\mathscr{C}^1$, de matrice jacobienne J_g alors $v_n(g(Z_n) - g(z))$ converge en loi vers la loi $\mathrm{J}_g(z)\ell$ sur $\mathbb{R}^p$.*

Ainsi, si la fonction g est de classe $\mathscr{C}^1$, on a, pour tout $\theta \in \Theta$,

$$v_n \left(g(\hat{\theta}) - g(\theta)\right) \xrightarrow{\mathscr{L}(P_\theta)} \mathrm{J}_g(\theta)\ell(\theta),$$

$J_g(\theta)$ désignant la matrice jacobienne de g évaluée en θ. De ce fait, $g(\hat{\theta})$ est, comme $\hat{\theta}$, un estimateur de vitesse v_n dès que la loi $\mathrm{J}_g(\theta)\ell(\theta)$ est différente de la loi de Dirac en 0.

Comme dans la partie précédente, on peut préciser l'erreur commise en approchant $g(\theta)$ par $g(\hat{\theta})$ au moins lorsqu'il existe une fonction $\sigma : \Theta \to \mathbb{R}^\star$ et une loi τ sur $\mathbb{R}^k$ telles que, pour chaque $\theta \in \Theta$, $\ell(\theta) = \sigma(\theta)\tau$. En effet, si $J_g(\theta)$ est une matrice carrée (donc $k = p$) inversible pour chaque $\theta \in \Theta$ et $\hat{\sigma}$ est un estimateur consistant de $\sigma(\theta)$, on déduit du lemme de Slutsky que

$$\frac{v_n}{\hat{\sigma}} J_g(\hat{\theta})^{-1} \left(g(\hat{\theta}) - g(\theta)\right) \xrightarrow{\mathscr{L}(P_\theta)} \tau, \quad \forall \theta \in \Theta,$$

car, g étant de classe $\mathscr{C}^1$, $J_g(\hat{\theta})$ est un estimateur consistant de $J_g(\theta)$. La loi de l'erreur renormalisée $(v_n/\hat{\sigma})J_g(\hat{\theta})^{-1}(g(\hat{\theta}) - g(\theta))$ est donc proche de celle de τ pour les grandes valeurs de n.

Remarque : Si $k \neq p$, on peut aussi se demander si la loi $\mathrm{J}_g(\theta)\ell(\theta)$ sur $\mathbb{R}^p$ s'écrit sous la forme $\sigma(\theta)\tau'$ où τ' est une loi sur $\mathbb{R}^p$ indépendant de θ et appliquer Slutsky.

Exemple. $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathscr{B}(\theta)$, $\theta \in (0,1)$ et $\theta \neq 1/2$. Si $g(\theta) = \theta(1-\theta)$ est le paramètre d'intérêt, la méthode des moments nous conduit à considérer l'estimateur $g(\bar{X}_n)$. Le théorème central limite et la δ-méthode donnent alors

$$\sqrt{n}\left(g(\bar{X}_n) - g(\theta)\right) \xrightarrow{\mathscr{L}(P_\theta)} (1-2\theta)\mathcal{N}\left(0, \theta(1-\theta)\right) \overset{\mathscr{L}(P_\theta)}{=} \mathcal{N}\left(0, \theta(1-\theta)(1-2\theta)^2\right),$$

pour tout $\theta \in (0,1)$. Puis, la loi des grands nombres et le lemme de Slutsky montrent que

$$\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)(1-2\bar{X}_n)^2}}\left(g(\bar{X}_n) - g(\theta)\right) \xrightarrow{\mathscr{L}(P_\theta)} \mathcal{N}(0,1), \quad \forall \theta \in (0,1) \setminus \{1/2\}.$$

Remarque. L'utilisation de la δ-méthode ne se limite pas à l'obtention de lois limites pour les estimateurs de $g(\theta)$. Pour s'en convaincre, considérons un n-échantillon $\mathbb{X} = (X_1, \dots, X_n)$ i.i.d., de loi commune $\mathscr{P}(\theta)$, $\theta > 0$. Dans ce contexte, la variance empirique $\hat{\theta} = S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2$ est un estimateur consistant de θ. Le théorème central limite multivarié (appliqué au couple de variables aléatoires $(\frac{1}{n}\sum_{i=1}^n X_i^2, \bar{X}_n)$) et la δ-méthode (appliquée avec la fonction $g(x,y) = x - y^2$) conduisent alors à

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathscr{L}(P_\theta)} \mathcal{N}(0, \theta + 2\theta^2), \quad \forall \theta > 0.$$

En effet $\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n X_i^2 - \mathbb{E}_\theta X_1^2,\ \frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}_\theta X_1\Big)$ converge en loi vers $\mathcal{N}_2(\mathbf{0}, B)$ la loi gaussienne dans $\mathbb{R}^2$ centrée, de matrice de variance-covariance B avec $b_{11} = \mathbb{V}X_1^2$, $b_{22} = \mathbb{V}X_1$, $b_{12} = b_{21} = \mathrm{cov}(X_1^2, X_1)$. D'après la δ-méthode appliqué avec $g(x,y) = x - y^2$, on a

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2 - (\mathbb{E}_\theta X_1^2 - (\mathbb{E}_\theta X_1)^2)\Big)$$

converge vers le produit de la matrice ligne $(\frac{\partial g}{\partial x}(\mathbb{E}_\theta X_1^2, \mathbb{E}_\theta X_1), \frac{\partial g}{\partial y}(\mathbb{E}_\theta X_1^2, \mathbb{E}_\theta X_1))$ et du vecteur gaussien colonne ci dessus. C'est donc le produit du vecteur ligne $(1, -2\mathbb{E}_\theta X_1)$ et du vecteur colonne gaussien (N_1, N_2) centré de matrice de variance-covariance B. La loi de ce produit est celle d'une variable aléatoire gaussienne centrée dont la variance est $b_{1,1} + 4(\mathbb{E}_\theta X_1)^2 b_{2,2} - 4(\mathbb{E}_\theta X_1) b_{1,2}$. Il reste à substituer les moments de la loi de Poisson pour obtenir que cette variance vaut $\theta + 2\theta^2$.

Démonstration du Théorème 15. Notons Z une variable aléatoire de loi ℓ sur $\mathbb{R}^k$ et ψ la matrice (de taille $p \times k$) définie pour tout $y \in \mathbb{R}^k$ par

$$\psi(y) = \int_0^1 \mathrm{J}_g\left(z + u(y - z)\right) \mathrm{d}u.$$

La formule de Taylor avec reste intégral nous donne, pour tout $y \in \mathbb{R}^k$:

$$g(y) - g(z) = \psi(y)(y - z).$$

En effet, en notant $g = (g^1, \dots, g^p)$ et ψ^i la i-ème ligne de la matrice ψ,

$$\begin{aligned}\psi^i(y)(y - z) &= \int_0^1 \sum_{j=1}^k \frac{\partial g^i}{\partial x_j}(z_1 + u(y_1 - z_1), \dots, z_k + u(y_k - z_k))(y_j - z_j) du \\ &= \int_0^1 \frac{\partial (u \mapsto g^i(z + u(y - z))}{\partial u} du = g^i(y) - g^i(z), \quad i = 1, \dots, p.\end{aligned}$$

Donc

$$v_n(g(Z_n) - g(z)) = v_n \psi(Z_n)(Z_n - z). \tag{5.8}$$

Dans cette égalité (5.8), $v_n(Z_n - z)$ converge en loi vers ℓ par l'énoncé. Il reste à prouver que $\psi(Z_n)$ converge en loi vers la matrice constante $J_g(z)$.

Tout d'abord, le vecteur $(1/v_n, v_n(Z_n - Z))$ converge en loi vers le couple $(0, \ell)$ par le lemme de Slutsky. Alors, la fonction produit étant continue, $Z_n - z = (1/v_n) \times v_n(Z_n - z)$ converge vers 0 en loi, i.e. (puisque z est une constante), Z_n converge en loi vers z.

Par ailleurs, l'élément (i, j) de la matrice ψ est la fonction $y \to \int_0^1 \frac{\partial g^i}{\partial x_j}(z + u(y - z))du$, qui est continue au point $y = z$ par le théorème de convergence dominée. En effet, g étant de classe $\mathscr{C}^1$, les fonctions $y \to \frac{\partial g^i}{\partial x_j}(z + u(y - z))$ sont continues et bornées dans une boule fermée au voisinage de z.

On en déduit que tous les éléments de la matrice $\psi(Z_n)$ convergent en loi vers ceux de la matrice $\psi(z) = J_g(z)$.

Finalement dans (5.8), $\psi(Z_n)$ converge en loi vers la matrice constante $J_g(z)$ et $v_n(Z_n - z)$ converge en loi vers ℓ. En appliquant encore une fois le lemme de Slutsky, on déduit que la partie droite de (5.8) converge en loi vers $J_g(z)\ell$. □

5.5 Tests asymptotiques

Dans cette section, on considère le problème de test d'hypothèse sur le paramètre θ. Ainsi, dans le cadre du modèle statistique $(\mathscr{H}^n, \{P_\theta\}_{\theta \in \Theta})$, on se donne deux sous-ensembles Θ_0 et Θ_1, disjoints et inclus dans Θ (on n'impose pas que leur union soit égale à Θ). Au vu d'une observation $\mathbb{X} = (X_1, \ldots, X_n) \sim P_\theta$, on veut décider si θ (le vrai paramètre) appartient à Θ_0, ce sera l'hypothèse dite *nulle* et notée H_0 ; ou s'il appartient à Θ_1, ce sera l'hypothèse dite *alternative*, notée H_1.

Dans le cadre du problème de test de H_0 contre H_1, un test est une statistique T à valeurs dans $\{0, 1\}$ associée à la stratégie suivante : pour l'observation $\mathbb{X} = (X_1, \ldots, X_n)$, l'hypothèse H_0 est conservée (respectivement rejetée) si $T(\mathbb{X}) = 0$ (respectivement $T(\mathbb{X}) = 1$).

Un test peut donc toujours s'écrire $T(\mathbb{X}) = \mathbb{1}_{\mathbb{X}\in R}$, où R est la région de rejet. Il est parfois plus naturel de l'écrire sous la forme $T(\mathbb{X}) = \mathbb{1}_{h(\mathbb{X})\in R'}$, où h est une fonction mesurable appelée statistique de test. On dit aussi souvent, en commettant un abus de langage, que R' est la région de rejet associée à la statistique de test $h(\mathbb{X})$.

Le risque de première espèce du test T est la fonction définie sur Θ_0 par

$$\begin{array}{rccl} \underline{\alpha}: & \Theta_0 & \to & [0,1] \\ & \theta & \mapsto & \mathbb{E}_\theta T = P_\theta(T(\mathbb{X}) = 1). \end{array}$$

La **taille** du test est le réel $\alpha^\star$ défini par

$$\alpha^\star = \sup_{\theta\in\Theta_0} \underline{\alpha}(\theta).$$

On dit que le test T est de **niveau** $\alpha \in (0,1)$ si sa taille est inférieure ou égale à α.

Le risque de seconde espèce du test T est l'application définie sur Θ_1 par

$$\begin{array}{rccl} \underline{\beta}: & \Theta_1 & \to & [0,1] \\ & \theta & \mapsto & 1 - \mathbb{E}_\theta T = P_\theta(T(\mathbb{X}) = 0). \end{array}$$

À partir de là, on définit la **puissance** du test comme la fonction $1 - \underline{\beta}$, c'est-à-dire l'application qui à chaque élément de Θ_1 associe la probabilité de prendre la bonne décision.

À défaut d'informations suffisantes ou appropriées sur la loi de la statistique de test, on est amenés à définir la notion de test asymptotique.

Définition 18. *Une suite de tests asymptotiques $(T_n)_{n\geq 1}$ de niveau $\alpha \in (0,1)$ est une suite de tests qui vérifient*

$$\limsup_{n\to+\infty} \sup_{\theta\in\Theta_0} \mathbb{E}_\theta T_n \leq \alpha.$$

La procédure de décision est calquée sur celle des tests à taille d'échantillon finie. La seule différence est qu'un test asymptotique est construit pour contrôler le risque de première espèce, mais seulement asymptotiquement. Dans ce contexte, il est raisonnable d'exiger une puissance asymptotique maximale. C'est le concept de convergence décrit ci-dessous.

Définition 19. *Une suite de tests $(T_n)_{n\geq 1}$ est dite convergente (ou consistante) au niveau $\alpha \in (0,1)$ si c'est une suite de niveau α telle que*

$$\lim_{n\to+\infty} \mathbb{E}_\theta T_n = 1, \quad \forall \theta \in \Theta_1.$$

Ainsi, une suite de tests convergente a une puissance qui tend vers 1 lorsque l'échantillon devient grand.

Exemple. (Test de signe). Soit $\mathbb{X} = (X_1, \dots, X_n)$ i.i.d. de loi commune Q_θ sans atomes et de fonction de répartition $F_\theta = F(\cdot - \theta)$. On suppose que la loi Q_θ admet pour médiane θ, i.e $F(0) = Q_\theta(X_i \leq \theta) = 1/2$. On veut réaliser le test (unilatère) de $H_0 : \theta = 0$ contre l'alternative $H_1 : \theta > 0$.

Une statistique naturelle est donnée par la statistique de signe $S_n = n^{-1}\sum_{i=1}^n \mathbb{1}_{X_i>0}$. Or $\mathbb{E}_\theta S_n = 1 - F(-\theta) := \mu(\theta)$ et $\mathbb{V}_\theta S_n = n^{-1}(1-F(-\theta))F(-\theta) := \sigma^2(\theta)/n$ donc par le théorème central limite, $\sqrt{n}(S_n - \mu(\theta))$ converge en loi sous P_θ vers $\mathcal{N}(0, \sigma^2(\theta))$.

Sous l'hypothèse nulle, on a $\mu(0) = 1/2$ et $\sigma^2(0) = 1/4$, donc $\sqrt{n}(S_n - 1/2)$ converge en loi sous $\mathbb{P}_{\theta=0}$ vers $\mathcal{N}(0, 1/4)$. La suite de tests $(T_n)_{n\geq 1}$ qui rejettent H_0 lorsque $\mathbb{X}$ appartient à la zone de rejet

$$R_n = \{\sqrt{n}(S_n - 1/2) > \frac{1}{2} q_{1-\alpha}\},$$

avec $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la $\mathcal{N}(0,1)$, est une suite de niveau asymptotique α.

Par ailleurs, la puissance π_n de cette suite de tests vérifie, pour tout $\theta > 0$,

$$\begin{aligned}
\pi_n(\theta) = \mathbb{E}_\theta(T_n) =& P_\theta(\sqrt{n}(S_n - 1/2) > \frac{1}{2} q_{1-\alpha}) \\
=& P_\theta\Big(\sqrt{n}\frac{S_n - \mu(\theta)}{\sigma(\theta)} > \frac{\frac{1}{2}q_{1-\alpha} - \sqrt{n}(\frac{1}{2} - \mu(\theta))}{\sigma(\theta)}\Big) \\
=& 1 - \Phi\Big(\frac{\frac{1}{2}q_{1-\alpha} - \sqrt{n}(\frac{1}{2} - \mu(\theta))}{\sigma(\theta)}\Big) + o(1)
\end{aligned}$$

où Φ est la fonction de répartition de la $\mathcal{N}(0,1)$. Le terme de droite converge vers 1 lorsque n tend vers l'infini car pour $\theta > 0$, on a $\mu(\theta) = Q_\theta(X_1 > 0) < 1/2$. La suite de tests est donc convergente.

Vitesse de convergence d'une suite de tests. En pratique, la notion de convergence est trop faible pour qualifier l'optimalité d'un test, car c'est une propriété qui s'avère assez répandue. Pour choisir entre deux suites de tests, une approche plus intéressante consiste à considérer la puissance calculée en des alternatives dont la difficulté augmente avec la taille de l'échantillon.

Exemple. (Test de signe). Dans l'exemple précédent, pour toute suite $(\theta_n)_{n\geq 1}$ de paramètres dans l'alternative (i.e. $\theta_n > 0$), la puissance du test vaut

$$\pi_n(\theta_n) = 1 - \Phi\Big(\frac{\frac{1}{2}q_{1-\alpha} - \sqrt{n}(F(0) - F(-\theta_n))}{\sigma(\theta_n)}\Big) + o(1).$$

Si θ_n tend vers 0 assez vite pour que $\sqrt{n}(F(0) - F(-\theta_n))$ tende vers 0, alors $\pi(\theta_n)$ converge vers α et le test de signe n'est pas capable de distinguer l'alternative de l'hypothèse nulle. Si θ_n tend vers 0 assez lentement pour qu'au contraire $\sqrt{n}(F(0) - F(-\theta_n))$ tende vers $+\infty$, la puissance tend vers 1 et l'alternative $\theta = \theta_n$ est facile à distinguer de l'hypothèse nulle. On comprend que le cas intéressant surgit dans un cadre intermédiaire, quand θ_n tend vers 0 mais que $\sqrt{n}(F(0) - F(-\theta_n))$ se « stabilise ». En particulier, si la fonction de répartition est différentiable au voisinage de 0, avec une dérivée (positive) $f(0)$, on a

$$\sqrt{n}(F(0) - F(-\theta_n)) = \sqrt{n}\theta_n f(0) + \sqrt{n}o(\theta_n).$$

La suite d'alternatives intéressante à considérer est donc $\theta_n = h/\sqrt{n}$ pour un $h > 0$. On obtient pour ces alternatives :

$$\pi(h/\sqrt{n}) = 1 - \Phi\Big(q_{1-\alpha} - 2hf(0)\Big) + o(1).$$

L'allure de cette fonction est donnée dans la figure 5.5.

Dans la suite, on fixe l'hypothèse nulle H_0 et on s'intéresse à la puissance d'une suite de tests pour des alternatives qui convergent vers l'hypothèse nulle. Pour simplifier les notations, on s'intéresse au cas d'un paramètre réel $\theta \in \mathbb{R}$ (i.e. $k = 1$) et (sans perte de généralité) $H_0 : \theta = 0$. On suppose que le test T_n rejette l'hypothèse nulle pour les grandes valeurs de la statistique $h_n(\mathbb{X})$ et que cette statistique est asymptotiquement normale pour les alternatives $\theta_n = h/\sqrt{n}$

$$\sqrt{n}\frac{(h_n(\mathbb{X}) - \mu(\theta_n))}{\sigma(\theta_n)} \xrightarrow{\mathscr{L}(\mathbb{P}_{\theta_n})} \mathcal{N}(0,1). \tag{5.9}$$

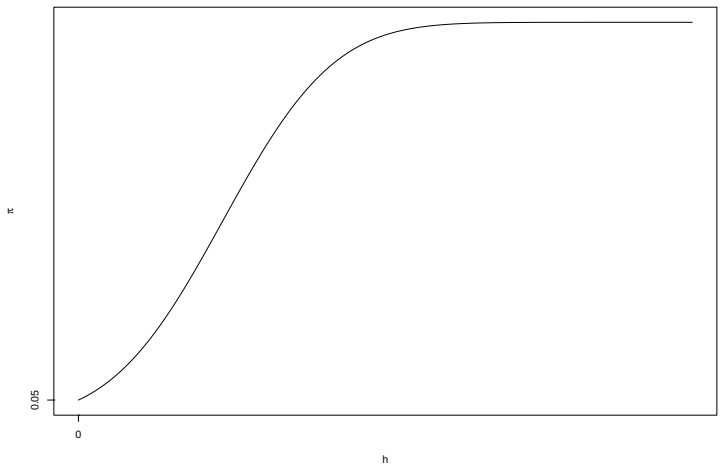

FIGURE 5.1 – Allure de la fonction puissance pour les alternatives de la forme $\theta_n = h/\sqrt{n}$ en fonction de h avec $\alpha = 0.05$.

Attention : la convergence doit avoir lieu sous la loi $\mathbb{P}_{\theta_n}$ qui est indexée par n. Cette convergence ne découle donc pas simplement de la convergence

$$\sqrt{n}\frac{(h_n(\mathbb{X}) - \mu(\theta))}{\sigma(\theta)} \xrightarrow{\mathscr{L}(P_\theta)} \mathcal{N}(0,1), \quad \forall \theta \in \Theta.$$

D'un autre côté, la convergence ci-dessus a lieu pour tous les $\theta \in \Theta$ tandis que (5.9) porte seulement sur un voisinage de $\theta = 0$. On parle de normalité asymptotique *localement uniforme*.

Théorème 16. *Soient $\mu : \Theta \mapsto \mathbb{R}$ et $\sigma : \Theta \mapsto (0; +\infty)$ des fonctions telles que (5.9) soit vérifiée pour toute suite $\theta_n = h/\sqrt{n}$. Supposons que μ (resp. σ) est dérivable (resp. continue) en $\theta = 0$. Alors la suite de tests T_n qui rejette $H_0 : \theta = 0$ contre $H_1 : \theta > 0$ lorsque $h_n(\mathbb{X}) \in R_n$ avec*

$$R_n = \{\sqrt{n}(h_n(\mathbb{X}) - \mu(0)) > \sigma(0)q_{1-\alpha}\},$$

(où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$) est asymptotiquement de niveau α. De plus, la puissance de cette suite vérifie

$$\pi_n\Big(\frac{h}{\sqrt{n}}\Big) \to 1 - \Phi\Big(q_{1-\alpha} - h\frac{\mu'(0)}{\sigma(0)}\Big),$$

pour tout h.

Démonstration. L'hypothèse (5.9) implique en particulier que $\sqrt{n}\frac{(h_n(\mathbb{X})-\mu(0))}{\sigma(0)}$ converge en loi sous l'hypothèse nulle ($\theta = 0$) vers $\mathcal{N}(0,1)$. Donc la suite de tests est asymptotiquement de niveau α.

Considérons à présent des alternatives qui convergent vers H_0 de la forme $\theta_n = h/\sqrt{n}$. La puissance vérifie

$$\begin{aligned}\pi_n(\theta) &= \mathbb{E}_{\theta_n} T_n = \mathbb{P}_{\theta_n}(\sqrt{n}(h_n(\mathbb{X}) - \mu(0)) > \sigma(0)q_{1-\alpha}) \\ &= \mathbb{P}_{\theta_n}\Big(\sqrt{n}\frac{(h_n(\mathbb{X}) - \mu(\theta_n))}{\sigma(\theta_n)} > \frac{\sigma(0)q_{1-\alpha} + \sqrt{n}(\mu(0) - \mu(\theta_n))}{\sigma(\theta_n)}\Big) \\ &= 1 - \Phi\Big(\frac{\sigma(0)q_{1-\alpha} + \sqrt{n}(\mu(0) - \mu(\theta_n))}{\sigma(\theta_n)}\Big) + o(1),\end{aligned}$$

d'après (5.9). Puisque σ est continue en 0, que μ est dérivable en ce même point et $\theta_n = h/\sqrt{n}$, on obtient

$$\sqrt{n}\frac{(\mu(0) - \mu(\theta_n))}{\sigma(\theta_n)} \to -h\frac{\mu'(0)}{\sigma(0)},$$

et la continuité de Φ achève la preuve. □

Sous l'hypothèse de normalité asymptotique localement uniforme, la puissance de la suite de tests dépend uniquement de la *pente* $\frac{\mu'(0)}{\sigma(0)}$. Deux suites de tests (de même niveau asymptotique α) peuvent donc être comparées via leur pente : pour le test de $H_0 : \theta = 0$ contre $\theta > 0$, celui qui a la plus grande pente aura la plus grande puissance et sera donc le meilleur.

Définition 20 (Efficacité relative asymptotique)**.** *Soient deux suites de test $(T_n^1)_{n\geq 1}$ et $(T_n^2)_{n\geq 1}$ de même niveau asymptotique α pour le test de $H_0 : \theta = 0$ et satisfaisant (5.9) (pour (μ_1, σ_1) et (μ_2, σ_2), respectivement. Alors, le rapport*

$$r = \Big(\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\Big)^2$$

est l'efficacité relative asymptotique de T_1 par rapport à T_2.

Exemple. [Test de signe versus t-test]. Soit $\mathbb{X} = (X_1, \ldots, X_n)$ i.i.d. de densité $f(\cdot - \theta)$ où f est une fonction paire et le moment d'ordre 2 noté $\sigma^2 = \int x^2 f(x)dx < +\infty$. Remarquez que la fonction de répartition correspondante $F(\cdot - \theta)$ vérifie $F(0) = 1/2$; on est donc dans un cas particulier

du contexte de l'exemple précédent sur le test de signe. On veut toujours réaliser le test de $H_0 : \theta = 0$, qui correspond à l'hypothèse que les observations ont une distribution symétrique par rapport à 0.

Le test de Student (t-test)[4] de niveau asymptotique α utilise la statistique de test $h_n(\mathbb{X}) = \sqrt{n}\bar{X}_n/S_n$ et rejette l'hypothèse nulle pour les grandes valeurs de $h_n(\mathbb{X})$ (en valeur absolue si on fait un test bilatère ; et pour les valeurs trop petites ou trop grandes en cas de test unilatère). (Attention sans hypothèse gaussienne sur les X_i, la statistique de Student ne suit pas une loi de Student !) Commençons par vérifier l'hypothèse (5.9) pour cette suite. On considère des alternatives $\theta_n = h/\sqrt{n}$ et on remarque

$$\sqrt{n}\Big(\frac{\bar{X}_n}{S_n} - \frac{h}{\sqrt{n}\sigma}\Big) = \frac{\sqrt{n}(\bar{X}_n - h/\sqrt{n})}{S_n} + h\Big(\frac{1}{S_n} - \frac{1}{\sigma}\Big).$$

Remarquons que sous $\mathbb{P}_{\theta_n}$, la loi de $\bar{X}_n - h/\sqrt{n}$ ne dépend plus de θ_n ; c'est la même que celle de $\bar{X}_n$ sous $\theta = 0$. De la même façon, la loi sous $\mathbb{P}_{\theta_n}$ de la variance empirique $S_n^2 = n^{-1}(\sum_{i=1}^n X_i^2 - \bar{X}_n^2)$ ne dépend pas de θ_n et a la même loi que sous $\theta = 0$. On en déduit que comme S_n converge en probabilité vers σ (sous toutes les lois P_θ, y compris $\mathbb{P}_{\theta_n}$), en combinant le lemme de Slutsky avec la loi des grands nombres pour la variable $\sqrt{n}(\bar{X}_n - h/\sqrt{n})/\sigma$ qui converge sous $\mathbb{P}_{\theta_n}$ vers une $\mathcal{N}(0,1)$, on obtient que le premier terme converge en loi sous $\mathbb{P}_{\theta_n}$ vers une $\mathcal{N}(0,1)$ et le second terme converge vers 0. Donc (5.9) est vérifiée avec $\mu(\theta) = \theta/\sigma$ et $\sigma(\theta) = 1$. La pente de cette suite de tests vaut $1/\sigma = (\int x^2 f(x)dx)^{-1/2}$.

La pente de la suite de tests de signe vaut (reprendre les calculs précédents) $2f(0)$. On en déduit que l'efficacité relative asymptotique de la suite de tests de signe par rapport à la suite de t-tests vaut

$$4f^2(0)(\int x^2 f(x)dx).$$

Ce rapport dépend de la densité f. Par exemple pour f une densité uniforme (sur un intervalle symétrique autour de 0) on trouve $1/3$ qui est inférieur à 1 (donc le t-test est meilleur en ce sens) alors que pour la densité de Laplace, on trouve 2 (et le test de signe est meilleur).

4. Voir le Chapitre 6.

Chapitre 6

Échantillons gaussiens et modèle linéaire

6.1 Notations

- χ^2_n : loi du *chi-deux* à $n \in \mathbb{N}^\star$ degrés de liberté. Densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$:

$$\frac{1}{2^{n/2}\Gamma(n/2)} x^{(n-2)/2} e^{-x/2} \mathbb{1}_{\mathbb{R}_+}(x), \quad x \in \mathbb{R}.$$

  Note : χ^2_n est la loi de $U_1^2 + \cdots + U_n^2$, où $U_1, \ldots, U_n$ sont des variables aléatoires indépendantes de même loi $\mathcal{N}(0,1)$.

- $\mathcal{T}_n$: loi de *Student* à $n \in \mathbb{N}^\star$ degrés de liberté. Densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$:

$$\frac{\Gamma\big((n+1)/2\big)}{\sqrt{n\pi}\Gamma(n/2)} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in \mathbb{R}.$$

  Note : $\mathcal{T}_n$ est la loi de $\sqrt{n}U/\sqrt{V}$, où U et V sont des variables aléatoires indépendantes de lois respectives $\mathcal{N}(0,1)$ et χ^2_n.

- $\mathcal{F}(n_1, n_2)$: loi de *Fisher* de paramètres $(n_1, n_2) \in (\mathbb{N}^\star)^2$. Densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$:

$$\frac{1}{B(n_1/2, n_2/2)} \frac{(n_1 x)^{n_1/2} n_2^{n_2/2}}{x(n_1 x + n_2)^{(n_1+n_2)/2}} \mathbb{1}_{\mathbb{R}_+}(x), \quad x \in \mathbb{R},$$

avec B la fonction bêta définie pour tout $(t_1, t_2) \in (\mathbb{R}_+^\star)^2$ par la relation $B(t_1, t_2) = \int_0^1 u^{t_1-1}(1-u)^{t_2-1}\mathrm{d}u$.

Note : $\mathcal{F}(n_1, n_2)$ est la loi de $(U/n_1)/(V/n_2)$, où U et V sont des variables aléatoires indépendantes de lois respectives $\chi^2_{n_1}$ et $\chi^2_{n_2}$.

6.2 Rappels sur les vecteurs gaussiens

Cas réel. Une variable aléatoire réelle X est dite gaussienne (ou de loi normale) de paramètres $m \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}_+$ ($\sigma \geq 0$) si sa fonction caractéristique s'écrit

$$\mathbb{E}\exp(iuX) = \exp\Big(ium - \frac{\sigma^2 u^2}{2}\Big), \quad \forall u \in \mathbb{R}.$$

La loi de X est notée $\mathcal{N}(m, \sigma^2)$, et l'on a $\mathbb{E}X = m$ et $\mathbb{V}X = \sigma^2$. Lorsque $\sigma = 0$, on dit que X est dégénérée ; dans le cas contraire, elle admet la densité par rapport à la mesure de Lebesgue

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R}.$$

Cas vectoriel. Plus généralement, une variable aléatoire X à valeurs dans $\mathbb{R}^d$ est un vecteur gaussien de $\mathbb{R}^d$ s'il existe $M \in \mathbb{R}^d$ et Σ une matrice $d \times d$ réelle, symétrique et positive, tels que la fonction caractéristique de X s'écrive

$$\mathbb{E}\exp(iu^\intercal X) = \exp\Big(iu^\intercal M - \frac{1}{2}u^\intercal \Sigma u\Big), \quad \forall u \in \mathbb{R}^d.$$

(Les vecteurs sont considérés comme des matrices colonnes.) La loi de X est notée $\mathcal{N}_d(M, \Sigma)$. Alors M est la moyenne de X, i.e. $\mathbb{E}X = M$, et Σ est la matrice de variance-covariance de X, i.e.

$$\Sigma = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\intercal.$$

Lorsque la matrice Σ est inversible (i.e., définie positive), X admet la densité par rapport à la mesure de Lebesgue

$$\frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}}\exp\Big(-\frac{1}{2}(x-M)^\intercal \Sigma^{-1}(x-M)\Big), \quad \forall x \in \mathbb{R}^d.$$

Lorsque Σ n'est pas inversible, on montre facilement que la loi de X est $\mathbb{P}$-p.s. concentrée sur le sous-espace affine de $\mathbb{R}^d$ d'origine M et engendré par les vecteurs propres correspondant aux valeurs propres non nulles de Σ.

Un moyen simple de montrer qu'un vecteur aléatoire est gaussien est d'utiliser la définition équivalente suivante :

Proposition 7. *Un vecteur aléatoire est gaussien si et seulement si toute combinaison linéaire de ses composantes est une variable aléatoire réelle gaussienne.*

Démonstration. Soit $X \in \mathbb{R}^d$ gaussien, $\lambda \in \mathbb{R}^d$ et $Y = \lambda^\intercal X$ une combinaison linéaire de ses composantes. Alors $\mathbb{E}\exp(itY) = \mathbb{E}\exp(it\lambda^\intercal X) = \mathbb{E}\exp(i(t\lambda)^\intercal X) = \exp(i(t\lambda)^\intercal M - (1/2)(t\lambda)^\intercal\Sigma(t\lambda)) = \exp(it\lambda^\intercal M - (t^2/2)\lambda^\intercal\Sigma\lambda)$ ce qui signifie que Y est une variable aléatoire gaussienne avec $\mathbb{E}Y = \lambda^\intercal M$ et $\mathbb{V}Y = \lambda^\intercal\Sigma\lambda$.

Réciproquement, soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$ avec le vecteur $\mathbb{E}X = M$ et la matrice $\mathbb{V}X = \mathbb{E}(X-M)(X-M)^\intercal = \Sigma$. Alors pour tout $u \in \mathbb{R}^d$, la variable aléatoire (réelle) $u^\intercal X$ a pour moyenne $\mathbb{E}(u^\intercal X) = u^\intercal M$ et sa variance vaut $\mathbb{E}(u^\intercal X - u^\intercal M)^2$. Comme $u^\intercal X - u^\intercal M$ est un scalaire alors $u^\intercal X - u^\intercal M = (u^\intercal X - u^\intercal M)^\intercal = X^\intercal u - M^\intercal u$. Donc sa variance $\mathbb{E}(u^\intercal X - u^\intercal M)^2 = \mathbb{E}(u^\intercal X - u^\intercal M)(X^\intercal u - M^\intercal u) = u^\intercal\mathbb{E}(X-M)(X-M)^\intercal u = u^\intercal\Sigma u$.

Supposons que cette variable aléatoire est gaussienne pour tout u. Alors sa fonction caractéristique prend la forme $\mathbb{E}\exp(itu^\intercal X) = \exp(itu^\intercal M - (t^2/2)u^\intercal\Sigma u)$. Pour $t = 1$, cela donne $\mathbb{E}\exp(iu^\intercal X) = \exp(iu^\intercal M - (1/2)u^\intercal\Sigma u)$. Comme c'est vrai pour tout $u \in \mathbb{R}^d$, le vecteur X est gaussien. □

Mentionnons pour finir deux résultats d'utilité constante dans la manipulation des vecteurs gaussiens.

Proposition 8.

(**i**) **Transformation affine.** *Si A est une matrice réelle de taille $k \times d$, et $b \in \mathbb{R}^k$ et X est un vecteur de loi $\mathcal{N}_d(M,\Sigma)$, alors $AX + b$ suit la loi $\mathcal{N}_k(AM + b, A\Sigma A^\intercal)$.*

(**ii**) **Caractérisation de l'indépendance.** *Soit X un vecteur gaussien. Les composantes de X sont des variables aléatoires réelles indépendantes si et seulement si la matrice de variance-covariance de X est diagonale.*

Ainsi, lorsque $X = (X_1, \dots, X_d)$ est un vecteur gaussien, on a l'équivalence :

$$\forall i \neq j,\ X_i \text{ et } X_j \text{ indépendantes} \Leftrightarrow \mathrm{Cov}(X_i, X_j) = 0.$$

Rappelons que seule l'implication $\Rightarrow$ est vraie dans le cas général.

Démonstration. (**i**) On a

$$\begin{aligned}
\mathbb{E}\exp(iu^\intercal(AX+b)) &= \exp(iu^\intercal b)\mathbb{E}\exp(i(A^\intercal u)^\intercal X) \\
&= \exp(iu^\intercal b + (A^\intercal u)^\intercal M - (1/2)(A^\intercal u)^\intercal \Sigma (A^\intercal u)) \\
&= \exp(iu^\intercal b + u^\intercal AM - (1/2)(u^\intercal A)\Sigma(A^\intercal u)) \\
&= \exp(iu^\intercal(b+AM) - (1/2)u^\intercal A\Sigma A^\intercal u).
\end{aligned}$$

(**ii**) L'implication est immédiate. Réciproque : si $\mathrm{Cov}(X_i, X_j) = 0$, la matrice Σ est diagonale, la fonction caractéristique de X se présente comme le produit de fonctions caractéristiques de composantes, ce qui est la fonction caractéristique du vecteur gaussien avec des composantes indépendantes. □

Exemple. Soit $Z \sim \mathcal{N}(0,1)$ et $\varepsilon \sim \mathscr{B}(1/2)$ deux variables aléatoires indépendantes. Alors $X_1 = Z$ et $X_2 = (2\varepsilon - 1)Z$ sont des variables aléatoires réelles gaussiennes (pourquoi ?), mais $X = (X_1, X_2)$ n'est pas un vecteur gaussien, puisque $X_1 + X_2 = 2\varepsilon Z$ prend avec probabilité $1/2$ la valeur 0. On notera que $\mathrm{Cov}(X_1, X_2) = 0$ (faites le calcul) mais que X_1 et X_2 ne sont pas indépendantes (si elles l'étaient, comme leurs lois sont gaussiennes, le vecteur X serait gaussien ...).

6.3 Théorème de Cochran

Dans le monde des vecteurs gaussiens, orthogonalité et indépendance se confondent. Ce lien entre la géométrie et les probabilités a pour conséquence le théorème de Cochran, qui constitue la pierre angulaire de toute la statistique des échantillons gaussiens.

Théorème 17 (Cochran). *Soit $\sigma > 0, X \sim \mathcal{N}_n(0, \sigma^2 \mathrm{Id}_n)$ et $V_1, \dots, V_p$ des sous-espaces vectoriels orthogonaux de dimensions respectives $r_1, \dots, r_p$ tels que*

$$V_1 \oplus \cdots \oplus V_p = \mathbb{R}^n.$$

Alors les projections orthogonales $\pi_1, \dots, \pi_p$ de X sur $V_1, \dots, V_p$ sont des vecteurs gaussiens indépendants et, pour chaque $i = 1, \dots, p$,

$$\frac{1}{\sigma^2}\|\pi_i\|^2 \sim \chi^2(r_i).$$

Démonstration. Soit $(e_j^i)_{i,j}$ une base orthonormée de $\mathbb{R}^n$ telle que pour chaque $i = 1, \dots, p$, $(e_j^i)_{1 \le j \le r_i}$ est une base orthonormée de V_i. Pour tout $i = 1, \dots, p$, on a $\pi_i = M_i X$, où M_i est la matrice symétrique de taille $n \times n$ définie par

$$M_i = (e_1^i \cdots e_{r_i}^i)\,(e_1^i \cdots e_{r_i}^i)^\intercal.$$

Noter que puisque les vecteurs $(e_j^i)_{i,j}$ sont normés et orthogonaux, M_i est idempotente ($M_i^2 = M_i$) et $M_i M_j = 0$ pour tout $i \neq j$.

Montrons la première assertion du théorème. Puisque X est gaussien, toute combinaison linéaire de ses composantes est gaussienne. En conséquence, toute combinaison linéaire de composantes du vecteur $(\pi_1, \dots, \pi_p)$ (est encore une combinaison linéaire des composantes de X et donc) est aussi une variable gaussienne. De plus, la covariance entre les vecteurs aléatoires π_i et π_j est nulle pour tout $i \neq j$. En effet, ces vecteurs aléatoires étant centrés,

$$\begin{aligned}\mathrm{Cov}(\pi_i, \pi_j) &= \mathbb{E}\,(\pi_i - \mathbb{E}\pi_i)\,\big(\pi_j - \mathbb{E}\pi_j\big)^\intercal = \mathbb{E}\pi_i \pi_j^\intercal \\ &= \mathbb{E} M_i X (M_j X)^\intercal = M_i \mathbb{E} X X^\intercal M_j \\ &= \sigma^2 M_i M_j,\end{aligned}$$

d'où $\mathrm{Cov}(\pi_i, \pi_j) = 0$. Par suite, $\pi_1, \dots, \pi_p$ sont des vecteurs gaussiens indépendants.

Pour montrer la seconde assertion, fixons $i = 1, \dots, p$ et remarquons que comme M_i est symétrique et idempotente :

$$\pi_i = M_i X \sim \mathcal{N}_n\big(0, \sigma^2 M_i \,\mathrm{Id}_n\, M_i\big) = \mathcal{N}_n(0, \sigma^2 M_i).$$

En notant E_i la matrice de taille $n \times r_i$ définie par $E_i = (e_1^i \cdots e_{r_i}^i)$, on a $M_i = E_i E_i^\intercal$ et donc

$$\pi_i \sim \sigma E_i \,\mathcal{N}_{r_i}(0, \mathrm{Id}_{r_i}).$$

Or, si Z est un vecteur aléatoire de loi $\mathcal{N}_{r_i}(0, \mathrm{Id}_{r_i})$, $\|E_i Z\|^2 = \|Z\|^2 \sim \chi^2_{r_i}$ car $E_i^\intercal E_i = \mathrm{Id}_{r_i}$, d'où le théorème.

□

6.4 Échantillons gaussiens

Rappelons que pour une suite $X_1, \dots, X_n$ de variables aléatoires réelles, on note

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{et} \quad \widetilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Rappelons que $S_n^2 = (1/n)\sum X_i^2 - 2\bar{X}_n(1/n)\sum X_i + \bar{X}_n^2 = 1/n\sum X_i^2 - \bar{X}_n^2$. Alors $\mathbb{E}S_n^2 = EX_1^2 - E\bar{X}_n^2 = \sigma^2 + m^2 - \mathbb{V}\bar{X}_n - (E\bar{X}_n)^2 = \sigma^2 + m^2 - n\sigma^2/n^2 - m^2 = \frac{n-1}{n}\sigma^2$. Ainsi S_n^2 est un estimateur de la variance σ^2 qui a un biais. On lui préfère donc $\widetilde{S}_n^2$ qui est sans biais.

Le théorème ci-dessous met en évidence le rôle tenu par la loi de Student et la loi du χ^2 lorsque $X_1, \dots, X_n$ sont indépendantes et de même loi gaussienne.

Théorème 18 (Fisher). *Soit $m \in \mathbb{R}, \sigma > 0$ et $X_1, \dots, X_n$ des variables aléatoires indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$. Alors :*

(**i**) *$\bar{X}_n$ et S_n^2 sont indépendantes.*
(**ii**) *$nS_n^2/\sigma^2 \sim \chi^2(n-1)$ (ou encore $(n-1)\widetilde{S}_n^2/\sigma^2 \sim \chi^2(n-1)$).*
(**iii**) *$\sqrt{n}(\bar{X}_n - m)/\widetilde{S}_n \sim \mathcal{T}(n-1)$.*

Remarque. Dans ce théorème, (**iii**) est à rapprocher de la propriété classique $\sqrt{n}(\bar{X}_n - m)/\sigma \sim \mathcal{N}(0,1)$ satisfaite par la suite de variables aléatoires indépendantes $X_1, \dots, X_n$ de même loi $\mathcal{N}(m, \sigma^2)$.

Démonstration. Soit V le sous-espace vectoriel de $\mathbb{R}^n$ engendré par $e = (1, \dots, 1)^\intercal$ et soit $\frac{X-m}{\sigma} = (\frac{X_1-m}{\sigma}, \dots, \frac{X_n-m}{\sigma})^\intercal \sim \mathcal{N}_n(0, \mathrm{Id}_n)$.

Le projecteur orthogonal P sur V est la matrice $n \times n$ dont tous les coefficients valent $1/n$. En effet, la matrice C est la matrice colonne qui se compose de coordonnées du vecteur e normalisé : $(1/\sqrt{n}, \dots, 1/\sqrt{n})$. Alors $P = CC^\intercal$ est la matrice de taille $n \times n$ avec tous les éléments $1/n$.

De ce fait,

$$P(\frac{X-m}{\sigma}) = (\frac{\bar{X}_n - m}{\sigma})e \quad \text{et} \quad (\mathrm{Id}_n - P)\frac{X-m}{\sigma} = \frac{1}{\sigma}\begin{pmatrix} X_1 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{pmatrix}.$$

Comme $(\mathrm{Id}_n - P)\frac{X-m}{\sigma}$ est la projection orthogonale de $\frac{X-m}{\sigma}$ sur l'orthogonal de V et $\frac{X-m}{\sigma}$ suit la loi $\mathcal{N}_n(0, \mathrm{Id}_n)$, on déduit du théorème de Cochran (Théorème 17) que $P\frac{X-m}{\sigma}$ est indépendant de $(\mathrm{Id}_n - P)\frac{X-m}{\sigma}$, et donc en particulier que $\bar{X}_n$ est indépendant de $S_n^2 = (\sigma^2/n)\|(\mathrm{Id}_n - P)X\|^2$, d'où (**i**). De plus, comme V est de dimension 1,

$$\frac{nS_n^2}{\sigma^2} = \|(\mathrm{Id}_n - P)X\|^2 \sim \chi^2(n-1)$$

d'après le théorème de Cochran, d'où (**ii**). Par la définition de $\widetilde{S}_n^2$

$$\frac{(n-1)\widetilde{S}_n^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Par ailleurs $P(\frac{X-m}{\sigma}) = \frac{\bar{X}_n - m}{\sigma}e$ est un vecteur gaussien $\mathcal{N}(0, PId_nP^\intercal) = \mathcal{N}(0, P)$ (car P étant la matrice d'un projecteur, $PP^\intercal = PP = P$), en particulier la variable aléatoire $\frac{\bar{X}_n - m}{\sigma}$ est gaussienne $\mathcal{N}(0, 1/n)$. Donc $\frac{\sqrt{n}(\bar{X}_n - m)}{\sigma}$ est gaussienne centrée réduite.

Enfin, (**iii**) se déduit des résultats précédents, car $\sqrt{n}(\frac{\bar{X}_n - m}{\sigma})$ et $(n-1)\widetilde{S}_n^2/\sigma^2$ sont indépendantes, et de lois respectives $\mathcal{N}(0,1)$ et $\chi^2(n-1)$. On a alors

$$\frac{\sqrt{n}(\bar{X}_n - m)}{\sqrt{\widetilde{S}_n^2}} = \frac{\sqrt{n-1}\sqrt{n}\frac{\bar{X}_n - m}{\sigma}}{\sqrt{\frac{(n-1)\widetilde{S}_n^2}{\sigma^2}}} = \frac{\sqrt{n-1}U}{\sqrt{V}}$$

de loi de Student à $n-1$ degrés de libertés.

□

Le Théorème de Fisher (Théorème 18) a des conséquences importantes pour le traitement des échantillons gaussiens i.i.d.. Nous détaillons dans les paragraphes qui suivent quatre exemples, mais bien d'autres extensions sont possibles. À partir de maintenant, on considère un n-échantillon $X = (X_1, \ldots, X_n)$ i.i.d., de loi commune $\mathcal{N}(m, \sigma^2)$, avec $m \in \mathbb{R}$ et $\sigma > 0$.

Intervalle de confiance pour m. Lorsque σ est connu, on utilise l'estimateur $\bar{X}_n$ et le fait que

$$\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \sim \mathcal{N}(0,1)$$

quelque soit la valeur de m. C'est une statistique pivotale, pour obtenir l'intervalle de confiance

$$\left[\bar{X}_n - q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

de niveau $(1-\alpha)$ pour le paramètre m (comme d'habitude, $q_{1-\frac{\alpha}{2}}$ désigne le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi $\mathcal{N}(0,1)$).

Lorsque σ est inconnu, cet intervalle n'est pas utilisable. On s'en sort grâce au Théorème 18, qui nous permet d'affirmer que

$$\sqrt{n}\frac{\bar{X}_n - m}{\widetilde{S}_n} \sim \mathcal{T}(n-1).$$

C'est une autre statistique pivotale. On en conclut que

$$\left[\bar{X}_n - t_{1-\frac{\alpha}{2}}^{(n-1)}\frac{\widetilde{S}_n}{\sqrt{n}}, \bar{X}_n + t_{1-\frac{\alpha}{2}}^{(n-1)}\frac{\widetilde{S}_n}{\sqrt{n}}\right]$$

est un intervalle de confiance pour m de niveau $(1-\alpha)$, où $t_{1-\frac{\alpha}{2}}^{(n-1)}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{T}(n-1)$. On notera le remplacement de $q_{1-\frac{\alpha}{2}}$ par $t_{1-\frac{\alpha}{2}}^{(n-1)}$ ainsi que celui de σ par $\widetilde{S}_n$: ce mécanisme est parfois appelé « studentisation » .

Intervalle de confiance pour σ^2. D'après **(ii)** du Théorème 18,

$$\frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1).$$

C'est la statistique pivotale pour σ^2. On en déduit que

$$\left[\frac{nS_n^2}{x_{1-\frac{\alpha}{2}}^{(n-1)}}, \frac{nS_n^2}{x_{\frac{\alpha}{2}}^{(n-1)}}\right]$$

est un intervalle de confiance pour σ^2 de niveau $(1-\alpha)$, où $x_\alpha^{(n-1)}$ est le quantile d'ordre α de la loi $\chi^2(n-1)$ (noter le sens des dénominateurs dans l'intervalle).

Test de Student. Construisons un test de niveau $\alpha \in (0,1)$ dans le problème de test de

$$H_0 : m \geq m_1 \quad \text{contre} \quad H_1 : m < m_1,$$

avec σ^2 inconnu et m_1 un réel fixé. Un protocole naturel de rejet pour ce problème est de la forme $\bar{X}_n < k_\alpha$, avec k_α un seuil à préciser, car H_0 est rejetée lorsque la moyenne des observations prend une valeur anormalement faible.

Si T_n est une variable aléatoire de loi $\mathcal{T}(n-1)$, on a, $\forall m \geq m_1$,

$$\begin{aligned}\mathbb{P}(\bar{X}_n < k_\alpha) &= \mathbb{P}\Big(\sqrt{n}\frac{\bar{X}_n - m}{\widetilde{S}_n} < \sqrt{n}\frac{k_\alpha - m}{\widetilde{S}_n}\Big) \\ &= \mathbb{P}\Big(T_n < \sqrt{n}\frac{k_\alpha - m}{\widetilde{S}_n}\Big),\end{aligned}$$

la dernière égalité découlant du Théorème 18. Du coup,

$$\sup_{m \geq m_1} \mathbb{P}(\bar{X}_n < k_\alpha) = \mathbb{P}\Big(T_n \leq \sqrt{n}\frac{k_\alpha - m_1}{\widetilde{S}_n}\Big).$$

Il suffit donc de choisir k_α tel que

$$k_\alpha = m_1 + t_\alpha^{(n-1)} \frac{\widetilde{S}_n}{\sqrt{n}},$$

où $t_\alpha^{(n-1)}$ est le quantile d'ordre α de la loi $\mathcal{T}(n-1)$. Ainsi, le test de région de rejet

$$R_{\text{Student}} = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \bar{x}_n < m_1 + t_\alpha^{(n-1)} \frac{\widetilde{S}_n}{\sqrt{n}} \right\},$$

appelé test de Student, est de niveau (de taille) α. La procédure de décision consiste donc à rejeter H_0 au niveau α lorsque $(X_1, \ldots, X_n) \in R_{\text{Student}}$. (On remarque que $t_\alpha^{(n-1)}$ remplace q_α et que $\widetilde{S}_n$ remplace σ.

On montre de même que le test bilatéral

$$H_0 : m = m_1 \quad \text{contre} \quad H_1 : m \neq m_1$$

est associé à la région de rejet

$$R_{\text{Student}} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : |\bar{x}_n - m_1| > t_{1-\frac{\alpha}{2}}^{(n-1)} \frac{\widetilde{S}_n}{\sqrt{n}} \right\}.$$

En effet, dans ce cas le risque de l'erreur de première espèce est égal à la taille du test

$$\mathbb{P}_{m_1}(\mathbb{X} \in R_{Student}) = \mathbb{P}\Big(\sqrt{n}|X_n - m|/\widetilde{S}_n > t_{1-\alpha/2}^{(n-1)}\Big)$$
$$= \mathbb{P}(|\mathcal{T}_{n-1}| > t_{1-\alpha/2}^{(n-1)}) = 1 - (1-\alpha/2) + 1 - (1-\alpha/2) = \alpha.$$

Test de Fisher. Construisons un test de niveau $\alpha \in (0,1)$ dans le problème de test de

$$H_0 : \sigma \geq \sigma_1 \quad \text{contre} \quad H_1 : \sigma < \sigma_1,$$

avec $\sigma_1 > 0$ fixé. Une région de rejet naturelle pour ce problème de test est de la forme $\hat{\sigma}_n^2 < k_\alpha$ avec k_α un seuil à préciser, car H_0 est rejetée lorsque la variance empirique prend une valeur anormalement faible. Sous H_0 (i.e. $\sigma \geq \sigma_1$), d'après le Théorème 18,

$$\mathbb{P}(\widetilde{S}_n^2 < k_\alpha) = \mathbb{P}\Big(\frac{(n-1)\widetilde{S}_n^2}{\sigma^2} < \frac{(n-1)k_\alpha}{\sigma^2}\Big)$$
$$= \mathbb{P}\Big(Z < \frac{(n-1)k_\alpha}{\sigma^2}\Big),$$

où $Z \sim \chi^2(n-1)$. Dès lors,

$$\sup_{\sigma \geq \sigma_1} \mathbb{P}(\widetilde{S}_n^2 < k_\alpha) = \mathbb{P}\Big(\chi^2(n-1) \leq \frac{(n-1)k_\alpha}{\sigma_1^2}\Big).$$

On choisit k_α tel que

$$k_\alpha = \frac{x_\alpha^{(n-1)}}{n-1}\sigma_1^2,$$

où $x_\alpha^{(n-1)}$ est le quantile d'ordre α de la loi $\chi^2(n-1)$. Le test de Fisher est le test de région de rejet

$$R_{\text{Fisher}} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \widetilde{S}_n^2 < \frac{x_\alpha^{(n-1)}}{n-1}\sigma_1^2 \right\}.$$

Ce test est de taille (et de niveau) α, et la procédure de décision consiste à rejeter H_0 au niveau α lorsque $(X_1, \dots, X_n) \in R_{\text{Fisher}}$.

Par ailleurs, sa fonction de puissance sur $(0, \sigma_1)$ est donnée par

$$\mathbb{P}(\widetilde{S}_n^2 < \frac{x_\alpha^{(n-1)}}{n-1}\sigma_1^2) = \mathbb{P}\Big(\frac{(n-1)\widetilde{S}_n^2}{\sigma^2} < x_\alpha^{(n-1)}\frac{\sigma_1^2}{\sigma^2}\Big)$$
$$= F_{\chi^2(n-1)}(x_\alpha^{(n-1)}\sigma_1^2/\sigma^2),$$

où $F_{\chi^2(n-1)}$ est la fonction de répartition de la loi de $\chi^2(n-1)$. Sa plus petite valeur sur $(0, \sigma_1)$ est α (atteinte pour $\sigma = \sigma_1$) et cette fonction tend vers 1 quand $\sigma \to 0$. C'est aussi un test sans biais.

6.5 Régression linéaire des moindres carrés

Modèle statistique. De manière générale, il s'agit de modéliser une expérience dont chaque observation $Y_i \in \mathbb{R}, 1 \leq i \leq n$, est influencée par des mesures (déterministes) connues $x_i^1, \dots, x_i^k$. On s'intéresse par exemple à l'effet pour un individu i de la concentration dans le sang de k marqueurs chimiques (les $x_i^1, \dots, x_i^k$) sur une certaine charge virale (Y_i). En désignant par $\mathbf{X}$ la matrice de taille $n \times k$ définie par $\mathbf{X} = (x_i^j)_{1 \leq i \leq n, 1 \leq j \leq k}$, le modèle de régression linéaire multiple admet la formulation suivante :

$$Y = \mathbf{X}\theta + \varepsilon,$$

avec $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathrm{Id}_n)$, pour des paramètres inconnus $\theta \in \mathbb{R}^k$ et $\sigma > 0$. Les k vecteurs formant les colonnes de $\mathbf{X}$ sont appelés régresseurs. Ce modèle s'écrit, de façon équivalente,

$$\Big(\mathbb{R}^n, \{\mathcal{N}_n(\mathbf{X}\theta, \sigma^2 \mathrm{Id}_n)\}_{\theta \in \mathbb{R}^k, \sigma > 0}\Big)$$

et l'observation associée n'est autre que $Y = (Y_1, \dots, Y_n)^\intercal$. On impose donc en particulier l'hypothèse d'homoscédasticité selon laquelle la matrice de variance-covariance de la loi dont l'observation est issue est proportionnelle à la matrice identité. Notons également que, pour tout i, si $\theta = (\theta_1, \dots, \theta_k)^\intercal$, on a

$$Y_i = \sum_{j=1}^{k} x_i^j \theta_j + \varepsilon_i,$$

avec $\varepsilon_1, \ldots, \varepsilon_n$ des variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$. En réduisant au besoin leur nombre, on peut toujours considérer que les régresseurs sont linéairement indépendants et que, par conséquent, la matrice $\mathbf{X}$ est de rang k. Cela implique en particulier que $k \leq n$.

Estimation des paramètres. Dans ce qui suit, E désigne l'espace vectoriel engendré par les colonnes de $\mathbf{X}$ et u_E désigne la projection orthogonale de $u \in \mathbb{R}^n$ sur E.

Théorème 19 (Estimation des moindres carrés). *Soit $\hat{\theta} \in \mathbb{R}^k$ tel que $Y_E = \mathbf{X}\hat{\theta}$ soit la projection orthogonale de Y sur E. Alors*

(a) $\hat{\theta} = Argmin_{\theta \in \mathbb{R}^k} \|Y - X\theta\|$

(b) $\hat{\theta} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal Y$

(c) $\hat{\theta} \sim \mathcal{N}_k(\theta, \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1})$. *C'est en particulier un estimateur sans biais de θ.*

(d) $\hat{\sigma}^2 = \frac{\|Y - \mathbf{X}\hat{\theta}\|^2}{n-k} \sim \frac{\sigma^2}{n-k} \chi^2_{(n-k)}$ *est un estimateur de σ^2.*

(e) $\hat{\theta}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

Démonstration. La matrice $\mathbf{X}$ est la matrice d'une application linéaire de $\mathbb{R}^k$ dans $E \subset \mathbb{R}^n$ avec $k \leq n$, application qui est injective (car matrice de rang plein) donc la projection orthogonale Y_E de Y sur E s'écrit $Y_E = \mathbf{X}\hat{\theta}$, où $\hat{\theta}$ existe et est unique. On choisit naturellement $\hat{\theta}$ comme estimateur de θ et, puisque par définition d'une projection orthogonale

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^k} \|Y - \mathbf{X}\theta\|,$$

on l'appelle l'estimateur des moindres carrés.

On peut décrire explicitement $\hat{\theta}$ en remarquant que, comme $Y - \mathbf{X}\hat{\theta}$ est dans l'orthogonal de E, pour tout $u \in \mathbb{R}^k$:

$$\langle \mathbf{X}u, Y - \mathbf{X}\hat{\theta} \rangle = 0.$$

Par suite, $\langle u, \mathbf{X}^\intercal Y - \mathbf{X}^\intercal \mathbf{X}\hat{\theta} \rangle = 0$ pour tout $u \in \mathbb{R}^k$ et donc $\mathbf{X}^\intercal Y = \mathbf{X}^\intercal \mathbf{X}\hat{\theta}$. Remarquez que la matrice $\mathbf{X}^\intercal \mathbf{X}$ (de taille $k \times k$) est symétrique et positive (pour tout $u \in \mathbb{R}^k$, on a $u^\intercal \mathbf{X}^\intercal \mathbf{X} u = \langle \mathbf{X}u, \mathbf{X}u \rangle \geq 0$). Comme par ailleurs $rang(\mathbf{X}) = k$, la matrice $\mathbf{X}^\intercal \mathbf{X}$ est définie positive (si $< \mathbf{X}u, \mathbf{X}u >= 0$, alors $\mathbf{X}u = 0$ donc $u = 0$). Ainsi $\mathbf{X}^\intercal \mathbf{X}$ est une matrice inversible et son inverse est aussi symétrique. On obtient

$$\hat{\theta} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal Y.$$

En conséquence, $\hat{\theta}$ est un vecteur gaussien et c'est un estimateur sans biais de θ car, si $\mathbb{E}_{\theta,\sigma}$ désigne l'espérance sous la loi de Y alors

$$\mathbb{E}_{\theta,\sigma}\hat{\theta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbb{E}_{\theta,\sigma}Y = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}\theta = \theta.$$

Par ailleurs,

$$\begin{aligned}\mathbb{V}_{\theta,\sigma}\hat{\theta} &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\sigma^2\mathrm{Id}_n[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal]^\intercal \\ &= \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1},\end{aligned}$$

ce qui montre le point (c).

Construisons maintenant un estimateur de σ^2. Comme Y_E est la projection de $Y = \mathbf{X}\theta + \varepsilon$ sur E et $\mathbf{X}\theta \in E$ on a $Y_E = \mathbf{X}\theta + \varepsilon_E$ et $Y - Y_E = \varepsilon - \varepsilon_E$, d'où $\|Y - Y_E\|^2 \sim \sigma^2\chi^2(n-k)$ d'après le théorème de Cochran (Théorème 17). La moyenne de la loi $\chi^2(n-k)$ valant $n-k$, l'estimateur

$$\hat{\sigma}^2 = \frac{\|Y - Y_E\|^2}{n-k} = \frac{\|Y - \mathbf{X}\hat{\theta}\|^2}{n-k}$$

de σ^2 est donc sans biais.

Par ailleurs, toujours d'après le théorème de Cochran, on a d'une part que

$$\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k)$$

et, d'autre part, que $\hat{\theta}$ et $\hat{\sigma}^2$ sont des vecteurs aléatoires indépendants (puisque Y_E et $Y_{E^\perp} = Y - Y_E$ le sont). □

Partant du théorème précédent, il est alors possible de construire des intervalles de confiance et des tests portant sur θ ou ses composantes. Nous en donnons quelques exemples ci-dessous.

Test de Wald. Le test de Wald est un test d'hypothèses sur le paramètre θ du modèle linéaire. On l'écrit de la façon générale suivante : On se donne une matrice C de taille $q \times k$, de rang $q \leq k$ (pour éviter les redondances) et un vecteur $a \in \mathbb{R}^q$. L'objectif est de tester

$$H_0 : C\theta = a \quad \text{contre} \quad H_1 : C\theta \neq a.$$

Par exemple si C est l'identité, on teste si la valeur de θ vaut un vecteur a fixé. Lorsque $q = 1$, on teste en fait une combinaison affine des composantes de θ, comme par exemple la nullité de l'un des θ_j (dans ce cas C est une matrice de taille $1 \times k$ et a est un nombre réel).

Le test utilise la statistique de Wald

$$W(Y) = \frac{(C\hat{\theta} - a)^\intercal \left(C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal\right)^{-1} (C\hat{\theta} - a)/q}{\|Y - \mathbf{X}\hat{\theta}\|^2/(n-k)}. \tag{6.1}$$

Théorème 20. *Pour tout paramètre $\theta \in \mathbb{R}^k$, la statistique de Wald définie par (6.1) suit, sous H_0, la loi de Fisher $\mathcal{F}(q, n-k)$. Par ailleurs, si $q = 1$ alors*

$$Z = \frac{C\hat{\theta} - a}{\sqrt{C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal}} \times \frac{\sqrt{n-k}}{\|Y - \mathbf{X}\hat{\theta}\|}$$

suit, sous H_0, la loi de Student à $n-k$ degrés de libertés.

Démonstration. 1) On commence par prouver qu'il existe une matrice Δ de taille $q \times q$ symétrique et définie positive telle que

$$C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal = \Delta^2.$$

Rappelons que $(\mathbf{X}^\intercal \mathbf{X})^{-1}$ est symétrique définie positive, d'où il vient que $C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal$ est également symétrique. Par ailleurs comme C est de rang q, la matrice $C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal$ est également définie positive, ce qui assure l'existence (et même l'unicité) de Δ (racine carrée). D'après le théorème 19,

$$C\hat{\theta} - a \sim \mathcal{N}_q(C\theta - a, \sigma^2 C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal) = \mathcal{N}_q(0, \sigma^2 \Delta^2).$$

Alors $\Delta^{-1}(C\hat{\theta} - a) \sim \mathcal{N}_q(0, \sigma^2 \mathrm{Id})$ et $\|\Delta^{-1}(C\hat{\theta} - a)\|^2 \sim \sigma^2 \chi^2(q)$. Par ailleurs,

$$\begin{aligned} \|\Delta^{-1}(C\hat{\theta} - a)\|^2 &= (C\hat{\theta} - a)^\intercal (\Delta^2)^{-1} (C\hat{\theta} - a) \\ &= (C\hat{\theta} - a)^\intercal C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal (C\hat{\theta} - a). \end{aligned}$$

Finalement

$$(C\hat{\theta} - a)^\intercal \left(C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal\right)^{-1} (C\hat{\theta} - a) \sim \sigma^2 \chi^2(q).$$

Toujours par le théorème 19, $\|Y - \mathbf{X}\hat{\theta}\|^2$ sont $\hat{\theta}$ indépendants, et $\|Y - \mathbf{X}\hat{\theta}\|^2$ est de loi $\sigma^2\chi^2(n-k)$. Il reste à écrire

$$\frac{(C\hat{\theta} - a)^\intercal \left(C(\mathbf{X}^\intercal\mathbf{X})^{-1}C^\intercal\right)^{-1}(C\hat{\theta} - a)}{\|Y - \mathbf{X}\hat{\theta}\|^2} \times \frac{n-k}{q}$$

qui est de loi $\mathcal{F}(q, n-k)$.

Lorsque $q = 1 : C(\mathbf{X}^\intercal\mathbf{X})^{-1}C^\intercal$ est de taille 1×1, c'est un réél. Donc $C\hat{\theta} - a \sim \mathcal{N}_1(0, \sigma^2 C(\mathbf{X}^\intercal\mathbf{X})^{-1}C^\intercal)$ d'où $\frac{C\hat{\theta}-a}{\sigma\sqrt{C(\mathbf{X}^\intercal\mathbf{X})^{-1}C^\intercal}}$ est de loi $\mathcal{N}_1(0,1)$. Par ailleurs $\|Y - \mathbf{X}\hat{\theta}\|^2/\sigma^2$ est indépendante de $\hat{\theta}$, et donc de la variable précédente, et suit la loi $\chi^2(n-k)$. □

Ainsi, si $f_{1-\alpha}^{(q,n-k)}$ désigne le quantile d'ordre $(1-\alpha)$ de la loi $\mathcal{F}(q, n-k)$, on a, sous H_0,

$$P_{H_0}\left(W(Y) > f_{1-\alpha}^{(q,n-k)}\right) = \alpha.$$

La région de rejet

$$R_{\text{Wald}} = \left\{y \in \mathbb{R}^n : W(y) > f_{1-\alpha}^{(q,n-k)}\right\}$$

nous donne donc un test (dit de Wald) de niveau (de taille) α pour le problème de test de $H_0 : C\theta = a$ contre $H_1 : C\theta \neq a$. La procédure de décision consiste à rejeter H_0 au niveau α si l'observation Y tombe dans R_{Wald}.

Pour $q = 1$, ce test reste valable. On peut cependant également utiliser le test dit de Student, de région de rejet

$$R_{\text{Student}} = \left\{y \in \mathbb{R}^n : \left|\frac{C\hat{\theta} - a}{\hat{\sigma}\sqrt{C(\mathbf{X}^\intercal\mathbf{X})^{-1}C^\intercal}}\right| > t_{1-\frac{\alpha}{2}}^{(n-k)}\right\},$$

où $t_{1-\frac{\alpha}{2}}^{(n-k)}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{T}(n-k)$.

En suivant ces mêmes principes, on montre facilement que

$$\left\{a \in \mathbb{R}^k : \frac{(C\hat{\theta} - a)^\intercal \left(C(\mathbf{X}^\intercal\mathbf{X})^{-1}C^\intercal\right)^{-1}(C\hat{\theta} - a)/q}{\|Y - \mathbf{X}\hat{\theta}\|^2/(n-k)} \leq f_{1-\alpha}^{(q,n-k)}\right\}$$

fournit un ellipsoïde de confiance de niveau $(1-\alpha)$ pour le paramètre $C\theta$. Lorsque $C = \text{Id}$, nous avons obtenu un ellipsoïde de confiance pour le vecteur θ à valeurs dans $\mathbb{R}^k$ et si $k = 1$ un intervalle de confiance pour θ.

Par ailleurs pour $q = 1$ on obtient l'intervalle de confiance pour la combinaison affine $C\theta$ suivant de niveau $1 - \alpha$:

$$\left[C\hat{\theta} \pm t_{1-\frac{\alpha}{2}}^{(n-k)} \hat{\sigma} \sqrt{C(\mathbf{X}^\intercal \mathbf{X})^{-1} C^\intercal} \right]$$

où pour rappel $\hat{\sigma} = \|Y - \mathbf{X}\hat{\theta}\| / \sqrt{n-k}$.

Test de Fisher de l'utilité des régresseurs. Dans le cadre d'une modélisation trop complète, tous les régresseurs n'ont pas la même influence et certains n'ont qu'une contribution mineure. Nous allons construire un test dans le but de supprimer ces régresseurs à l'influence réduite.

Fixons $q = 0, \dots, k-1$. S'interroger sur l'utilité des $(k-q)$ derniers régresseurs mène au problème de test suivant :

$$H_0 : \forall i = q+1, \dots, k,\ \theta_i = 0 \quad \text{contre} \quad H_1 : \exists i = q+1, \dots, k, \theta_i \neq 0.$$

Noter que c'est un cas particulier du test de Wald avec la matrice C de taille $(k-q) \times k$ dont les q premières colonnes sont nulles et les suivantes forment Id_{k-q} (cette matrice est de rang $q' = k - q$). Mais nous allons présenter différemment ce cas particulier (pour le même résultat au final).

Notons $R_1, \dots, R_k$ les k vecteurs (régresseurs) formant les colonnes de $\mathbf{X}$, de sorte que $\mathbf{X} = (R_1 \dots R_k)$. Sous H_0, la matrice des régresseurs utiles $\bar{\mathbf{X}} = (R_1 \dots R_q)$ est la restriction de $\mathbf{X}$ à ses q premiers régresseurs. L'effet moyen $\mathbf{X}\theta$ se trouve alors dans l'espace vectoriel V engendré par $R_1, \dots, R_q$, dont la dimension est q (car $R_1, \dots, R_q$ sont linéairement indépendants par hypothèse). Avec ces notations, le problème de test se réécrit de la manière suivante :

$$H_0 : \mathbf{X}\theta \in V \quad \text{contre} \quad H_1 : \mathbf{X}\theta \in E \setminus V.$$

(On rappelle que E est l'espace vectoriel engendré par les colonnes de $\mathbf{X}$). Le principe de construction du test est de rejeter H_0 lorsque les projections orthogonales de l'observation Y sur E et sur V sont significativement différentes. Selon ce principe, une région de rejet naturelle est de la forme $\{y \in \mathbb{R}^n : \|y_E - y_V\| > s\}$ avec s un seuil à préciser. Mais la loi de $\|Y_E - Y_V\|$ dépend du paramètre inconnu σ. En effet, sous H_0, $Y_V = \mathbf{X}\theta + \varepsilon_V$

car $\mathbf{X}\theta \in V$ et donc, d'après le théorème de Cochran appliqué au vecteur gaussien ε,

$$\|Y_E - Y_V\|^2 = \|\varepsilon_E - \varepsilon_V\|^2 \sim \sigma^2\chi^2(k-q).$$

Or, le théorème de Cochran montre aussi que sous H_0, le vecteur aléatoire $\varepsilon_E - \varepsilon_V = Y_E - Y_V$ est indépendant de $\varepsilon - \varepsilon_E = Y - Y_E$. Enfin, $\|Y - Y_E\|^2 \sim \sigma^2\chi^2(n-k)$. En réunissant ces observations et en notant pour $y \in \mathbb{R}^n$:

$$F(y) = \frac{\|y_E - y_V\|^2/(k-q)}{\|y - y_E\|^2/(n-k)},$$

on trouve $F(Y) \sim \mathscr{F}(k-q, n-k)$ sous H_0. Si $f_{1-\alpha}^{(k-q,n-k)}$ désigne le quantile d'ordre $(1-\alpha)$ de la loi $\mathscr{F}(k-q, n-k)$ alors, sous H_0,

$$\mathbb{P}\left(F(Y) > f_{1-\alpha}^{(k-q,n-k)}\right) = \alpha.$$

La région de rejet

$$R_{\text{Fisher}} = \left\{y \in \mathbb{R}^n : F(y) > f_{1-\alpha}^{(k-q,n-k)}\right\}$$

nous donne donc un test (dit de Fisher) de niveau (de taille) α pour le problème de test de H_0 contre H_1 : on rejette H_0 au niveau α si l'observation $Y = (Y_1, \ldots, Y_n)^\intercal$ tombe dans R_{Fisher}. Notons que $F(y)$ se calcule très facilement comme on l'a vu dans la preuve du théorème 19 :

$$y_E = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal y \quad \text{et} \quad y_V = V(V^\intercal V)^{-1}V^\intercal y,$$

où V est la matrice des q premières colonnes de $\mathbf{X}$.

Régression linéaire simple et prévision. Lorsque

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

et $\theta = (\mu, \beta)^\intercal$, le modèle s'écrit $Y_i = \mu + \beta x_i + \varepsilon_i$ pour $1 \leq i \leq n$, avec $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d., de loi commune $\mathcal{N}(0,1)$. On parle alors de régression linéaire simple.

On calcule

$$\mathbf{X}^\intercal\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, (\mathbf{X}^\intercal\mathbf{X})^{-1} = \frac{1}{\sum(x_i - \bar{x}_n)^2}\begin{pmatrix} (1/n)\sum x_i^2 & -\bar{x}_n \\ -\bar{x}_n & 1 \end{pmatrix}$$
$$\mathbf{X}^\intercal Y = \begin{pmatrix} \sum Y_i \\ < Y, x > \end{pmatrix},$$

(avec la notation $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$) et on obtient (puisque $\hat{\theta} = (\hat{\mu}, \hat{\beta})^\intercal = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal Y$) tout d'abord que

$$\hat{\beta} = \frac{-\bar{x}_n \sum Y_i + < Y, x >}{\sum(x_i - \bar{x}_n)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Par ailleurs,

$$\begin{aligned}
\hat{\mu} &= \frac{(1/n)(\sum x_i^2)(\sum Y_i) - \bar{x}_n < Y, x >}{\sum(x_i - \bar{x}_n)^2} \\
&= \frac{(1/n)(\sum Y_i)(\sum x_i^2 - n(\bar{x}_n)^2)}{\sum(x_i - \bar{x}_n)^2} + \frac{(\sum Y_i)(\bar{x}_n)^2 - \bar{x}_n < Y, x >}{\sum(x_i - \bar{x}_n)^2} \\
&= \bar{Y}_n + \bar{x}_n \frac{(\sum Y_i)(\bar{x}_n) - < Y, x >}{\sum(x_i - \bar{x}_n)^2} \\
&= \bar{Y}_n - \hat{\beta}\bar{x}_n.
\end{aligned}$$

On obtient également

$$\hat{\theta} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu \\ \beta \end{pmatrix}, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x}_n \\ -\bar{x}_n & 1 \end{pmatrix}\right). \tag{6.2}$$

Enfin, la variance est estimée sans biais par

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\mu} - \hat{\beta}x_i)^2,$$

quantité indépendante de $\hat{\theta}$ et telle que $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$.

Lorsque l'on dispose d'une nouvelle observation $x^\star$ de la variable explicative et que l'on souhaite prédire l'espérance $m_{x^\star} = \mu + \beta x^\star$, l'estimateur obtenu par *plug-in* est $\hat{m}_{x^\star} = \hat{\mu} + \hat{\beta}x^\star$. Comme $(\hat{\mu}, \hat{\beta})$ est un vecteur gaussien de paramètres donnés par (6.2), la variable aléatoire

$$\hat{m}_{x^\star} - m_{x^\star} = \hat{\mu} - \mu + x^\star(\hat{\beta} - \beta)$$

est aussi gaussienne de moyenne zero et de variance (le calcul demande quelques lignes)

$$\begin{aligned}
&\mathbb{V}(\hat{\mu}-\mu)+(x^\star)^2\mathbb{V}(\hat{\beta}-\beta)+2x^\star\mathrm{Cov}(\hat{\mu}-\mu,\hat{\beta}-\beta)\\
&=\frac{\sigma^2}{\sum(x_i-\bar{x}_n)^2}\Big(\frac{1}{n}\sum x_i^2+x^2+2x^\star(-\bar{x}_n)\Big)\\
&=\cdots\\
&=\sigma^2\Big(\frac{1}{n}+\frac{(x^\star-\bar{x}_n)^2}{\sum(x_i-\bar{x}_n)^2}\Big).
\end{aligned}$$

Par ailleurs $\hat{\sigma}=\frac{1}{n-2}\sum_{i=1}^n(y_i-\hat{\mu}-\hat{\beta}x_i)^2$ est de loi $\sigma^2\chi^2(n-2)$ et indépendante de $(\hat{\mu},\hat{\beta})$. En utilisant les résultats ci-dessus, on conclut que

$$\frac{\hat{m}_{x^\star}-m_{x^\star}}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x^\star-\bar{x}_n)^2}{\sum_{i=1}^n(x_i-\bar{x}_n)^2}}}\sim\mathcal{T}(n-2),$$

d'où l'intervalle de confiance pour la valeur moyenne $m_{x^\star}$ de niveau $(1-\alpha)$:

$$\left[\hat{m}_{x^\star}\pm t_{1-\frac{\alpha}{2}}^{(n-2)}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x^\star-\bar{x}_n)^2}{\sum_{i=1}^n(x_i-\bar{x}_n)^2}}\right].$$

On peut également observer que la projection de ε sur cette nouvelle observation $x^\star$ vérifie $\varepsilon_{x^\star}\sim\mathcal{N}(0,\sigma^2)$ et est indépendante de Y. Il en découle

$$\frac{\hat{m}_{x^\star}-(m_{x^\star}+\varepsilon_{x^\star})}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^\star-\bar{x}_n)^2}{\sum_{i=1}^n(x_i-\bar{x}_n)^2}}}\sim\mathcal{T}(n-2).$$

On peut donc donner un intervalle de **prévision** dans lequel la nouvelle observation $Y_{x^\star}=m_{x^\star}+\varepsilon_{x^\star}$ appartiendra avec probabilité $1-\alpha$:

$$\left[\hat{m}_{x^\star}\pm t_{1-\frac{\alpha}{2}}^{(n-2)}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^\star-\bar{x}_n)^2}{\sum_{i=1}^n(x_i-\bar{x}_n)^2}}\right].$$

(Noter l'accroissement de la variance, donc de l'imprécision.)