Statistiques 2017-19

FIMFA ENS ULM

S. Boucheron, C. Boyer et R. Ryder

11 décembre 2018

TABLE DES MATIÈRES

1.1 OBJECTIFS

Ce premier cours introduit autour d'un exemple élémentaire les principaux thèmes de la statistique dite inférentielle (qu'on distingue de la statistique dite descriptive). Ces trois thèmes, l'estimation ponctuelle, la construction de régions de confiance et la construction de procédures de décision (les tests), suppose un effort préalable de modélisation stochastique. Sur l'exemple élémentaire, on peut mener ce travail de modélisation. Cela nous conduit à une première formulation de ce qu'est une expérience ou un modèle statistique. Dans le cadre le plus simple, une expérience statistique est une collection de lois de probabilités. On observe une ou des réalisations d'une de ces lois (sans savoir à laquelle on a affaire). On cherche à estimer, inférer des propriétés de cette loi, peut être pour prendre une décision. Nous passerons en revue des définitions qui nous seront utiles pendant toute la suite du cours (statistique, estimateur, biais, risque, ...) et surtout nous verrons à cette occasion comment les théorèmes limites du calcul des probabilités, loi des grands nombres, théorème central limite, nous guident dans la construction et la justification des méthodes d'estimation et de décision. Nous verrons aussi que ces théorèmes limites sont complétés par des résultats non-asymptotiques appelés inégalités de concentration. Nous terminerons ce cours par une première version du résultat fondateur de la thérie des tests, le lemme de Neyman et Pearson.

1.2 UN PROBLÈME JOUET, TROIS QUESTIONS

On s'apprête à jouer à pile ou face avec une pièce de monnaie. Mais on soupçonne que cette pièce n'est pas parfaitement équilibrée, que la probabilité d'obtenir « face » (θ) n'est pas $1/2$. Avant de jouer avec un adversaire, on veut estimer cette probabilité θ ou encore le ratio $\theta/(1-\theta)$. Pour estimer cette probabilité durant un jeu futur (et peut être ajuster une stratégie), on réalise n lancers aléatoires indépendants. On note les résultats $x_1, x_2, \ldots, x_n$. Ces résultats constituent « les données » ou l'« échantillon ». On construit à partir de ces données une « estimation » $\widehat{\theta}_n$ de θ, cette estimation est une fonction des données, pas de l'estimande θ qui reste inconnue. On espère que $\widehat{\theta}_n$ sera proche de θ. Nous avons affaire là à un problème d'« estimation ponctuelle ».

Le résultat d'une estimation ponctuelle est une valeur. Savoir que cette valeur est probablement proche de l'estimande est satisfaisant mais d'un intérêt limité. Pour envisager l'avenir, il est plus utile de construire un intervalle de confiance, c'est-à-dire deux fonctions des données $\underline{\theta}_n, \overline{\theta}_n$ telles qu'avec une forte probabilité, l'estimande θ appartient à l'intervalle aléatoire

$$\left[\underline{\theta}_n(x_1, \ldots, x_n), \overline{\theta}_n(x_1, \ldots, x_n)\right] =: \left[\underline{\theta}_n, \overline{\theta}_n\right].$$

Ce problème est celui de la construction de *régions de confiance*. Il faut réaliser un bon compromis entre la précision de l'intervalle de confiance $\overline{\theta}_n - \underline{\theta}_n$ et la probabilité de couverture, c'est à dire la probabilité que $\theta \in [\underline{\theta}_n, \overline{\theta}_n]$.

Enfin on peut se poser un problème de « décision ». Si par exemple, on est prêt à jouer avec une pièce biaisée en faveur de « face », mais pas avec une pièce biaisée en faveur de « pile », comment décider à partir des données si on est prêt à jouer ou non (comment décider entre l'hypothèse $\theta > 1/2$ et l'hypothèse $\theta < 1/2$) ? C'est le problème des « tests ».

1.3 QUELQUES DÉFINITIONS : EXPÉRIENCE STATISTIQUE, ÉCHANTILLON, STATISTIQUE, ESTIMATEUR

La notion d'« expérience statistique » est une formalisation dans le langage du calcul des probabilités du jeu que nous venons d'évoquer.

Au départ, on dispose d'un espace probabilisable $(\Omega, \mathcal{F})$ (l'univers et une tribu de parties). Ici $\Omega = \{\text{pile}, \text{face}\}$ et $\mathcal{F} = 2^\Omega$. C'est en général plus riche, avec $\Omega = \mathbb{R}^d$ et $\mathcal{F}$ les boréliens de $\mathbb{R}^d$. On peut aussi rencontrer des situations où Ω est un espace de fonctions (statistique des processus), le choix de la tribu n'est plus tout à fait évident.

Sur cet espace probabilisable, on considère un *ensemble de lois de probabilités* $\mathcal{P}$. Chaque loi de $\mathcal{P}$ est susceptible de régir le phénomène que le statisticien cherche à étudier. Dans le cadre du problème jouet de la section précédente, on peut choisir $\mathcal{P}$ comme l'ensemble de lois non-dégénérées sur $\{\text{pile}, \text{face}\}$ (la probabilité d'obtenir « face » $\theta \in]0, 1[$).

On peut munir $\mathcal{P}$ d'un « système de coordonnées », d'une *paramétrisation*, c'est à dire d'une fonction d'un ensemble Θ (souvent une partie de $\mathbb{R}^d$) dans $\mathcal{P}$. On note génériquement P_θ l'élément de $\mathcal{P}$ associé à θ. Dans le cas de notre problème jouet, nous avons implicitement paramétrisé les lois de Bernoulli par les probabilités de succès. Une paramétrisation est un choix de convenance.

Une paramétrisation est dite *identifiable* si $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$.

Dans notre problème jouet, les paramétrisations (par la probabilité de « face », par le ratio des probabilités « face »/« pile », ou son logarithme) sont identifiables. L'identifiabilité est une propriété désirable mais ce n'est pas indispensable (les modèles de « mélange » sont utiles mais rarement identifiables).

Il est possible que le statisticien n'ait pas directement accès aux réalisations des tirages selon P (la loi de la nature), c'est à dire aux éléments de Ω. Par exemple, lorsque Ω est un espace de fonctions (les trajectoires d'un processus), il est sans doute trop couteux d'observer l'infinité de points qui forment la trajectoire, on se contente d'observer la trajectoire périodiquement, on « échantillonne ». Pour formaliser ce genre de situations, on ajoute à l'expérience un espace d'observations $\mathcal{X}$ (muni d'une tribu $\mathcal{G}$) et une fonction $X : \Omega \longrightarrow \mathcal{X}$ qu'on suppose $\mathcal{G}/\mathcal{F}$ mesurable. Toute loi $P \in \mathcal{P}$ définit une loi image $P \circ X^{-1}$. Au lieu d'observer $\omega \in \Omega$, on observe $x = X(\omega)$.

Une expérience statistique générale est donc définie par $(\Omega, \mathcal{F}, \mathcal{P}, \Theta, \mathcal{X}, \mathcal{G}, X)$. Dans les situations dites canoniques, $\Omega = \mathcal{X}$ et $X = \mathrm{Id}$.

Dans ce cours, nous nous concentrons sur les expériences dites *produit*, construites à partir de répétitions indépendantes d'une expérience de base. Ces expériences sont de la forme $(\Omega^n, \sigma\left(\times_{i=1}^n \mathcal{F}\right), \mathcal{P}_n := \{P^{\otimes n}, P \in \mathcal{P}\}, \Theta, \mathcal{X}^n, \sigma\left(\times_{i=1}^n \mathcal{G}\right), X)$.

On dit que x_i est la réalisation de X_i (variable aléatoire). La loi jointe de $X_1, \ldots, X_n$ est une loi produit de la forme $(P_\theta \circ X^{-1})^{\otimes n}$ pour $\theta \in \Theta$: pour $B_1, \ldots, B_n \in \mathcal{G}$,

$$P_\theta^{\otimes n}\left(\cup_{i=1}^n \{X_i \in B_i\}\right) = \prod_{i=1}^n P_\theta\{X_i \in B_i\}.$$

On parle d'*expérience échantillonnée*. Très souvent, on se contente de rappeler $(P_\theta, \theta \in \Theta)$, le reste étant sous-entendu. Par exemple, dans notre problème jouet, $(B_\theta, \theta \in]0,1[)$ où B_θ est la loi de Bernoulli de probabilité de succès θ.

Toute fonction mesurable sur l'espace des observations $(\mathcal{X}^n)$ définit ce qu'on nomme une *statistique*.

Exemple 1.1 La moyenne empirique

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^n x_i,$$

la variance empirique

$$S^2 := \frac{1}{n}\sum_{i=1}^n \left(x_i - \overline{X}_n\right)^2$$

sont des statistiques. Dans le langage des statistiques descriptives, la moyenne empirique décrit la localisation de l'échantillon, la variance empirique décrit la dispersion.

Un *estimateur* n'est qu'une statistique censée estimer une caractéristique (inconnue) de la loi inconnue qui sous-tend l'échantillonage.

Par exemple, dans notre problème jouet, on peut chercher à estimer $P_\theta\{\text{Face}\} = \theta$, par $\overline{X}_n$ en utilisant la convention $X(\text{Face}) = 1 = 1 - X(\text{Pile})$.

Attention : un estimateur est une fonction de l'échantillon, et non pas une fonction de la loi de l'échantillonnage. La loi de l'estimateur dépend (en général) de la loi de l'échantillonnage.

Quand le paramètre à estimer s'appelle $\theta, \psi, \ldots$, on utilise souvent le raccourci $\widehat{\theta}$ ou $\widehat{\theta}_n, \widehat{\psi}_n, \ldots$ pour désigner l'estimateur (plutôt que $\widehat{\theta}(X_1, \ldots, X_n)$ ou $\widehat{\psi}(X_1, \ldots, X_n)$).

Un estimateur est une variable aléatoire. On peut visualiser ses fluctuations à l'aide de maintes techniques graphiques comme les histogrammes, voir Figure 1.3.

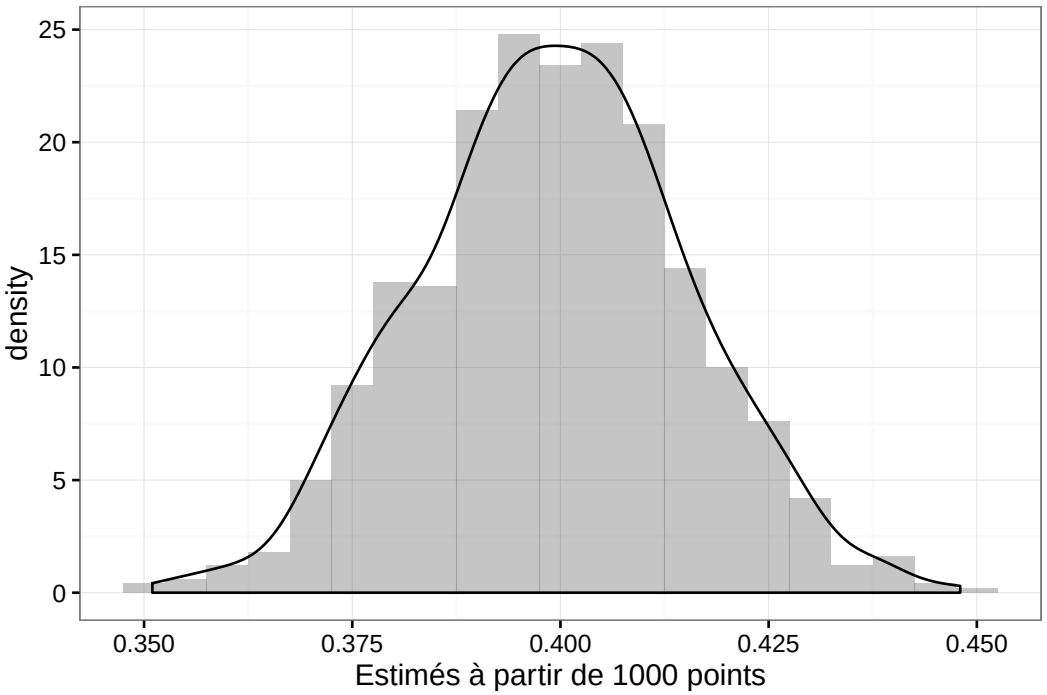

FIG. 1.1 : Histogramme d'un échantillon de 1000 estimations obtenues à partir de 1000 échantillons de taille 1000 de tirages de Bernoulli de probabilité de succès $\theta = .4$.

1.4 PROPRIÉTÉS DES ESTIMATEURS

La plupart des expériences/modèles statistiques que nous rencontrerons dans ce cours, seront de nature paramétrique, autrement dit indexés par des parties de $\mathbb{R}^d$. Dans de nombreux développements des statistiques, par exemple en estimation de densité, on travaille sur des modèles plus riches qui n'admettent pas de paramétrisation « naturelle » par une partie d'un espace euclidien de dimension finie. On parle pourtant de paramètre d'une distribution pour désigner ce qui devrait plutôt s'appeler une fonctionnelle. Par exemple, la moyenne, la covariance d'une distribution sur $\mathbb{R}^d$ sont des paramètres de cette distribution. Les quantiles, l'asymétrie, la kurtosis sont d'autres paramètres.

DÉFINITION 1.2 (BIAIS) Soit $\psi(P)$ un paramètre à estimer, et $\hat{\psi}$ un estimateur, on appelle *biais* (ou biais moyen) sous la loi P de l'estimateur $\hat{\psi}$, la quantité

$$\mathbb{E}_P\left[\hat{\psi} - \psi\right].$$

C'est l'écart entre la valeur moyenne de $\hat{\theta}$ et la valeur visée θ.

L'estimateur est dit *sans biais* s'il est de biais nul.

EXEMPLE 1.3 Si on se place dans le modèle binomial et qu'on cherche à estimer la probabilité de succès θ, la fréquence empirique des succès est un estimateur sans biais de θ (la fréquence empirique d'un événement est toujours un estimateur sans biais de la probabilité de cet événement).

En revanche, on peut vérifier qu'il n'existe pas d'estimateur sans biais de $1/\theta$ ou de $\theta/(1-\theta)$.

EXEMPLE 1.4 Si $\psi(P)$ désigne la variance de la loi P sur $\mathbb{R}$, la variance empirique S^2 définie plus haut est un estimateur biaisé de $\psi(P)$:

$$\mathbb{E}_P\left[S^2\right] = \frac{n-1}{n}\mathbb{E}_P\left[(X - \mathbb{E}_P X)^2\right].$$

DÉFINITION 1.5 (RISQUE QUADRATIQUE) Soit ψ une paramètre à estimer, et $\hat{\psi}$ un estimateur, on appelle *écart quadratique moyen* sous la loi P de l'estimateur $\hat{\psi}$ la quantité

$$\mathbb{E}_P\left[(\hat{\psi} - \psi)^2\right].$$

Exemple 1.6 Dans le cas du problème jouet, le risque quadratique de l'estimateur $\overline{X}_n$ de θ n'est autre que la variance de l'estimateur :

$$\mathbb{E}_\theta\left[\left(\overline{X}_n-\theta\right)^2\right]=\frac{\theta(1-\theta)}{n}\,.$$

La *décomposition biais-variance* du risque quadratique est une relation pythagoricienne.

$$\mathbb{E}_P\left[(\hat{\psi}-\psi)^2\right]=\underbrace{\mathrm{Var}_P[\hat{\psi}]}_{\text{variance}}+\underbrace{\left(\mathbb{E}_P[\hat{\psi}]-\psi\right)^2}_{\text{carré du biais}}\,.$$

La dépendance du risque quadratique vis à vis de la taille de l'échantillon est une question importante en statistique mathématique. Elle concerne la *vitesse d'estimation* (pour une suite d'expériences donnée, quelles sont les meilleures vitesses envisageables, et comment les obtenir ?).

Pour introduire la notion de consistance d'une suite d'estimateurs, nous aurons besoin des notions de convergence en probabilité et de convergence presque sûre.

Définition 1.7 Une suite de variables aléatoires X_n à valeurs dans $\mathbb{R}^k$, vivant sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$ *converge en probabilité* vers une variable aléatoire X à valeurs dans $\mathbb{R}^k$, vivant sur cet espace probabilisé si et seulement si, pour tout $\epsilon>0$

$$\lim_n \mathbb{P}\{\|X_n-X\|>\epsilon\}=0\,.$$

Définition 1.8 (consistance) Dans une suite d'expériences statistiques échantillonnées, une suite d'estimateurs $(\widehat{\theta}_n)$ est *consistante* (pour l'estimation de θ) si

$$\forall\theta\in\Theta,\forall\epsilon>0,\qquad \lim_n P_\theta^{\otimes n}\left\{\|\widehat{\theta}-\theta\|>\epsilon\right\}=0\qquad\text{(convergence en probabilité).}$$

La suite est *fortement consistante* si

$$\forall\theta\in\Theta,\forall\epsilon>0,\qquad P_\theta^{\otimes\mathbb{N}}\left\{\lim_n\|\widehat{\theta}-\theta\|=0\right\}=1\qquad\text{(convergence presque sûre).}$$

Pour notre problème jouet, la suite d'estimateurs $(\overline{X}_n)$ est fortement consistante pour l'estimation de θ (loi forte des grands nombres). On peut aussi vérifier que la suite $(\overline{X}_n/(1-\overline{X}_n))$ est fortement consistante pour l'estimation de $\theta/(1-\theta)$.

Ces suites d'estimateurs répondent aux questions d'estimation ponctuelle. On peut toutefois se demander s'il s'agit des meilleures réponses possibles. On peut par exemple se demander s'il n'y a pas d'information inexploitée dans l'échantillon. On peut se rassurer en remarquant que pour tout θ

$$\begin{aligned}P_\theta\{x_1,\ldots,x_n\} &= \theta^{n\overline{X}_n}(1-\theta)^{n(1-\overline{X}_n)}\\ &= \left(\frac{\theta}{1-\theta}\right)^{n\overline{X}_n}(1-\theta)^n\\ &= \exp\left(n\overline{X}_n\log\left(\frac{\theta}{1-\theta}\right)-n\log(1-\theta)\right)\,,\end{aligned}$$

et que donc

$$P_\theta\{x_1,\ldots,x_n\mid\overline{X}_n\}=\frac{\mathbb{I}_{n\overline{X}_n=\sum_{i=1}^n x_i}}{\binom{n}{n\overline{X}_n}}$$

autrement dit que conditionnellement à $\overline{X}_n$, la probabilité de l'échantillon ne dépend pas de θ (est « libre de θ »). Dans ce modèle jouet, $\overline{X}_n$ est une *statistique suffisante* ou *exhaustive*.

1.5 intervalles de confiance

Définition 1.9 (intervalle de niveau de confiance $1-\alpha$) Lorsque l'espace des paramètres Θ est inclus dans $\mathbb{R}$, un intervalle de niveau de confiance $1-\alpha$ ($\alpha\in]0,1[$) est un couple de statistiques $\underline{\theta}_n,\overline{\theta}_n$ telles que

$$\forall\theta\in\Theta,\qquad P_\theta^{\otimes n}\left\{\theta\in[\underline{\theta}_n,\overline{\theta}_n]\right\}\geq 1-\alpha\,.$$

Il ne faut jamais perdre de vue que l'intervalle de confiance est une statistique, il doit être calculable à partir des données accessibles au statisticien (l'échantillon, y compris sa taille, α, le cadre de l'expérience statistique).

Il n'est pas toujours évident de construire un intervalle de niveau de confiance exactement $1-\alpha$. On est très souvent amené à proposer des solutions très conservatrices (des intervalles trop larges). En revanche, le calcul des probabilités nous fournit des constructions assez simples d'intervalles de niveau de confiance asymptotique prescrit.

Définition 1.10 (intervalle de niveau de confiance asymptotique $1-\alpha$) Lorsque l'espace des paramètres Θ est inclus dans $\mathbb{R}$, un intervalle de niveau de confiance asymptotique $1-\alpha$ ($\alpha \in]0,1[$) est un couple de statistiques $\underline{\theta}_n, \overline{\theta}_n$ telles que

$$\forall \theta \in \Theta, \qquad \lim_n P_\theta^{\otimes n} \left\{ \theta \in [\underline{\theta}_n, \overline{\theta}_n] \right\} = 1-\alpha \,.$$

Construction naïve

Si les X_i sont des variables de Bernoulli indépendantes et si $Z = \sum_{i=1}^n X_i$ alors l'inégalité de Chebychev implique

$$\mathbf{P}\left\{ |Z - \mathbf{E}Z| \geq \sqrt{\frac{n}{4\alpha}} \right\} \leq \alpha \,.$$

On en déduit un intervalle de niveau de confiance $1-\alpha$:

$$\left[\widehat{\theta} - \sqrt{\frac{1}{4n\alpha}}, \widehat{\theta} + \sqrt{\frac{1}{4n\alpha}} \right] .$$

Si on cherche à évaluer le taux de couverture de l'IC déduit de l'inégalité de Bienaymée-Chebychev lorsque la taille de l'échantillon n'est que $N = 1000$, en visant un niveau de confiance $1-\alpha$ avec $\alpha = .25$, on constate que ce taux évalué à partir de 1000 essais est largement supérieur au taux de couverture ciblé. Cet intervalle manque définitivement de précision.

Construction asymptotique

Dans cette section nous ne considérons que des probabilités sur $\mathbb{R}$. Celles-ci sont complètement caractérisées par leur fonction de répartition. Les livres d'introduction aux probabilités contiennent souvent la définition suivante.

Définition 1.11 Une suite $(P_n)_{n\in\mathbb{N}}$ de probabilités sur $\mathbb{R}$ (de fonctions de répartition $(F_n)_{n\in\mathbb{N}}$) converge étroitement/faiblement vers une loi de probabilité P de fonction de répartition F si et seulement si, pour tout x où F est continue, on a

$$\lim_n F_n(x) = F(x) \,.$$

La situation des points où F est discontinue est la suivante.

Proposition 1.12 *Si une suite de fonctions de répartition $(F_n)_{n\in\mathbb{N}}$ converge simplement vers une fonction de répartition F en tout point de continuité de F, alors en tout x de $\mathbb{R}$*

$$\limsup_n F_n(x) \leq F(x) \,.$$

Convention : Pour $\alpha \in]0,1[$, on note z_α le quantile d'ordre $1-\alpha$ de la gaussienne centrée réduite (standard),

C'est la solution de l'équation en x :

$$1-\alpha = \int_{-\infty}^{x} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}} \mathrm{d}u =: \Phi(x) \,.$$

Dans la suite du texte, on utilise la notation $\rightsquigarrow$ pour désigner la convergence en loi/distribution.

Le théorème central limite dans sa version la plus simple (De Moivre-Laplace) nous indique que si les $\widehat{\theta}_n$ sont distribués selon $P_\theta^{\otimes n}$,

$$\frac{\sqrt{n}}{\sqrt{\theta(1-\theta)}}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0,1)\,,$$

ce qui se traduit (entre autres) par la convergence simple des fonctions de répartitions, soit pour tout $\alpha \in]0,1[$

$$\lim_n \mathbf{P}_\theta^{\otimes n}\left\{\frac{\sqrt{n}}{\sqrt{\theta(1-\theta)}}\left(\widehat{\theta}_n - \theta\right) \leq z_\alpha\right\} = 1-\alpha\,.$$

Le lemme de Slutsky (voir Appendice C.3), et le fait que $\widehat{\theta}_n/\theta$ converge en probabilité vers 1 lorsque $n \to \infty$, permet d'écrire pour tout $\alpha \in]0,1[$,

$$\lim_n \mathbf{P}_\theta^{\otimes n}\left\{\frac{\sqrt{n}}{\sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)}}\left(\widehat{\theta}_n - \theta\right) \leq z_\alpha\right\} = 1-\alpha\,.$$

Cela conduit à proposer l'intervalle de niveau de confiance asymptotique $1-\alpha$:

$$\left[\widehat{\theta}_n - z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}, \widehat{\theta}_n + z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}\right]\,.$$

Un raffinement du théorème central limite, le théorème de Berry-Esseen (voir Appendice C.8), nous indique que le niveau de confiance est $1-\alpha+O(1/\sqrt{n})$.

Intervalle non-asymptotique construit à partir de l'inégalité de Hoeffding

L'inégalité de Hoeffding est la plus simple des inégalités exponentielles qui fournissent des bornes non-asymptotiques sur les probabilités de queue des sommes de variables aléatoires indépendantes.

LEMME 1.13 (LEMME DE HOEFFDING) *Si X est une variable aléatoire qui prend ses valeurs dans $[a,b]$, alors pour tout $\lambda \geq 0$,*

$$\log \mathbb{E}\mathrm{e}^{\lambda(X-\mathbb{E}X)} \leq \frac{\lambda^2(b-a)^2}{8}\,.$$

PREUVE. Sans perdre en généralité, on suppose X centrée (au pire cela revient à translater l'intervalle $[a,b]$, ce qui ne change pas sa longueur). On note Q la loi (implicite) de la variable aléatoire X. Observons d'abord que la variance de toute variable aléatoire qui prend ses valeurs dans $[a,b]$ est majorée par $(b-a)^2/4$. Considérons maintenant la fonction F de λ définie par

$$F(\lambda) = \log \mathbb{E}_Q \mathrm{e}^{\lambda X}\,.$$

Et notons Q_λ la loi de densité $\exp\left(\lambda x - F(\lambda)\right)$ par rapport à Q. On vérifie que

$$F'(\lambda) = \mathbb{E}_{Q_\lambda} X \qquad \text{et} \qquad F''(\lambda) = \mathrm{var}_{Q_\lambda}(X)\,.$$

Comme Q_λ est absolument continue par rapport à Q, sous Q_λ, X est à valeur dans $[a,b]$, et donc

$$F''(\lambda) \leq \frac{(b-a)^2}{4}\,.$$

On peut intégrer cette inégalité différentielle en notant au passage que $F(0) = F'(0) = 0$, et vérifier

$$F(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}\,.$$

□

LEMME 1.14 (INÉGALITÉ DE HOEFFDING) *Si les $(X_i)_{i\leq n}$ sont des variables aléatoires indépendantes à valeur dans $[a_i, b_i]$ et si $Z = \sum_{i=1}^n X_i$ alors pour tout $t > 0$*

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + t\right\} \leq \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}.$$

PREUVE. La preuve se réduit à une invocation de l'inégalité de Markov exponentielle et du lemme de Hoeffding. □

Si les X_i sont des variables de Bernoulli indépendantes et si $Z = \sum_{i=1}^n X_i$ alors l'inégalité de Hoeffding implique

$$\mathbf{P}\left\{|Z - \mathbf{E}Z| \geq \sqrt{\frac{n\log(2/\alpha)}{2}}\right\} \leq \alpha\,.$$

On en déduit un intervalle de niveau de confiance $1-\alpha$:

$$\left[\widehat{\theta} - \sqrt{\frac{\log(2/\alpha)}{2n}}, \widehat{\theta} + \sqrt{\frac{\log(2/\alpha)}{2n}}\right].$$

Dans toutes ces constructions on retrouve deux ingrédients, l'intervalle est d'une largeur proportionnelle à $\sqrt{1/n}$ et à un facteur qui dépend du niveau de couverture recherché. Plus nos renseignements sur les fluctuations de $\overline{X}_n$ autour de son espérance sont précis, plus petit est l'intervalle de confiance.

Construction calculatoire

Dans cette section, on note $\mathrm{qb}(\alpha, n, \theta)$ le quantile d'ordre α de la loi binomiale de paramètres n, θ (cela correspond à la fonction `qbinom()` de `R`). On définit la région empirique

$$\left\{\theta' : \mathrm{qb}(\alpha/2, n, \theta') \leq n\widehat{\theta}_n \leq \mathrm{qb}(1-\alpha/2, n, \theta')\right\}$$

Cette région est un intervalle. On peut le vérifier à l'aide d'un argument de *domination stochastique* : si $0 < \theta < \theta' < 1$ et si $F_{n,\theta}$, $F_{n,\theta'}$ désignent les fonctions de répartition des binomiales de paramètres (n, θ) et (n, θ'), alors pour tout x

$$F_{n,\theta'}(x) \leq F_{n,\theta}(x)\,.$$

Cette dernière relation se vérifie par un argument de couplage.

La région de confiance est délimitée par

$$\underline{\theta} = \inf\{\theta' : \mathrm{qb}(1-\alpha/2, n, \theta') \geq n\widehat{\theta}_n\}$$

et

$$\overline{\theta} = \sup\{\theta' : \mathrm{qb}(\alpha/2, n, \theta') \leq n\widehat{\theta}_n\}\,.$$

C'est aussi une région de niveau de confiance $1 - \alpha + O(1/\sqrt{n})$.

1.6 TESTS

Une *hypothèse* est une collection de loi de probabilités. La collection peut être réduite à une seule loi, on parle alors d'*hypothèse simple*, sinon on parle d'*hypothèse composée ou composite*.

Ici, on veut tester

1. H_0 : l'hypothèse nulle, $\theta \leq \theta_0 = .5$ contre
2. H_1 : l'alternative $\theta > .5$.

Un test binaire est une fonction des données qui vaut 1 (on rejette l'hypothèse nulle H_0) ou 0 (on ne rejette pas H_0). Dans la suite on notera T le test binaire.

On peut se demander pourquoi on emploie l'expression « on ne rejette pas l'hypothèse nulle H_0 », plutôt que « on accepte l'hypothèse nulle ». Ce n'est pas par goût des formes négatives. C'est parce que dans les usages historiques qui ont conduit à la construction de la notion de test, l'hypothèse nulle et l'alternative ne jouent pas le même rôle. L'hypothèse nulle correspond à une position conservatrice. Lorsqu'on procède à des essais cliniques, pour évaluer l'intérêt de mettre sur le marché un nouveau médicament, l'hypothèse nulle affirme que ce nouveau traitement ne vaut pas mieux que l'existant, l'alternative affirme qu'au contraire ce nouveau traitement est meilleur. Ne pas rejeter l'hypothèse nulle, cela ne veut pas dire accepter l'existant pour l'éternité, mais s'y tenir jusqu'à l'apparition d'éléments nouveaux.

On note $\mathcal{P}_0$ la collection des lois de probabilité qui définissent l'hypothèse nulle et $\mathcal{P}_1$ la collection des lois de probabilité qui définissent l'alternative.

Définition, types d'erreur

De même que le risque quadratique nous permet de quantifier les performances d'un estimateur, les notions d'erreur de première et de seconde espèce nous permettent de quantifier les performances d'un test binaire. Notez qu'il nous faut introduire deux quantités pour quantifier les performances d'un test.

L'*erreur de première espèce* consiste à rejeter H_0 à tort lorsque les données sont des tirages selon une loi appartenant à l'hypothèse nulle (les données sont tirées sous l'hypothèse nulle).

L'*erreur de seconde espèce* consiste à ne pas rejeter H_0 à tort lorsque les données sont des tirages selon une loi appartenant à l'hypothèse alternative (les données sont tirées sous l'alternative).

On appelle *niveau* du test T,

$$\sup_{P\in\mathcal{P}_0} P\{T=1\}$$

(le supremum de l'erreur de première espèce).

On appelle *puissance* du test T sous $P \in \mathcal{P}_1 \cup \mathcal{P}_0$, la probabilité que T rejette H_0 sous P :

$$\beta_T(P) = P\{T=1\}\,.$$

Sous l'alternative, la puissance est le complément à un de l'erreur de seconde espèce.

On veut à la fois un test de petit niveau et de grande puissance sous l'alternative. Ces deux souhaits sont antagonistes. Dans le cas où on teste deux hypothèses simples, il existe une méthodologie qui réalise le meilleur compromis possible.

TESTS DITS DE RAPPORT DE VRAISEMBLANCE On peut associer à chaque $\theta \in]0,1[$ et à chaque échantillon $x_1,\dots,x_n$, une *vraisemblance* qui n'est autre que la probabilité de $x_1,\dots,x_n$ sous $P_\theta^{\otimes n}$:

$$P_\theta^{\otimes n}\{x_1,\dots,x_n\} = \left(\frac{\theta}{1-\theta}\right)^{n\overline{X}_n}(1-\theta)^n\,.$$

DÉFINITION 1.15 (TEST DE RAPPORT DE VRAISEMBLANCE ENTRE HYPOTHÈSES SIMPLES) Un test de rapport de vraisemblance de H_1 contre H_0 consiste à comparer le rapport $P_{\theta_1}^{\otimes n}\{x_1,\dots,x_n\}/P_{\theta_0}^{\otimes n}\{x_1,\dots,x_n\}$ à un seuil, à rejeter H_0 si le seuil est dépassé, à ne pas rejeter H_0 si le seuil n'est pas dépassé.

Ici, le rapport de vraisemblance est une fonction de $\overline{X}_n = \sum_{i=1}^n X_i/n = \widehat{\theta}_n$ (ce n'est pas du tout une simple coïncidence).

$$\left(\frac{1-\theta_1}{1-\theta_0}\right)^n\left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{n\widehat{\theta}_n}$$

Comparer le rapport de vraisemblance à un seuil, c'est ici équivalent à comparer $\widehat{\theta}_n$ à un seuil, à rejeter H_0 lorsque $\widehat{\theta}_n$ dépasse le seuil.

Optimalité des tests dits de rapport de vraisemblance

LEMME 1.16 (VERSION PRÉLIMINAIRE DU LEMME DE NEYMAN-PEARSON) *S'il existe un test de rapport de vraisemblance T_0 de niveau $\alpha > 0$ et de fonction puissance β_{T_0}, alors pour tout test T de niveau inférieur ou égal à α, la fonction puissance β_T de T vérifie*

$$\beta_T(P_1) \leq \beta_{T_0}(P_1).$$

PREUVE. On note $p_0()$ et $p_1()$ les versions des densités utilisées dans la définition du test T_0. Il existe une valeur $\tau < \infty$, telle que

$$P_0\{p_0(X)/p_1(X) > \tau\} = \alpha.$$

Et T_0 est défini par

$$T_0(x) = \mathbb{1}_{p_1(x)/p_0(x)>\tau.}$$

La preuve du lemme 1.16 se réduit alors à :

$$\begin{aligned}
\beta_{T_0}(P_1) - \beta_T(P_1) &= \mathbb{E}_{P_1}[T_0 - T] \\
&= \mathbb{E}_{P_0}\left[\frac{p_1(X)}{p_0(X)}(T_0 - T)\right] + \mathbb{E}_{P_1}\left[(T_0 - T)\mathbb{1}_{p_0(X)=0}\right] \\
&\quad \text{sur l'événement } p_0(X) = 0, T_0 = 1, \text{car le rapport} \\
&\quad \text{de vraisemblance est infini} \\
&\geq \mathbb{E}_{P_0}\left[\frac{p_1(X)}{p_0(X)}(T_0 - T)\right] \\
&= \mathbb{E}_{P_0}\left[\left(\frac{p_1(X)}{p_0(X)} - \tau\right)(T_0 - T)\right] + \tau\mathbb{E}_{P_0}[T_0 - T] \\
&\quad \text{comme } \left(\frac{p_1(X)}{p_0(X)} - \tau\right)(T_0 - T) \geq 0, \\
&\geq \tau\mathbb{E}_{P_0}[T_0 - T] \\
&\geq 0.
\end{aligned}$$

□

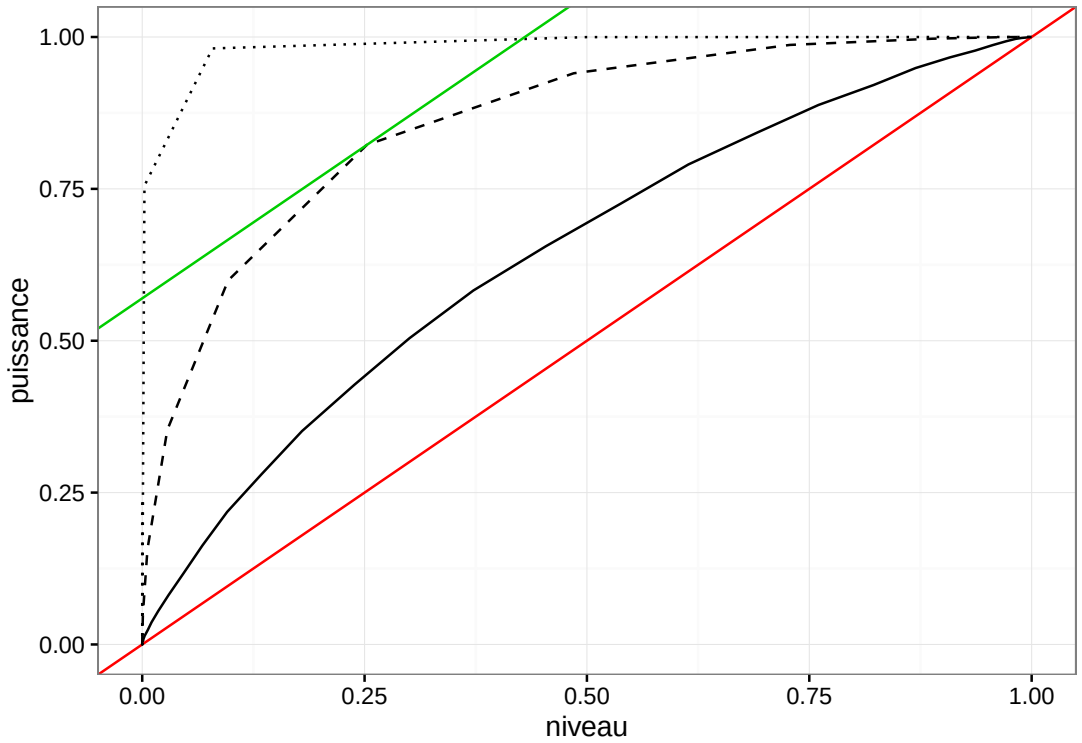

FIG. 1.2 : Pour une grille de seuils, on évalue niveau et puissance pour différentes tailles d'échantillon (100, 1000, 5000). Sur un même graphique on représente la courbe niveau/puissance pour ces trois tailles d'échantillon (lignes brisées noires). Pour chaque courbe, le meilleur compromis erreur de première espèce/erreur de seconde espèce est la distance ℓ_1 au point $(0, 1)$. On constate (sans surprise) que cette distance diminue lorsque la taille de l'échantillon augmente. Les compromis optimaux (au sens de la minimisation de $\alpha_T(P_0) - \beta_T(P_1)$) peuvent être visualisés en traçant des parallèles à la diagonale principale comme la ligne verte, en choisissant comme `intercept` l'écart maximal entre puissance et niveau. On observe que les compromis optimaux ne sont pas obtenus en égalisant les erreurs de première et de seconde espèce. Pour les tailles d'échantillons (1000,5000), au compromis optimal, l'erreur de première espèce est sensiblement plus importante que l'erreur de seconde espèce.

1.7 RÉFÉRENCES

Pour une introduction puissante mais d'un formalisme minimal à la modélisation statistique, on pourra lire *Statistical models* de FREEDMAN [2] et le volume compagnon [3]. Il s'agit de livres écrits par un mathématicien engagé, pour un public cultivé mais large. Une discussion critique et érudite de l'usage de l'inférence statistique dans la vie.

Les ouvrages de LEHMANN et ROMANO [6] LEHMANN et CASELLA [5] constituent toujours une somme sur les problèmes fondamentaux de la statistique.

L'inégalité de Hoeffding est la plus simple des inégalités de concentration. Voir [4] pour une perspective générale sur le phénomène de concentration, [1] pour un exposé tourné vers les applications.

Références

[1] S. BOUCHERON, G. LUGOSI et P. MASSART. **Concentration inequalities**. Oxford University Press, 2013.

[2] D. FREEDMAN. **Statistical models : theory and practice**. Cambridge University Press, 2005, p. x+414.

[3] D. FREEDMAN et al. **Statistical models and causal inference : a dialogue with the social sciences**. Cambridge University Press, 2009.

[4] M. LEDOUX. **The concentration of measure phenomenon**. Providence, RI : American Mathematical Society, 2001.

[5] E. L. Lehmann et G. Casella. **Theory of point estimation**. Second. Springer Texts in Statistics. Springer-Verlag, New York, 1998, p. xxvi+589.

[6] E. L. Lehmann et J. P. Romano. **Testing statistical hypotheses**. Third. Springer Texts in Statistics. Springer, New York, 2005, p. xiv+784.

2.1 MOTIVATIONS

Un cours de statistique classique consacre classiquement une leçon aux vecteurs gaussiens. Il y a plusieurs raisons à cela. La première est didactique. Dans les modèles gaussiens les calculs exacts sont souvent possibles et même faciles. Beaucoup de questions se réduisent à des problèmes d'algèbre linéaire. Cela permet d'illustrer à moindre frais les notions de statistique. Les modèles linéaires gaussiens (voir leçon 3) constituent ainsi la partie facile des modèles linéaires généralisés [4]. Les modèles gaussiens sont les plus simples des modèles exponentiels (voir leçon 4).

La seconde raison tient au théorème central limite et aux différents principes d'invariance qui s'y rattachent. Beaucoup de modèles échantillonnés, lorsque la taille de l'échantillon tend vers l'infini, tendent à ressembler à les modèles de translation gaussienne. L'étude de ces comportements limites fournit une grille de lecture pour les questions de statistiques inférentielles (au moins dans le cas dit paramétrique). Cette perspective, celle de la théorie de la convergence d'expériences de Le Cam, est exposée dans [8]. Pour l'aborder il est nécessaire de connaître quelques résultats de la théorie des vecteurs gaussiens.

2.2 LOI(S) NORMALE(S)

La densité de la *gaussienne centrée réduite (standard)* est notée ϕ, elle vaut

$$\phi(x) = \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\,.$$

La fonction de répartition est notée Φ, la fonction de survie $1-\Phi$ est notée $\overline{\Phi}$.

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)\mathrm{d}u\,.$$

On note $\mathcal{N}(0,1)$ (espérance 0, variance 1) la loi de probabilité définie par la densité ϕ.

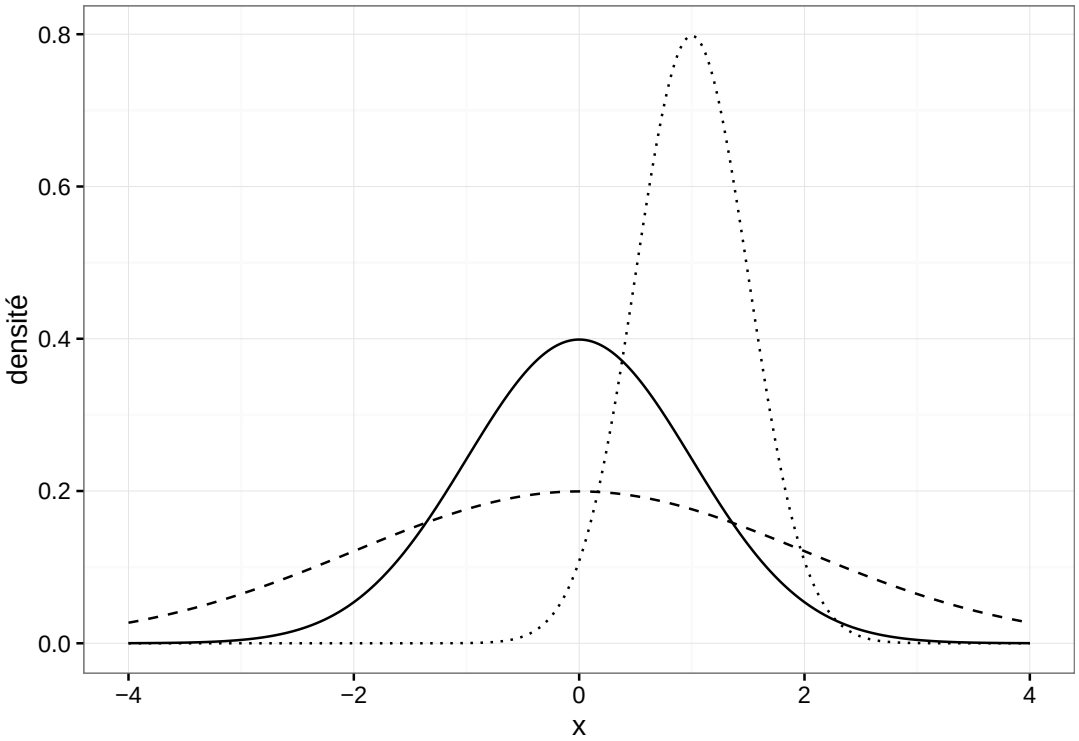

FIG. 2.1 : Densités de $\mathcal{N}(0,1)$, $\mathcal{N}(0,2)$ (tirets), $\mathcal{N}(1,1/4)$ (pointillés)

Toute transformation affine d'une gaussienne centrée réduite suit une loi gaussienne. Si $X \sim \mathcal{N}(0,1)$ alors $\sigma X + \mu \sim \mathcal{N}\left(\mu, \sigma^2\right)$ de densité $\frac{1}{\sigma}\phi\left(\frac{\cdot-\mu}{\sigma}\right)$, de fonction de répartition $\Phi\left(\frac{\cdot-\mu}{\sigma}\right)$.

La loi normale standard est caractérisée par l'identité suivante.

LEMME 2.1 (LEMME DE STEIN) *Soit $X \sim \mathcal{N}(0,1)$, soit g une fonction absolument continue de dérivée g' telle que $\mathbb{E}[|Xg(X)|] < \infty$, alors $g'(X)$ est intégrable et*

$$\mathbb{E}[g'(X)] = \mathbb{E}[Xg(X)]\,.$$

PREUVE. La preuve consiste en une intégration par partie. On observe d'abord qu'en remplacant g par $g - g(0)$ on ne modifie ni la dérivée, ni $\mathbb{E}[Xg(X)]$. On suppsera donc que $g(0) = 0$.

$$\begin{aligned}
\mathbb{E}[Xg(X)] &= \int_{\mathbb{R}} xg(x)\phi(x)\mathrm{d}x \\
&= \int_0^\infty xg(x)\phi(x)\mathrm{d}x + \int_{-\infty}^0 xg(x)\phi(x)\mathrm{d}x \\
&= \int_0^\infty x\int_0^\infty g'(y)\mathbb{I}_{y\le x}\mathrm{d}y\phi(x)\mathrm{d}x - \int_{-\infty}^0 x\int_{-\infty}^0 g'(y)\mathbb{I}_{y\ge x}\mathrm{d}y\phi(x)\mathrm{d}x \\
&= \int_0^\infty g'(y)\int_0^\infty \mathbb{I}_{y\le x}x\phi(x)\mathrm{d}x\mathrm{d}y - \int_{-\infty}^0 g'(y)\int_{-\infty}^0 x\phi(x)\mathbb{I}_{y\ge x}\mathrm{d}x\mathrm{d}y \\
&= \int_0^\infty g'(y)\int_y^\infty x\phi(x)\mathrm{d}x\mathrm{d}y - \int_{-\infty}^0 g'(y)\int_{-\infty}^y x\phi(x)\mathrm{d}x\mathrm{d}y \\
&= \int_0^\infty g'(y)\phi(y)\mathrm{d}y - \int_{-\infty}^0 -g'(y)\phi(y)\mathrm{d}y \\
&= \int_{-\infty}^\infty g'(y)\phi(y)\mathrm{d}y\,.
\end{aligned}$$

La quatrième inégalité est justifiée par le théorème de Tonelli-Fubini. On utilise ensuite $\phi'(x) = -x\phi(x)$. □

La fonction caractéristique est un outil extraordinaire lorsqu'il s'agit d'étudier les lois gaussiennes et plus généralement les vecteurs gaussiens.

PROPOSITION 2.2 *La fonction caractéristique de $\mathcal{N}(\mu,\sigma^2)$ est*

$$\widehat{\Phi}(t) := \mathbb{E}\left[\mathrm{e}^{\imath tX}\right] = \mathrm{e}^{\imath t\mu - \frac{t^2\sigma^2}{2}}\,.$$

PREUVE. Il suffit de vérifier l'égalité dans le cas $\mathcal{N}(0,1)$. Parce que la densité ϕ est une fonction paire,

$$\begin{aligned}
\widehat{\Phi}(t) &= \int_{-\infty}^\infty \mathrm{e}^{\imath tx}\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x \\
&= \int_{-\infty}^\infty \cos(tx)\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x\,.
\end{aligned}$$

On dérive sous le signe somme par rapport à t (pourquoi est-ce justifié ?)

$$\widehat{\Phi}'(t) = \int_{-\infty}^\infty -x\sin(tx)\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x\,.$$

On peut maintenant invoquer l'identité de Stein avec $g(x) = -\sin(tx)$ et $g'(x) = -t\cos(tx)$

$$\begin{aligned}\widehat{\Phi}'(t) &= -t\int_{-\infty}^{\infty}\cos(tx)\phi(x)\mathrm{d}x \\ &= -t\widehat{\Phi}(t)\,.\end{aligned}$$

On a immédiatement $\widehat{\Phi}(0) = 1$, et la résolution de l'équation différentielle conduit à

$$\log\widehat{\Phi}(t) = -\frac{t^2}{2}\,.$$

□

La fonction qui à une loi associe sa fonction caractéristique est injective. Cela permet d'établir une réciproque au lemme 2.1.

LEMME 2.3 (LEMME DE STEIN BIS) *Si X est une variable aléatoire de loi quelconque telle que l'identité*

$$\mathbb{E}[g'(X)] = \mathbb{E}[Xg(X)]$$

est vérifiée pour toute fonction g dérivable telle que g' et $x \mapsto xg(x)$ sont intégrables alors la loi de X est normale standard.

PREUVE. Si on considère la partie réelle $\widehat{F}$ et la partie imaginaire $\widehat{G}$ de la fonction caractéristique de la loi de X, on vérifie à l'aide de l'identité que $\widehat{F}'(t) = -t\widehat{F}(t)$ et $\widehat{G}'(t) = -t\widehat{G}(t)$ avec $\widehat{F}(0) = 1$ et $\widehat{G}(0) = 0$. En intégrant ces deux équations différentielles on obtient $\widehat{F}(t) = \mathrm{e}^{-t^2/2}$ et $\widehat{G}(t) = 0$. On vient de vérifier que la fonction caractéristique de la loi de X est celle de $\mathcal{N}(0,1)$. □

On en déduit immédiatement que la somme de deux gaussiennes indépendantes est une gaussienne.

PROPOSITION 2.4 *Si X et Y sont deux variables aléatoires indépendantes de lois $\mathcal{N}(\mu,\sigma^2)$ et $\mathcal{N}(\mu',\sigma'^2)$ alors $X+Y$ est une variable aléatoire gaussienne de loi $\mathcal{N}(\mu+\mu',\sigma^2+\sigma'^2)$.*

La fonction génératrice des moments d'une gaussienne centrée réduite est donnée par :

$$s \mapsto \mathbb{E}\left[\mathrm{e}^{sX}\right] = \mathrm{e}^{\frac{s^2}{2}}\,.$$

Via l'inégalité de Markov, on peut en déduire des majorants intéressants de la fonction de survie. Un peu de calcul permet d'améliorer ces majorants.

PROPOSITION 2.5 (PROBABILITÉS DE QUEUE) *Pour $x \geq 0$,*

$$\frac{\phi(x)}{x}\left(1-\frac{1}{x^2}\right) \leq \overline{\Phi}(x) \leq \min\left(\mathrm{e}^{-\frac{x^2}{2}}, \frac{\phi(x)}{x}\right)\,. \tag{2.1}$$

PREUVE. Essentiellement une intégration par parties

$$\begin{aligned}\overline{\Phi}(x) &= \int_x^{\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u \\ &= \left[-\frac{1}{\sqrt{2\pi}u}\mathrm{e}^{-\frac{u^2}{2}}\right]_x^{\infty} - \int_x^{\infty}\frac{1}{\sqrt{2\pi}}\frac{1}{u^2}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u.\end{aligned}$$

Comme le second terme est négatif, on a :

$$\overline{\Phi}(x) \leq \left[-\frac{1}{\sqrt{2\pi}u} \mathrm{e}^{-\frac{u^2}{2}} \right]_x^\infty = \frac{\phi(x)}{x}.$$

Ceci nous donne la partie droite de l'encadrement (2.1) (l'autre morceau de cette partie droite provient immédiatement de l'inégalité de Markov et des calculs effectués sur la fonction génératrice des moments).

Pour la partie gauche de (2.1), il faut majorer $\int_x^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{u^2} \mathrm{e}^{-\frac{u^2}{2}} \mathrm{d}u$.

$$\begin{aligned} \int_x^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{u^2} \mathrm{e}^{-\frac{u^2}{2}} \mathrm{d}u &= \left[\frac{-1}{\sqrt{2\pi}} \frac{1}{u^3} \mathrm{e}^{-\frac{u^2}{2}} \right]_x^\infty - \int_x^\infty \frac{1}{\sqrt{2\pi}} \frac{3}{u^4} \mathrm{e}^{-\frac{u^2}{2}} \mathrm{d}u \\ &\leq \frac{1}{\sqrt{2\pi}} \frac{1}{x^3} \mathrm{e}^{-\frac{x^2}{2}}. \end{aligned}$$

□

PROPOSITION 2.6 (MOMENTS) *Pour une variable gaussienne standard,*

$$\mathbb{E}\left[X^k\right] == \begin{cases} 0 & \text{si } k \text{ est impair} \\ \frac{k!}{2^{k/2}(k/2)!} = \frac{\Gamma(k+1)}{2^{k/2}\Gamma(k/2+1)} & \text{si } k \text{ est pair.} \end{cases}$$

PREUVE. Par symétrie, $\mathbb{E}\left[X^k\right] = 0$ pour tout k impair. Pour traiter les moments d'ordre pair, on observe par intégration par parties,

$$\mathbb{E}\left[X^{k+2}\right] = (k+1)\mathbb{E}\left[X^k\right].$$

D'où, par récurrence sur k,

$$\mathbb{E}\left[X^{2k}\right] = \prod_{j=1}^{k} (2j-1) = \frac{(2k)!}{2^k k!}.$$

□

On note au passage que $(2k)!/(2^k k!)$ est égal au nombre de partitions de $\{1, \ldots, 2k\}$ en sous-ensembles à exactement 2 éléments.

L'asymétrie (*skewness*) est nulle, le kurtosis (ratio du moment centré d'ordre 4 sur le carré de la variance) vaut 3 :

$$\mathbb{E}[X^4] = 3 \times \mathbb{E}[X^2]^2.$$

2.3 VECTEURS GAUSSIENS

Un vecteur gaussien est une collection de variables aléatoires gaussiennes qui vérifie une propriété particulière :

DÉFINITION 2.7 Une vecteur aléatoire $(X_1, \ldots, X_n)^t$ est un *vecteur gaussien* si et seulement si pour toute collection de réels $(\lambda_1, \lambda_2, \ldots, \lambda_n)$, la loi de $\sum_{i=1}^n \lambda_i X_i$ est une loi gaussienne.

Une collection de variables gaussiennes n'est pas toujours un vecteur gaussien ($(X, \epsilon X)$ avec $X \sim \mathcal{N}(0,1)$, indépendante de ϵ qui vaut ± 1 avec probabilité $1/2$, n'est pas un vecteur gaussien).

Il existe des vecteurs gaussiens ! Un moyen simple d'obtenir un vecteur gaussien est fourni par la proposition suivante (vérifiée par un argument de fonction caractéristique).

PROPOSITION 2.8 *Si $X_1, \ldots, X_n$ est une suite de variables gaussiennes indépendantes, alors $(X_1, \ldots, X_n)^t$ est un vecteur gaussien.*

Dans la suite, on appellera *vecteur gaussien stantard*, un vecteur aléatoire dont les coordonnées sont indépendantes et dont chaque coordonnée est distribuée selon $\mathcal{N}(0,1)$. Avant de voir comment fabriquer des vecteurs gaussiens, nous vérifions qu'ils sont caractérisés par leurs espérances et leurs covariances.

On appelle *covariance* du vecteur aléatoire $X = (X_1, \ldots, X_n)^t$ la matrice K de dimensions $n \times n$ dont les coefficients sont donnés par :

$$K[i,j] = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

On peut supposer sans perdre en généralité que le vecteur aléatoire X est centré. On vérifie que pour tout $\lambda = (\lambda_1, \ldots, \lambda_n)^t$:

$$\mathrm{var}(\langle \lambda, X \rangle) = \lambda^t K \lambda = \mathrm{trace}(K \lambda \lambda^t)$$

(ce n'est pas un résultat gaussien).

En effet

$$\begin{aligned} \mathrm{var}(\langle \lambda, X \rangle) &= \mathbb{E}\left[\left(\sum_{i=1}^n \lambda_i X_i \right)^2 \right] \\ &= \sum_{i,j=1}^n \mathbb{E}\left[\lambda_i \lambda_j X_i X_j \right] \\ &= \sum_{i,j=1}^n \lambda_i \lambda_j K[i,j] \\ &= \lambda^t K \lambda \,. \end{aligned}$$

On observe que la fonction caractéristique d'un vecteur gaussien X d'espérance μ et de covariance K vaut en λ :

$$\mathbb{E}\mathrm{e}^{\imath \langle \lambda, X \rangle} = \mathrm{e}^{\imath \langle \lambda, \mu \rangle - \frac{\lambda^t K \lambda}{2}} \,.$$

Une transformation linéaire d'un vecteur gaussien est un vecteur gaussien.

PROPOSITION 2.9 *Si $Y = (Y_1, \ldots, Y_n)^t$ est un vecteur gaussien de covariance K et A une matrice $p \times n$, alors AY est un vecteur gaussien de matrice de covariance*

$$AKA^t.$$

PREUVE. Sans perdre en généralité, on suppose Y centré.

Pour tout $\lambda \in \mathbb{R}^p$, $\langle \lambda, AY \rangle = \langle A^t \lambda, Y \rangle$, donc AY est gaussien et sa variance est donnée par

$$\lambda^t A K A^t \lambda \,,$$

on en déduit la matrice de covariance de AY. □

Pour fabriquer des vecteurs gaussiens de matrice de covariance pas nécessairement diagonale, nous allons utiliser quelques outils d'analyse matricielle.

DÉFINITION 2.10 (MATRICES DÉFINIES POSITIVES) Une matrice symétrique M de dimensions $k \times k$ est définie positive (respectivement semi-définie positive) si et seulement si, pour tout vecteur v non nul de $\mathbb{R}^k$,

$$v^t M v > 0 \qquad (\text{resp.} \qquad v^t M v \geq 0) \,.$$

On notera $\mathrm{DP}(k)$ et $\mathrm{SDP}(k)$ les cônes des matrices définies positives et semi-définies positives.

PROPOSITION 2.11 *Si K est la matrice de covariance d'un vecteur aléatoire, K est symétrique et semi-définie positive.*

PREUVE. Si X est un vecteur aléatoire à valeur dans $\mathbb{R}^k$, de covariance K, pour tout vecteur $\lambda \in \mathbb{R}^n$,

$$\lambda^t K \lambda = \sum_{i,j \leq k} K_{i,j} \lambda_i \lambda_j = \mathrm{cov}(\langle \lambda, X \rangle, \langle \lambda, X \rangle)$$

soit $\lambda^t K \lambda = \mathrm{var}(\langle \lambda, X \rangle)$. Cette variance est toujours positive ou nulle. □

Cette observation va nous fournir la clé de la construction de vecteurs gaussiens de covariance arbitraire.

PROPOSITION 2.12 *Si A est une matrice symétrique semi-définie positive alors il existe (au moins) une matrice B telle que $A = B^t B$.*

Nous ne vérifions pas immédiatement cette proposition. On peut l'établir à partir du théorème de décomposition spectrale des matrices symétriques. On peut aussi l'établir par une démarche constructive simple : une matrice définie positive K admet une *décomposition de Cholesky*, autrement dit, il existe une matrice triangulaire inférieure L telle que $K = L \times L^t$.

On obtient le corollaire suivant en invoquant la formule de calcul des densités des lois image.

PROPOSITION 2.13 *Si A est une matrice symétrique définie positive ($A \in \mathrm{DP}(n)$), alors la loi du vecteur gaussien centré de covariance A admet une densité par rapport à la mesure de Lebesgue :*

$$\frac{1}{(2\pi)^{n/2} \det(A)^{1/2}} \exp\left(-\frac{x^t A^{-1} x}{2}\right) .$$

PREUVE. La formule de densité est trivialement correcte pour les vecteurs gaussiens standards. Pour le cas général, il suffit d'invoquer la formule de calcul de la densité image et d'appliquer au vecteur gaussien standard la transformation linéaire définie par le facteur de Cholesky de A. Le déterminant de ce facteur est la racine carrée du déterminant de A. □

DÉFINITION 2.14 ESPACE GAUSSIEN Si $(X_1, \ldots, X_n)^t$ est un vecteur gaussien centré (de matrice de covariance K), l'ensemble des variables aléatoires (gaussiennes) de la forme $\sum_{i=1}^n \lambda_i X_i$ est appelé espace gaussien (engendré par $(X_1, \ldots, X_n)^t$).

L'espace gaussien est un espace vectoriel (sur $\mathbb{R}$). Si on note $(\Omega, \mathcal{F}, P)$ l'espace probabilisé sur lequel vit le vecteur gaussien $(X_1, \ldots, X_n)^t$, l'espace Gaussien est un sous-espace de $L^2_{\mathbb{R}}(\Omega, \mathcal{F}, P)$. Il hérite du produit intérieur associé à $L^2_{\mathbb{R}}(\Omega, \mathcal{F}, P)$.

Ce produit scalaire est complètement défini par la matrice de covariance K du vecteur gaussien. Si K est définie positive, l'espace est muni d'un produit scalaire :

$$\begin{aligned}
\left\langle \sum_{i=1}^n \lambda_i X_i, \sum_{i=1}^n \lambda'_i X_i \right\rangle &\equiv \mathbb{E}_P \left[\left(\sum_{i=1}^n \lambda_i X_i \right) \left(\sum_{i=1}^n \lambda'_i X_i \right) \right] \\
&= \sum_{i,i'=1}^n \lambda_i \lambda'_{i'} K[i, i'] \\
&= (\lambda_1, \ldots, \lambda_n) K \begin{pmatrix} \lambda'_1 \\ \vdots \\ \lambda'_n \end{pmatrix} \\
&= \mathrm{trace} \left(K \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \begin{pmatrix} \lambda'_1 & \ldots & \lambda'_n \end{pmatrix} \right) .
\end{aligned}$$

Sur cette expression, on lit la relation entre la matrice de covariance du vecteur utilisé pour définir l'espace gaussien et le produit scalaire (intrinsèque) défini sur cet espace gaussien.

Les espaces gaussiens possèdent des propriétés remarquables.

PROPOSITION 2.15 *Deux variables aléatoires centrées Z et Y, éléments d'un espace gaussien, sont indépendantes si et seulement si elles sont orthogonales (ou décorrélées), c'est-à-dire si et seulement si*

$$\mathrm{Cov}_P[Y,Z] = \mathbb{E}_P[YZ] = 0.$$

Sans perdre en généralité, on suppose la matrice de covariance K définie positive.

PREUVE. L'indépendance implique toujours l'orthogonalité.

Sans perdre en généralité on suppose que l'espace gaussien est engendré par une collection de gaussiennes indépendantes de variance 1, $X_1, \ldots, X_n$ (si ce n'est pas d'emblée le cas, on peut se ramener à cette situation en utilisant la décomposition de Cholesky de la matrice de covariance). On écrit $Z = \sum_{i=1}^n \lambda_i X_i$ et $Y = \sum_{i=1}^n \lambda'_i X_i$.

Si Z et Y sont orthogonales (ou décorrélées)

$$\mathbb{E}[ZY] = \sum_{i=1}^n \lambda_i \lambda'_i = 0.$$

Pour montrer que Z et Y sont indépendantes, il suffit de montrer que pour tous μ et μ' réels

$$\mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\mathrm{e}^{\imath\mu' Y}\right] = \mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\right] \times \mathbb{E}\left[\mathrm{e}^{\imath\mu' Y}\right].$$

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\mathrm{e}^{\imath\mu' Y}\right] &= \mathbb{E}\left[\mathrm{e}^{\imath\mu \sum_i \lambda_i X_i}\mathrm{e}^{\imath\mu' \sum_i \lambda'_i X_i}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^n \mathrm{e}^{\imath(\mu\lambda_i + \mu'\lambda'_i)X_i}\right] \\
&\quad \text{les } X_i \text{ sont indépendantes}\ldots \\
&= \prod_{i=1}^n \mathbb{E}\left[\mathrm{e}^{\imath(\mu\lambda_i + \mu'\lambda'_i)X_i}\right] \\
&= \prod_{i=1}^n \mathrm{e}^{-(\mu\lambda_i + \mu'\lambda'_i)^2/2} \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^n \mu^2\lambda_i^2 + 2\mu\mu'\lambda_i\lambda'_i + \mu'^2\lambda_i'^2\right) \\
&\quad \text{orthogonalité} \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^n \mu^2\lambda_i^2 + \mu'^2\lambda_i'^2\right) \\
&\quad \ldots \\
&= \mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\right] \times \mathbb{E}\left[\mathrm{e}^{\imath\mu' Y}\right].
\end{aligned}$$

□

La proposition suivante est une conséquence directe de la précédente.

COROLLAIRE 2.16 *Si E et E' sont deux sous-espaces vectoriels de l'espace gaussien engendré par le vecteur de gaussiennes indépendantes $X_1, \ldots, X_n$, les variables aléatoires (gaussiennes) appartenant à l'espace E et les variables aléatoires (gaussiennes) appartenant à l'espace E' sont indépendantes si et seulement ces deux sous espaces sont orthogonaux.*

2.4 CONVERGENCE DE VECTEURS GAUSSIENS

On rappelle un critère de convergence très utile.

THÉORÈME 2.17 (CRITÈRE DE LÉVY-THÉORÈME DE CONTINUITÉ) *Une suite de lois de probabilités $(P_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^k$ converge faiblement vers une probabilité P si et seulement si pour tout $\vec{s}\in\mathbb{R}^k$:*

$$\mathbb{E}_{P_n}\left[\mathrm{e}^{\imath\langle\vec{s},\vec{X}\rangle}\right]\to\mathbb{E}_P\left[\mathrm{e}^{\imath\langle\vec{s},\vec{X}\rangle}\right].$$

REMARQUE 2.18 Pour chaque $\vec{s}\in\mathbb{R}^k$, les fonctions $\vec{x}\mapsto\cos(\langle\vec{s},\vec{x}\rangle)$ et $\vec{x}\mapsto\sin(\langle\vec{s},\vec{x}\rangle)$ sont des fonctions non seulement continues et bornées, mais aussi infiniment dérivables. Il est remarquable que la convergence en loi puisse être vérifiée sur ce seul ensemble de fonctions.

THÉORÈME 2.19 (CRITÈRE DE LÉVY-THÉORÈME DE CONTINUITÉ-BIS) *Une suite de lois de probabilités $(P_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^k$ converge faiblement vers une loi de probabilité si et seulement si il existe une fonction f définie sur $\mathbb{R}^k$, continue en $\vec{0}$, telle que pour tout $\vec{s}\in\mathbb{R}^k$:*

$$\mathbb{E}_{P_n}\left[\mathrm{e}^{\imath\langle\vec{s},\vec{X}\rangle}\right]\to f(\vec{s}).$$

La fonction f est alors la fonction caractéristique d'une loi P (qu'elle définit).

La condition de continuité en 0 de la fonction f est indispensable. Toute fonction caractéristique de loi de probabilité est continue en 0. La continuité en 0 garantit la *tension de la suite de lois.*

PROPOSITION 2.20 *Une suite de gausiennes multidimensionnelles (vecteurs gaussiens) (X_n) est définie par une suite $(\vec{\mu}_n)$ de $\mathbb{R}^k$ et une suite de matrices semi-définies positives (K_n) de dimensions $k\times k$. Si*

$$\begin{aligned}\lim_n\vec{\mu}_n &= \mu\\ \lim_n K_n &= K\end{aligned}$$

où K est une matrice $k\times k$ alors K est une matrice semi-définie positive. La suite $(X_n)_n$ converge en loi vers une gausienne d'espérance $\vec{\mu}$ et de covariance K (si K est nulle, on interprète la limite comme la loi concentrée en μ).

2.5 CONDITIONNEMENT GAUSSIEN

Soit $(X_1,\ldots,X_n)^t$ un vecteur gaussien de loi $\mathcal{N}(\mu,K)$ où K est non-singulière. La matrice de covariance K est partitionnée en blocs

$$K=\left[\begin{array}{c:c}A & B^t\\ \hdashline B & W\end{array}\right]$$

où A est de dimension $k\times k$, $1\le k<n$. Notons au passage que A et W sont non-singulières.

On s'intéresse à l'espérance conditionnelle de $(X_1,\ldots,X_k)^t$ sachant $(X_{k+1},\ldots,X_n)$ et à la loi conditionnelle de $(X_1,\ldots,X_k)^t$ sachant $(X_{k+1},\ldots,X_n)$.

Pour décrire ces deux entités, on utilisera la notion de *complément de Schur* :

$$W - BA^{-1}B^t$$

est le complément de Schur de A dans K.

Nous vérifierons que si K est définie positive, alors le complément de Schur de A dans K est une matrice définie positive.

Nous noterons dans l'énoncé des théorèmes $A^{-1/2}$ la matrice triangulaire inférieure issue de la factorisation de Cholesky de A^{-1} : $A^{-1} = A^{-1/2} \times (A^{-1/2})^t$.

THÉORÈME 2.21 *L'espérance conditionnelle de $(X_{k+1}, \ldots, X_n)^t$ sachant $(X_1, \ldots, X_k)^t$ est une transformation affine de $(X_1, \ldots, X_k)^t$:*

$$\mathbb{E}\left[\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} \mid \begin{matrix} X_1 \\ \vdots \\ X_k \end{matrix}\right] = \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix} + \left(BA^{-1}\right) \times \left(\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}\right) .$$

THÉORÈME 2.22 *La loi conditionnelle de $(X_{k+1}, \ldots, X_n)^t$ sachant $(X_1, \ldots, X_k)$ est une loi gaussienne dont l'espérance est l'espérance conditionnelle et dont la variance est le complément de Schur de la covariance de $(X_1, \ldots, X_k)^t$ dans la matrice de covariance de $(X_1, \ldots, X_n)^t$.*

REMARQUE 2.23 Nous allons étudier d'emblée la densité conditionnelle, établir qu'elle est gaussienne avec un minimum de calcul. L'espérance conditionnelle se calculera comme l'espérance sous la loi conditionnelle. Pour caractériser la densité conditionnelle, nous allons utiliser un argument de représentation en loi (un vecteur gaussien quelconque est distribué comme l'image d'une vecteur gaussien standard par une transformation affine) et un résultat d'analyse matricielle qui est au coeur de la factorisation de Cholesky des matrices symétriques définies positives.

Nous allons invoquer une identité matricielle, qui est à la base de la factorisation de Cholesky, vérifier que $(X_1, \ldots, X_n)^t$ est distribué comme la multiplication d'un vecteur gaussien standard par une matrice triangulaire par blocs, et utiliser des propriétés des lois conditionnelles pour établir à la fois les deux résultats.

PROPOSITION 2.24 *Soit une matrice symétrique définie positive de dimensions $n \times n$*

$$K = \left[\begin{array}{c:c} A & B^t \\ \hdashline B & W \end{array}\right]$$

où A est de dimension $k \times k$, $1 \leq k < n$.

Alors, le complément de Schur de A dans K

$$W - BA^{-1}B^t$$

est défini positif. Les sous-matrices A et $W - BA^{-1}B^t$ admettent chacune une décomposition de Cholesky $A = L_1 L_1^t$, $W - BA^{-1}B^t = L_2 L_2^t$ où L_1, L_2 sont triangulaires inférieures, et la factorisation de Cholesky de K s'écrit :

$$K = \left[\begin{array}{c:c} L_1 & 0 \\ \hdashline B(L_1^t)^{-1} & L_2 \end{array}\right] \times \left[\begin{array}{c:c} L_1^t & L_1^{-1}B^t \\ \hdashline 0 & L_2^t \end{array}\right] .$$

On trouvera une preuve de cette proposition dans l'appendice A.1.

PREUVE.[Théorème 2.22] Sans perdre en généralité, on se contente de vérifier le résultat pour des vecteurs gaussiens centrés. On peut utiliser la factorisation de Cholesky de la matrice de covariance pour établir que

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \left[\begin{array}{c:c} L_1 & 0 \\ \hdashline B(L_1^t)^{-1} & L_2 \end{array}\right] \times \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

où $(Y_1, \ldots, Y_n)^t$ est un vecteur gaussien standard.

Dans la suite, on suppose que $(X_1, \ldots, X_n)^t$ et $(Y_1, \ldots, Y_n)^t$ vivent sur le même espace probabilisé. Comme L_1 est inversible, les tribus engendrées par $(X_1, \ldots, X_k)^t$ et $(Y_1, \ldots, Y_k)^t$ sont égales. On note cette tribu $\mathcal{G}$. L'espérance conditionnelle et la loi conditionnelle sachant ces deux tribus coïncident aussi.

$$\begin{aligned} \mathbb{E}\left[\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} \mid \mathcal{G}\right] &= \mathbb{E}\left[B(L_1^t)^{-1}\begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} \mid \mathcal{G}\right] + \mathbb{E}\left[L_2\begin{pmatrix} Y_{k+1} \\ \vdots \\ Y_n \end{pmatrix} \mid \mathcal{G}\right] \\ &= B(L_1^t)^{-1}L_1^{-1}\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} = BA^{-1}\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}, \end{aligned}$$

car $(Y_{k+1}, \ldots, Y_n)^t$ est centré et indépendant de $\mathcal{G}$.

On note au passage que le vecteur des résidus

$$\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} - \mathbb{E}\left[\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} \mid \mathcal{G}\right] = L_2\begin{pmatrix} Y_{k+1} \\ \vdots \\ Y_n \end{pmatrix}$$

est indépendant de $\mathcal{G}$. C'est une particularité gaussienne. Dans le cas de variables de carré intégrable quelconques, on peut seulement affirmer que les résidus sont orthogonaux à toute variable $\mathcal{G}$-mesurable.

La loi conditionnelle de $(X_{k+1}, \ldots, X_n)^t$ sachant $(X_1, \ldots, X_k)^t$ est la même que la loi conditionnelle de

$$B(L_1^t)^{-1} \times \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} + L_2 \times \begin{pmatrix} Y_{k+1} \\ \vdots \\ Y_n \end{pmatrix}$$

sachant $(Y_1, \ldots, Y_k)^t$. Comme $(Y_1, \ldots, Y_k)^t = L_1^{-1}(X_1, \ldots, X_k)^t$, la loi conditionnelle recherchée est une loi gaussienne d'espérance

$$BA^{-1} \times \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$$

(l'espérance conditionnelle) et de variance $L_2 \times L_2^t = W - BA^{-1}B^t$. □

EXEMPLE 2.25 Si $(X, Y)^t$ est un vecteur gaussien centré avec $\mathrm{var}(X) = \sigma_x^2$, $\mathrm{var}(Y) = \sigma_y^2$ et $\mathrm{cov}(X, Y) = \rho\sigma_x\sigma_y$, la loi conditionnelle de Y sachant X est

$$\mathcal{N}\left(\rho\sigma_y/\sigma_x X, \sigma_y^2(1-\rho^2)\right).$$

La quantité ρ est appelée *coefficient de corrélation linéaire* entre X et Y. On vérifie (à l'aide de l'inégalité de Cauchy-Schwarz) que $\rho \in [-1, 1]$.

REMARQUE 2.26 On aborde habituellement ces deux résultats dans l'ordre où ils sont énoncés. On caractérise l'espérance conditionnelle en adoptant le point de vue L^2 : l'espérance conditionnelle du vecteur aléatoire Y sachant X est définie comme le meilleur prédicteur du vecteur Y X-mesurable au sens de l'erreur quadratique (le vecteur aléatoire Z, X-mesurable qui minimise $\mathbb{E}\left[\|Y - Z\|^2\right]$).

Pour caractériser cette espérance conditionnelle, on calcule d'abord le meilleur prédicteur *affine* de $(X_{k+1}, \ldots, X_n)^t$ à partir de $(X_1, \ldots, X_k)^t$, qui s'avère être

$$\begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix} + \left(BA^{-1}\right) \times \left(\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}\right),$$

(si les vecteurs gaussiens sont centrés, le calcul revient à déterminer la matrice P de dimensions $(n-k)\times k$ qui minimise trace$(PAP^t - 2BP^t)$). Il s'agit d'un vecteur gaussien, et on vérifie que le vecteur gaussien dit des *résidus*

$$\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} - \left\{ \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix} + \left(BA^{-1}\right) \times \left(\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \right) \right\}$$

est décorrélé, donc indépendant du prédicteur linéaire. Ceci suffit à garantir que le prédicteur linéaire est la projection orthogonale de $(X_{k+1},\dots,X_n)^t$ sur le sous-espace fermé des fonctions $(X_1,\dots,X_k)^t$-mesurables de carré intégrable. Et cela démontre que le prédicteur linéaire est l'espérance conditionnelle.

Cette démarche s'appuie sur l'habitude de manipuler l'espérance conditionnelle dans sa définition abstraite.

Le conditionnement décrit ici est un cas particulier de conditionnement linéaire. On peut se donner $X \sim \mathcal{N}(0,K)$ (on suppose le vecteur centré pour alléger calculs et notations, la translation ne change rien au rôle des différentes tribus), avec K une matrice définie positive de dimensions $n\times n$, et une transformation linéaire définie par une matrice H de dimensions $m\times n$ qu'on supposera de rang m. On note $Y = HX$. En considérant le vecteur gaussien $[X^t : Y^t]$ de matrice de covariance

$$\left[\begin{array}{c:c} K & KH^t \\ \hdashline HK & HKH^t \end{array}\right]$$

et en adaptant le calcul précédent (la matrice de covariance n'est plus définie positive), on peut établir que la loi conditionnelle de X sachant Y est gaussienne d'espérance

$$KH^T(HKH^T)^{-1}$$

et de variance

$$K - KH^t(HKH^T)^{-1}HK\,.$$

La linéarité de l'espérance conditionnelle est une propriété des vecteurs gaussiens et du contionnement linéaire. Si on conditionne par $\|X\|_2$, la loi conditionnelle n'est pas gaussienne.

2.6 COMPLÉMENTS SUR LES LOIS GAMMA

L'étude de la norme euclidienne des vecteurs gaussiens nous conduira à introduire des lois particulières, les lois du χ^2 qui sont des cas particuliers de lois Gamma.

DÉFINITION 2.27 Une LOI GAMMA DE PARAMÈTRE (p,λ) avec $\lambda \in \mathbb{R}_+$ et $p \in \mathbb{R}_+$, est une loi sur $\mathbb{R}_+$ de densité

$$g_{p,\lambda}(x) \equiv \frac{\lambda^p}{\Gamma(p)} \mathbf{1}_{x\geq 0} x^{p-1} e^{-\lambda x}$$

où $\Gamma(p) \equiv \int_0^\infty t^{p-1}e^{-t}\mathrm{d}t$.

Le paramètre p est appelé paramètre de *forme*, λ est appelé paramètre d'*intensité*, $1/\lambda$ est appelé paramètre d'*échelle*.

Si $X \sim \text{Gamma}(p,1)$ alors $\sigma X \sim \text{Gamma}(p, 1/\sigma)$ pour $\sigma > 0$.

REMARQUE 2.28 La fonction $\Gamma()$ interpole la factorielle. On vérifie que pour tout réel p, $\Gamma(p+1) = p\Gamma(p)$. Si p est entier, $\Gamma(p+1) = p!$

Si X est distribuée selon une loi Gamma de paramètres $(p,1)$, alors λX est distribuée selon une loi Gamma de paramètres $(p, 1/\lambda)$.

PROPOSITION 2.29 *L'espérance d'une loi Gamma de paramètres (p,λ) est égale à p/λ. La variance d'une loi Gamma de paramètres (p,λ) est égale à p/λ^2 .*

Les lois Gamma possèdent une propriété de clôture intéressante.

PROPOSITION 2.30 *Si X et Y sont deux variables aléatoires indépendantes distribuées respectivement selon des lois Gamma de paramètres (p,λ) et (q,λ) alors $X+Y$ est distribuée selon une loi Gamma de paramètre $(p+q,\lambda)$.*

PREUVE. La densité de la loi de $X+Y$ s'écrit comme le produit de convolution des densités $g_{p,\lambda}$ et $g_{q,\lambda}$ qui est une densité de loi de probabilité.

$$\begin{aligned}
g_{p,\lambda} * g_{q,\lambda}(x) &= \int_{\mathbb{R}} g_{p,\lambda}(z) g_{q,\lambda}(x-z)\mathrm{d}z \\
&= \int_0^x g_{p,\lambda}(z) g_{q,\lambda}(x-z)\mathrm{d}z \\
&= \int_0^x \frac{\lambda^p}{\Gamma(p)} z^{p-1}\mathrm{e}^{-\lambda z}\frac{\lambda^q}{\Gamma(q)}(x-z)^{q-1}\mathrm{e}^{-\lambda(x-z)}\mathrm{d}z \\
&= \frac{\lambda^{p+q}}{\Gamma(p)\Gamma(q)}\mathrm{e}^{-\lambda x}\int_0^x z^{p-1}(x-z)^{q-1}\mathrm{d}z \\
&\quad \text{changement de variable } z = xu \\
&= \frac{\lambda^{p+q}}{\Gamma(p)\Gamma(q)}\mathrm{e}^{-\lambda x}x^{p+q-1}\int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u \\
&= g_{p+q,\lambda}(x)\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u\,.
\end{aligned}$$

On a établi au passage que

$$B(p,q) := \int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u$$

vérifie $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$. □

Les lois Gamma de paramètres $(k/2, 1/2)$ pour $k \in \mathbb{N}$ apparaissent assez souvent pour mériter un nom particulier : ce sont les *lois du χ^2 à k degrés de liberté.*

DÉFINITION 2.31 [LOI DU CHI-DEUX] La loi du χ^2 à k degrés de liberté (notée χ^2_k) a pour densité sur $[0,\infty)$,

$$\frac{x^{\frac{1}{2}(k-2)}e^{-\frac{x}{2}}}{2^{k/2}\Gamma(k/2)}.$$

PROPOSITION 2.32 *La somme de k carrés de variables gaussiennes indépendantes standard suit une loi du χ^2 à k degrés de liberté.*

PREUVE. D'après la proposition 2.30. Il suffit de démontrer la proposition pour $k=1$. Si X est une gaussienne centrée réduite, la densité de $|X|$ est $2\mathrm{e}^{-x^2/2}/\sqrt{2\pi}$. La transformation $x \mapsto x^2$ est bijective et différentiable sur $\mathbb{R}^+$. La dérivée de l'inverse $u \mapsto \sqrt{u}$ vaut $1/(2\sqrt{u})$. Et la densité de X^2 vaut donc

$$\frac{1}{\sqrt{\pi}}\left(\frac{1}{2}\right)^{1/2} u^{-1/2}\mathrm{e}^{-u/2}.$$

□

2.7 NORMES DE VECTEURS GAUSSIENS CENTRÉS

La distribution du carré de la norme euclidienne d'un vecteur gaussien centré ne dépend que du spectre de sa matrice de covariance.

THÉORÈME 2.33 *Si $X := (X_1, X_2, \dots, X_n)^t$ est un vecteur gaussien centré de matrice de covariance $A = LL^t$ (avec L triangulaire inférieure), si M est une matrice symétrique semi-définie positive sur $\mathbb{R}^n$, alors la variable aléatoire $X^t M X$ est distribuée comme $\sum_{i=1}^n \lambda_i Z_i$ où $(\lambda_i)_{i\in\{1,\dots,n\}}$ désigne la suite des valeurs propres de la matrice $L^t \times M \times L$ et où les Z_i sont des variables indépendantes distribuées selon une loi χ_1^2.*

C'est un corollaire d'une propriété importante des vecteurs gaussiens standards : leur loi est invariante par transformation orthogonale (une matrice O est orthogonale ssi $OO^t = \mathrm{Id}$).

PREUVE. La matrice A admet une factorisation $A = LL^t$ (par exemple la factorisation de Cholesky), et X est distribué comme LY où Y est distribué selon une gaussienne standard $\mathcal{N}(0, I_n)$. La forme quadratique $X^t M X$ est donc distribuée comme $Y^t L^t M L Y$. Il existe une transformation orthogonale O telle que $L^t M L = O^t \operatorname{diag}(\lambda_i) O$. Le vecteur OY est distribué selon $\mathcal{N}(0, I_n)$ car les lois gaussiennes standards sont invariantes par transformation orthogonale. □

2.8 NORMES DE VECTEURS GAUSSIENS DÉCENTRÉS

La loi du carré de la norme d'un vecteur gaussien général de matrice de covariance $\sigma^2 \mathrm{Id}$, ne dépend que de la norme de l'espérance (et pas de sa direction) et de σ^2. De plus, cette loi se compare simplement à celle de la loi du carré de la norme d'un vecteur gaussien centrée de même covariance.

DÉFINITION 2.34 (ORDRE STOCHASTIQUE) Dans un espace probabilisé muni d'une loi $\mathbb{P}$, une variable aléatoire réelle X est stochastiquement inférieure à une variable aléatoire réelle Y, si

$$\mathbb{P}\{X \le Y\} = 1\,.$$

On dit que la loi de Y domine stochastiquement celle de X.

Si X est stochastiquement inférieure à Y et si on note F et G les fonctions de répartition des lois de X et Y, alors pour tout $x \in \mathbb{R}$, $F(x) \ge G(x)$. Les fonctions quantiles $F^{\leftarrow}, G^{\leftarrow}$ vérifient $F^{\leftarrow}(p) \le G^{\leftarrow}(p)$ pour $p \in (0,1)$.

Réciproquement.

PROPOSITION 2.35 *Si F et G sont deux fonctions de répartition qui vérifient pour tout $x \in \mathbb{R}$ $F(x) \ge G(x)$ alors il existe un espace probabilisé muni d'une loi $\mathbb{P}$ où deux variables aléatoires X et Y de fonctions de répartition F et G vérifient :*

$$\mathbb{P}\{X \le Y\} = 1\,.$$

La proposition se démontre par un argument dit de « couplage quantile ».

PREUVE. Il suffit de munir $[0,1]$ de la loi uniforme, de noter U une variable aléatoire de densité uniforme sur $[0,1]$ et de définir $X = F^{\leftarrow}(U)$ et $Y = G^{\leftarrow}(U)$, on a alors X (resp. Y) de fonction de répartition F (resp. G) et

$$\mathbb{P}\{X \le Y\} = \mathbb{P}\{F^{\leftarrow}(U) \le G^{\leftarrow}(U)\} = 1\,.$$

□

THÉORÈME 2.36 *Si* $X \sim \mathcal{N}\left(0, \sigma^2 \operatorname{Id}\right)$ *et* $Y \sim \mathcal{N}\left(\theta, \sigma^2 \operatorname{Id}\right)$ *avec* $\theta \in \mathbb{R}^d$ *alors*

$$\|Y\|^2 \sim \left((Z_1 + \|\theta\|_2)^2 + \sum_{i=1}^{d} Z_i^2 \right)$$

où les Z_i *sont i.i.d.* $\mathcal{N}(0, \sigma^2)$.

Par ailleurs, pour tout $x \geq 0$,

$$\mathbb{P}\left\{\|Y\| \leq x\right\} \leq \mathbb{P}\left\{\|X\| \leq x\right\} .$$

On dit que la loi $\|Y\|^2/\sigma^2$ *(*χ^2 *décentrée de paramètre de décentrage* $\|\theta\|_2/\sigma$*)* domine stochastiquement *celle de* $\|X\|^2/\sigma^2$ *(*χ^2 *centrée).*

PREUVE. Le vecteur gaussien Y est distribué comme $\theta + X$. Il existe une transformation orthogonale O telle que

$$O\theta = \begin{pmatrix} \|\theta\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} .$$

Les normes de OY et de OX sont distribuées comme celles de X et Y.

Le carré de la norme de Y est distribué comme celui de la norme de OY, c'est-à-dire comme $(Z_1 + \|\theta\|_2)^2 + \sum_{i=2}^{d} Z_i^2$. Ceci prouve la première partie de l'énoncé.

Pour établir la seconde partie, il suffit d'établir que pour tout $x \geq 0$,

$$\mathbb{P}\left\{(Z_1 + \|\theta\|_2)^2 \leq x\right\} \leq \mathbb{P}\left\{X_1^2 \leq x\right\} ,$$

soit

$$\mathbb{P}\left\{|Z_1 + \|\theta\|_2| \leq \sqrt{x}\right\} \leq \mathbb{P}\left\{|X_1| \leq \sqrt{x}\right\} ,$$

ou encore

$$\Phi(\sqrt{x} - \|\theta\|_2) - \Phi(-\sqrt{x} - \|\theta\|_2) \leq \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) .$$

Pour $y > 0$, la fonction sur $[0, \infty)$, définie par $a \mapsto \Phi(y - a) - \Phi(-y - a)$ est décroissante en a : sa dérivée par rapport à a vaut $-\phi(y-a) + \phi(-y-a) = \phi(y+a) - \phi(y-a) \leq 0$. Ceci permet de conclure. □

REMARQUE 2.37 La dernière étape de la preuve du théorème peut se lire comme

$$\mathbb{P}\left\{X \in \theta + C\right\} \leq \mathbb{P}\left\{X \in C\right\}$$

où $X \sim \mathcal{N}(0, \mathrm{Id}_1)$, $\theta \in \mathbb{R}$ et $C = [-\sqrt{x}, \sqrt{x}]$. Cette inégalité reste vraie en dimension $d \geq 1$, si on impose que C est compact, convexe, symétrique. Ce résultat (délicat) s'appelle le lemme d'Anderson. La relation de domination stochastique décrite dans le théorème est aussi un cas particulier du lemme d'Anderson.

2.9 THÉORÈME DE COCHRAN ET CONSÉQUENCES

THÉORÈME 2.38 (THÉORÈME DE COCHRAN) *Soit* $X \sim \mathcal{N}(0, I_n)$ *et* $\mathbb{R}^n = \oplus_{j=1}^{k} E_j$ *où les* E_j *sont des sous-espaces deux à deux orthogonaux de* $\mathbb{R}^n$. *On note* π_{E_j} *la projection orthogonale sur* E_j.

La famille de vecteurs gaussiens $\left(\pi_{E_j} X\right)_{j \leq k}$ *est une famille indépendante et pour chaque* j

$$\|\pi_{E_j} X\|_2^2 \sim \chi^2_{dim(E_j)} .$$

PREUVE. La matrice de covariance de $\pi_{E_j} X$ est $\pi_{E_j}\pi_{E_j}^t = \pi_{E_j}$. Les valeurs propres de π_{E_j} sont 1 avec multiplicité $\dim(E_j)$ et 0. L'assertion sur la loi de $\|\pi_{E_j} X\|_2^2$ est une conséquence immédiate des propositions 2.32 et 2.33.

Pour établir l'indépendance, considérons $\mathcal{I}, \mathcal{J} \subset \{1, \dots, k\}$ avec $\mathcal{I} \cap \mathcal{J} = \emptyset$. Il suffit de vérifier que pour tous $(\alpha)_{j\in\mathcal{I}}, (\beta_j)_{j\in\mathcal{J}}$, la fonction caractéristique de

$$\left(\sum_{j\in\mathcal{I}} \langle \alpha_j, \pi_{E_j} X\rangle, \sum_{j\in\mathcal{J}} \langle \beta_j, \pi_{E_j} X\rangle \right)$$

se factorise. Il suffit de vérifier que ces deux gaussiennes sont orthogonales.

$$\mathbb{E}\left[\left(\sum_{j\in\mathcal{I}} \langle \alpha_j, \pi_{E_j} X\rangle \right) \times \left(\sum_{j'\in\mathcal{J}} \langle \beta_{j'}, \pi_{E_{j'}} X\rangle \right) \right] = \sum_{j\in\mathcal{I}, j'\in\mathcal{J}} \alpha_j^t \pi_{E_j}\pi_{E_{j'}}\beta_{j'} = 0\,.$$

□

Le résultat suivant est capital pour comprendre l'estimation des paramètres d'une loi gaussienne en dimension 1. C'est une conséquence directe du théorème de Cochran.

THÉORÈME 2.39 (THÉORÈME DE STUDENT)
Si $(X_1, \dots, X_n)$ sont des variables indépendantes distribuées selon $\mathcal{N}(\mu, \sigma^2)$, si $\overline{X}_n := \sum_{i=1}^n X_i/n$ et $V := \sum_{i=1}^n (X_i - \overline{X}_n)^2$, alors

1. *$\overline{X}_n$ est distribuée selon $\mathcal{N}(\mu, \sigma^2/n)$,*
2. *V est indépendante de $\overline{X}_n$*
3. *V/σ^2 est distribuée selon χ^2_{n-1}.*

PREUVE. Sans perdre en généralité, on peut supposer que $\mu = 0$ et $\sigma = 1$.

Comme

$$\begin{pmatrix} \overline{X}_n \\ \vdots \\ \overline{X}_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \times \begin{pmatrix} 1 & \dots & 1 \end{pmatrix} X$$

le vecteur $(\overline{X}_n, \dots, \overline{X}_n)^t$ est la projection orthogonale sur la droite engendrée par $(1, \dots, 1)^t$ du vecteur gaussien standard X.

Le vecteur $(X_1 - \overline{X}_n, \dots, X_n - \overline{X}_n)^t$ est la projection orthogonale du vecteur gaussien X sur l'hyperplan orthogonal à $(1, \dots, 1)^t$.

D'après le théorème de Cochran (2.38), les vecteurs $(\overline{X}_n, \dots, \overline{X}_n)^t$, et $(X_1 - \overline{X}_n, \dots, X_n - \overline{X}_n)^t$ sont indépendants.

La loi de $\overline{X}_n$ a déjà été établie.

La loi de V se déduit elle aussi du théorème de Cochran. □

DÉFINITION 2.40 (LOI DE STUDENT) Si $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2_p$ et si X et Y sont indépendantes, alors la variable aléatoire $Z = X/\sqrt{Y/p}$ suit une loi de Student (centrée) à p degrés de liberté.

EXEMPLE 2.41 Application à la construction d'un intervalle de confiance pour l'espérance d'une gaussienne de variance inconnue.

2.10 CONCENTRATION GAUSSIENNE

La définition même des vecteurs gaussiens nous indique ce qu'est la distribution de toute fonction affine d'un vecteur gaussien standard. Si la partie linéaire de la fonction affine est définie par un vecteur λ, nous savons que la variance sera $\|\lambda\|_2^2$. Que se passe-t-il si on s'intéresse à des fonctions assez régulières d'un vecteur gaussien standard ? par exemple si on considère des fonctions L-lipschitziennes ? Ce sont

des généralisations des fonctions affines. Nous ne pouvons donc pas espérer une majoration générale de la variance des fonctions L-Lipschitziennes d'un vecteur gaussien standard meilleure que L^2 (dans le cas linéaire la constante de Lipschitz est la norme euclidienne de λ). Il est remarquable que la borne fournie par les fonctions linéaires se généralise aux fonctions lipschitziennes. Il est encore plus remarquable que cette majoration ne fasse pas intervenir la dimension de l'espace ambiant.

THÉORÈME 2.42 *Soit $X \sim \mathcal{N}(0, Id_d)$.*

1. *si f est différentiable sur $\mathbb{R}^d$,*

$$\mathrm{var}(f(X)) \leq \mathbb{E}\|\nabla f\|^2 \qquad \textit{(inégalité de Poincaré)}$$

2. *si f est L-Lipschitzienne sur $\mathbb{R}^d$,*

$$\mathrm{var}(f(X)) \leq L^2$$

et pour $\lambda > 0$

$$\log \mathbb{E}\mathrm{e}^{\lambda(f(X)-\mathbb{E}f)} \leq \frac{\lambda^2 L^2}{2}\,.$$

Pour tout $t \geq 0$,

$$\mathbb{P}\left\{f(X) - \mathbb{E}f(X) \geq t\right\} \leq \mathrm{e}^{-\frac{t^2}{2L^2}}\,.$$

La clé de la preuve repose sur l'identité suivante.

PROPOSITION 2.43 (IDENTITÉ DE COVARIANCE) *Soit X, Y deux vecteurs gaussiens standards indépendants à valeur dans $\mathbb{R}^d$, f, g deux fonctions différentiables de $\mathbb{R}^d$ dans $\mathbb{R}$.*

$$\mathrm{cov}(f(X), g(X)) = \int_0^1 \mathbb{E}\left\langle \nabla f(X), \nabla g\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\right\rangle \mathrm{d}\alpha$$

Pour vérifier cette proposition, on peut commencer par la vérifier pour les fonctions de la forme $x \mapsto \mathrm{e}^{\imath\langle\lambda, x\rangle}, x \in \mathbb{R}^d$, et étendre le résultat par un argument de densité.

PREUVE.[Théorème 2.42] On commence par établir l'inégalité de Poincaré.

On choisit $f = g$. A partir de l'identité de covariance, on a, en invoquant l'inégalité de Cauchy-Schwarz :

$$\begin{aligned}
\mathrm{var}(f(X)) &= \mathrm{cov}(f(X), f(X)) \\
&= \int_0^1 \mathbb{E}\left\langle \nabla f(X), \nabla f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\right\rangle \mathrm{d}\alpha \\
&\leq \int_0^1 \left(\mathbb{E}\|\nabla f(X)\|^2\right)^{1/2} \times \left(\mathbb{E}\|\nabla f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\|^2\right)^{1/2} \mathrm{d}\alpha\,.
\end{aligned}$$

On obtient le résultat désiré en observant que X et $\alpha X + \sqrt{1-\alpha^2}Y$ sont $\mathcal{N}(0, \mathrm{Id})$ distribués.

Pour obtenir l'inégalité exponentielle, on choisit f différentiable et 1-Lipschitzienne, et on choisit $g = \exp(\lambda f)$ pour $\lambda \geq 0$. On suppose sans perdre en généralité que $\mathbb{E}f(X) = 0$ (f est centrée). L'identité de covariance et la règle de chainage nous conduisent à

$$\begin{aligned}
\mathrm{cov}\left(f(X), \mathrm{e}^{\lambda f(X)}\right) &= \lambda \int_0^1 \mathbb{E}\left[\left\langle \nabla f(X), \nabla f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\right\rangle \mathrm{e}^{\lambda f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)}\right] \mathrm{d}\alpha \\
&\leq \lambda L^2 \int_0^1 \mathbb{E}\left[\mathrm{e}^{\lambda f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)}\right] \mathrm{d}\alpha \\
&= \lambda L^2 \mathbb{E}\left[\mathrm{e}^{\lambda f(X)}\right]
\end{aligned}$$

Si on pose $F(\lambda) := \mathbb{E}\left[\mathrm{e}^{\lambda f(X)}\right]$, on remarque que nous venons d'établir une inégalité différentielle pour F, en vérifiant $\mathrm{cov}(f, \mathrm{e}^{\lambda f}) = F'(\lambda)$ car f est centrée :

$$F'(\lambda) \leq \lambda L^2 F(\lambda)\,.$$

On résout cette inégalité sous la condition initiale $F(0) = 1$, pour $\lambda \geq 0$

$$F(\lambda) \leq \mathrm{e}^{\frac{\lambda^2 L^2}{2}}\,.$$

On peut développer la même démarche pour $\lambda < 0$. Il suffit ensuite d'invoquer l'inégalité de Markov exponentielle et d'optimiser en $\lambda = t/L^2$. □

Les inégalités de concentration décrivent facilement le comportement de la norme des vecteurs gaussiens en grande dimension.

COROLLAIRE 2.44 *Soit X un vecteur gaussien standard à valeur dans $\mathbb{R}^d$. On a*

$$\mathrm{var}(\|X\|_2) \leq 1$$

et

$$\sqrt{d-1} \leq \mathbb{E}\|X\|_2 \leq \sqrt{d}\,.$$

PREUVE. La norme euclidienne est une fonction 1-Lipschitzienne (inégalité triangulaire). La première inégalité est une conséquence de l'inégalité de Poincaré.

La majoration de l'espérance est une conséquence directe de l'inégalité de Jensen. La minoration de l'espérance découle de $(\mathbb{E}\|X\|_2)^2 = \mathbb{E}\|X\|_2^2 - \mathrm{var}(\|X\|_2) = d - \mathrm{var}(\|X\|_2)$ et de la majoration de variance obtenue de l'inégalité de Poincaré. □

EXERCICE 2.45 Soit $X \sim \mathcal{N}(0, K)$ où K est une matrice symétrique DP de dimensions $d \times d$ et $Z = \max_{i \leq d} X_i$.

Montrer que

$$\mathrm{Var}(Z) \leq \max_{i \leq d} K_{i,i} := \max_{i \leq d} \mathrm{Var}(X_i).$$

EXERCICE 2.46 Soit $X, Y \sim \mathcal{N}(0, \mathrm{Id}_n)$ avec X, Y indépendantes. Montrer que

$$\sqrt{2n-1} \leq \mathbb{E}[\|X - Y\|] \leq \sqrt{2n}$$

et que

$$\mathbb{P}\left\{\|X - Y\| - \mathbb{E}[\|X - Y\|] \geq t\right\} \leq \mathrm{e}^{-t^2}\,.$$

2.11 REMARQUES BIBLIOGRAPHIQUES

La littérature gaussienne est très abondante, voir par exemple [6]. Une grande partie de cette littérature intéresse les statistiques.

Les lemmes 2.1 et 2.3 qui caractérisent la gaussienne standard constituent le point de départ de la méthode de (Charles) Stein pour démontrer le théorème central limite (et bien d'autres résultats). Ce développement relativement récent est décrit dans [1].

L'analyse et l'algorithmique matricielles jouent un rôle important en analyse gaussienne et en statistique. Les ouvrages [5], et si on souhaite aller plus loin [3], fournissent une présentation des notions et techniques de factorisation matricielle et des éléments de la théorie de la perturbation.

Il existe une version multi-dimensionnelle des lois du χ^2 qui apparaissent lors de la détermination de la loi de la variance empirique. Il s'agit des lois de Wishart. Elles ont fait l'objet d'études intensives en théorie des matrices aléatoires, voir par exemple [2].

La concentration gaussienne joue un rôle important en statistique non-paramétrique et c'est une source d'inspiration en apprentissage statistique. Le livre de M. Ledoux [7] fournit une perspective élégante sur cette question.

Références

[1] N. Ross. « Fundamentals of Stein's method ». In : *ArXiv e-prints* (sept. 2011).

[2] G. Anderson, A. Guionnet et O. Zeitouni. **An introduction to random matrices**. T. 118. Cambridge Studies in Advanced Mathematics. Cambridge : Cambridge University Press, 2010.

[3] R. Bhatia. **Matrix analysis**. Springer-Verlag, 1997.

[4] P. Bickel et K. Doksum. **Mathematical statistics**. San Francisco, Calif. : Holden-Day Inc., 1976. mr : 56\#1513.

[5] R. Horn et C. Johnson. **Matrix analysis**. Cambridge University Press, 1990.

[6] S. Janson. **Gaussian Hilbert spaces**. T. 129. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1997, p. x+340.

[7] M. Ledoux. **The concentration of measure phenomenon**. AMS, 2001.

[8] A. van der Vaart. **Asymptotic statistics**. Cambridge University Press, 1998.

3.1 EXEMPLES DE MODÈLES GAUSSIENS

Les modèles (ou expériences statistiques) gaussiens sont au centre à la fois de la statistique classique et de la statistique moderne dite en grande dimension. Les propriétés des vecteurs gaussiens rendent certains calculs exacts possibles, ce sont des modèles didactiques. Mais en plus, le théorème central limite, et une multitude de principes d'invariance font des modèles gaussiens la limite de nombreuses suites d'expériences statistiques, dans un sens éventuellement très formel.

i) Le plus simple des modèles gaussiens est celui du décalage (*shift*) où

$$\mathcal{P} = \left\{\mathcal{N}(\theta, \Sigma) : \theta \in \Theta = \mathbb{R}^d\right\} \qquad \Sigma \text{ matrice définie positive supposée connue.}$$

L'espace des paramètres est un espace vectoriel, on parle de *modèle linéaire*.

ii) On peut étudier les modèles gaussiens sans se contenter des modèles linéaires. En visant éventuellement des modèles en dimension infinie, on peut s'intéresser à $\Theta = B_q^d(r)$ boule ℓ_q de rayon r dans $\mathbb{R}^d$,

$$\mathcal{P} = \left\{\mathcal{N}(\theta, \sigma^2) : \theta \in \Theta = B_p^d(r)\right\} \qquad \sigma \text{ supposée connue.}$$

On convient de noter $B_p^d(r) = \left\{\theta : \theta \in \mathbb{R}^d, \sum_{i=1}^d |\theta_i|^p \leq r^p\right\}$. C'est un cas particulier de *modèle courbe*.

iii) Enfin on peut s'intéresser dans le cadre gaussien, à l'inférence sous contrainte de *parcimonie*. Pour $0 \leq s \leq d$, $\Theta_s \subset \mathbb{R}^d$ est l'ensemble des vecteurs à au plus s coordonnées non-nulles. On se propose de reconstruire $\theta \in \Theta_s$ inconnu à partir de $(\langle X_i, \theta\rangle)_{i\leq n}$ où les X_i sont des vecteurs gaussiens standards indépendants (à valeur dans $\mathbb{R}^d$). On note Y le vecteur aléatoire dont les coordonnées sont les $\langle X_i, \theta\rangle$. Les observations sont les vecteurs $X_1, \ldots, X_n$ et la suite $\langle X_1, \theta\rangle, \ldots, \langle X_n, \theta\rangle$. C'est un domaine (*Compressed sensing*) où les propriétés de concentration gaussienne facilitent grandement les calculs.

3.2 MODÈLES DOMINÉS ET VRAISEMBLANCE, ESTIMATION DE VECTEURS GAUSSIENS (DÉCALAGE GAUSSIEN)

On se concentre ici sur les modèles gaussiens définis par les lois $\mathcal{N}(\mu, \Sigma)$ où $\mu \in \mathbb{R}^d$ et $\Sigma \in \text{DP}(d)$ ($\text{DP}(d)$ désignant le cône des matrices (symétriques) définies positives de dimension d). Le problème est d'estimer μ et Σ à partir d'un échantillon. Quand Σ est connue, on parle de problème de décalage gaussien.

Ces modèles gaussiens sont des cas particuliers de *modèles dominés*.

DÉFINITION 3.1 On dit qu'un modèle $(\Omega, \mathcal{F}, \mathcal{P}, \ldots)$ est *dominé* si toutes les lois de $\mathcal{P}$ sont absolument continues par rapport à une mesure σ-finie sur $(\Omega, \mathcal{F})$.

Toutes les lois $\mathcal{N}(\mu, \Sigma)$ avec Σ définie positive sont en effet mutuellement absolument continues et absolument continues par rapport à la mesure de Lebesgue.

Nous utiliserons souvent dans la suite la notion de *vraisemblance*. Dans le cadre des modèles gaussiens, la (fonction de) vraisemblance est une fonction des paramètres et de l'échantillon (domaine $\Theta \times \mathcal{X}^{d\times n}$) qui à $\theta = (\mu, \Sigma), x = (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n}$ fait correspondre la densité de la loi $\mathcal{N}(\mu, \Sigma)$ en x soit :

$$\prod_{i=1}^n \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{(x_i-\mu)^t\Sigma^{-1}(x_i-\mu)}{2}\right).$$

La *log-vraisemblance* est le logarithme de la vraisemblance.

Dans les modèles gaussiens en dimension d, si on s'intéresse aux problèmes d'inférence (estimation, régions de confiance, tests), on peut résumer un échantillon de $n \geq d$ observations par un vecteur aléatoire de dimension d (la moyenne empirique $\overline{X}_n$)

$$\overline{X}_n = \sum_{i=1}^n \frac{1}{n} X_i$$

et une matrice aléatoire à valeur dans $\mathrm{DP}(d)$ (la matrice de covariance empirique $\widehat{\Sigma}$),

$$\widehat{\Sigma} = \sum_{i=1}^{n} \frac{1}{n}(X_i - \overline{X}_n)(X_i - \overline{X}_n)^t$$

sans perdre d'information.

Définition 3.2 (statistique suffisante) Dans un modèle paramétré, une statistique T est dite *suffisante* si la loi conditionnelle de l'échantillon sachant la statistique T est libre du paramètre.

D'un point de vue opérationnel, si T est une statistique suffisante, le rapport entre les vraisemblances évaluées en deux points de l'espace des paramètres peut se calculer à partir de la statistique suffisante T sans connaître le détail de l'échantillon.

On peut démontrer la suffisance de $(\overline{X}_n, \widehat{\Sigma})$ dans les modèles gaussiens à partir de l'observation : la vraisemblance d'un échantillon gaussien ne dépend de l'échantillon qu'au travers de la moyenne empirique et de la (co)variance empirique. Dans le cas univarié, pour la log-vraisemblance :

$$\ell_n(\mu, \sigma^2) := \ell(\mu, \sigma^2, x_1, \ldots, x_n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{n}{2}\frac{(\overline{X}_n - \mu)^2}{\sigma^2} - \frac{\sum_{i=1}^{n}(x_i - \overline{X}_n)^2}{2\sigma^2}\,.$$

Dans cadre multivarié ($d > 1$), pour un n-échantillon, on obtient pour la log-vraisemblance :

$$\begin{aligned}
\ell_n(\mu, \Sigma) &= -\frac{n \times d}{2}\log(2\pi) - \frac{n}{2}\log\det(\Sigma) \\
&\quad -\frac{n}{2}(\overline{X}_n - \mu)^t\Sigma^{-1}(\overline{X}_n - \mu) \\
&\quad -\frac{1}{2}\mathrm{trace}\left(\Sigma^{-1}\sum_{i=1}^{n}(x_i - \overline{X}_n)(x_i - \overline{X}_n)^t\right) \\
&= -\frac{n \times d}{2}\log(2\pi) - \frac{n}{2}\log\det(\Sigma) \\
&\quad -\frac{n}{2}\mathrm{trace}\left(\Sigma^{-1}(\overline{X}_n - \mu)(\overline{X}_n - \mu)^t\right) \\
&\quad -\frac{n}{2}\mathrm{trace}\left(\Sigma^{-1}\widehat{\Sigma}\right)\,.
\end{aligned}$$

L'échantillon lui même constitue toujours une statistique suffisante. Ce n'est pas un résumé impressionnant de l'information fournie par l'échantillon. Les résumés intéressants sont les *statistiques suffisantes minimales*.

Définition 3.3 (statistique suffisante minimale) Une statistique suffisante T est minimale si pour tout autre statistique suffisante T', T est une fonction de T'.

Exemple 3.4 Dans le modèle gaussien $\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathrm{DP}(d)\}$, la moyenne empirique et la covariance empirique forment une statistique suffisante minimale. Si la covariance est connue (modèle $\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d\}$), la moyenne empirique est une statistique suffisante minimale.

Si les statistiques suffisantes semblent en apparence les ingrédients nécessaires à la construction des bons estimateurs, les statistiques dites *ancillaires* semblent superflues.

Définition 3.5 (statistique ancillaire) Dans un modèle, une statistique dont la loi ne dépend pas d'un paramètre est dite ancillaire par rapport à ce paramètre.

Exemple 3.6 Dans le modèle gaussien $\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathrm{DP}(d)\}$, la covariance empirique est ancillaire par rapport à l'espérance. De fait, l'estimateur de choix $(\overline{X}_n)$ suffit pour l'estimation ponctuelle de μ. Mais si on s'intéresse à la construction de régions de confiance pour μ, la statistique ancillaire $\widehat{\Sigma}$ intervient naturellement (via le théorème de Student).

Remarque 3.7 Dans un modèle, une variable aléatoire est dite *pivotale* si sa loi est libre des paramètres du modèle. Attention, une variable aléatoire pivotale n'est pas nécessairement une statistique ! Ce n'est pas toujours une quantité calculable à partir des seules données (c'est même ce qui fait son intérêt). En revanche, une statistique ancillaire est une statistique, c'est-à-dire une fonction de l'échantillon. Et sa loi doit être libre des paramètres ou d'au moins d'une partie d'entre eux.

Le théorème de Student, qui implique l'indépendance de $\overline{X}_n$ et de $\widehat{\Sigma}$, suggère qu'une statistique suffisante minimale est indépendante d'une statistique ancillaire. Ce n'est pas toujours le cas. Pour caractériser les cas d'indépendance, on introduit la notion de *complétude*.

DÉFINITION 3.8 (MODÈLE COMPLET POUR UNE STATISTIQUE) Un modèle $(P_\theta, \theta \in \Theta)$ est dit complet pour une statistique T si pour toute fonction mesurable g,

$$\mathbb{E}_\theta[g(T(X))] = 0 \quad \text{pour tout } \theta \in \Theta \qquad \Rightarrow \qquad P_\theta\{g(T(X)) = 0\} = 1 \quad \text{pour tout } \theta \in \Theta\,.$$

Nous admettrons le théorème suivant.

THÉORÈME 3.9 (BASU) *Si T est une statistique suffisante minimale et si le modèle est complet pour la statistique alors T est indépendante de toute statistique ancillaire.*

EXEMPLE 3.10 Si P_θ est la loi uniforme sur $[\theta - 1, \theta + 1]$, on peut vérifier que

$$\big((X_{(n)} + X_{(1)})/2, (X_{(n)} - X_{(1)})/2\big)$$

est une statistique suffisante minimale ($X_{(i)}$ est la i^{eme} statistique d'ordre de l'échantillon) et que $(X_{(n)} - X_{(1)})/2$ est une statistique ancillaire.

Le modèle n'est pas complet pour cette statistique suffisante minimale.

Le théorème dit de Student qui affirme que dans un échantillon gaussien, la covariance empirique est indépendante de la moyenne empirique peut être vu comme un corollaire du théorème de Basu. Si on suppose la covariance connue, la moyenne empirique est une statistique suffisante complète et minimale pour l'espérance alors que la covariance empirique est ancillaire.

L'estimation au *maximum de vraisemblance* consiste à estimer les paramètres en recherchant les valeurs de μ, σ^2 qui maximisent $\ell_n(\mu, \sigma^2)$ (cas univarié) ou $\ell_n(\mu, \Sigma)$ (cas multivarié).

Dans le cas univarié, on obtient directement les estimateurs $\big(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{X}_n)^2\big)$ (c'est-à-dire la moyenne empirique et la variance empirique).

Dans le cas général, un peu plus de travail conduit à la version multivariée des estimateurs précédents

$$\left(\overline{X}_n, \widehat{\Sigma}\right) := \left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{X}_n)(x_i - \overline{X}_n)^t\right).$$

3.3 RÉGRESSION AVEC DESIGN FIXE ET BRUIT GAUSSIEN HOMOSCHÉDASTIQUE

Le problème de régression gaussienne avec *design fixe* et *bruit homoschédastique* est le plus simple des problèmes de régression. C'est une généralisation du problème de *décalage gaussien* (*Gaussian shift*). On dispose d'une matrice $n \times p$ le *design* $\mathbf{X}$, on observe Y à valeur dans $\mathbb{R}^n$ obtenu par

$$Y = \mathbf{X}\theta_0 + \sigma\epsilon \qquad \text{avec } \epsilon \sim \mathcal{N}(0, \mathrm{Id}_n)$$

où le paramètre inconnu est θ_0 dans $\mathbb{R}^p$. Selon les circonstances, on connait ou on ne connait pas σ.

En statistique classique, on suppose que $\mathbf{X}$ est de plein rang (ce qui implique $p \leq n$). En « grande dimension », on suppose au contraire que $n \ll p$ mais on suppose alors que θ_0 vérifie une hypothèse de *parcimonie* (θ_0 possède peu de coordonnées non-nulles).

On estime θ_0 en cherchant à minimiser l'erreur quadratique de prédiction (d'où le nom de *méthode des moindre carrés ordinaires*, MCO)

$$\widehat{\theta}_n := \arg\min \|y - \mathbf{X}\theta\|^2 \qquad \text{où } y \text{ est la réalisation de } Y$$

On note $\mathcal{L}(\mathbf{X})$ le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}$. La relation pythagoricienne suivante est la clé de l'analyse :

$$\|y - \mathbf{X}\theta\|^2 = \|y - \widehat{y}\|^2 + \|\widehat{y} - \mathbf{X}\theta\|^2$$

où $\widehat{y}$ est la projection orthogonale de y sur $\mathcal{L}(\mathbf{X})$.

Comme $\mathbf{X}$ est de plein rang, la matrice de la projection orthogonale sur le sous-espace $\mathcal{L}(\mathbf{X})$ est

$$\mathbf{H} := \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,.$$

En effet, cette matrice laisse invariant les colonnes de $\mathbf{X}$, donc le sous-espace $\mathcal{L}(\mathbf{X})$, et tout vecteur orthogonal aux colonnes de $\mathbf{X}$ appartient au noyau de matrice $\mathbf{H}$.

Dans le monde de la régression (gaussienne ou non) la matrice $\mathbf{H}$ est souvent appelée *hat matrix*, ou *matrice d'influence*. Ses coefficients diagonaux sont appelés (en anglais) *leverage* et parfois traduits en français par *influence* ou *effet levier*.

Comme $\mathbf{X}$ est de plein rang, il existe un unique optimum, la solution de $\widehat{y} = \mathbf{X}\theta$. Comme

$$\widehat{y} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y$$

cette solution est

$$\widehat{\theta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\widehat{y}\,.$$

On note $\mathbf{X}^+ := (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$. C'est la *pseudo-inverse* de Moore-Penrose de $\mathbf{X}$ (voir Appendice A.4). On a donc

$$\widehat{\theta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y = \mathbf{X}^+ y\,,$$

et

$$\widehat{\theta} - \theta_0 = \sigma(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\epsilon = \sigma\mathbf{X}^+\epsilon\,.$$

Comme

$$\mathbb{E}_{\theta_0}\left[\widehat{\theta} - \theta_0\right] = \sigma\mathbf{X}^+\mathbb{E}[\epsilon] = 0\,,$$

on constate que l'estimateur des moindres carrés $\widehat{\theta}$ est un estimateur sans biais.

Remarque 3.11 Jusqu'à maintenant, nous n'avons pas utilisé l'hypothèse gaussienne mais seulement le fait que le bruit ϵ est centré. L'estimateur des moindre carrés $\mathbf{X}^+ y$ peut être utilisé chaque fois que l'on utilise un modèle du type

$$Y = \mathbf{X}\theta_0 + \epsilon$$

où ϵ est un bruit centré $\mathbb{E}\epsilon = 0$. L'estimateur des moindre carrés restera sans biais. En revanche les résultats qui suivent, les estimations du risque, les constructions de régions de confiance ne seront plus valables. Ces résultats reposent sur l'hypothèse de normalité du bruit.

Dans le cadre des modèles linéaires gaussiens avec bruit homoschédastique, l'estimation de la variance du bruit σ^2 est grandement facilitée. Et la construction de région de confiance pour θ_0 et σ^2 est immédiate.

Proposition 3.12 *Dans les modèles linéaires gaussiens avec bruit homoschédastique,*

$$\widehat{\theta} - \theta_0 \sim \mathcal{N}\left(0, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\right)\,.$$

Si l'écart-type du bruit σ est connu, on dispose d'une quantité pivotale :

$$\frac{1}{\sigma}(\mathbf{X}^t\mathbf{X})^{1/2}\left(\widehat{\theta} - \theta_0\right) \sim \mathcal{N}\left(0, \mathrm{Id}_p\right)\,,$$

preuve. On part de l'identité $\widehat{\theta} - \theta_0 = \sigma\mathbf{X}^+\epsilon$. Sous l'hypothèse $\epsilon \sim \mathcal{N}(0, \mathbf{Id}_n)$, on a

$$\widehat{\theta} - \theta_0 \sim \mathcal{N}\left(0, \sigma^2\mathbf{X}^+(\mathbf{X}^+)^t\right) = \mathcal{N}\left(0, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\right)\,.$$

La preuve se termine en utilisant la décomposition spectrale de la matrice symétrique définie positive $\mathbf{X}^t\mathbf{X}$:

$$\mathbf{X}^t\mathbf{X} = \mathbf{O}\mathbf{D}\mathbf{O}^t\,,$$

avec $\mathbf{O}$ orthogonale et $\mathbf{D}$ diagonale à coefficients diagonaux positifs (voir Appendice A). On multiplie $\widehat{\theta} - \theta_0$ par $\mathbf{O}\mathbf{D}^{1/2}\mathbf{O}^t = (\mathbf{X}^t\mathbf{X})^{1/2}$. □

On en déduit une région de confiance pour θ lorsque l'intensité (l'écart-type) du bruit σ est connue : l'ellipsoïde

$$\left\{\theta' : \left\|\frac{1}{\sigma}(\mathbf{X}^t\mathbf{X})^{1/2}\left(\widehat{\theta}-\theta'\right)\right\|^2 \leq q_{p,1-\alpha}\right\} \qquad \text{où } q_{p,1-\alpha} \text{ est le quantile d'ordre } 1-\alpha \text{ de } \chi^2_p\,.$$

Lorsque l'écart-type du bruit σ n'est pas connu, cette région aléatoire n'est plus une région de confiance.

Pour construire une région de confiance, il faut estimer σ. On estime σ à partir de l'analyse des résidus. On appelle *résidus* les coefficients de $\widehat{\epsilon} := \widehat{y} - y$. La loi du vecteur des résidus est facilement déduite de l'observation :

$$\widehat{Y} - Y = (\mathbf{H}-\mathbf{Id})Y = (\mathbf{H}-\mathbf{Id})(\mathbf{X}\theta_0 + \sigma\epsilon) = \sigma(\mathbf{H}-\mathbf{Id})\epsilon\,.$$

Comme $\mathbf{Id}-\mathbf{H}$ est la matrice de la projection orthogonale sur le sous-espace orthogonal à $\mathcal{L}(\mathbf{X})$, on peut invoquer le théorème de Cochran pour justifier la proposition suivante.

PROPOSITION 3.13 *Dans le modèle de régression linéaire avec bruit gaussien homoschédastique,*

i) $\widehat{Y} - Y$ *est indépendant de* $\mathbf{X}(\widehat{\theta}-\theta_0)$.
ii) $\frac{\|\widehat{Y}-Y\|^2}{\sigma^2} \sim \chi^2_{n-p}$.
iii) $\frac{\|\mathbf{X}(\widehat{\theta}-\theta_0)\|^2}{\sigma^2} \sim \chi^2_p$.

PREUVE. On observe $\widehat{Y} - Y = \sigma(\mathbf{H}-\mathbf{Id})\epsilon$ et $\mathbf{X}(\widehat{\theta}-\theta_0) = \sigma\mathbf{H}\epsilon$. Au facteur σ près, on obtient $\widehat{Y}-Y$ et $\mathbf{X}(\widehat{\theta}-\theta_0)$ en projetant le vecteur gaussien standard ϵ sur deux sous espaces orthogonaux de $\mathbb{R}^n$. La proposition est donc une conséquence immédiate du théorème de Cochran. □

Nous utiliserons une nouvelle famille de lois de probabilité dans la suite.

DÉFINITION 3.14 (DISTRIBUTION DE FISHER) Si $U \sim \chi^2_j$ est indépendante de $V \sim \chi^2_k$ alors $\frac{U/j}{V/k}$ est distribuée selon une loi de Fisher à j,k degrés de liberté. On note cette loi $F_{j,k}$.

La quantité aléatoire $\frac{\|H\epsilon\|^2}{\|(\mathrm{Id}-H)\epsilon\|^2}$ est distribuée selon $F_{p,n-p}$. Le fait que cette quantité aléatoire soit pivotale conduit à une région de confiance pour θ_0.

On note $f_{p,n-p,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de $F_{p,n-p}$.

PROPOSITION 3.15 *Dans le modèle linéaire gaussien avec bruit homoschédastique, l'ellipsoïde*

$$\left\{\theta' : \frac{\left\|\mathbf{X}\left(\widehat{\theta}-\theta'\right)\right\|^2/p}{\|\widehat{\epsilon}\|^2/(n-p)} \leq f_{p,n-p,1-\alpha}\right\}$$

est une région de confiance pour θ_0 *de taux de couverture* $1-\alpha$ *(*$\alpha \in]0,1[$*).*

Le volume de cet ellipsoïde est proportionnel à (une puissance de) l'estimateur de l'écart type du bruit. La forme de cet ellipsoide est dictée par les valeurs propres de $\mathbf{X}^t\mathbf{X}$, autrement dit par les valeurs singulières non nulles de $\mathbf{X}$.

3.4 TESTS D'HYPOTHÈSES LINÉAIRES

La régression linéaire est un outil de modélisation. Le statisticien peut hésiter entre plusieurs modèles et se poser des questions de test. On peut par exemple se demander si les observations $y_1,\ldots,y_n$ ne sont pas simplement un échantillon d'une loi $\mathcal{N}(0,\sigma^2)$. Cela revient à poser le problème de test suivant :

i) $\text{Hyp}_0 : \theta = 0$;
ii) $\text{Hyp}_1 : \theta \neq 0$.

L'hypothèse nulle Hyp_0 est simple mais l'alternative Hyp_1 ne l'est pas.

D'une manière générale, on peut chercher à tester :

i) Hyp_0 : les $p-k$ dernières coordonnées de θ sont nulles ;
ii) Hyp_1 : les $p-k$ dernières coordonnées de θ ne sont pas nulles ;

Dans tous les cas l'hypothèse nulle est indexée par un sous-espace vectoriel de l'ensemble des paramètres.

On notera $\mathbf{X}^0$ la matrice formée par les k premières colonnes de $\mathbf{X}$ et $\mathcal{L}(\mathbf{X}^0)$ le sous-espace de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}^0$. On notera $\mathbf{H}^0$ la matrice de projection orthogonale sur $\mathcal{L}(\mathbf{X}^0)$, $\widehat{y}^0 = \mathbf{H}^0 y = \mathbf{H}^0 \times \mathbf{H} y$. On note θ^0 le vecteur de dimension k formé par les k premiers coefficients de θ.

Sous l'hypothèse nulle Hyp_0,

$$\widehat{y} - \widehat{y}^0 = \left(\mathbf{H} - \mathbf{H}^0\right)\left(\mathbf{X}\begin{bmatrix}\theta^0 \\ \cdots \\ 0\end{bmatrix} + \sigma\epsilon\right) = \sigma\left(\mathbf{H} - \mathbf{H}^0\right)\epsilon\,.$$

Donc $\widehat{Y} - \widehat{Y}^0$ et $Y - \widehat{Y}$ sont indépendantes (Théorème de Cochran), et, sous Hyp_0, la statistique

$$S := \frac{\|\widehat{Y} - \widehat{Y}^0\|^2/(p-k)}{\|\widehat{Y} - Y\|^2/(n-p)}$$

est distribuée selon une loi de Fisher $F_{p-k,n-p}$.

Le test qui rejette Hyp_0 lorsque la statistique S est plus grande que le quantile d'ordre $1-\alpha$ de la loi $F_{p-k,n-p}$ est de niveau α (la probabilité d'une erreur de première espèce est égale à α).

L'étude de la puissance de ce test (la probabilité de rejeter Hyp_0 lorsque les données sont tirées sous l'alternative Hyp_1) est plus délicate. On peut cependant vérifier que le test est *sans biais*, c'est à dire que la probabilité de rejeter Hyp_0 dans ce cas est supérieure à α.

Choisissons une loi relevant de l'alternative. Cela revient à choisir $\theta^0 \in \mathbb{R}^k$ et $\theta^1 \in \mathbb{R}^{p-k}$, avec $\theta^1 \neq 0$. On note $\mathbf{X}^1$ la matrice formée par les $p-k$ dernières colonnes de $\mathbf{X}$. Sous cette loi relevant de Hyp_1, les observations sont distribuées selon $\mathcal{N}\left(\mathbf{X}^0\theta^0 + \mathbf{X}^1\theta^1, \sigma^2\mathbf{Id}_n\right)$.

Les vecteurs aléatoires $\widehat{Y} - \widehat{Y}^0$ et $Y - \widehat{Y}$ restent indépendants, mais la statistique

$$\|\widehat{Y} - \widehat{Y}^0\|^2/\sigma_0^2$$

est maintenant distribuée selon une loi du χ^2 à $p-k$ degrés de liberté décentrée de $\|(\mathbf{Id} - \mathbf{H}^0)\mathbf{X}^1\theta^1\|_2$. La quantité aléatoire $\|\widehat{Y} - Y\|^2/\sigma_0^2$ reste distribuée selon χ^2_{n-p}. Pour établir que la loi de S sous Hyp_1 domine stochastiquement la loi de S sous Hyp_0, il suffit de se rappeler que à nombres de degrés de liberté égaux, la loi du χ^2 décentrée domine stochastiquement la loi du χ^2 centrée (voir Théorème 2.36).

3.5 DONNÉES WHITESIDE

Les données `whiteside` sont issues du package `MASS` du logiciel `R`. Il s'agit d'une étude sur la relation entre consommation de gaz et température extérieure menée à son domicile par l'ingénieur Whiteside. Cette étude visait à évaluer l'impact de mesures d'isolation. Deux hivers de suite, Whiteside a enregistré chaque semaine la consommation hebdomadaire de gaz et la température extérieure dans une maison. Il a disposé ainsi de

- 26 enregistrements réalisés avant l'isolation ;
- 30 enregistrements après .

On dispose de 56 observations arrangées en un `data.frame` comprenant trois colonnes, `Gas` (volume de gaz consommé durant la semaine), `Temp` (température extérieure moyenne durant la semaine d'observation), et `Insul` (observation effectuée avant (`Before`) ou après (`After`) l'isolation). La colonne `Insul` forme ce que l'on appelle une *variable qualitative* ou un *facteur*.

La première tentative de modélisation postule une dépendance linéaire entre la variable à expliquer ($Y =$ `Gas`) et la variable explicative `Temp`. On se place dans le cadre de la régression linéaire gaussienne. Le design $\mathbf{X}$ est formé par une colonne de 1 et par la colonne `Temp` de l'échantillon. La fonction `lm` de `R` construit le design et invoque la méthode des moindres carrés (MCO). Le résumé du résultat suit.

```
##
## Call:
## lm(formula = Gas ~ Temp, data = whiteside)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6324 -0.7119 -0.2047  0.8187  1.5327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4862     0.2357  23.275  < 2e-16 ***
## Temp         -0.2902     0.0422  -6.876 6.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8606 on 54 degrees of freedom
## Multiple R-squared:  0.4668, Adjusted R-squared:  0.457
## F-statistic: 47.28 on 1 and 54 DF,  p-value: 6.545e-09
```

La dernière ligne décrit un test d'hypothèses linéaires de l'espèce présentée dans la section précédente. Sous H_0, on postule que la pente est nulle. La F-statistique calculée doit être comparée aux quantiles d'une loi de Fisher à 1 et 54 degrés de libertés $n = 56, p = 2, k = 1$. La dernière information, la p-value ou *degré de signification atteint*, est de 6.545×10^{-9}, ce qui signifie qu'on a atteint le quantile d'ordre $1 - 6.545 \times 10^{-9}$ de la loi de Fisher à 1 et 54 degrés de liberté $F_{1,54}$. Un statisticien raisonnable qui croit en la modélisation d'ensemble sera conduit à rejeter l'hypothèse nulle.

Mais cette modélisation est très critiquable. Pourquoi la consommation de gaz ne dépendrait-elle que de la température extérieure moyenne ? et pas du vent ? de l'humidité ? de la durée du jour ? des aléas de la vie en société (vacances, congés de maladie, ...), et bien sûr de l'état de la maison ? On n'est pas non plus obligé de croire que les écarts à la linéarité sont dus à un bruit gaussien homoschédastique. Nous ne disposons pas d'informations pour explorer toutes ces voies, mais on peut tout de même se demander si l'isolation modifie la sensibilité de la consommation de gaz à la température extérieure moyenne. C'est ce que suggère la figure 3.5. On y a reporté les observations et la droite de régression issue de la modélisation naïve. La forme des points indique s'ils proviennent d'une observation effectuée avant ou après l'isolation.

Une modélisation plus ambitieuse envisage que la droite de régression peut être modifiée par l'isolation. L'invocation suivante de `lm` construit un design à quatre colonnes : une colonne de 1 correspondant au coefficient appelé `Intercept`, une colonne comportant des 0 pour les observations effectuées avant isolation, des 1 pour les autres (coefficient `InsulAfter`, une colonne formée par la colonne `Temp` du `data.frame` (coefficient `Temp`) et enfin une colonne formée en multipliant les colonnes `InsulAfter` et `Temp` (`InsulFater :Temp`). On peut intrepréter le résultat comme la production de deux droites de régression, avec des intercepts différents et des pentes différentes (pour la seconde année, l'intercept est `Intercept` + `InsulAfter`, la pente est `Temp` + `InsulAfter :Temp`).

On constate que la distribution des résidus est moins asymétrique (la médiane est plus proche de 0), moins dispersée (l'écart interquartile est divisé par quatre). La somme des carrés des résidus (la variance inexpliquée) qui est égale au carré de la `Residual standard error` multiplié par le nombre de degrés de liberté est divisé par à peu près quatre. L'adéquation aux données (le *fit*) est bien meilleure, en apparence au moins (voir figure 3.5).

```
lmgood <- lm(data=whiteside, Gas ~ Insul * Temp)
summary(lmgood)
##
## Call:
## lm(formula = Gas ~ Insul * Temp, data = whiteside)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97802 -0.18011  0.03757  0.20930  0.63803
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

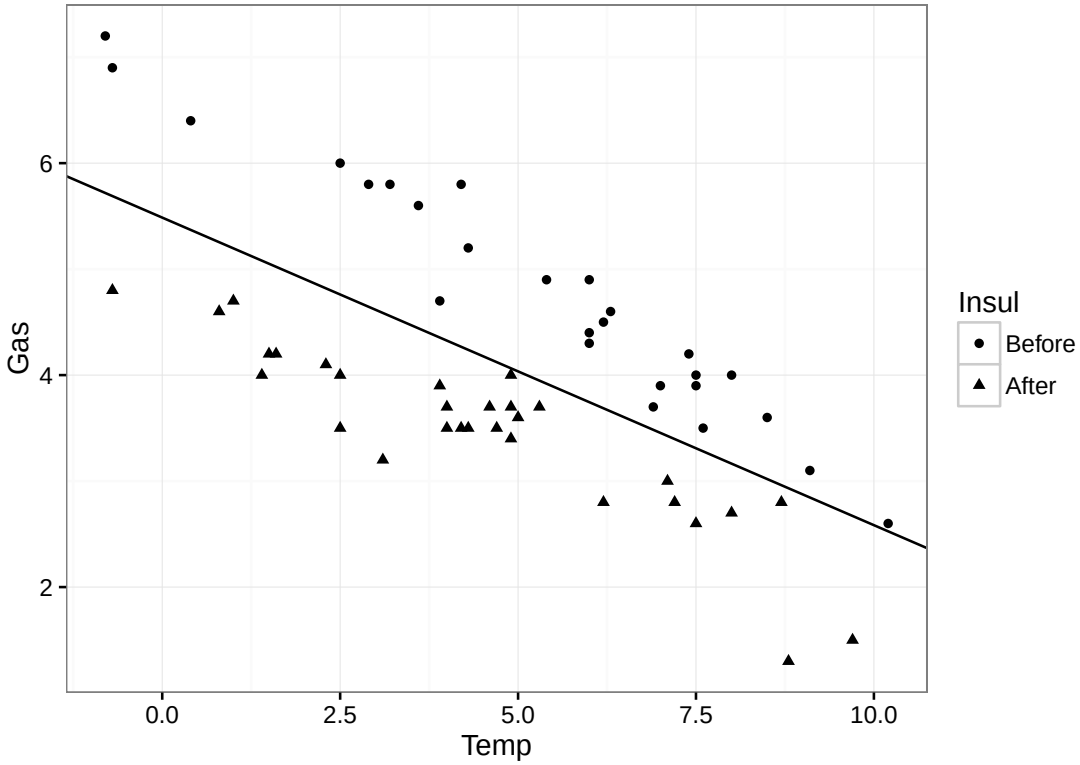

FIG. 3.1 : Les points correspondant à l'hiver précédent l'isolation se situent au dessus de la droite de régression et les autres au dessous.

```
## (Intercept)       6.85383    0.13596   50.409  < 2e-16 ***
## InsulAfter       -2.12998    0.18009  -11.827 2.32e-16 ***
## Temp             -0.39324    0.02249  -17.487  < 2e-16 ***
## InsulAfter:Temp   0.11530    0.03211    3.591 0.000731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 52 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9235
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

On peut effectuer un test d'hypothèses linéaires pour comparer les performances de la modélisation naïve et celles de la modélisation qui l'est moins. La fonction `anova` (*analysis of variance*) prend en argument les deux modèles renvoyés par `lm` et elle calcule la statistique de Fisher. Le degré de signification atteint est inférieur à la précision des calculs numériques sur la machine : la statistique de Fisher dépasse le quantile d'ordre $1 - 2.2 \times 10^{-16}$ de $F_{2,52}$.

```
anova(lm(data=whiteside, Gas ~ Temp),
lm(data=whiteside, Gas ~ Insul*Temp))
```

```
## Analysis of Variance Table
##
## Model 1: Gas ~ Temp
## Model 2: Gas ~ Insul * Temp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     54 39.995
## 2     52  5.425  2     34.57 165.67 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

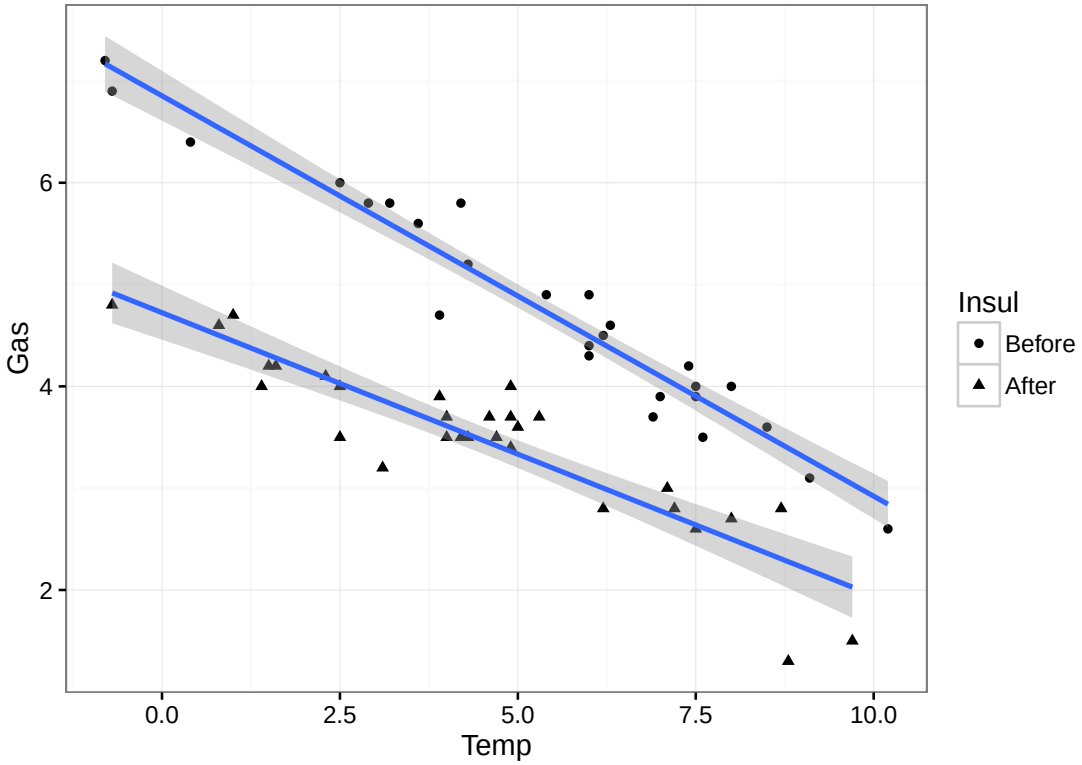

FIG. 3.2 : Visualition des droites de régression correspondant aux deux conditions d'isolation.

3.6 REMARQUES BIBLIOGRAPHIQUES

Les notions de statistique suffisante ou exhaustive, de statistique suffisante minimale, de statistique ancillaire, complète sont discutées en détail dans [6].

Le théorème de Basu est prouvé dans [6].

L'inférence dans les modèles gaussiens est le problème de référence en statistique mathématiques. Le modèle des suites gaussiennes est étudié en profondeur dans la monographie de [5]. Il existe une abondante littérature appliquée sur la régression linéaire ou non, par exemple le livre de Fox et Weisberg [3] qui accompagne le paquet `R` nommé `car`.

Pour les limites de suites d'expériences statistiques, la théorie de la normalité asymptotique locale (LAN), ce que l'on appelle la théorie de Le Cam, voir [8].

Sur la sélection de modèles gaussiennes, voir [7] et références.

Sur la statistique en grandes dimensions voir [4] ou [1]. Sur le *compressed sensing* voir [2].

Sur les modèles graphiques gaussiens, voir [10].

Les données `whiteside` sont analysées dans [9].

Références

[1] P. BÜHLMANN et S. VAN DE GEER. **Statistics for high-dimensional data**. Springer, Heidelberg, 2011.

[2] S. FOUCART et H. RAUHUT. **A mathematical introduction to compressive sensing**. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[3] J. FOX et S. WEISBERG. **An R companion to applied regression**. Sage, 2010.

[4] C. GIRAUD. **Introduction to high-dimensional statistics**. T. 139. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2015, p. xvi+255. ISBN : 978-1-4822-3794-8. MR : 3307991.

[5] I. JOHNSTONE. **Function Estimation and Gaussian Sequence Models**. Manuscript. Santford University : Dept. Statistics, 2002, p. 343.

[6] E. L. Lehmann et G. Casella. **Theory of point estimation**. Second. Springer Texts in Statistics. Springer-Verlag, New York, 1998, p. xxvi+589.

[7] P. Massart. **Concentration inequalities and model selection**. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006.

[8] A. van der Vaart. **Asymptotic statistics**. Cambridge University Press, 1998.

[9] W. N. Venables et B. D. Ripley. **Modern applied statistics with S-Plus**. Statistics and Computing. With 1 IBM-PC floppy disk (3.5 inch ; HD). New York : Springer-Verlag, 1994, p. xiv+462. isbn : 0-387-94350-1. mr : MR1337030(97c:62003).

[10] M. J. Wainwright et M. I. Jordan. **Graphical models, exponential families, and variational inference**. T. 1. 1-2. Now Publishers Inc., 2008, p. 1-305.

4.1 INTRODUCTION

A propos des modèles binomiaux et gaussiens, nous avons vu des méthodes d'estimation motivées par l'intuition. A chaque fois, nous avons utilisé la possibilité d'identifier les lois par quelques moments. Dans le cas gaussien, nous avons aussi noté que la méthode des moindre carrés correspond à une maximisation de vraisemblance.

Nous revenons sur ces méthodes en essayant de dégager une démarche. Les méthodes que nous abordons ne sont pas exclusives les unes des autres. Ce que nous appelons méthode, il faudrait peut être l'appeler point de vue. Nous verrons les aspects suivants.

i) Méthode des moments (Z-estimation)
ii) Maximisation de la vraisemblance
iii) Minimisation de contraste (M-estimation) qui généralise les méthodes de maximisation de la vraisemblance.

4.2 MÉTHODE DES MOMENTS

Nous avons déjà mis à profit la « méthode des moments » dans le cadre des modèles multinomiaux et des modèles gaussiens. L'idée générale est la suivante. Si dans une expérience statistique échantillonnée $(\mathcal{X}, \mathcal{F}, \{P_\theta : \theta \in \Theta\})$, on dispose d'une collection de fonctions intégrables $(T_1, \dots, T_k)$ telles que

$$\theta \mapsto \mathbb{E}_\theta \left[(T_1(X), \dots, T_k(X))^t \right] =: \mathbb{E}_\theta[T(X)] \,.$$

soit injective, c'est-à-dire si les lois P_θ peuvent être identifiées par une collection de moments, on peut envisager une re-paramétrisation de (P_θ) en désignant chaque loi par un élément de $\mathbb{R}^k$, la collection des moments. On désigne par Ψ le nouvel espace de paramètres, par $\psi(\theta)$ le nouveau paramètre correspondant à θ. La méthode des moments consiste à estimer le paramètre par la solution (si elle existe) de l'équation en ϑ :

$$\psi(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} (T_1(X_i), \dots, T_k(X_i))^t =: \overline{T}_n \,.$$

EXEMPLE 4.1 Si $\Theta =]0, +\infty)$, et $P_\theta\{[x, +\infty)\} = \exp(-\theta x)$ (les lois exponentielles paramétrées par leur intensité), on vérifie que $\mathbb{E}_\theta X = 1/\theta$. On peut identifier θ à partir de l'espérance de $\mathrm{Id}(X)$. La nouvelle paramétrisation correspond au paramètre d'échelle $\psi(\theta) = 1/\theta$. On peut estimer sans biais ce paramètre d'échelle à partir de la moyenne empirique $\overline{X}_n$ de l'échantillon. Pour estimer le paramètre originel, on inverse ψ : on estime θ par $\widehat{\theta} = 1/\overline{X}_n$. L'estimateur est alors biaisé (inégalité de Jensen).

On peut (et doit) alors se poser quelques questions :

i) dans quelles conditions, l'équation $\psi(\vartheta) = \overline{T}_n$ a-t-elle une solution ?
ii) dans quelles conditions, $\psi^{-1}(\overline{T}_n)$ définit-elle une suite consistante d'estimateurs ?
iii) dans quelles conditions, $\psi^{-1}(\overline{T}_n)$ définit-elle une suite d'estimateurs asymptotiquement normale ?

Si ces réponses sont positives, peut-on s'assurer que les estimateurs construits de cette façon ont de bonnes propriétés ? Qu'ils sont de risque sinon minimal, du moins raisonnable ?

La première question est la plus délicate. Nous verrons comment y répondre dans le cadre des modèles exponentiels, où la méthode dite du maximum de vraisemblance coïncide avec une méthode de moments.

Pour le moment, nous allons supposer que nous sommes dans une situation comparable à celle rencontrée dans le cadre des modèles gaussiens, c'est à dire que $\overline{T}_n$ prend (avec probabilité 1) ses valeurs dans $\psi(\Theta) =: \Psi$.

L'appendice B résume les outils probabilistes utiles à l'étude de la méthode des moments.

Conditions suffisantes de consistance et de normalité pour la méthode des moments

Si on souhaite construire un estimateur pour la paramétrisation originelle (par exemple par l'intensité dans le cadre des lois exponentielles), il est tentant d'utiliser l'estimateur $\psi^{-1}(\overline{T}_n)$ pour peu qu'il soit bien défini.

La proposition suivante décrit des conditions suffisantes qui garantissent les performances asymptotiques de la méthode des moments. Ces conditions ne sont pas toujours faciles à vérifier. Et meme si elles sont vérifiées, elles ne garantissent pas que la méthode des moments soit praticable.

PROPOSITION 4.2 *Si Θ est un ouvert de $\mathbb{R}^k$, si la fonction*

$$\begin{aligned}\psi :&\Theta \to \mathbb{R}^k\\ &\theta \mapsto (\mathbb{E}_\theta T_1(X), \dots, \mathbb{E}_\theta T_k(X))^t\end{aligned}$$

est bijective, continument différentiable et d'inverse continuement différentiable de Θ dans (l'ouvert) $\psi(\Theta)$, et si pour tout $\theta \in \Theta$, pour $j \leq k$, $\mathbb{E}_\theta T_j(X)^2 < \infty$ alors

i) avec probabilité qui tend vers 1 (sous $P_\theta^{\otimes \mathbb{N}}$), l'équation en ϑ

$$\psi(\vartheta) = \overline{T}_n$$

admet une solution $\widehat{\theta}_n = \psi^{-1}(\overline{T}_n) \in \Theta$.

ii) La suite $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ est une suite consistante d'estimateurs de θ.

iii) La suite $\left(\sqrt{n}\big(\widehat{\theta}_n - \theta\big)\right)_{n\in\mathbb{N}}$ est asymptotiquement normale centrée. La covariance de la loi limite est

$$(D\psi_{|\theta})^{-1} \operatorname{cov}_\theta(T)(D\psi_{|\theta}^t)^{-1} .$$

Les conditions décrites dans l'énoncé de la proposition sont si fortes que la preuve en est trivialisée.

PREUVE.

i) L'image de Θ par ψ est un ouvert de $\mathbb{R}^k$. Pour tout $\theta \in \Theta$, sous $P_\theta^{\otimes \mathbb{N}}$, $\overline{T}_n$ converge presque sûrement vers $\psi(\theta)$. Presque sûrement, pour n assez grand, $\overline{T}_n$ appartient à un voisinage ouvert de $\psi(\theta)$ qui est inclus dans l'ouvert $\Psi = \psi(\Theta)$. L'équation $\psi = \overline{T}_n$ a alors une solution.

Le point ii) est une conséquence de la loi des grands nombres.

Le point iii) résulte d'une application de la méthode delta.

□

La discussion sur l'approche des moments reste informelle, parce qu'il est difficile de caractériser les situations où la question i) (voir plus haut) possède une réponse satisfaisante.

4.3 ENTROPIE RELATIVE

Nous aurons besoin dans la suite de ce cours d'outils permettant quantifier la similitude entre deux lois de probabilités. Nous aurons recours à des *divergences d'information*. Une divergence d'information entre deux lois est une quantité intrinsèque, elle n'est pas définie relativement à une paramétrisation. Une divergence d'information peut être calculée en choisissant des densités par rapport à une mesure de référence. Mais le résultat du calcul ne doit dépendre que des lois, et pas du choix d'un mesure dominante ou du choix des densités.

Nous avons déjà vu une divergence d'information, la distance en variation. L'entropie relative, qui n'est pas une distance est une autre diveregence. Elle est particulièrement adaptée à l'étude des modèles exponentiels.

Dans la suite, on notera $P \not\ll Q$ le fait que la mesure P ne soit pas absolument continue par rapport à Q, c'est-à-dire qu'il existe un ensemble mesurable A tel que $P(A) > 0 = Q(A)$.

Définition 4.3 (Entropie relative) Soit P et Q deux lois de probabilité sur $(\Omega, \mathcal{F})$, l'entropie relative de P par rapport à Q, notée $D(P\|Q)$ est définie par

$$D(P\|Q) = \begin{cases} +\infty & \text{si } P \not\trianglelefteq Q \\ \mathbb{E}_P\left[\log \frac{dP}{dQ}\right] & \text{sinon.} \end{cases}$$

La notion d'entropie relative s'est imposée dans différentes disciplines indépendamment (statistiques, théorie de l'information, mécanique statistique, théorie des grandes déviations). On l'appelle aussi *information de Kullback-Leibler*, *divergence d'information*.

i) Si $P \trianglelefteq Q$ (P est absolument continue par rapport à Q),

$$\mathbb{E}_P\left[\log \frac{dP}{dQ}\right] = \mathbb{E}_Q\left[\frac{dP}{dQ}\log\frac{dP}{dQ}\right]$$

est toujours bien défini, même si cette quantité n'est pas finie. En effet, $x\log x \geq -\mathrm{e}^{-1}$ sur $]0,\infty)$.

ii) Si $P \trianglelefteq Q$, l'identité du i) et la convexité de $x \mapsto x\log x$ nous indiquent que

$$D(P\|Q) \geq \mathbb{E}_Q\left[\frac{dP}{dQ}\right]\log \mathbb{E}_Q\left[\frac{dP}{dQ}\right] = 0\,.$$

Le seul cas d'égalité possible est $P = Q$ (stricte convexité de $x \mapsto x\log x$).

4.4 Modèles exponentiels : définition, identifiabilité, propriétés générales

Les modèles exponentiels sont définis par une collection de densités par rapport à une mesure de référence. Rappelons qu'une mesure ν sur $(\Omega, \mathcal{F})$ est dite σ-finie si Ω est réunion dénombrable d'ensembles de ν-mesure finie (la mesure de Lebesgue est σ-finie, la mesure de comptage sur $\mathbb{R}$ n'est pas σ-finie). En statistique, les mesures dominantes sont toujours des mesures σ-finies. Cette propriété permet de garantir sous conditions d'absolue continuité l'existence de dérivées de Radon-Nikodym.

Définition 4.4 (forme canonique) Une mesure σ-finie ν sur $(\Omega, \mathcal{F})$ et une famille $T_1, \ldots, T_k$ de fonctions mesurables de Ω dans $\mathbb{R}$ définissent le modèle exponentiel en forme canonique indexé par $\Theta = \widetilde{\Theta}^\circ$ (l'intérieur de $\widetilde{\Theta}$) avec

$$\widetilde{\Theta} := \left\{\theta : \theta \in \mathbb{R}^k, \quad Z(\theta) := \int_\Omega \mathrm{e}^{\langle\theta, T(x)\rangle}\mathrm{d}\nu(x) < \infty\right\}$$

où $T(x) = (T_1(x), \ldots, T_k(x))^t$. On choisit pour densité de P_θ par rapport à ν :

$$p_\theta(x) := \mathrm{e}^{\langle\theta, T(x)\rangle - \log Z(\theta)}\,.$$

On peut appeler $\tilde{\Theta}$ le domaine du modèle. On appelle la fonction $\theta \mapsto Z(\theta)$ la fonction de partition. L'ensemble $\tilde{\Theta}$ est en fait le domaine de définition de la fonction de partition.

Exemple 4.5 En utilisant une re-paramétrisation, on peut vérifier que les gaussiennes univariées s'inscrivent dans ce schéma :

$$\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right)\,,$$

La mesure dominante ν est la mesure de Lebesgue, $T_1(x) = x, T_2(x) = -x^2$, $\theta = (\mu/\sigma^2, 1/(2\sigma^2))^t$, et

$$Z(\theta) = \exp\left(\frac{\mu^2}{2\sigma^2}\right)\sqrt{2\pi\sigma^2}\,.$$

Exemple 4.6 Les lois Gamma sont paramétrées par $]0,\infty)^2$, la loi $P_{p,\lambda}$ est définie par la densité sur $]0,\infty$:

$$\begin{aligned} p_{p,\lambda}(x) &= \frac{\lambda^p x^{p-1}\mathrm{e}^{-\lambda x}}{\Gamma(p)} \\ &= \exp\left(-\lambda x + (p-1)\log x - (\log\Gamma(p) - \lambda\log(p))\right)\,. \end{aligned}$$

On est dans le cadre des modèles exponentiels avec

$$\theta = \begin{pmatrix} -\lambda \\ p-1 \end{pmatrix} \quad \text{et} \quad T(x) = \begin{pmatrix} x \\ \log x \end{pmatrix}\,.$$

Dans ce chapitre, et le suivant nous collerons à la paramétrisation classique, avec p paramètre de forme, et λ paramètre d'intensité.

On peut formuler dans le cadre des modèles exponentiels en forme canonique, les gaussiennes multivariées, les lois gamma, les lois de Poisson, les lois géométriques, et bien d'autres familles de lois usuelles.

Convexité du modèle exponentiel

La convexité joue un rôle essentiel dans l'étude des modèles exponentiels, mais aussi dans la mise en oeuvre des méthodes d'estimation.

THÉORÈME 4.7 *Soient ν une mesure σ-finie dominante sur $(\mathcal{X}, \mathcal{F})$, et $T : \mathcal{X} \to \mathbb{R}^d$ $\mathcal{F}$ mesurable. S'il n'est pas vide, l'ensemble*

$$\widetilde{\Theta} = \left\{ \theta : \theta \in \mathbb{R}^d, Z(\theta) := \int_{\mathcal{X}} \mathrm{e}^{\langle \theta, T(x) \rangle} \nu(\mathrm{d}x) < \infty \right\}$$

(le domaine de la fonction de partition) est convexe.

La fonction $\theta \mapsto \log Z(\theta)$ est convexe sur le domaine $\widetilde{\Theta}$.

PREUVE. La preuve s'appuie sur l'inégalité de Hölder. Soit θ et θ', tels que $Z(\theta) < \infty$ et $Z(\theta') < \infty$. Soit $\lambda \in [0, 1]$

$$\begin{aligned} \int \mathrm{e}^{\langle \lambda\theta + (1-\lambda)\theta', T(x) \rangle} \nu(\mathrm{d}x) &= \int \left(\mathrm{e}^{\langle \theta, T(x) \rangle} \right)^{\lambda} \left(\mathrm{e}^{\langle \theta', T(x) \rangle} \right)^{1-\lambda} \nu(\mathrm{d}x) \\ &\leq \left(\int \mathrm{e}^{\langle \theta, T(x) \rangle} \nu(\mathrm{d}x) \right)^{\lambda} \times \left(\int \mathrm{e}^{\langle \theta', T(x) \rangle} \nu(\mathrm{d}x) \right)^{1-\lambda} . \end{aligned}$$

On a donc

$$\log Z\left(\lambda\theta + (1-\lambda)\theta'\right) \leq \lambda \log Z(\theta) + (1-\lambda) \log Z(\theta') .$$

□

Si Θ est non-vide, on peut choisir $\theta_0 \in \Theta$ et utiliser P_{θ_0} comme dominante. La densité de P_θ par rapport à P_{θ_0} est alors

$$\exp\left(\langle \theta - \theta_0, T(X) \rangle - \log \frac{Z(\theta)}{Z(\theta_0)} \right) ,$$

on peut alors changer de paramétrisation et utiliser $\theta - \theta_0$ à la place de θ. Le nouvel espace de paramètres $\Theta - \theta_0$ contient alors 0.

Dans la suite, nous supposerons que la dominante est une loi du modèle, et donc que $0 \in \Theta$, on aura $\nu = P_0$.

Identifiabilité du modèle exponentiel

Rappelons qu'un modèle ou une expérience statistique est dite *identifiable* si deux paramètres distincts définissent des lois distinctes. Dans les modèles exponentiels, l'identifiabilité est liée à l'irredondance des statistiques $T(X)$.

DÉFINITION 4.8 (MODÈLE MINIMAL OU DE PLEIN RANG) Un modèle exponentiel est minimal ou de plein rang si et seulement si pour tout $c \in \mathbb{R}^k$, le vecteur aléatoire $\langle c, T(X) \rangle$ n'est pas ν-presque partout constant.

PROPOSITION 4.9 *Un modèle exponentiel en forme canonique est minimal (ou de plein rang) si et seulement si il est identifiable.*

PREUVE.
$\Rightarrow$) Si $\theta \neq \theta'$ et $p_\theta(x) = p_{\theta'}(x)$ ν-presque partout, alors $\langle \theta - \theta', T(x)\rangle = \log Z(\theta)/Z(\theta')$ ν-pp. Ce qui signifie que ν-pp, $\langle \theta - \theta', T(x)\rangle$ est constant, ce qui contredit l'hypothèse de minimalité.
$\Leftarrow$) Si un modèle n'est pas minimal, il existe $c \neq 0$ et $b \in \mathbb{R}$, tel que $P_0\{\langle c, T(x)\rangle = b\} = 1$. Dans ce cas θ et $\theta + c$ désignent la même loi de probabilité. Le modèle n'est donc pas identifiable. □

Régularité de la fonction de partition

Si l'ouvert Θ est non vide, la fonction de partition est infiniment différentiable sur Θ. Et les différentielles de $\log Z$ en θ correspondent aux moments de P_θ. Ces propriétés, résumées dans la proposition suivante, s'avèrent très utiles pour aborder les problèmes d'estimation, elles motivent les méthodes de moments ou de vraisemblance (qui coïncident ici) et permettent d'en étudier les propriétés.

THÉORÈME 4.10 *La fonction $\theta \mapsto \log Z(\theta)$ est C^∞ sur $\Theta = \widetilde{\Theta}^\circ$.*
De plus

$$\nabla \log Z(\theta) = \mathbb{E}_{P_\theta} T(X)$$

et

$$\nabla^2 \log Z(\theta) = \mathrm{var}_{P_\theta}(T(X))\,.$$

Les preuves sont des exercices de convergence dominée. Pour donner une idée des arguments, on vérifie la première identité en dimension $d = 1$.

La proposition suivante, intéressante en soi, jouera un rôle dans la preuve du théorème.

PROPOSITION 4.11 *Dans un modèle exponentiel canonique, si $\theta \in \Theta$ (θ appartient à l'intérieur du domaine de la fonction de partition), alors pour tout $i \leq d$, pour tout $k \in \mathbb{N}$,*

$$\mathbb{E}_\theta \left[|T_i(X)|^k \right] < \infty\,.$$

PREUVE. Sans perdre en généralité, on se contente de vérifier la proposition en dimension 1.

On suppose $\theta \in \Theta$. Donc pour $|h| \leq h_0$, $\theta + h, \theta - h \in \Theta$ aussi. Cela entraine que sous P_θ, $\exp(|hT(X)|)$ est intégrable, et que donc toutes les puissances de $|T(x)|$ le sont aussi. On peut donc affirmer que

$$\int_{\mathcal{X}} \mathrm{e}^{\theta T(x)} |T(x)|^k \nu(\mathrm{d}x) < \infty$$

pour tout $\theta \in \Theta$ et $k \in \mathbb{N}$. □

PREUVE. Pour établir la première identité en dimension 1, il suffit de s'intéresser à la dérivée de $Z(\theta)$.

$$\frac{Z(\theta + h) - Z(\theta)}{h} = \int_{\mathcal{X}} \mathrm{e}^{\theta T(x)} \frac{\mathrm{e}^{hT(x)} - 1}{h} \nu(\mathrm{d}x$$

Pour x, h fixés avec $|h| < h_0$, avec $[\theta - h_0, \theta + h_0] \subset \Theta$,

$$\begin{aligned} \frac{\left|\mathrm{e}^{hT(x)} - 1\right|}{|h|} &\leq |T(x)| \mathrm{e}^{|h_0 T(x)|} \\ &\leq |T(x)| \left(\mathrm{e}^{h_0 T(x)} + \mathrm{e}^{-h_0 T(x)} \right)\,. \end{aligned}$$

La proposition technique montre que

$$\int_{\mathcal{X}} \mathrm{e}^{\theta T(x)} |T(x)| \left(\mathrm{e}^{h_0 T(x)} + \mathrm{e}^{-h_0 T(x)} \right) \nu(\mathrm{d}x) < \infty\,,$$

on peut donc invoquer le théorème de convergence dominée et $\lim_{h\to 0} \frac{\mathrm{e}^{hT(x)}-1}{h} = T(x)$ pour conclure que

$$\lim_{h\to 0} \frac{Z(\theta+h)-Z(\theta)}{h} = \int_{\mathcal{X}} \mathrm{e}^{\theta T(x)} T(x) \nu(\mathrm{d}x)\,.$$

Ce qui établit $\nabla \log Z(\theta) = \mathbb{E}_\theta T(X)$ (en dimension 1). □

La combinaison du théorème 4.10 et de la proposition 4.9 nous livre une autre caractérisation de l'identifiabilité.

PROPOSITION 4.12 *Un modèle exponentiel est identifiable si et seulement si en tout $\theta \in \Theta$, la différentielle seconde de $\log Z$ est inversible (le Hessien est défini positif).*

PREUVE. Supposons qu'en $\theta \in \Theta$, le Hessien de $\log Z$ ne soit pas défini positif. Il existe alors $c \in \mathbb{R}^k \setminus \{0\}$ tel que

$$c^t \nabla^2 \log Z(\theta) c = 0\,.$$

D'après le Théorème 4.10, cela signifie que

$$c^t \operatorname{cov}_\theta(T) c = \operatorname{var}_\theta(\langle c, T\rangle) = 0\,.$$

Cela entraine que $\langle c, T\rangle$ est P_θ-presque surement constant, que le modèle n'est pas minimal et donc pas identifiable (Proposition 4.9).

La réciproque se démontre en notant que la chaine d'implications peut etre renversée. □

Entropie relative dans un modèle exponentiel

La paramétrisation semble si naturelle dans les modèles exponentiels qu'on est tenté de cacher les lois derrière les paramètres. Il est pourtant très utile de disposer d'outils qui permettent de comparer les lois indépendamment de la paramétrisation. Les divergences d'information sont l'outil adéquat. Nous avons déjà utilisé la distance en variation.

L'*entropie relative* permet elle aussi de comparer des lois plutôt que des paramètres. Ce n'est pas une distance entre lois mais elle possède des propriétés remarquables. L'entropie relative entre deux éléments d'une famille exponentielle se lit directement sur la fonction $\log Z(\cdot)$:

$$D(P_\theta \| P_{\theta'}) = \log Z(\theta') - \log Z(\theta) + \langle \nabla \log Z(\theta), \theta - \theta' \rangle\,.$$

Lorsque θ' tend vers θ,

$$D(P_\theta \| P_{\theta'}) \sim \frac{1}{2}(\theta-\theta')^t \nabla^2 \log Z(\theta)(\theta-\theta')\,.$$

Ce calcul fournit une preuve concise de la relation entre identifiabilité et injectivite de $\theta \mapsto \nabla \log Z(\theta)$.

PROPOSITION 4.13 *Si un modèle exponentiel est identifiable, alors*

$$\Lambda : \theta \mapsto \Lambda(\theta) = \nabla \log Z(\theta)$$

est injective sur Θ.

PREUVE. On vérifie la contraposée.

D'après l'expression de l'entropie relative dans les modèles exponentiels,

$$D(P_\theta \| P_{\theta'}) + D(P_{\theta'} \| P_\theta) = \langle \nabla \log Z(\theta) - \nabla \log Z(\theta'), \theta - \theta' \rangle\,.$$

Si

$$\nabla \log Z(\theta) = \nabla \log Z(\theta')$$

alors

$$D(P_\theta \| P_{\theta'}) + D(P_{\theta'} \| P_\theta) = 0\,.$$

Comme ces deux quantités sont positives ou nulles, elles sont toutes les deux nulles et donc $P_\theta = P_{\theta'}$.

Le fait que l'injectivité implique l'identifiabilité est trivial (compte tenu de l'observation $\nabla \log Z(\theta) = \mathbb{E}_\theta[T(X)]$). □

Modèles exponentiels et méthode des moments

Nous sommes maintenant outillés pour vérifier que la méthode des moments fournit une recette pour construire des estimateurs dans le cadre des modèles exponentiels.

THÉORÈME 4.14 *Soit $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \Theta \subseteq \mathbb{R}^d)$ un modèle exponentiel en forme canonique de mesure dominante ν, défini par les vraisemblances*

$$p_\theta(x) = \exp\left(\langle \theta, T(x)\rangle - \log Z(\theta)\right)\,,$$

où T est une fonction mesurable de $\mathbb{R}^p$ dans $\mathbb{R}^d$, et où

$$\Theta := \left\{\theta : \theta \in \mathbb{R}^d, \quad \int_{\mathcal{X}} \exp(\langle \theta, T(x)\rangle)\nu(dx) < \infty\right\}^\circ\,.$$

Le modèle est identifiable si et seulement si

$$\Lambda : \theta \mapsto \Lambda(\theta) = \nabla \log Z(\theta)$$

est un difféomorphisme de Θ dans $\Lambda(\Theta)$.

Si le modèle est identifiable, $\Lambda(\Theta)$ est un ouvert de $\mathbb{R}^d$.

Nous avons vérifié que Λ est continuement différentiable et meme plus (Théorème 4.10). Nous avons vérifié que le modèle est identifiable si et seulement si la différentielle de Λ est inversible en tout $\theta \in \Theta$ (proposition 4.12). Nous pouvons donc invoquer le théorème d'inversion locale (rappelé plus bas) en tout $\theta \in \Theta$ pour affirmer que pour tout $\theta \in \Theta$, il existe un voisinage ouvert V de θ et un voisinage ouvert W de $\Lambda(\theta)$, tel que Λ soit bijective de V dans W, et d'inverse différentiable en $\Lambda(\theta)$. On note au passage que $\Lambda(\Theta)$ est ouvert.

THÉORÈME 4.15 (THÉORÈME D'INVERSION LOCALE) *Soit E un espace euclidien, $U \subseteq E$ ouvert, $f : U \to E$ continument différentiable, et $y_0 \in U$.*

Si la différentielle de f en y_0 $(Df(y_0))$ est inversible, d'inverse $(Df(y_0))^{-1}$ alors

i) il existe un voisinage ouvert $V \subseteq U$ de y_0 et un voisinage ouvert W de $x_0 = f(y_0)$, tel que f est bijective de V vers W,

ii) l'inverse g de f sur W est différentiable en x_0, de différentielle $(Df(y_0))^{-1}$.

Voir l'appendice E pour quelques rappels d'analyse.

Nous avons fait un peu plus que vérifier que les conditions du théorème d'inversion locale en tout $\theta \in \Theta$, nous avons aussi établi que Λ est injective sur Θ. Nous somme donc en mesure d'invoquer le théorème d'inversion globale ci-dessous.

THÉORÈME 4.16 (THÉORÈME D'INVERSION GLOBALE) *Soit E un espace euclidien, $U \subseteq E$ ouvert, $f : U \to E$ continument différentiable (de classe C^1) et injective. Si la différentielle de f est inversible en tout point de U, alors $f(U)$ est ouvert dans E, et f admet une inverse continument différentiable sur $f(U)$ (f est un difféomorphisme).*

Le théorème 4.14, justifie l'emploi de la méthode des moments : un modèle exponentiel minimal en forme canonique remplit les conditions suffisantes discutées à la fin de la sous-section 4.2.

Sur le papier, nous disposons d'une méthode d'estimation : il « suffit » de résoudre l'équation : en ϑ

$$\overline{T}_n = \nabla \log Z(\vartheta)\,,$$

soit de définir l'estimateur comme $\Lambda^{-1}(\overline{T}_n)$.

Il peut s'agir d'un problème calculatoire délicat, voire intraitable. Non seulement, il se peut que l'équation n'ait pas de solution explicite (le cas gaussien est plutôt un heureux accident), mais le simple calcul de $Z(\theta)$ pour θ peut s'avérer un problème difficile. De fait, les modèles graphiques (*graphical models*) qui peuvent être considérés comme un regard particulier sur les modèles exponentiels forment une branche très active de la statistique computationnelle.

EXEMPLE 4.17 On considère encore le modèle exponentiel défini par les lois Gamma :

$$p_{p,\lambda}(x) = \exp(-\lambda x + (p-1)\log(x) - (\log\Gamma(p) - p\log\lambda))$$

pour $x \in]0,\infty)$. Il s'agit d'un modèle exponentiel en forme (presque) canonique avec pour statistiques suffisantes $(\log X, -X)^t$. On a $\Lambda(p,\lambda)^t = \log Z(p,\lambda)^t = \log\Gamma(p) - p\log\lambda$ et

$$\nabla\Lambda\begin{pmatrix}p\\ \lambda\end{pmatrix} = \nabla \log Z\begin{pmatrix}p\\ \lambda\end{pmatrix} = \begin{pmatrix}\frac{\Gamma'(p)}{\Gamma(p)} - \log(\lambda)\\ -\frac{p}{\lambda}\end{pmatrix}.$$

Le système d'équations en p, λ :

$$\begin{aligned}\frac{\Gamma'(p)}{\Gamma(p)} - \log(\lambda) &= (\overline{\log X})_n\\ \frac{p}{\lambda} &= \overline{X}_n\end{aligned}$$

n'admet pas de solution explicite.

Dans la section suivante, nous allons voir la résolution de l'équation $\overline{T}_n = \nabla \log Z(\vartheta)$ comme un cas de maximisation de la vraisemblance. Comme la vraisemblance est ici une fonction concave de θ qui appartient lui-même à un ensemble convexe, cette façon de voir permet d'utiliser les outils de l'optimisation convexe. Elle ouvre la voie aux techniques de résolution, ou si nécessaire, aux techniques de relaxation.

4.5 PERFORMANCE DE LA MÉTHODE DES MOMENTS DANS LES MODÈLES EXPONENTIELS

On peut résumer les performances des moments décrites ici en rappelant que si un modèle exponentiel canonique est minimal, alors la suite des estimateurs des moments $(\Lambda^{-1}(\overline{T}_n))$ est consistante.

On peut raffiner par un résultat de normalité asymptotique. Pour décrire la covariance asymptotique des estimateurs au maximum de vraisemblance, nous introduisons les *fonctions score*, il s'agit du gradient de la log-vraisemblance par rapport à θ, $\nabla\ell_n(\theta)$. On note au passage que pour tout $\vartheta \in \Theta$,

$$\nabla\ell_n(\vartheta) = n(\overline{T}_n - \nabla \log Z(\vartheta))\,.$$

S'il existe ϑ tel que $\nabla\ell_n(\vartheta) = 0$, alors par convexité de $-\ell_n(\cdot)$, ce ϑ est un maximum global de la vraisemblance. Si $\log Z(.)$ est strictement convexe, ce maximum de vraisemblance est unique. Le maximum de vraisemblance est solution de l'équation $\overline{T}_n = \nabla \log Z(\vartheta)$. La méthode des moments que nous venons de décrire, dans ces modèles exponentiels minimaux en forme canonique, coïncide avec la méthode du maximum de vraisemblance.

Dans un modèle exponentiel minimal, pour tout $\theta \in \Theta$, la fonction score en θ est centrée :

$$\mathbb{E}_\theta \left[\nabla \ell_n(\theta)\right] = \mathbb{E}_\theta \left[n(\overline{T}_n - \nabla \log Z(\theta))\right] = 0\,.$$

On peut calculer la covariance de la fonction score, qu'on appelle l'*information de Fisher* pour le modèle échantillonné n fois en θ :

$$I_n(\theta) := \mathbb{E}_\theta \left[\nabla \ell_n(\theta) \nabla \ell_n(\theta)^t\right] = n I_1(\theta)\,.$$

Dans la suite, on notera $I(\theta) := I_1(\theta)$, l'information de Fisher du modèle échantillonné une fois. On a

$$I_1(\theta) = \nabla^2 \log Z(\theta)$$

Du théorème central limite, se déduit la normalité asymptotique de $\sqrt{n}\left(\overline{T}_n - \nabla \log Z(\theta)\right)$, sous $P_\theta^{\otimes \mathbb{N}}$:

$$\sqrt{n}\left(\overline{T}_n - \nabla \log Z(\theta)\right) \rightsquigarrow \mathcal{N}\left(0, I(\theta)\right)\,.$$

Si on note ψ la fonction $\theta \mapsto \nabla \log Z(\theta)$, l'estimateur au maximum de vraisemblance de θ est donné par $\psi^{-1}(\overline{T}_n)$. Le fait que ψ soit un difféomorphisme et la méthode delta (voir Appendice C.3) indiquent que $\sqrt{n}\big(\widehat{\theta}_n - \theta\big)$ est asymptotiquement normale de matrice de covariance

$$(\nabla^2 \log Z(\theta))^{-1} I(\theta) (\nabla^2 \log Z(\theta))^{-1} = I(\theta)^{-1}\,.$$

On peut conclure.

THÉORÈME 4.18 (NORMALITÉ ASYMPTOTIQUE) *Si un modèle exponentiel canonique est minimal, alors l'estimateur du maximum de vraisemblance est aussi un estimateur des moments. La suite des estimateurs au maximum de vraisemblance est consistante et après centrage et normalisation, elle est asymptotiquement normale :*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}\left(0, I(\theta)^{-1}\right)\,.$$

4.6 REMARQUES BIBLIOGRAPHIQUES

L'étude de la méthode des moments décrite ici est inspirée de l'ouvrage de VAN DER VAART [4]. On y trouve en particulier une démonstration du théorème d'inversion locale.

L'entropie relative est un cas particulier de divergence d'information. Parmi les autres divergences très utiles en statistique figurent la distance de Hellinger ou son carré, la distance en variation, la divergence χ^2. On trouve dans l'ouvrage de CSISZÁR et SHIELDS [2] une exploration en profondeur des liens entre ces divergences d'information.

L'entropie relative joue un rôle essentiel en théorie des grandes déviations. Une étude détaillée de l'entropie relative et de ses relations avec la topologie de la convergence en loi est présentée par DEMBO et ZEITOUNI [3].

L'introduction aux modèles exponentiels est inspirée de BICKEL et DOKSUM [1].

Références

[1] P. BICKEL et K. DOKSUM. **Mathematical statistics**. San Francisco, Calif. : Holden-Day Inc., 1976. MR : 56\#1513.

[2] I. CSISZÁR et P. SHIELDS. **Information theory and statistics : A tutorial**. Now Publishers Inc, 2004.

[3] A. DEMBO et O. ZEITOUNI. **Large deviation techniques and applications**. New York : Springer, 1998.

[4] A. VAN DER VAART. **Asymptotic statistics**. Cambridge University Press, 1998.

5.1 MOTIVATIONS

Dans les modèles exponentiels, nous avons vu que la méthode du maximum de vraisemblance est aussi une méthode de moments. Cette méthode du maximum de vraisemblance est étudiée un peu plus en détail dans ce chapitre. On précise en particulier quelques conditions portant sur le support convexe de la loi de la statistique suffisante T qui permettent de montrer que dans certains cas, le maximum de vraisemblance est bien défini avec probabilité 1.

Le point de vue du maximum de vraisemblance nous conduit à étudier le phénomène de Wilks et à construire des régions de confiance de taux de couverture asymptotique garanti, à partir de la normalité asymptotique de l'estimateur du maximum de vraisemblance.

Nous terminons ce chapitre en examinant les rapports entre statistique et optimisation. Il est tentant de dire que le travail du statisticien consiste à réduire un problème d'inférence à un problème d'optimisation à confier aux numériciens. La méthode à un pas décrite en fin de chapitre montre que les problèmes d'optimisation issus de la statistique sont de nature particulière et qu'ils doivent etre étudié d'une manière adaptée.

5.2 MODÈLES DOMINÉS ET VRAISEMBLANCE (BIS)

Dans toute cette section Ω est un espace métrique complet séparable, et $\mathcal{F}$ la tribu des boréliens.

Une mesure ν' sur $(\Omega, \mathcal{F})$ est dite *absolument continue* par rapport à une autre mesure σ-finie ν, si et seulement si pour tout $A \in \mathcal{F}$, $\nu(A) = 0 \Rightarrow \nu'(A) = 0$. On notera cela $\nu' \trianglelefteq \nu$.

L'absolue continuité garantit l'existence d'une densité (ou dérivée de Radon-Nykodym) notée $\frac{\mathrm{d}\nu'}{\mathrm{d}\nu}$ sur Ω qui vérifie

$$\forall A \in \mathcal{F}, \qquad \nu'(A) = \int \mathbb{I}_A(x) \frac{\mathrm{d}\nu'}{\mathrm{d}\nu}(x) \mathrm{d}\nu(x) \,.$$

Une mesure et en particulier une loi de probabilité est complètement définie par une densité (mais une mesure peut avoir plusieurs densités, deux densités peuvent représenter la même mesure, elles diffèrent éventuellement sur un sous-ensemble de ν-mesure nulle).

DÉFINITION 5.1 Une collection $\mathcal{P}$ de lois de probabilités sur $(\Omega, \mathcal{F})$ est *dominée* si il existe une mesure *σ-finie ν* sur $(\Omega, \mathcal{F})$ telle que tout $P \in \mathcal{P}$ soit *absolument continue* par rapport à ν ($\forall P \in \mathcal{P}, P \trianglelefteq \nu$).

Certaines collections de lois ne sont pas pas dominées. L'ensemble de toutes les lois de probabilité sur $\mathbb{R}$ par exemple n'est pas dominé.

EXEMPLE 5.2 L'ensemble des lois gaussiennes sur $\mathbb{R}^d$ avec une matrice de covariance définie positive est dominé (par exemple par la mesure de Lebesgue sur $\mathbb{R}^d$). Si on ajoute les lois gaussiennes à matrice de covariance pas toujours définie positive, le modèle n'est plus dominé.

Dans le cadre des modèles/expériences statistiques, on parle de mesure dominante. Un modèle dominé admet en fait une infinité de mesures dominantes possibles, il faut en choisir une. Ce choix est un choix de convenance. On peut vérifier qu'il n'a pas d'effet sur les méthodes d'inférence qui utilisent des méthodes de vraisemblance.

Une fois le choix de la mesure dominante arrêté, il faut « choisir » pour chaque loi de $\mathcal{P}$ une densité par rapport à la dominante. Le choix d'une version de la densité plutôt que d'une autre peut modifier les performances de la méthode. Un choix pathologique peut dégrader ces performances de manière spectaculaire. Si le modèle est paramétré par Θ, on notera (souvent) p_θ la densité de la loi $P_\theta, \theta \in \Theta$. La vraisemblance est une fonction de $\Theta \times \Omega$ qui à (θ, x) associe $p_\theta(x)$. Dans une expérience produit, la vraisemblance est une fonction sur $\Theta \times \Omega \times \ldots \times \Omega$ qui à $\theta, x_1, \ldots, x_n$ associe $\prod_{i=1}^n p_\theta(x_i)$.

EXEMPLE 5.3 Pour les gaussiennes univariées on peut prendre comme dominante la mesure de Lebesgue, la vraisemblance en $(\theta = (\mu, \sigma), x_1, \ldots, x_n)$ s'écrira

$$\prod_{i=1}^n \frac{\phi\left(\frac{x_i - \mu}{\sigma}\right)}{\sigma} = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{n(\overline{X}_n - \mu)^2}{2\sigma^2} - \frac{\sum_{i=1}^n \left(x_i - \overline{X}_n\right)^2}{2\sigma^2} \right)$$

ou encore

$$\frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}+2\frac{\mu\sum_{i=1}^n x_i}{2\sigma^2}-\frac{n\mu^2}{2\sigma^2}\right).$$

On peut aussi choisir comme mesure dominante la gaussienne standard $\mathcal{N}(0,1)$.

Exemple 5.4

Exemple 5.5

5.3 maximisation de la vraisemblance

L'inférence par la méthode du *maximum de vraisemblance* (la maximisation de la vraisemblance) consiste, étant donné l'échantillon $x_1,\dots,x_n$, à choisir $\widehat{\theta}$ qui maximise $p_\theta(x_1,\dots,x_n)$ dans Θ :

$$\widehat{\theta} := \arg\max\{p_\theta(x_1,\dots,x_n), \theta\in\Theta\}$$

Cette démarche doit d'abord être motivée. Pourquoi devrait-elle être prometteuse ? On doit ensuite en étudier les propriétés. Est-elle bien définie ? (Le maximum existe-t-il ? Est-il unique ?) Le maximum est-il calculable efficacement ?

Si le maximum de vraisemblance est bien défini (au moins avec une probabilité qui tend vers 1 quand la taille de l'échantillon tend vers l'infini), s'agit-il d'une méthode consistante d'estimation ? Est-elle (au moins) asymptotiquement gaussienne ? Comment se comporte son risque quadratique ? Cette méthode possède-t-elle des vertus particulières ?

Les remarques sur l'entropie relative au chapitre précédent fournissent une motivation informelle de l'estimation au maximum de vraisemblance. Supposons les données collectées indépendamment sous P_θ. Soit $P_{\theta'}$, une autre loi. On suppose $D(P_\theta\|P_{\theta'})<\infty$, c'est-à-dire que $\log p_\theta(X)/p_{\theta'}(X)$ est intégrables sous P_θ. La différence entre les log vraisemblances notées

$$\ell_n(\theta,X_1,\dots,X_n) \qquad \text{et} \qquad \ell_n(\theta',X_1,\dots,X_n)$$

est alors une somme de variables aléatoires indépendantes, intégrables sous P_θ, d'espérance $D(P_\theta\|P_{\theta'})$. La loi des grands nombres nous indique que P_θ presque sûrement,

$$\frac{1}{n}\left(\ell_n(\theta,X_1,\dots,X_n)-\ell_n(\theta',X_1,\dots,X_n)\right)\to D(P_\theta\|P_{\theta'})>0$$

si $P_\theta\neq P'_\theta$. Si le modèle Θ est dénombrable, cela fournit une preuve de consistance. Dans le cas général il faut développer un argument formel.

Nous allons le faire dans le cadre des modèles exponentiels.

5.4 maximum de vraisemblance dans les modèles exponentiels canoniques

L'estimation par maximisation de vraisemblance est invoquée dans des cadres beaucoup plus généraux que celui des modèles exponentiels. Son étude et sa justification sont grandement facilités dans le cadre des modèles exponentiels. Elle permet de revisiter la méthode des moments décrites dans le chapitre précédent.

Le *support d'une loi de probabilité* sur $\mathbb{R}^k$ est le plus petit fermé de mesure 1 sous cette loi. Dans la suite, on définit le *support convexe d'une loi de probabilité* sur $\mathbb{R}^k$, comme le plus petit convexe fermé de probabilité 1 sous cette loi (est-ce une partie mesurable ? pourquoi ?).

Dans un modèle exponentiel de plein rang/minimal, toutes les lois $P_\theta,\theta\in\Theta$ ont même support, et donc meme support convexe, (en raison de l'absolue continuité des lois les unes par rapport aux autres). Ici nous nous intéresserons au support convexe de la loi de $T(X)$ plutot qu'à ce lui de la loi de X. Nos statistiques sont construites à partir de $T(X_1),\dots,T(X_n)$. On pourra donc parler de support convexe du modèle exponentiel.

Nous avons vu dans la section précédente que l'ensemble $\{\mathbb{E}_\theta T(X):\theta\in\Theta\}$ est un ouvert, et que c'est l'image de Θ par la fonction $\theta\mapsto\nabla\log Z(\theta)$. Nous allons voir dans cette section un autre point de vue et développer un argument *ad hoc* pour montrer que $\theta\mapsto\nabla\log Z(\theta)$ est un difféomorphisme.

D'une manière générale si la log-vraisemblance d'un n-échantillon s'écrit comme

$$\vartheta\mapsto\ell_n(\vartheta)=n\left(\langle\vartheta,\overline{T}_n(X)\rangle-\log Z(\vartheta)\right),$$

le problème de la maximisation de la vraisemblance se décrit comme

$$\begin{aligned} &\text{Entrée : } t \in \mathbb{R}^k \\ &\text{Résultat : } \widehat{\theta} = \arg\max_{\vartheta\in\Theta} [\langle \vartheta, t\rangle - \log Z(\vartheta)] \qquad \text{si existe.} \end{aligned}$$

Il faut avoir en tête que meme dans un modèle exponentiel en forme canonique, le maximum de vraisemblance n'est pas toujours bien défini.

EXEMPLE 5.6 Considérons le modèle des lois géométriques avec une paramétrisation canonique :

$$p_\theta(k) = \mathrm{e}^{-\theta k}(\mathrm{e}^\theta - 1) = \exp\left(-\theta k + \log(\mathrm{e}^\theta - 1)\right) \quad \text{pour} \quad k = 1, \ldots, \quad \theta \in \Theta =]0, \infty)\,.$$

Si l'échantillon est une suite de 1 (ce qui se produit avec probabilité > 0 sous $P_\theta, \theta \in \Theta$), la log-vraisemblance s'écrit

$$\ell_n(\vartheta) = n \log\left(\mathrm{e}^\vartheta - 1\right) - n\vartheta = n \log\left(1 - \mathrm{e}^{-\vartheta}\right)\,.$$

Elle est concave et croissante sur Θ. Elle tend vers son supremum (fini) lorsque $\vartheta \to \infty$, c'est à dire lorsque ϑ tend vers la frontière de Θ. Le maximum de vraisemblance n'est pas défini.

Cet exemple jouet est plein d'enseignements. Dans ce modèle exponentiel, $\log Z(\vartheta) = -\log(\mathrm{e}^\vartheta - 1)$ et $\nabla \log Z(\vartheta) = -\mathrm{e}^\vartheta/(\mathrm{e}^\vartheta - 1)$. La moyenne empirique de la statistique suffisante $-\overline{X}_n$ appartient à $(-\infty, -1]$. Il existe $\vartheta \in \Theta$ tel que $\nabla \log Z(\vartheta) = -\overline{X}_n$ si et seulement si $\overline{X}_n > 1$. Lorsqu'elle existe la solution de $\nabla \log Z(\vartheta) = -\overline{X}_n$ réalise alors le maximum de vraisemblance.

Nous verrons qu'il s'agit d'un phénomène général. Lorsque la statistique suffisante prend une valeur située sur la frontière de l'enveloppe convexe du support de la mesure dominante, il se peut que le maximum de vraisemblance ne soit pas défini. La probabilité de ce genre d'événement tend vers 0 lorsque la taille de l'échantillon tend vers l'infini.

Notons aussi que cet échantillon extrême met la méthode des moments en défaut.

Nos objectifs sont :

i) caractériser les situations où le maximum de vraisemblance n'est pas défini ;
ii) montrer que la probabilité de ces situations tend vers 0 lorsque la taille de l'échantillon tend vers l'infini ;
iii) montrer que dans certains modèles, cette probabilité est nulle.

Nous débutons par une observation élémentaire mais importante.

PROPOSITION 5.7 *Dans une famille exponentielle canonique, sur tout échantillon ω, la log-vraisemblance $\ell_n(\theta)$ est concave par rapport au paramètre θ. Si le modèle est identifiable, $\ell_n(\cdot)$ est strictement concave.*

PREUVE. Immédiate à partir de l'expression et de la convexité de $\theta \mapsto \log Z(\theta)$ sur Θ. □

Est-ce suffisant pour garantir que le maximum de vraisemblance existe ? Non, une fonction linéaire sur $\mathbb{R}$ est concave et pourtant elle n'admet pas de maximum (sauf si elle est nulle). Cela garantit tout de même que la région de $\mathbb{R}^k$ où le maximum de vraisemblance est éventuellement atteint est convexe.

PROPOSITION 5.8

i) Si l'équation $t = \nabla \log Z(\theta)$ admet une solution $\widetilde{\theta}$ dans Θ, cette solution maximise $\langle t, \theta\rangle - \log Z(\theta)$.
ii) Si $\widehat{\theta}$ maximise $\theta \mapsto \langle \theta, t\rangle - \log Z(\theta)$ à l'intérieur de Θ alors $t = \nabla \log Z(\widehat{\theta})$.

Dans l'étude des fonctions convexes, la *sous-différentielle* d'une fonction convexe f en $x \in \mathbb{R}^p$ est un sous-ensemble (non vide, compact, convexe) de $\mathbb{R}^p$ de points λ tels que pour tout y dans le domaine de f,

$$f(y) \geq f(x) + \langle \lambda, (y - x) \rangle \, .$$

On note $\partial f(x)$ la sous-différentielle de f en x. Les éléments de la sous-différentielle sont appelés *sous-gradients*.

Tout minimum local d'une fonction convexe est un minimum global. Un point x est un minimum de la fonction convexe f si et seulement si $0 \in \partial f(x)$.

Une fonction convexe f est différentiable en un point $x \in \mathbb{R}^p$ de l'intérieur de son domaine de définition si et seulement si sa sous-différentielle est réduite à un point.

PREUVE.

i) Pour tout θ,

$$\langle t, \widetilde{\theta} \rangle - \log Z(\widetilde{\theta}) - \langle t, \theta \rangle + \log Z(\theta) = \log Z(\theta) - \log Z(\widetilde{\theta}) - \left\langle \nabla \log Z(\widetilde{\theta}), \theta - \widetilde{\theta} \right\rangle \, .$$

Le membre droit est positif en raison de la convexité de $\theta \mapsto \log Z(\theta)$.

ii) Si $\widehat{\theta} \in \Theta$ vérifie $\langle \widehat{\theta}, t \rangle - \log Z(\widehat{\theta}) \geq \langle \theta, t \rangle - \log Z(\theta)$ pour tout $\theta \in \Theta$, on a pour tout $\theta \in \Theta$

$$\log Z(\theta) \geq \langle t, \theta - \widehat{\theta} \rangle + \log Z(\widehat{\theta}) \, .$$

Ceci signifie que t est un sous-gradient de $\theta \mapsto \log Z(\theta)$ en $\widehat{\theta}$, qui est réduit à $\nabla \log Z(\widehat{\theta})$ en raison de la différentiabilité dans Θ. □

Avec cette proposition, nous venons de vérifier que $\theta \mapsto \langle \theta, t \rangle - \log Z(\theta)$ atteint son supremum dans Θ si et seulement s'il existe $\widehat{\theta} \in \Theta$ tel que $t = \nabla \log Z(\widehat{\theta})$, autrement dit que la méthode du maximum de vraisemblance fonctionne exactement là où fonctionne la méthode des moments (lorsque le choix des moments est dicté par les statistiques suffisantes dans le formalisme canonique).

Dans la suite, on définit $\partial\Theta = \overline{\Theta} - \Theta$ la frontière de Θ. Notons que $\partial\Theta$ peut éventuellement contenir des points à l'infini.

Pour une suite (θ_n) d'éléments de Θ ouvert de $\mathbb{R}^d$, on s'accorde pour interpréter $\theta_n \to \partial\Theta$ de la façon suivante. On a soit $\theta_n \to \theta$ avec $\theta \notin \Theta$, soit $\|\theta_n\| \to \infty$.

Rappelons que si $\Theta \subseteq \mathbb{R}^d$ est ouvert et si $f : \Theta \to \mathbb{R}$ est continue et vérifie $\lim_{\theta \to \partial\Theta} f(\theta) = -\infty$, alors f atteint son maximum dans Θ. La proposition suivante est alors une conséquence immédiate de la concavité et de la régularité de la log-vraisemblance dans les modèles exponentiels de plein rang.

PROPOSITION 5.9 *Si la fonction log-vraisemblance $\theta \mapsto \ell_n(\theta)$ est strictement concave et si elle tend vers $-\infty$ lorsque θ tend vers $\partial\Theta = \overline{\Theta} \setminus \tilde{\Theta}$, alors le maximum de vraisemblance est atteint en un point unique de Θ.*

EXERCICE 5.10 Vérifier les conditions d'existence et d'unicité du maximum de vraisemblance dans les modèles Poisson, géométrique, multinomiaux.

Quels sont les supports convexes de la mesure dominante dans ces différents cas ?

Nous utiliserons le résultat suivant de convexité qui est un cas particulier du théorème de Hahn-Banach géométrique. En dimension finie, ce résultat peut etre vérifié par des arguments simples [1].

THÉORÈME 5.11 (HYPERPLAN SÉPARATEUR) *Soient A, B deux sous-ensembles convexes de $\mathbb{R}^d$. Si $A \cap B = \emptyset$ alors il existe $c \in \mathbb{R}^d$, tel que*

$$\sup_{x \in A} \langle x, c \rangle \leq \inf_{x \in B} \langle x, c \rangle \, .$$

THÉORÈME 5.12 *Dans un modèle exponentiel canonique minimal où $\Theta = \widetilde{\Theta}$ (le domaine $\widetilde{\Theta}$ de la fonction de partition est un ouvert), si $t_0 \in \mathbb{R}^k$ appartient à l'intérieur du support convexe de la loi de $T(X)$ pour le modèle, la fonction*

$$\theta \mapsto \langle \theta, t_0 \rangle - \log Z(\theta)$$

est strictement concave et elle tend vers $-\infty$ lorsque θ tend vers $\partial\Theta = \overline{\Theta} \setminus \Theta$.

REMARQUE 5.13 Dans un modèle exponentiel canonique minimal, $\widetilde{\Theta}$ n'est pas toujours ouvert. On peut choisir comme dominante une loi de densité proportionnelle à $\exp(-|x|)/(1+|x|^3)$ et comme statistique suffisante $|x|$. L'ensemble $\widetilde{\Theta}$ est alors $(-\infty, 1]$. Le support de la dominante est $\mathbb{R}$. L'ensemble des valeurs possibles de $\frac{\mathrm{d} \log Z(\theta)}{\mathrm{d}\theta}$ est borné.

Commençons par expliciter la condition « t appartient à l'intérieur C° du support convexe C de la mesure dominante ».

PROPOSITION 5.14 *Dans les conditions du théorème 5.12, si t appartient à l'intérieur C°, du support convexe C de la loi de $T(X)$ por le modèle, alors pour tout $c \in \mathbb{R}^k \setminus \{0\}$, pour tout $\theta \in \Theta$,*

$$P_\theta \left\{ \langle c, t \rangle < \langle c, T(X) \rangle \right\} > 0 \,.$$

PREUVE. Soit t un point de l'intérieur du support convexe C°, et $\theta \in \Theta$, Supposons qu'il existe $c \in \mathbb{R}^k \setminus \{0\}$, avec

$$P_\theta \left\{ \langle c, t \rangle \geq \langle c, T(X) \rangle \right\} = 1 \,.$$

Cela implique que le support convexe de la loi de T sous P_θ est inclus dans le demi-espace

$$\left\{ y : y \in \mathbb{R}^k, \langle c, t \rangle \geq \langle c, y \rangle \right\} \quad .$$

D'après le théorème de séparation, ceci contredit l'hypothèse selon laquelle t appartient à l'intérieur du support convexe de la loi de $T(X)$ sous P_θ. □

PREUVE.[Preuve du théorème 5.12] La stricte concavité a déjà été démontrée.

Comme t_0 appartient à l'intérieur du support convexe C de la mesure dominante, pour toute direction c et tout $\theta \in \Theta$,

$$P_\theta \left\{ \langle c, t_0 \rangle < \langle c, T(X) \rangle \right\} > 0$$

Considérons une suite $(\theta_m)_{m \in \mathbb{N}}$ d'éléments de $\Theta \subset \mathbb{R}^k$, qui tend vers $\partial\Theta$.

a) Si $\theta_n \to \eta \in \partial\Theta \cap \mathbb{R}^k$ ($\eta \notin \Theta$, et donc comme $\widetilde{\Theta}$ est supposé ouvert, $Z(\eta) = \infty$). Comme $\log Z$ est continue, on en conclut que $\log Z(\theta_m)$ tend vers l'infini lorsque θ_m tend vers η, et que donc

$$\langle \theta_m, t_0 \rangle \, - \log Z(\theta_m)$$

tend vers $-\infty$ lorsque θ_m tend vers η.

b) Considérons maintenant le cas où $\|\theta_m\|$ tend vers ∞. On suppose sans perdre en généralité que la mesure dominante est une probabilité P_0 (voir page 44).

Comme la suite $\theta_m / \|\theta_m\|$ est bornée (confinée dans la boule unité de $\mathbb{R}^k$), elle possède au moins un point d'accumulation. Appelons le η (encore), et supposons que $\log Z(\eta) < \infty$ (si nécessaire, prendre un multiple de η). Il existe une sous-suite de la suite (θ_m) qui se dirige vers l'infini dans la direction de η. Sans perdre en généralité, admettons que c'est toute la suite (θ_m) qui se dirige vers l'infini dans la direction de η.

Dans le calcul qui suit, $\delta > 0$ est tel que

$$\mathbb{P}_0 \left\{ \langle \eta, T(X) \rangle \geq \langle \eta, t_0 \rangle + \delta \right\} > 0 \,.$$

L'existence de δ est garantie par l'hypothèse que $t_0 \in C^\circ$.

$$\begin{aligned}\log Z(\theta_m) &= \log \mathbb{E}_0\left[\mathrm{e}^{\langle \theta_m, T(X)\rangle}\right] \\ &\geq \log \mathbb{E}_0\left[\mathrm{e}^{\|\theta_m\|\langle \frac{\theta_m}{\|\theta_m\|}, T(X)\rangle} \mathbb{I}_{\langle \frac{\theta_m}{\|\theta_m\|}, T\rangle \geq \langle \frac{\theta_m}{\|\theta_m\|}, t_0\rangle + \delta}\right] \\ &\geq \langle \theta_m, t_0\rangle + \|\theta_m\|\delta + \log P_0\left\{\langle \frac{\theta_m}{\|\theta_m\|}, T\rangle \geq \langle \frac{\theta_m}{\|\theta_m\|}, t_0\rangle + \delta\right\}.\end{aligned}$$

Comme

$$\limsup_m \log P_0\left\{\langle \frac{\theta_m}{\|\theta_m\|}, T\rangle \geq \langle \frac{\theta_m}{\|\theta_m\|}, t_0\rangle + \delta\right\}$$

est finie (convergence dominée, $\frac{\theta_m}{\|\theta_m\|} \to \eta \in \Theta$ et $t_0 \in C^\circ$), on en déduit que $\ell_n(\theta_m)$ tend aussi vers $-\infty$ lorsque (θ_m) tend vers $\partial\Theta$ dans la direction η.

□

Le théorème 5.12 entraîne le corollaire suivant.

COROLLAIRE 5.15 *Dans un modèle exponentiel minimal dont le domaine est ouvert, si la frontière du support convexe est de mesure nulle sous la dominante, alors le maximum de vraisemblance est presque sûrement bien défini.*

5.5 PHÉNOMÈNE DE WILKS ET RÉGIONS DE CONFIANCE

Dans toute cette section, on travaille sur un modèle exponentiel minimal en forme canonique de dimension k ($\Theta \subseteq \mathbb{R}^k$ où Θ est l'intérieur du domaine de la fonction de partition Z du modèle. Le vecteur des statistiques suffisantes est noté T).

Nous aurons encore recours à la notion d'*information de Fisher*. La matrice d'information de Fisher $I_n(\theta)$ est définie par

$$I_n(\theta) := \mathbb{E}_\theta\left[\nabla \ell_n(\theta) \nabla \ell_n(\theta)^t\right]$$

où le gradient de la log-vraisemblance $\ell_n(\vartheta, x_1, \ldots, x_n)$ est pris par rapport à ϑ en θ. Ce gradient est une fonction de $\overline{T}_n$, il vaut

$$n\left(\overline{T}_n - \nabla \log Z(\theta)\right).$$

Comme $\mathbb{E}_\theta[\overline{T}_n] = \nabla \log Z(\theta)$,

$$I_n(\theta) := n^2 \operatorname{cov}_\theta\left(\overline{T}_n\right) = n \operatorname{cov}_\theta(T) = n\nabla^2 \log Z(\theta) = nI(\theta).$$

Lorsqu'il s'agit de construire des régions de niveau de confiance asymptotique garanti, ou des tests de niveau asymptotique garanti, on s'intéresse à la construction de quantités pivotales, c'est à dire de quantités qui font intervenir à la fois l'estimande et des quantités empiriques mais dont la loi asymptotique est libre de l'estimande. Les résultats sur la normalité asymptotique du maximum de vraisemblance dans les modèles exponentiels nous fournissent déjà des éléments : le fait que

$$\sqrt{n} I(\theta)^{1/2}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0, \mathrm{Id}_k)$$

conduit naturellement à une construction de région de confiance,

$$\left\{\theta' : n\left\|I(\theta')\left(\widehat{\theta}_n - \theta'\right)\right\|^2 \leq q_{k,1-\alpha}\right\}$$

où $q_{k,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi χ^2_k. Cette région de confiance peut être rendue plus transparente en substituant à l'information $I(\theta')$ la quantité empirique $\widehat{I}_n = I(\widehat{q}_n)$. La région de confiance devient alors un ellipsoïde centré en $\widehat{\theta}_n$. Cette construction n'est pas la seule possible et pour des tailles d'échantillon modérées, ce n'est pas la plus recommandée : la convergence en loi du maximum de vraisemblance recentré vers une gaussienne n'est pas toujours rapide.

Le résultat suivant qui reste valable pour des modèles plus généraux, offre une technique simple et souple de construction de régions de confiance. Les régions de confiance ne sont pas nécessairement des ellipsoïdes, elles ne sont pas nécessairement symétriques autour du maximum de vraisemblance.

THÉORÈME 5.16 (PHÉNOMÈNE DE WILKS) *Dans un modèle exponentiel canonique minimal* $(P_\theta, \theta \in \Theta \subseteq \mathbb{R}^k)$*, pour tout* $\theta \in \Theta$*, sous* $P_\theta^{\otimes \mathbb{N}}$*,*

$$2nD\left(P_\theta \| P_{\widehat{\theta}_n}\right) \rightsquigarrow \chi^2_k$$

et

$$nD\left(P_\theta \| P_{\widehat{\theta}_n}\right) - \left(\ell_n(\widehat{\theta}_n) - \ell_n(\theta)\right) \xrightarrow{P} 0\,.$$

L'ingrédient essentiel de la preuve est une variante de la méthode delta.

THÉORÈME 5.17 (MÉTHODE DELTA AU SECOND ORDRE) *Soient*

i) $(a_n)_n$ *une suite positive qui tend vers l'infini.*
ii) $(X_n)_n$ *une suite de variables aléatoires à valeur dans* $\mathbb{R}^k$*.*
iii) X *une variable aléatoire à valeur dans* $\mathbb{R}^k$*.*
iv) $x \in \mathbb{R}^k$*.*
v) f *une fonction de* $\mathbb{R}^k$ *dans* $\mathbb{R}$*, deux fois différentiable en* x*, de différentielle nulle en* x*, et dont le Hessien en* x *est noté* $\nabla^2 f$*.*

Si

$$a_n\left(X_n - x\right) \rightsquigarrow X$$

alors

i) $X_n \xrightarrow{P} x$ *;*
ii) $a_n^2\left(f(X_n) - f(x)\right) - a_n^2 \frac{1}{2}\left(X_n - x\right)^t \nabla^2 f\left(X_n - x\right) \xrightarrow{P} 0$ *;*
iii) $a_n^2\left(f(X_n) - f(x)\right) \rightsquigarrow \frac{1}{2} X^t \nabla^2 f X$*.*

PREUVE.[Preuve du Théorème 5.17] La preuve est calquée sur la preuve de la méthode delta au premier ordre. La clause i) a déjà été établie.

Les hypothèses sur f entrainent l'existence d'une fonction R de $\mathbb{R}^k$ dans $\mathbb{R}$ telle que

$$f(y) = f(x) + \frac{1}{2}(y-x)^t \nabla^2 f(y-x) + R(y-x)$$

avec $|R(h)| = o(\|h\|^2)$. Soit,

$$\left(f(X_n) - f(x)\right) - \frac{1}{2}\left(X_n - x\right)^t \nabla^2 f\left(X_n - x\right) \quad = \quad R(X_n - x)$$

La preuve de la clause ii) consiste à vérifier que $a_n^2 R(X_n - x)$ converge en loi (et donc en probabilité) vers 0.

La tension uniforme (voir Théorème C.13) de la suite $\left(a_n(X_n - x)\right)_{n\in\mathbb{N}}$ entraine que pour tout $\eta > 0$, il existe $M(\eta) > 0$, tel que pour tout n

$$\mathbb{P}\left\{a_n^2 \|X_n - x\|^2 \geq M(\eta)\right\} \leq \eta\,.$$

Pour tout $\delta > 0$, il existe $\epsilon(\delta) > 0$ tel que

$$\|h\|^2 \leq \epsilon(\delta) \Rightarrow R(h) \leq \delta^2 \|h\|^2\,.$$

Pour n assez grand, $M(\eta) \leq a_n^2 \epsilon(\delta)$, d'où

$$\mathbb{P}\left\{a_n^2 R(X_n - x) \geq \delta^2 M(\eta)\right\} \leq \eta\,.$$

Si on choisit δ de façon à ce que $\delta^2 M(\eta) < \eta$, on a donc pour n assez grand

$$\mathbb{P}\left\{a_n^2 R(X_n - x) \geq \eta\right\} \leq \eta\,.$$

Comme on peut choisir η arbitrairement petit, on a établi la convergence en probabilité de $a_n^2 R(X_n - x)$ vers 0.

La clause iii) est là encore une conséquence directe du Lemme de Slutsky. Si la différence de deux suites aléatoires converge en probabilité vers 0, elles ont même limite en loi. □

Nous pouvons maintenant passer à la preuve du Théorème 5.16.

PREUVE. La fonction $\theta' \mapsto D(P_\theta \| P_{\theta'})$ possède un gradient nul en θ et son Hessien est égal à $I(\theta)$.

La normalité asymptotique du maximum de vraisemblance et la méthode delta au second ordre entrainent que sous $P_\theta^{\otimes \mathbb{N}}$

$$nD\left(P_\theta \| P_{\widehat{\theta}_n}\right) - \frac{n}{2}\left(\widehat{\theta}_n - \theta\right)^t I(\theta)\left(\widehat{\theta}_n - \theta\right) \xrightarrow{P} 0\,.$$

Ceci établit la première partie du théorème.

Notons tout d'abord que lorsque le maximum de vraisemblance est bien défini,

$$\ell_n(\widehat{\theta}_n) - \ell_n(\theta) = nD\left(P_{\widehat{\theta}_n} \| P_\theta\right)\,.$$

La fonction $\theta' \mapsto D(P_{\theta'} \| P_\theta)$ est C^∞ en θ, de gradient nul et de Hessien (lui aussi) égal à $I(\theta)$ en $\theta' = \theta$. Les mêmes arguments entrainent que sous $P_\theta^{\otimes \mathbb{N}}$

$$nD\left(P_{\widehat{\theta}_n} \| P_\theta\right) - \frac{n}{2}\left(\widehat{\theta}_n - \theta\right)^t I(\theta)\left(\widehat{\theta}_n - \theta\right) \xrightarrow{P} 0\,.$$

Ceci établit la seconde partie du théorème. □

Considérons maintenant la région aléatoire définie par

$$\widehat{A}_n(\alpha) := \left\{\theta' : \theta' \in \Theta, \ell_n(\widehat{\theta}_n) - \ell_n(\theta') \leq \frac{1}{2} q_{k,1-\alpha}\right\}\,.$$

Un corollaire immédiat du Théorème 5.16 suit.

COROLLAIRE 5.18 *Dans un modèle exponentiel canonique minimal $(P_\theta, \theta \in \Theta \subseteq \mathbb{R}^k)$, pour tout $\theta \in \Theta$, sous $P_\theta^{\otimes \mathbb{N}}$, la suite de régions $\widehat{A}_n(\alpha)$ est de niveau de confiance asymptotique $1 - \alpha$:*

$$\lim_n P_\theta^n\left\{\theta \in \widehat{A}_n(\alpha)\right\} = 1 - \alpha\,.$$

5.6 MÉTHODE À UN PAS (ONE-STEP) DANS LES MODÈLES EXPONENTIELS

Dans cette section, nous considérons encore un modèle exponentiel identifiable, en forme canonique. Nous conservons les notations des sections précédentes. L'espace des paramètres est noté $\Theta \subset \mathbb{R}^k$. On suppose $0 \in \Theta$. $\theta^0 \in \Theta$ désigne la valeur du paramètre sous laquelle on effectue l'échantillonnage (« la vraie »). La statistique suffisante est dénotée par $T(X)$. Elle est à valeur dans $\mathbb{R}^k$. On note $\ell_n(\vartheta)$ la log-vraisemblance en ϑ pour un n-échantillon (elle n'est pas normalisée). Lorsqu'on écrit $\nabla \ell_n(\vartheta)$, on désigne le gradient de la log-vraisemblance par rapport au paramètre (la fonction *score*) pris en ϑ. On note $\nabla^2 \ell_n(\vartheta)$ le Hessien de la log-vraisemblance par rapport au paramètre (l'information empirique) pris en ϑ.

En étudiant la méthode du maximum de vraisemblance dans les modèles exponentiels, nous avons réduit le problème de la construction d'un estimateur ponctuel au problème de l'optimisation d'une

fonction concave et très régulière (infiniment différentiable) sur un domaine convexe. On peut en retirer l'impression que le problème de l'estimation dans ces modèles exponentiels n'est plus une question de statistique, mais un problème technique de numéricien. Si le domaine d'optimisation n'est pas trop compliqué, l'analyse numérique met à notre disposition une quantité de méthodes éprouvées pour approcher efficacement le maximum de vraisemblance. Ces méthodes sont (en général) itérative. À chaque pas, on calcule le gradient de la log-vraisemblance, éventuellement son Hessien ou une approximation, et on met à jour une estimée courante. On peut se demander si le fait de ne calculer qu'une approximation du maximum de vraisemblance dégrade la performance statistique. Quand on pose cette question, on réalise que le statisticien qui utilise une méthode numérique n'est pas dans la même situation que l'analyste. Ce dernier optimise une fonction en principe parfaitement connue. Le statisticien doit maximiser

$$\vartheta \mapsto \langle \vartheta, \overline{T}_n \rangle - \log Z(\vartheta)$$

alors qu'il souhaiterait optimiser

$$\vartheta \mapsto \langle \vartheta, \nabla \log Z(\theta) \rangle - \log Z(\vartheta)\,.$$

Dans ces conditions, il n'est peut être pas utile d'optimiser complètement la log-vraisemblance et on peut se demander si un critère d'arrêt bien choisi pour les algorithmes itératifs ne permettrait pas d'épargner des ressources calculatoires sans dégrader les performances statistiques. Cette question se pose de manière répétée en statistique mais aussi en apprentissage.

Dans le reste de cette section, nous allons vérifier que convenablement initialisée, une méthode itérative classique, la méthode de Newton peut s'avérer très performante : il suffit d'une itération (un pas, *one step*) pour obtenir un estimateur, dont les performances sont asymptotiquement de même qualité que l'estimateur du maximum de vraisemblance.

En optimisation convexe, la méthode de Newton peut être brièvement décrite comme suit. Supposons f une fonction strictement convexe deux fois différentiable sur $\mathbb{R}^k$ (pas de problème de contraintes) et supposons que nous disposons de boites noires (oracles) capables en chaque $x \in \mathbb{R}^k$ de renvoyer gradient $\nabla f(x)$ et Hessien $\nabla^2 f(x)$. Le problème consiste à déterminer x^* qui minimise f sur $\mathbb{R}^k$, c'est-à-dire la solution de $\nabla f(x) = 0$. Pour une estimation initiale x^0, l'algorithme de Newton définit une suite $(x^k)_k$ par la récurrence :

$$x^{k+1} = x^k - \gamma_k (\nabla^2 f(x^k))^{-1} \nabla f(x^k)\,,$$

où $\gamma_k > 0$ est taille du k^{eme} pas, à déterminer par une optimisation approximative en dimension 1 (on peut se contenter de convenir $\gamma_k = 1$). Un critère d'arrêt suggéré dans les livres d'optimisation est

$$\nabla f(x^k)^T (\nabla^2 f(x^k))^{-1} \nabla f(x^k) \leq \text{tol}$$

où tol > 0 est une tolérance fixée a priori (en fonction de la précision machine par exemple).

L'étude de la convergence de la méthode de Newton postule des conditions sur le conditionnement du Hessien à l'optimum (ici la matrice d'information de Fisher). Lorsque ces conditions sont remplies, la méthode de Newton converge très rapidement : il existe une constante $\kappa > 0$ telle qu'à chaque itération,

$$\kappa(f(x^{k+1}) - f(x^*)) \leq \big(\kappa(f(x^k) - f(x^*))\big)^2\,.$$

Si $\kappa(f(x^0) - f(x^*)) < 1$, l'erreur décroît doublement exponentiellement avec le nombre d'itérations. Nous n'allons pas prendre cette direction, mais examiner le résultat du premier pas de la méthode de Newton (avec $t_0 = 1$).

Dans la suite $\widetilde{\theta}_n$ désigne un estimateur (pas forcément un estimateur au maximum de vraisemblance). On suppose que la suite $(\widetilde{\theta}_n)_n$ est consistante et asymptotiquement normale. On note $J(\theta^0)$ la covariance asymptotique de $\sqrt{n}(\widetilde{\theta}_n - \theta^0)$.

On définit l'estimateur à un pas (*one-step*) $\check{\theta}_n$ comme l'estimateur obtenu en appliquant un pas de la méthode de Newton pour approcher le maximum de vraisemblance en partant de $\widetilde{\theta}_n$:

$$\check{\theta}_n := \widetilde{\theta}_n - (\nabla^2 \ell_n(\widetilde{\theta}_n))^{-1} \nabla \ell_n(\widetilde{\theta}_n)\,.$$

On continue de noter l'estimateur au maximum de vraisemblance $\widehat{\theta}_n$. On note $I(\theta^0)$ la matrice d'information de Fisher en θ^0 ($I(\theta^0) = \nabla^2 \log Z(\theta^0)$).

PROPOSITION 5.19 *Dans un modèle exponentiel identifiable, pour tout $\theta \in \Theta$, sous $P_\theta^{\otimes \mathbb{N}}$, la suite des estimateurs à un pas $(\check{\theta}_n)_n$ est consistante, après centrage et normalisation, elle a le même comportement asymptotique que la suite des estimateurs au maximum de vraisemblance,*

$$\sqrt{n}\left(\check{\theta}_n - \theta^0\right) \rightsquigarrow \mathcal{N}(0, I(\theta^0)^{-1})\,.$$

REMARQUE 5.20 Au prix de plus d'efforts, on peut établir cette propriété dans une classe de modèles beaucoup plus large.

La proposition suivante est triviale. C'est néanmoins une observation qui sert à orienter beaucoup d'études. Dans de nombreuses situations, on cherche à montrer qu'une suite d'estimateurs convenablement recentrée et renormalisée se comporte comme la fonction score évaluée au paramètre qui définit l'échantillonnage.

PROPOSITION 5.21 *Pour tout $\theta \in \Theta$, sous $P_\theta^{\otimes \mathbb{N}}$, la fonction score en θ est asymptotiquement normale de matrice de covariance égale à la matrice d'information de Fisher :*

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta) \rightsquigarrow \mathcal{N}(0, I(\theta))\,.$$

Une conséquence immédiate est : sous $P_\theta^{\otimes \mathbb{N}}$,

$$\frac{1}{\sqrt{n}} I(\theta)^{-1/2} \nabla \ell_n(\theta) \rightsquigarrow \mathcal{N}(0, \mathrm{Id}_k)\,,$$

avec la convention habituelle que $A = I(\theta)^{1/2}$ vérifie $AA^t = I(\theta)$ (on peut choisir un facteur de Cholesky ou construire A à partir de la décomposition en valeurs singulières).

PREUVE.[Proposition 5.21] C'est une conséquence immédiate du théorème central limite multivarié et des observations :

$$\begin{aligned} \nabla \ell_n(\theta) &= \sum_{i=1}^n \left(T(X_i) - \nabla \log Z(\theta)\right) \\ \mathbb{E}T(X_i) &= \nabla \log Z(\theta) && \text{Sous} \quad P_\theta \\ \mathrm{cov}(T(X_i)) &= I(\theta) && \text{Sous} \quad P_\theta\ . \end{aligned}$$

□

PREUVE. Pour établir la proposition 5.19, nous allons vérifier que sous $P_\theta^{\otimes \mathbb{N}}$,

$$\sqrt{n}\left(\check{\theta}_n - \theta\right) - I(\theta)^{-1} \frac{1}{\sqrt{n}} \nabla \ell_n(\theta) \xrightarrow{P} 0\,.$$

Le résultat recherché suit de la proposition 5.21.

Pour toute constante $M > 0$,

$$\sup_{\vartheta : \vartheta \in \Theta, \sqrt{n}\|\vartheta - \theta\| \le M} \frac{1}{\sqrt{n}} \left\| \nabla \ell_n(\theta) - \nabla \ell_n(\vartheta) + \nabla^2 \ell_n(\theta)\,(\vartheta - \theta) \right\| \longrightarrow 0\,.$$

En effet dans un modèle exponentiel

$$\nabla \ell_n(\vartheta) = n\widehat{T}_n - n \nabla \log Z(\vartheta) \qquad \text{et} \qquad \nabla^2 \ell_n(\vartheta) = -n \nabla^2 \log Z(\vartheta)\,.$$

On injecte dans l'expression à étudier

$$\begin{aligned}&\sup_{\vartheta:\sqrt{n}\|\vartheta-\theta\|\leq M}\frac{1}{\sqrt{n}}\left\|\nabla\ell_n(\theta)-\nabla\ell_n(\vartheta)+\nabla^2\ell_n(\theta)\,(\vartheta-\theta)\right\|\\ &=\sup_{\vartheta:\sqrt{n}\|\vartheta-\theta\|\leq M}\sqrt{n}\left\|\nabla\log Z(\theta)-\nabla\log Z(\vartheta)+\nabla^2\log Z(\theta)\,(\vartheta-\theta)\right\|.\end{aligned}$$

Pour tout θ^0,

$$\nabla\log Z(\theta)-\nabla\log Z(\vartheta)+\nabla^2\log Z(\theta)\,(\vartheta-\theta) \quad = \quad R(\vartheta-\theta)$$

avec $|R(\vartheta-\theta)|=o(\|\vartheta-\theta\|)$.

$$\begin{aligned}&\sup_{\vartheta:\sqrt{n}\|\vartheta-\theta\|\leq M}\sqrt{n}\left\|\nabla\log Z(\theta)-\nabla\log Z(\vartheta)+\nabla^2\log Z(\theta)\,(\vartheta-\theta)\right\|\\ &=\sup_{\vartheta:\sqrt{n}\|\vartheta-\theta\|\leq M}\sqrt{n}\,|R(\vartheta-\theta)|\\ &=\sqrt{n}o(M/\sqrt{n})\,.\end{aligned}$$

La convergence du supremum est vérifiée pour toute suite d'échantillons dont la taille tend vers l'infini

La suite $\frac{1}{n}\nabla^2\ell_n(\widetilde{\theta}_n)$ admet une limite en probabilité/presque sûre. En effet,

$$\frac{1}{n}\nabla^2\ell_n(\widetilde{\theta}) \quad = \quad -\nabla^2\log Z(\widetilde{\theta}_n)$$

qui tend vers $-\nabla^2\log Z(\theta)=-I(\theta)$ en probabilité si $(\widetilde{\theta}_n)$ est une suite consistante, presque sûrement si $(\widetilde{\theta}_n)$ est une suite forcément consistante d'estimateurs. En effet, la fonction $\vartheta\mapsto\log Z(\vartheta)$ est C^∞ sur l'intérieur du domaine de définition Θ.

$$\begin{aligned}&\sqrt{n}\left(\check{\theta}_n-\theta\right)-I(\theta)^{-1}\frac{1}{\sqrt{n}}\nabla\ell_n(\theta)\\ &=\sqrt{n}\left(\widetilde{\theta}_n-\theta\right)\\ &\quad-\left(\left(\frac{\nabla^2\ell_n(\widetilde{\theta}_n)}{n}\right)^{-1}+(I(\theta))^{-1}\right)\frac{1}{\sqrt{n}}\nabla\ell_n(\widetilde{\theta}_n)\\ &\quad+(I(\theta))^{-1}\frac{1}{\sqrt{n}}\left(\nabla\ell_n(\widetilde{\theta}_n)-\nabla\ell_n(\theta)\right)\\ &=\sqrt{n}\left(\widetilde{\theta}_n-\theta\right)\\ &\quad+\left((\nabla^2\log Z(\widetilde{\theta}_n))^{-1}-(\nabla^2\log Z(\theta))^{-1}\right)\frac{1}{\sqrt{n}}\nabla\ell_n(\widetilde{\theta}_n)\\ &\quad+(I(\theta))^{-1}\sqrt{n}\left(\nabla\log Z(\theta)-\nabla\log Z(\widetilde{\theta}_n)\right)\end{aligned}$$

La méthode δ au premier ordre, la somme du premier et du troisième terme converge en probabilité vers 0.

Le deuxième terme s'écrit

$$\begin{aligned}&\left((\nabla^2\log Z(\widetilde{\theta}_n))^{-1}-(\nabla^2\log Z(\theta))^{-1}\right)\frac{1}{\sqrt{n}}\nabla\ell_n(\widetilde{\theta}_n)\\ &=\left((\nabla^2\log Z(\widetilde{\theta}_n))^{-1}-(\nabla^2\log Z(\theta))^{-1}\right)\sqrt{n}\left(\overline{T}_n-\nabla\log Z(\theta)+\nabla\log Z(\theta)-\nabla\log Z(\widetilde{\theta}_n)\right)\\ &=\left((\nabla^2\log Z(\widetilde{\theta}_n))^{-1}-(\nabla^2\log Z(\theta))^{-1}\right)\sqrt{n}\left(\overline{T}_n-\nabla\log Z(\theta)\right)\\ &\quad+\left((\nabla^2\log Z(\widetilde{\theta}_n))^{-1}-(\nabla^2\log Z(\theta))^{-1}\right)\sqrt{n}\left(\nabla\log Z(\theta)-\nabla\log Z(\widetilde{\theta}_n)\right).\end{aligned}$$

Le facteur $\left((\nabla^2\log Z(\widetilde{\theta}_n))^{-1}-(\nabla^2\log Z(\theta))^{-1}\right)$ converge en probabilité vers 0. Le facteur $\sqrt{n}\left(\overline{T}_n-\nabla\log Z(\theta)\right)$ converge en loi vers $\mathcal{N}(0,I(\theta))$. Le premier terme converge donc en probabilité vers 0. Le facteur

$\sqrt{n}\left(\nabla \log Z(\theta) - \nabla \log Z(\widetilde{\theta}_n)\right)$ converge en loi vers $\mathcal{N}(0, I(\theta)J(\theta)I(\theta))$. Le second terme du membre droit converge donc lui aussi vers 0 en probabilité.

La suite $\sqrt{n}\left(\tilde{\theta}_n - \theta\right)$ admet donc une limite en loi $\mathcal{N}(0, I(\theta)^{-1})$. C'est aussi la limite en loi de l'estimateur du maximum de vraisemblance recentré et renormalisé. □

Remarque 5.22 Pour compléter l'argument, il faudrait se convaincre que $I(\theta)^{-1}$ est la plus petite covariance possible (dans l'ordre semi-défini positif) pour une suite d'estimateurs asymptotiquement normale.

Par ailleurs, la mise en oeuvre de la méthode à un pas peut elle-même être problématique. Il faut disposer d'un estimateur initial. Et il faut pouvoir calculer le gradient et le Hessien de la fonction de partition.

Enfin, la proposition 5.19 est un énoncé asymptotique. Qu'apporte-t-il au praticien qui doit travailler sur un échantillon fini ?

5.7 remarques bibliographiques

Au delà des modèles exponentiels les plus simples et les plus classiques, les modèles graphiques (*graphical models*) [5] forment un outil de modélisation très expressif. Ces modèles exponentiels permettent d'aborder des questions de causalité. Ils forment des défis calculatoires. Leur étude est un des moteurs de la statistique computationnelle.

L'optimisation convexe joue un rôle croisant en statistique computationnelle et en théorie de l'apprentissage statistique [1]. Un panorama récent des résultats et méthodes d'optimisation utiles aux statisticiens peut être lu dans [2].

La méthode delta remonte à [3], elle est exposée, illustrée et généralisée dans [4].

Le livre de Boyd et Vandenberghe [1] est une introduction à l'optimisation convexe qui illustre, motive et analyse la méthode de Newton. Une version beaucoup plus générale des propriétés asymptotiques la méthode à un pas est établie dans [4]. Le problème posé par l'arrêt des méthodes itératives en statistique et en apprentissage continue de susciter beaucoup d'intérêt.

Références

[1] S. Boyd et L. Vandenberghe. **Convex optimization**. Cambridge University Press, 2005.

[2] S. Bubeck. **Theory of Convex Optimization for Machine Learning**. T. 8. Foundations and Trends in Machine Learning. 2015.

[3] H. Cramér. **Mathematical Methods of Statistics**. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946, p. xvi+575. mr : 0016588.

[4] A. van der Vaart. **Asymptotic statistics**. Cambridge University Press, 1998.

[5] M. J. Wainwright et M. I. Jordan. **Graphical models, exponential families, and variational inference**. T. 1. 1-2. Now Publishers Inc., 2008, p. 1-305.

6.1 INTRODUCTION

Dans les trois chapitres précédents, nous avons considéré des modèles outrageusement simples, régulièrement paramétrés par des parties de $\mathbb{R}^k$. Nous nous sommes concentrés sur l'estimation des paramètres, ou sur des questions de décision concernant ces paramètres. Les modèles que nous avons étudiés étaient des modèles dominés. Nous avons pu choisir des fonctions de vraisemblance continues en les paramètres et mêmes plusieurs fois différentiables. Ce fut l'occasion d'établir quelques propriétés des méthodes de maximisation de la vraisemblance. L'estimation au maximum de vraisemblance n'est pas le seul usage qu'un statisticien peut faire d'une fonction de vraisemblance. Dans ce chapitre nous allons aborder les questions de test avec des méthodes qui utilisent une fonction de vraisemblance. Nous n'allons pas nous restreindre aux modèles paramétriques, mais envisagé des modèles dominés où les lois sont paramétrées par leur densité ou la racine carré de leur densité. L'écart entre lois sera évalué par des divergence d'information (distance en variation, entropie relative et distance de Hellinger).

Nous revenons d'abord au problème de la conception de tests, en particulier sur la conception de tests d'hypothèses binaires. La section 6.2 présente une preuve complète du Lemme de Neyman-Pearson (évoqué en Section 1.6). Ce lemme nous impose une méthode pour construire un test entre deux hypothèses simples : comparer le rapport de vraisemblance à un seuil. La portée de ce résultat va bien au delà du problème du test d'hypothèses simples. Il nous fournit dans de nombreux cas un étalon-or. Par exemple, dans le cas d'hypothèses composites, on peut éprouver une méthode de test ad hoc (tests du χ^2, tests de rapport de vraisemblance généralisés, tests de rang, de Student, de Kolmogorov-Smirnov, etc) en comparant son comportement à celui d'un test de rapport de vraisemblance entre une loi de l'hypothèse nulle et une loi de l'alternative. On peut aussi examiner le comportement des tests ad hoc pour distinguer un mélange de lois issues de l'hypothèse nulle d'un mélange de lois issues de l'alternative.

Dans la section 6.3, nous relions les probabilités d'erreur de première et de seconde espèce entre deux hypothèses simples à la distance en variation entre les lois qui définissent les hypothèses. Cette relation est fine mais elle nous aide pas beaucoup à comprendre comment la taille de l'échantillon influe sur notre capacité à distinguer deux lois. Nous ne disposons pas d'expressions simples de la distance en variation entre deux lois produits.

En revanche, l'entropie relative et l'affinité de Hellinger entre lois produits se déduisent immédiatement de l'entropie relative et de l'affinité entre les facteurs. La distance en variation est elle-même liée à l'entropie relative par l'inégalité de Pinsker. Cette relation importante (équivalente au lemme de Hoeffding 1.13) permet de minorer les probabilités d'erreur dans les tests entre lois « produit ».

La distance et l'affinité de Hellinger permettent non seulement d'étudier l'influence de la taille de l'échantillon sur notre capacité à distinguer deux hypothèses simples, mais le point de vue « Hilbertien » est crucial dans la construction de tests de rapport de vraisemblance robustes entre hypothèses composites. La construction explicite de ces tests robustes a récemment ouvert de nouvelles portes à la statistique non-paramétrique.

6.2 LEMME DE NEYMAN-PEARSON

Dans notre version initiale du Lemme de Neyman-Pearson, nous avons supposé l'existence d'un test de rapport de vraisemblance de niveau spécifié. Nous n'avons pas établi l'existence de ce test. Nous comblons ici cette lacune en construisant un test randomisé.

LEMME 6.1 (LEMME DE NEYMAN-PEARSON) *Pour tout niveau $\alpha \in]0,1[$,*

i) il existe un test de rapport de vraisemblance T éventuellement randomisé, de niveau α. Ce test consiste à comparer le rapport de vraisemblance à un seuil τ_α, à rejeter H_0 au dessus du seuil, à ne pas rejeter H_0 en dessous du seuil, à prendre une décision aléatoire lorsque le rapport de vraisemblance est exactement égal à τ_α.

ii) Sous l'alternative, le test de rapport de vraisemblance est de puissance maximale parmi les tests de niveau inférieur ou égal.

iii) Tout test T' de niveau exactement α et de même puissance qu'un test de rapport de vraisemblance T de niveau α, coïncide avec le test de rapport de vraisemblance lorsque le rapport de vraisemblance est différent du seuil τ_α qui définit T.

PREUVE.

i) On note G la fonction de répartition du rapport de vraisemblance sous P_0 et $G^{\leftarrow}$ la fonction quantile associée :

$$G^{\leftarrow}(p) = \inf\{x : G(x) \geq p\} \ .$$

Le seuil τ_α est défini par

$$\tau_\alpha = G^{\leftarrow}(1-\alpha)\,.$$

Si $G(\tau_\alpha) > 1-\alpha$, cela signifie que G est discontinue en τ_α. Sous P_0, la probabilité que le rapport de vraisemblance soit égal à $\{\tau_\alpha\}$ est positive. Et d'autre part

$$G(\tau_\alpha-) := \lim_{t \nearrow \tau_\alpha} G(t) \leq 1-\alpha < G(\tau_\alpha)\,.$$

Lorsque le rapport de vraisemblance est égal à τ_α, le test de rapport de vraisemblance effectue un tirage uniforme U sur $[0,1]$ et il rejette H_0 si

$$U \leq \frac{G(\tau_\alpha)-(1-\alpha)}{G(\tau_\alpha)-G(\tau_\alpha-)}\,.$$

Ce test (éventuellement) randomisé est de niveau exactement α. Nous le nommons T par la suite.

ii) On note p_0 et p_1 les versions des densités utilisée dans la définition du test T. Soit T' un test de niveau inférieur ou égal à α.

La différence entre les puissances vaut

$$\begin{aligned}
P_1\{T=1\} - P_1\{T'=1\} &= \mathbb{E}_{P_1}\left[T-T'\right] \\
&= \mathbb{E}_{P_0}\left[\frac{p_1(X)}{p_0(X)}(T-T')\right] + \mathbb{E}_{P_1}\left[(T-T')\mathbb{I}_{p_0(X)=0}\right] \\
&\quad \text{sur l'événement } p_0(X)=0, T-T'\geq 0, \text{car le rapport} \\
&\quad \text{de vraisemblance est infini} \\
&\geq \mathbb{E}_{P_0}\left[\frac{p_1(X)}{p_0(X)}(T-T')\right] \\
&= \mathbb{E}_{P_0}\left[\left(\frac{p_1(X)}{p_0(X)}-\tau_\alpha\right)(T-T')\right] + \tau_\alpha \mathbb{E}_{P_0}\left[T-T'\right] \\
&\geq \mathbb{E}_{P_0}\left[\left(\frac{p_1(X)}{p_0(X)}-\tau_\alpha\right)(T-T')\right] \ \text{comme } \mathbb{E}_{P_0}\left[T-T'\right]\geq 0 \\
&\geq 0 \qquad \text{car } \left(\frac{p_1(X)}{p_0(X)}-\tau_\alpha\right)(T-T') \geq 0\,.
\end{aligned}$$

iii) Pour établir le dernier point il suffit de réexaminer la preuve du second point et de noter que pour avoir l'égalité entre puissances, il faut $\left(\frac{p_1(X)}{p_0(X)}-\tau_\alpha\right)(T-T')=0$ soit vraie P_0 presque sûrement. □

6.3 SÉPARATION, INÉGALITÉ DE PINSKER

Le fait de savoir comment tester deux hypothèses simples de façon optimale ne nous donne pas un accès simple à la somme des erreurs de première et de seconde espèce commises par le meilleur test possible. Cette somme est reliée très directement à une distance entre lois.

DÉFINITION 6.2 (DISTANCE EN VARIATION) La distance en variation entre deux lois P et Q définies sur le même espace probabilisable $(\Omega,\mathcal{F})$ est définie par

$$\mathrm{d}_{\mathrm{TV}}(P,Q) := \sup_{A\in\mathcal{F}} P(A)-Q(A)\,.$$

Nous laissons en exercice la proposition qui motive la dénomination.

PROPOSITION 6.3 *La distance en variation est une distance sur l'ensemble des lois de probabilités sur* $(\Omega, \mathcal{F})$.

La distance en variation est exigeante. Elle définit en général une topologie plus fine que la topologie de la convergence en loi. Là encore nous laissons cette proposition en exercice (facile voir Théorème C.4).

PROPOSITION 6.4 *Si une suite de lois* (P_n) *sur* $(\Omega, \mathcal{F})$ *vérifie* $\lim_n \mathrm{d_{TV}}(P_n, Q) = 0$ *où* Q *est une loi sur* $(\Omega, \mathcal{F})$ *alors* (P_n) *converge étroitement vers* Q.

La réciproque est fausse : si P_n est obtenue en recentrant et en standardisant la loi binomiale de paramètres n et $1/2$, (P_n) converge étroitement vers $\mathcal{N}(0,1)$, et pourtant $\mathrm{d_{TV}}(P_n, \mathcal{N}(0,1)) = 1$.

Comme l'entropie relative, la distance en variation se définit à la fois comme une intégrale et comme un supremum.

PROPOSITION 6.5 *Si* ν *est une mesure* σ*-finie sur* $(\Omega, \mathcal{F})$ *qui domine les lois* P *et* Q, *alors*

$$\mathrm{d_{TV}}(P, Q) = \frac{1}{2} \int_\Omega \left| \frac{\mathrm{d}P}{\mathrm{d}\nu}(\omega) - \frac{\mathrm{d}Q}{\mathrm{d}\nu}(\omega) \right| \mathrm{d}\nu(\omega)$$

$$\mathrm{d_{TV}}(P, Q) = 1 - \int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\nu}(\omega) \wedge \frac{\mathrm{d}Q}{\mathrm{d}\nu}(\omega) \mathrm{d}\nu(\omega) \,.$$

PREUVE. On peut prendre comme mesure dominante $\nu = (P + Q)/2$.

$$P(A) - Q(A) = \int \mathbb{I}_A \left(\frac{\mathrm{d}P}{\mathrm{d}\nu} - \frac{\mathrm{d}Q}{\mathrm{d}\nu} \right) \mathrm{d}\nu \,.$$

Le membre droit est maximisé par le choix

$$A := \left\{ \omega : \frac{\mathrm{d}P}{\mathrm{d}\nu}(\omega) > \frac{\mathrm{d}Q}{\mathrm{d}\nu}(\omega) \right\} .$$

On a établi

$$\mathrm{d_{TV}}(P, Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}\nu} - \frac{\mathrm{d}Q}{\mathrm{d}\nu} \right)_+ \mathrm{d}\nu \,,$$

la proposition découle de

$$\int \left(\frac{\mathrm{d}P}{\mathrm{d}\nu} - \frac{\mathrm{d}Q}{\mathrm{d}\nu} \right)_+ \mathrm{d}\nu = \int \left(\frac{\mathrm{d}P}{\mathrm{d}\nu} - \frac{\mathrm{d}Q}{\mathrm{d}\nu} \right)_- \mathrm{d}\nu \,.$$

On a utilisé ici la convention

$$(x)_+ = \max(x, 0)$$
$$(x)_- = \max(-x, 0) \,.$$

La dernière partie de la proposition vient de

$$\frac{1}{2}|a-b| = \frac{a+b}{2} - a \wedge b\,.$$

□

La connexion entre erreurs de test et distance en variation est résumée ici.

PROPOSITION 6.6 (AFFINITÉ DE TEST)

$$\inf_T (P_0\{T=1\} + P_1\{T=0\}) = 1 - \mathrm{d_{TV}}(P_0, P_1) = \int_\Omega \frac{\mathrm{d}P_1}{\mathrm{d}\nu}(\omega) \wedge \frac{\mathrm{d}P_0}{\mathrm{d}\nu}(\omega)\mathrm{d}\nu(\omega)\,.$$

L'expression $1 - \mathrm{d_{TV}}(P_0, P_1)$ *est appelée affinité de test.*

PREUVE. Soit A la région critique d'un test T : $A := \{\omega : T(\omega) = 1\}$.

$$P_0\{T=1\} + P_1\{T=0\} \quad = \quad 1 - (P_1\{A\} - P_0\{A\})\ .$$

On obtient la proposition en choisissant une région critique qui témoigne de la distance en variation. □

Cette proposition est simple mais difficile à utiliser telle quelle. Si on veut étudier le comportement de la somme des deux erreurs lorsqu'on cherche à distinguer des lois « produit », on peut être tenté d'utiliser la borne suivante :

$$\mathrm{d_{TV}}(P^{\otimes n}, Q^{\otimes n}) \leq n\mathrm{d_{TV}}(P, Q)$$

pour se convaincre que deux hypothèses simples séparées d'au plus $1/(2n)$ en distance en variation, ne peuvent pas être distinguées avec des probabilités d'erreur de première et de seconde espèce toutes les deux inférieures à $1/4$ à partir d'échantillons de taille n.

Nous allons voir dans la section suivante, qu'on peut fournir une borne inférieure de bien meilleure qualité.

Une inégalité de transport

La distance en variation et l'entropie relative sont deux divergences d'information. On suppose ici que $P \trianglelefteq Q$. Si ν est une mesure σ-finie qui domine P et Q et f une fonction convexe qui s'annule en 1, en notant p et q les densités des lois P et Q par rapport à ν, on définit la f-divergence entre P et Q par

$$I_f(P\|Q) := \int_\Omega f\left(\frac{p(\omega)}{q(\omega)}\right) q(\omega)\mathrm{d}\nu(\omega)\,.$$

La distance en variation s'obtient en choisissant $f(x) = |x-1|$, l'entropie relative en choisissant $f(x) = x\log(x)$. En choisissant $f(x) = (\sqrt{x}-1)^2$, on obtient le carré de la distance de Hellinger, en choisissant $f(x) = (x-1)^2$ on obtient la « distance du χ^2 ».

Ces deux divergences d'information sont liées par une inégalité qui s'avère être une conséquence du lemme de Hoeffding.

THÉORÈME 6.7 (INÉGALITÉ DE PINSKER)

$$\mathrm{d_{TV}}(P, Q) \leq \sqrt{\frac{1}{2}D(P\|Q)}\,.$$

La représentation variationnelle de l'entropie relative permet d'étudier facilement de nombreuses propriétés de l'entropie relative. Elle nous permettra d'établir l'inégalité de Pinsker de façon modulaire.

THÉORÈME 6.8 (REPRÉSENTATION VARIATIONNELLE) *Soit P et Q deux lois de probabilité sur $(\Omega, \mathcal{F})$, l'entropie relative de P par rapport à Q vérifie*

$$D(P\|Q) = \sup_{f:\mathbb{E}_Q \mathrm{e}^f < \infty} \left\{ \mathbb{E}_P f - \log \mathbb{E}_Q \mathrm{e}^f \right\} .$$

PREUVE. Si $P \not\trianglelefteq Q$, il existe A telle que $P(A) > 0$ et $Q(A) = 0$, en considérant $f_n = n\mathbb{I}_A$, on observe $\mathbb{E}_P f_n - \log \mathbb{E}_Q \mathrm{e}^{f_n} = nP(A)$ qui tend vers ∞ avec n. Si $P \not\trianglelefteq Q$, la représentation variationnelle est vérifiée.

Si $P \trianglelefteq Q$, on vérifie d'abord que

$$D(P\|Q) = \sup_{U:\mathbb{E}_Q \mathrm{e}^U = 1} \mathbb{E}_P U . \tag{6.1}$$

En effet, pour toute variable aléatoire U vérifiant $\mathbb{E}_Q \mathrm{e}^U = 1$, on peut définir une autre loi Q' de densité e^U par rapport à Q, et $\frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{e}^{-U}$ est une dérivée de Radon-Nykodym de P par rapport à Q'. On a alors

$$D(P\|Q) - \mathbb{E}_P U = \mathbb{E}_{Q'} \left[\frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{e}^{-U} \log \left(\frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{e}^{-U} \right) \right] \geq 0$$

où la dernière inégalité suit de l'inégalité de Jensen. Cela prouve que lorsque $P \trianglelefteq Q$

$$D(P\|Q) \geq \sup_{U:\mathbb{E}_Q \mathrm{e}^U = 1} \mathbb{E}_P U .$$

Comme il y a égalité lorsque $U = \log \frac{\mathrm{d}P}{\mathrm{d}Q}$, la première formule (6.1) est établie.

Si $\mathbb{E}_Q \mathrm{e}^f < \infty$,

$$\mathbb{E}_P f - \log \mathbb{E}_Q \mathrm{e}^f \quad = \quad \mathbb{E}_P \left[\log \frac{\mathrm{e}^f}{\mathbb{E}_Q \mathrm{e}^f} \right] .$$

Le supremum dans l'énoncé de la formulation variationnelle n'est pas plus grand que celui de (6.1). □

A partir de la représentation variationnelle, on vérifie (très) facilement des propriétés importantes de l'entropie relative.

COROLLAIRE 6.9 *L'entropie relative est*

i) convexe vis à vis de ses deux arguments ;
ii) l'entropie relative entre lois « image » est inférieure ou égale à l'entropie relative entre les lois

$$D(P\|Q) \geq D(P \circ f^{-1} \| Q \circ f^{-1}) .$$

PREUVE.[Inégalité de Pinsker] Pour tout $A \in \mathcal{F}$, tout $\lambda \geq 0$, d'après la formulation variationnelle de l'entropie relative

$$D(P\|Q) \geq \mathbb{E}_P \left[\lambda(\mathbb{I}_A - \mathbb{E}_Q \mathbb{I}_A) \right] - \log \mathbb{E}_Q \left[\mathrm{e}^{\lambda(\mathbb{I}_A - \mathbb{E}_Q \mathbb{I}_A)} \right] .$$

D'après le lemme de Hoeffding,

$$\log \mathbb{E}_Q \left[\mathrm{e}^{\lambda(\mathbb{I}_A - \mathbb{E}_Q \mathbb{I}_A)} \right] \leq \frac{\lambda^2}{8} .$$

On a donc :

$$D(P\|Q) \geq \sup_{A \in \mathcal{F}, \lambda \geq 0} \left\{ \lambda(P(A) - Q(A)) - \frac{\lambda^2}{8} \right\} .$$

Pour A fixé, l'optimum en λ est atteint en $\lambda = 4(P(A) - Q(A))$, d'où

$$D(P\|Q) \geq \sup_{A\in\mathcal{F}} 2(P(A) - Q(A))^2 = 2\mathrm{d}_{\mathrm{TV}}(P,Q)^2 \,.$$

□

Comme $D(P^{\otimes n}\|Q^{\otimes n}) = nD(P\|Q)$, on peut en déduire une borne inférieure pour la somme des erreurs d'un test opérant sur des n-échantillons.

$$P_0^{\otimes n}\{T=1\} + P_1^{\otimes n}\{T=0\} \geq 1 - \sqrt{\frac{n}{2}D(P_0\|P_1)}\,.$$

Exemple 6.10 Si on souhaite décider entre deux lois de Bernoulli $P_0 = \mathcal{B}(\theta)$ et $P_1 = \mathcal{B}(\theta + h/\sqrt{n})$, on peut effectuer quelques calculs qui révèlent

$$D\left(P_1\|P_0\right) \leq \frac{1}{2n}\frac{h^2}{\theta(1-\theta)}\,.$$

Ceci implique :

$$P_0^{\otimes n}\{T=1\} + P_1^{\otimes n}\{T=0\} \geq 1 - \frac{h}{2\sqrt{\theta(1-\theta)}}\,.$$

Cette application illustre un phénomène plus général : si deux lois sont à distance en variation de l'ordre de $1/\sqrt{n}$, il est illusoire de chercher à les séparer à l'aide d'échantillons de taille sensiblement inférieure à n.

Application aux bornes inférieures en estimation

Dans les modèles exponentiels nous avons montré que les estimateurs au maximum de vraisemblance convergent à vitesse $\frac{1}{\sqrt{n}}$ (le risque quadratique décroit comme $\frac{1}{n}$). L'inégalité de Cramer-Rao nous indique que les estimateurs sans biais ne peuvent pas être plus rapides. Comme il n'y a pas de raison de ne considérer que les estimateurs sans biais, on peut se demander s'il n'est pas possible d'établir des bornes inférieures plus générales.

Nous nous plaçons dans un modèle exponentiel minimal ($P_\theta, \theta \in \Theta \subseteq \mathbb{R}^k$). Dans ce modèle, nous avons vu que

$$D(P_\theta\|P_{\theta'}) = \frac{1}{2}(\theta'-\theta)^t I(\theta)(\theta'-\theta) + R(\theta'-\theta)$$

avec $|R(\theta'-\theta)| = o(\|\theta'-\theta\|^2)$.

Soit $\widetilde{\theta}_n$ un estimateur à valeur dans $\mathbb{R}^k$. Cet estimateur permet de définir un test T entre $P_\theta^{\otimes n}$ et $P_{\theta'}^{\otimes n}$: on rejette H_0 (constituée par $P_\theta^{\otimes n}$) si $\|\widetilde{\theta}_n - \theta\| > \|\widetilde{\theta}_n - \theta'\|$.

$$\begin{aligned}
&\mathbb{E}_\theta\left[\|\widetilde{\theta}_n - \theta\|^2\right] + \mathbb{E}_{\theta'}\left[\|\widetilde{\theta}_n - \theta'\|^2\right] \\
&\quad\geq\quad \mathbb{E}_\theta\left[\|\widetilde{\theta}_n - \theta\|^2 \mathbb{1}_{\|\widetilde{\theta}_n-\theta\|>\|\widetilde{\theta}_n-\theta'\|}\right] + \mathbb{E}_{\theta'}\left[\|\widetilde{\theta}_n - \theta'\|^2 \mathbb{1}_{\|\widetilde{\theta}_n-\theta\|<\|\widetilde{\theta}_n-\theta'\|}\right] \\
&\quad\geq\quad \frac{\|\theta'-\theta\|^2}{4}\left(P_\theta^{\otimes n}\{T=1\} + P_{\theta'}^{\otimes n}\{T=0\}\right) \\
&\quad\geq\quad \frac{\|\theta'-\theta\|^2}{4}\left(1 - \sqrt{\frac{n}{2}D(P_\theta\|P_{\theta'})}\right)
\end{aligned}$$

Maintenant, choisissons $h \in \mathbb{R}^k$ et n_0 tels que $\theta + h/\sqrt{n_0} \in \Theta$ et tel que pour $n \geq n_0$, $|R(h/\sqrt{n})| \leq \frac{1}{4n}h^t I(\theta)h$. On suppose aussi que $h^t I(\theta)h < 8$.

Pour $n \geq n_0$, en combinant les calculs précédents avec l'inégalité de Pinsker, on obtient

$$\begin{aligned}
&\mathbb{E}_\theta\left[\|\widetilde{\theta}_n - \theta\|^2\right] + \mathbb{E}_{\theta+h/\sqrt{n}}\left[\|\widetilde{\theta}_n - \theta - \frac{h}{\sqrt{n}}\|^2\right] \\
&\quad\geq\quad \frac{\|h\|^2}{4n}\left(1 - \sqrt{\frac{h^t I(\theta)h}{8}})\right)\,.
\end{aligned}$$

Ce résultat peut être affiné. Les développements les plus intéressants s'appuient sur une autre f-divergence, la distance de Hellinger.

Moins intuitive que la distance en variation, moins naturelle que l'entropie relative, la distance de Hellinger est une divergence d'information (commme la distance en variation et l'entropie relative), elle est au centre non seulement de la théorie de Le Cam (version presque finale de la statistique paramétrique asymptotique) mais aussi de dévelopements actuels en statistique non-paramétrique.

DÉFINITION 6.11 (DISTANCE ET AFFINITÉ DE HELLINGER) Soient P et Q deux lois de probabilités sur $(\Omega, \mathcal{F})$, ν une mesure qui domine P, Q et p, q deux densités de P, Q par rapport à ν. La distance de Hellinger $H(P,Q)$ est définie par

$$H(P,Q)^2 := \frac{1}{2}\int_\Omega (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\nu \, .$$

L'affinité de Hellinger $\rho(P,Q)$ est définie par

$$\rho(P,Q) := 1 - H(P,Q)^2 = \int_\Omega \sqrt{pq} \mathrm{d}\nu \, .$$

On peut se demander si cette définition est cohérente : est ce que le choix de la mesure dominante et celui des densités sont importants ou pas ? On vérifie que ces choix n'ont pas d'influence sur la valeur de la distance de Hellinger.

On verifie immédiatement que

$$H(P,Q)^2 \leq \mathrm{d}_{\mathrm{TV}}(P,Q) \, .$$

L'affinité de Hellinger entre produits se calcule immédiatement :

$$\rho(P^n, Q^n) = (\rho(P,Q))^n \, .$$

Si on considère un modèle dominé $\mathcal{P}$, et un choix des densités par rapport à une dominante ν, on peut paramétriser le modèle $\mathcal{P}$ par une collection d'éléments de l'espace de Hilbert $L_2(\nu)$.

La distance et l'affinité de Hellinger sont deux notions très pratiques pour étudier les méthodes de vraisemblance, qu'il s'agisse des tests ou des techniques d'estimation.

Test entre hypothèses simples

La possibilité de distinguer deux hypothèses simples définies par des lois produits à partir d'un n-échantillon est facilement quantifiée à l'aide de la distance de Hellinger.

THÉORÈME 6.12 *Soit P, Q deux lois sur $(\Omega, \mathcal{F})$, L_n la différence des log-vraisemblances calculées sur un n-échantillon $L_n = \sum_{i=1}^n \log q(X_i)/p(X_i)$, pour tout $z \in \mathbb{R}$*

$$P^n \{L_n \geq z\} \leq \exp\left(-\frac{z}{2} - nH(P,Q)^2\right)$$

et

$$Q^n \{L_n \leq z\} \leq \exp\left(+\frac{z}{2} - nH(P,Q)^2\right) \, .$$

PREUVE. A partir de

$$L_n \geq z \qquad \Longleftrightarrow \qquad \prod_{i=1}^n \sqrt{\frac{q(X_i)}{p(X_i)}} \geq \mathrm{e}^{z/2}$$

l'inégalité de Markov entraine

$$\begin{aligned} P^n \{L_n \geq z\} &\leq \mathrm{e}^{-z/2} \mathbb{E}_{P^n}\left[\prod_{i=1}^n \sqrt{\frac{q(X_i)}{p(X_i)}}\right] \\ &= \mathrm{e}^{-z/2} \rho(P,Q)^n \\ &\leq \exp\left(-\frac{z}{2} - nH(P,Q)^2\right) \, . \end{aligned}$$

La seconde inégalité se démontre de manière semblable. □

En choisissant $z = 0$, et en définissant le test T_n par

$$T_n = \mathbb{I}_{L_n \geq 0}\,,$$

on a simultanément

$$\max\left(\mathbb{E}_{P^n} T_n, \mathbb{E}_{Q^n}(1 - T_n)\right) \leq \exp\left(-nH(P,Q)^2\right)\,.$$

Pour tester P^n contre Q^n, il suffit que $H(P,Q) \gg 1/\sqrt{n}$. Ce résultat peut être complété par une réciproque.

THÉORÈME 6.13 *Soit P, Q deux lois sur $(\Omega, \mathcal{F})$, pour toute région critique d'un test de P^n contre Q^n (tout événement A dans $\mathcal{F}^{\otimes n}$),*

$$\max\left(P^n(A), Q^n(A^c)\right) \geq \frac{1}{4} - \frac{nH(P,Q)^2}{2}\,.$$

PREUVE.

$$\begin{aligned}
\rho(P^n, Q^n) &= \int_{\Omega^n} \sqrt{p^n q^n}\mathbb{I}_A \mathrm{d}\nu + \int_{\Omega^n} \sqrt{p^n q^n}\mathbb{I}_{A^c} \mathrm{d}\nu \\
&\leq \left(\int_{\Omega^n} p^n \mathbb{I}_A \mathrm{d}\nu\right)^{1/2} \left(\int_{\Omega^n} q^n \mathrm{d}\nu\right)^{1/2} + \left(\int_{\Omega^n} p^n \mathrm{d}\nu\right)^{1/2} \left(\int_{\Omega^n} \mathbb{I}_{A^c} q^n \mathrm{d}\nu\right)^{1/2} \\
&= P^n(A)^{1/2} + Q^n(A^c)^{1/2}\,.
\end{aligned}$$

Par ailleurs, $\rho(P^n, Q^n) = (1 - H(P,Q)^2)^n$. Finalement

$$\max\left(P^{\otimes n}(A), Q^{\otimes n}(A^c)\right) \geq \frac{(1 - H(P,Q)^2)^{2n}}{4} \geq \frac{1}{4} - \frac{nH(P,Q)^2}{2}\,.$$

□

Si $H(P,Q) \ll 1/\sqrt{n}$, soit la probabilité d'erreur de première soit la probabilité d'erreur de seconde espèce ne peut pas être arbitrairement petite.

Ces observations élémentaires permettent non seulement d'estimer des vitesses de séparation mais aussi de quantifier rapidement les vitesses d'estimation dans les problèmes paramétriques.

Tests entre boules de Hellinger

La construction de bons tests entre hypothèses composées est (en général) un problème difficile si on cherche à disposer de garanties uniformes sous les deux hypothèses.

Lorsque les deux hypothèses sont des boules de Hellinger

$$\mathcal{B}(P_0, \epsilon) = \{P : H(P, P_0) \leq \epsilon\} \qquad \text{et} \qquad \mathcal{B}(Q_0, \epsilon) = \{Q : H(Q, Q_0) \leq \epsilon\}$$

dont les centres sont assez bien séparés $H(P_0, Q_0) \geq 3\epsilon$, un test de rapport de vraisemblance bien construit permet de majorer les erreurs de première et de seconde espèce sur la réunion des deux boules.

Ce test de rapport de vraisemblance ne compare pas les vraisemblances sous P_0 et Q_0 mais sous deux lois P_1 et Q_1 que nous allons spécifier.

Arc de Hellinger. Les deux lois centrales P_0 et Q_0 sont identifiées à des points de la sphère unité de $L_2(\nu)$. On note $\omega \in [0, \pi/2]$ la distance angulaire entre ces deux points :

$$\cos(\omega) = \rho(P_0, Q_0) \qquad \text{soit} \qquad H(P_0, Q_0)^2 = 2\sin(\omega/2)^2\,.$$

L'arc de cercle qui relie les points correspondants à P_0 et Q_0 sur la sphère unité de $L_2(\nu)$ unité peut être paramétré par l'angle $\beta\omega$ avec le rayon défini par P_0 ($\beta \in [0,1]$). La densité correspondante est v_β^2 où

$$v_\beta = \frac{1}{\sin(\omega)} \left(\sin(\beta\omega)\sqrt{q_0} + \sin((1-\beta)\omega)\sqrt{p_0}\right) .$$

Cet arc et les lois définies sont appelés *arc de Hellinger* engendré par P_0 et Q_0.

L'affinité entre P_0 et une loi de l'arc de Hellinger P_β se déduit de l'écart angulaire :

$$\cos(\beta\omega) = \rho(P_0, P_\beta) = \sin(\beta\omega)\cot(\omega) + \frac{\sin((1-\beta)\omega)}{\sin(\omega)} .$$

Les lois $P_1 \in \mathcal{B}(P_0, \epsilon)$ et $Q_1 \in \mathcal{B}(Q_0, \epsilon)$ correspondent à $\beta = 1/3$ et $\beta = 2/3$. Notons que

$$\rho(P_1, Q_1) = \rho(P_0, P_1) = \rho(Q_1, Q_0) = \cos(\omega/3) .$$

On note p_1 (resp. q_1) la densité de P_1 (resp. Q_1).

Le test. Le rapport de vraisemblance utilisé pour construire le test entre les boules de Hellinger $\mathcal{B}(P_0, H(P_0,Q_0)/4)$ et $\mathcal{B}(Q_0, H(Q_0,P_0)/4)$ est

$$\psi(X_i) = \sqrt{\frac{q_1(X_i)}{p_1(X_i)}} .$$

Le test sur un n-échantillon est construit sur la statistique

$$\prod_{i=1}^n \sqrt{\frac{q_1(X_i)}{p_1(X_i)}} .$$

Notons que le rapport de vraisemblance n'est pas construit à partir des centres de boules de Hellinger, mais à partir de points situés sur l'arc de Hellinger.

Les performances du test.

THÉORÈME 6.14 *Soit P_0, Q_0 deux lois de probabilités sur $\mathcal{X}$, il existe une statistique $\psi : \mathcal{X} \to [0,\infty)$ telle que pour toute loi R sur $\mathcal{X}$,*

$$\begin{aligned} \mathbb{E}_R\psi(X) &\leq 1 - \frac{5}{24}H(P_0,Q_0)^2 \qquad \text{si } H(P_0,R) \leq \frac{1}{4}H(P_0,Q_0) \\ \mathbb{E}_R\frac{1}{\psi(X)} &\leq 1 - \frac{5}{24}H(P_0,Q_0)^2 \qquad \text{si } H(Q_0,R) \leq \frac{1}{4}H(P_0,Q_0) . \end{aligned}$$

PREUVE. On note $\epsilon = H(P_0,P_1) = H(Q_0,Q_1)$.

Soit $R \in \mathcal{B}(P_0,\epsilon)$. Les lois R, P_0, Q_0 définissent une sphère de dimension 2 dans la sphère unité de $L_2(\nu)$. Les lois P_1 et Q_1 appartiennent aussi à cette sphère puisqu'elles appartiennent au cercle engendré par P_0 et Q_0.

On note s, t et r les densités de Q_1, P_1 et R par rapport à ν.

$$\mathbb{E}_R\psi(X) = \int_{\mathcal{X}} r\sqrt{\frac{s}{t}}\mathrm{d}\nu = \int_{\mathcal{X}} \sqrt{\frac{s}{t}}\left(\sqrt{r} - \sqrt{t}\right)^2 \mathrm{d}\nu + 2\int_{\mathcal{X}} \sqrt{sr}\mathrm{d}\nu - \int_{\mathcal{X}} \sqrt{st}\mathrm{d}\nu .$$

On note au passage que par construction de s et t

$$\sqrt{\frac{s}{t}} \leq \frac{\sin(2\omega/3)}{\sin(\omega/3)} = 2\cos(\omega/3) ,$$

ce qui entraine

$$\mathbb{E}_R\psi(X) \leq 4\cos(\omega/3)H(R,P_1)^2 + 2\rho(R,Q_1) - \rho(P_1,Q_1) .$$

On peut profiter du fait que $\sqrt{r} \in L_2(\nu)$: on peut décomposer $\sqrt{r}$ en une combinaison linéaire d'un élément de l'arc de Hellinger v_γ ($\gamma \in [0, 2\pi/\omega)$) et $u \in L_2(\nu), u \perp \sqrt{s}, u \perp \sqrt{t}$:

$$\sqrt{r} = u + \theta v_\gamma$$

avec $\int u^2 \mathrm{d}\nu + \theta^2 = 1$. Cette décomposition permet de réécrire

$$\rho(R, Q_1) = \theta \cos((\gamma - 2/3)\omega) \qquad \rho(R, P_1) = \theta \cos((\gamma - 1/3)\omega)\,.$$

Dans la suite on utilisera

$$H(P_1, P_0)^2 = 1 - \cos(\omega/3) = 2\sin(\omega/6)^2 \qquad H(P_0, Q_0)^2 = 2\sin(\omega/2)^2$$

et le fait que $\sin(x)/x$ est décroissant sur $[0, \pi/2]$, pour justifier

$$H(P_1, P_0)^2 \geq \frac{1}{9} H(P_0, Q_0)^2\,.$$

En combinant ces observations :

$$\begin{aligned}
\mathbb{E}_R \psi(X) &\leq 4\cos(\omega/3) H(R, P_1)^2 + 2\rho(R, Q_1) - \rho(P_1, Q_1) \\
&= 4\cos(\omega/3)(1 - \theta\cos((\gamma - 1/3)\omega)) + 2\theta\cos((\gamma - 2/3)\omega) - \cos(\omega/3) \\
&\leq 3\cos(\omega/3) - 2\theta\cos(\gamma\omega) \\
&= 3\rho(P_1, P_0) - 2\rho(R, P_0) \\
&= 1 - 3H(P_1, P_0)^2 + 2H(R, P_0)^2 \\
&\leq 1 - \tfrac{1}{3} H(Q_0, P_0)^2 + 2H(R, P_0)^2\,.
\end{aligned}$$

Si $H(R, P_0) \leq \frac{1}{4} H(Q_0, P_0)$,

$$\mathbb{E}_R \psi(X) \leq 1 - \tfrac{5}{24} H(Q_0, P_0)^2$$

Sous R^n,

$$\begin{aligned}
&\mathbb{P}\left\{ \prod_{i=1}^n \psi(X_i) \geq 1 \right\} \\
&\leq \quad \left(\mathbb{E}_R \psi(X)\right)^n \\
&\leq \quad \exp\left(-n \frac{5}{24} H(Q_0, P_0)^2 \right)\,.
\end{aligned}$$

□

Pour séparer $\mathcal{B}(P_0, H(P_0, Q_0)/4)$ de $\mathcal{B}(Q_0, H(P_0, Q_0)/4)$ à partir d'un n-échantillon, on utilise le test T_n défini par

$$T_n(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \prod_{i=1}^n \psi(X_i) \geq 1 \\ 0 & \text{sinon.} \end{cases}$$

La construction de ce test de rapport de vraisemblance original et son analyse conduisent à un résultat facile à résumer.

COROLLAIRE 6.15 *Soit P_0, Q_0 deux lois telles que $H(P_0, Q_0) = \epsilon$, il existe un test de rapport de vraisemblance T_n tel que*

$$\max\left(\sup_{R \in \mathcal{B}(P_0, \frac{\epsilon}{4})} \mathbb{E}_R T_n, \sup_{R \in \mathcal{B}(Q_0, \frac{\epsilon}{4})} \mathbb{E}_R (1 - T_n) \right) \leq \exp\left(-n \frac{5\epsilon^2}{24} \right)\,.$$

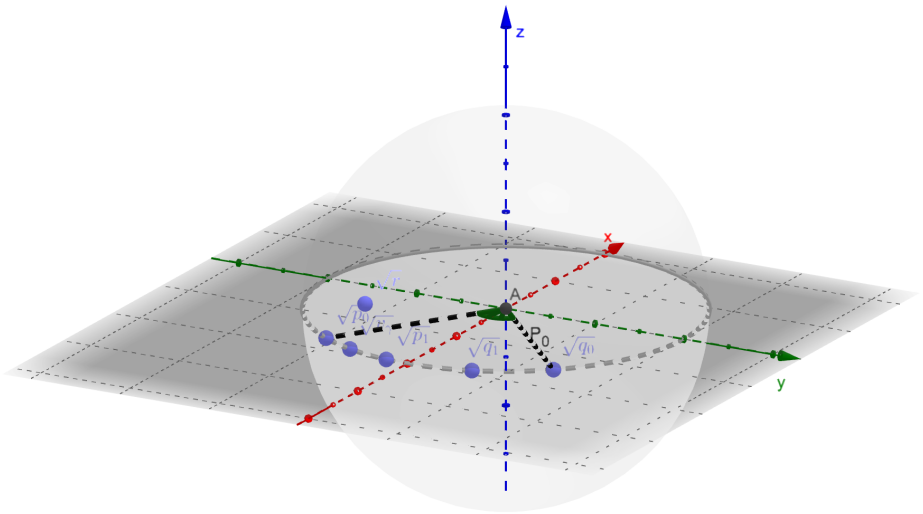

FIG. 6.1 : Arc de Hellinger engendré par $\sqrt{p_0}$ et $\sqrt{q_0}$. L'élément de l'arc de Hellinger v_γ représente la direction du plan équatorial sur laquelle se projette $\sqrt{r}$ élément de la boule de Hellinger centrée en $\sqrt{p_0}$.

6.5 REMARQUES BIBLIOGRAPHIQUES

Les mesures d'information (entropie relative, distance en variation, ...) sont discutées dans :

I. CSISZÁR et P. SHIELDS. **Information theory and statistics : A tutorial**. Now Publishers Inc, 2004.

On trouvera un model d'emploi raisonné et des comparaisons systématiques dans :

A. TSYBAKOV. **Introduction à l'estimation non-paramétrique**. T. 41. Mathématiques & Applications. Berlin : Springer-Verlag, 2004, p. x+175. MR : MR2013911(2005a:62007).

Le rôle essentiel de la distance de Hellinger dans l'analyse des modèles paramétriques est exposé dans :

A. VAN DER VAART. **Asymptotic statistics**. Cambridge University Press, 1998

La construction de tests robustes entre boules de Hellinger ouvre la voie à des méthodes d'estimations intéressantes. Cette idée remonte aux travaux de Lucien Le Cam durant les années 60 et 70. Aprés simplification, clarification et généralisation, grâce notamment à :

L. BIRGÉ. « Robust tests for model selection ». In : *From probability to statistics and back : high-dimensional models and processes*. T. 9. Inst. Math. Stat. (IMS) Collect. Inst. Math. Statist., Beachwood, OH, 2013, p. 47-64. MR : 3186748.

Ces tests robustes ont trouvé un usage dans la construction de méthodes d'estimation remarquables :

Y. BARAUD, L. BIRGÉ et M. SART. « A new method for estimation and model selection : ρ-estimation ». In : *Inventiones mathematicae* (2016), p. 1-93.

Dans le chapitre *Likelihood-based procedures* de :

E. GINÉ et R. NICKL. **Mathematical Foundations of Infinite-Dimensional Statistical Models**. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015. ISBN : 9781107043169,

on trouve des applications de la construction de tests entre boules de Hellinger à l'estimation non-paramétrique.

7.1 MOTIVATIONS

Dans la leçon précédente, nous avons adopté un point de vue abstrait et général sur les tests. Ici, nous revenons sur des constructions très classiques, qui remontent aux débuts de la statistique, à la seconde moitié du XIX$^{\text{eme}}$ siècle. Il s'agit de répondre à des questions simples : Un échantillon de valeurs issues d'un ensemble fini (les modalités) peut-il raisonnablement avoir été produit par des tirages indépendants selon une loi donnée à l'avance ? C'est le problème de l'adéquation *(goodness of fit)* à une loi fixée. D'une manière plus générale, on veut tester l'adéquation à une loi issue d'un modèle réputé régulier. Cela recouvre en particulier les tests d'indépendance, très utilisés dans les tentatives d'analyse causale. Tous les tests que nous verrons dans cette leçon relèvent de la famille des tests de type chi-deux. Ce sont des tests dont le niveau asymptotique est garanti. Leur construction passe par la construction d'une suite de vecteurs aléatoires dont la distribution limite est pivotale sous l'hypothèse nulle. Techniquement, tout repose sur le théorème central limite multivarié et la méthode Delta de Cramer.

7.2 NOTATIONS

Dans ce chapitre nous travaillons sur des tests dans les modèles multinomiaux, une variété de modèles exponentiels particulièrement simples. Ces tests sont historiquement importants. Faciles à calculer, ils ont longtemps constitué la technique de choix pour étudier les données de comptage.

Pour chaque $p \geq 1$, le modèle multinomial de dimension p est formé par les lois sur $\{0, \dots, p\}$. La paramétrisation naturelle est fournie par $\theta \in]0,1[^p$ qui vérifie $\sum_{\alpha=1}^p \theta[\alpha] < 1$ avec la convention

$$\mathbf{p}_\theta(\alpha) = \theta[\alpha] \text{ pour } \alpha > 0 \text{ et } \mathbf{p}_\theta(0) = 1 - \sum_{\alpha>0} \theta[\alpha]\,.$$

Pour rendre les notations plus transparentes, nous utiliserons la convention

$$\theta[0] = 1 - \sum_{\alpha=1}^p \theta[\alpha].$$

Un modèle multinomial est un modèle exponentiel minimal. Nous n'utilisons pas ici la paramétrisation canonique qui serait

$$\left(\log \frac{\theta[\alpha]}{\theta[0]}\right)_{\alpha\in\{1,\dots,p\}}\,.$$

Pour $\alpha \in \{0, \dots, p\}$, on notera $N^n(\alpha)$ le nombre d'occurrences de la *modalité* α dans l'échantillon de taille n.

Dans ces modèles, la log-vraisemblance s'écrit

$$\ell_n(\theta) = \sum_{\alpha=0}^p N^n(\alpha) \log \theta[\alpha] = \sum_{\alpha=1}^p N^n(\alpha) \log \theta[\alpha] + \left(n - \sum_{\alpha=1}^p N^n(\alpha)\right) \log\left(1 - \sum_{\alpha=1}^p \theta[\alpha]\right)\,.$$

La fonction score est comme d'habitude le gradient de la log-vraisemblance par rapport à θ.

On vérifie que, sous $\mathbf{p}_\theta$,

$$\frac{1}{n}\operatorname{cov}(N^n(\alpha), N^n(\beta)) = \begin{cases} \mathbf{p}_\theta(\alpha) - \mathbf{p}_\theta(\alpha)^2 & \text{si} \quad \alpha = \beta \\ -\mathbf{p}_\theta(\alpha)\mathbf{p}_\theta(\beta) & \text{sinon.} \end{cases}$$

Nous noterons $\operatorname{diag}\left(\frac{1}{\sqrt{n\mathbf{p}_\theta}}\right)$ la matrice diagonale dont les coefficients diagonaux sont les $1/\sqrt{n\mathbf{p}_\theta(\alpha)}$ pour $\alpha \in \{0, \dots p\}$.

On résume cela de façon matricielle en

$$\frac{1}{n}\operatorname{cov}(N^n) = \operatorname{diag}(\mathbf{p}_\theta) - \begin{pmatrix} \mathbf{p}_\theta(0) \\ \vdots \\ \mathbf{p}_\theta(p) \end{pmatrix} \times \big(\mathbf{p}_\theta(0), \dots, \mathbf{p}_\theta(p)\big)\,.$$

En multipliant le vecteur aléatoire $(N^n - n\mathbf{p}_\theta)/\sqrt{n}$ par $\mathrm{diag}(1/\sqrt{\mathbf{p}_\theta})$ on obtient un vecteur aléatoire de matrice de covariance

$$\mathrm{cov}\left(\left(\frac{N^n(\alpha) - n\mathbf{p}_\theta(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}}\right)_{\alpha\leq p}\right) = \mathrm{Id}_{p+1} - \begin{pmatrix}\sqrt{\mathbf{p}_\theta(0)}\\ \vdots \\ \sqrt{\mathbf{p}_\theta(p)}\end{pmatrix} \times \left(\sqrt{\mathbf{p}_\theta(0)}, \ldots, \sqrt{\mathbf{p}_\theta(p)}\right) .$$

On note $\Gamma(\theta)$ cette matrice de covariance. On reconnaît au passage qu'il s'agit de la matrice de projection sur le sous-espace orthogonal à la droite engendrée par $(\sqrt{\mathbf{p}_\theta(\alpha)})_{0\leq\alpha\leq p}$.

La matrice d'information de Fisher est la matrice de covariance sous $\mathbf{p}_\theta$ de la fonction score évaluée en θ, elle est définie par :

$$I(\theta) = \mathbb{E}_\theta\left[\nabla \log \mathbf{p}_\theta \times \nabla^t \log \mathbf{p}_\theta\right] = \left[\mathbb{E}_\theta\left[\frac{\partial_\alpha p_\theta}{p_\theta}\frac{\partial_{\alpha'} p_\theta}{p_\theta}\right]\right]_{1\leq\alpha,\alpha'\leq p} = \sum_{\alpha=0}^{p}\frac{\nabla\mathbf{p}_\theta(\alpha)}{\sqrt{\mathbf{p}_\theta(\alpha)}}\frac{\nabla\mathbf{p}_\theta(\alpha)^t}{\sqrt{\mathbf{p}_\theta(\alpha)}}.$$

La matrice d'information de Fisher en θ s'écrit

$$I(\theta) = \mathrm{diag}\begin{pmatrix}\frac{1}{\theta[1]}\\ \vdots \\ \frac{1}{\theta[p]}\end{pmatrix} + \frac{1}{\theta[0]}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \times \begin{pmatrix}1 & \vdots & 1\end{pmatrix} .$$

Remarque 7.1 La relation entre la matrice d'information et la matrice de covariance est facilement vérifiée à l'aide de la formule de Shermann-Morrison : Soit $\mathbf{A}$ une matrice inversible et u, v deux vecteurs colonnes tels que $1 + v^t\mathbf{A}^{-1}u \neq 0$. Alors

$$(\mathbf{A} + uv^t)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}uv^t\mathbf{A}^{-1}}{1 + v^t\mathbf{A}^{-1}u} .$$

(Si une perturbation de rang un d'une matrice inversible est inversible, alors l'inverse de la perturbation est une perturbation de rang un de l'inverse).

La matrice d'information est de rang p et

$$I(\theta)^{-1} = \frac{1}{n}\,\mathrm{cov}\left((N^n(\alpha))_{1\leq\alpha\leq p}\right) .$$

7.3 problèmes

Nous allons envisager plusieurs sortes de tests de type χ^2. Nous commençons par les plus simples : les tests d'adéquation à une hypothèse simple.

Définition 7.2 Test d'adéquation (ajustement) à une hypothèse simple. Soit P une loi de probabilités sur un ensemble fini Ω et $(x_1, \ldots, x_n)$ un échantillon de n points de Ω.

L'échantillon $(x_1, \ldots, x_n)$ a-t-il été engendré par n tirages indépendants selon la loi P ?

Exemple 7.3 En 1889, le docteur Geissler étudia les registres des hôpitaux de Saxe et nota le nombre de garçons dans 6115 familles de 12 enfants.

Si les sexes des enfants dans une même famille sont le résultat d'épreuves aléatoires indépendantes et identiquement distribuées, la loi du nombre de garçons dans une famille de 12 enfants est une loi binomiale de paramètres 12 et θ où $\theta \in (0, 1)$.

Les démographes nous disent que la fréquence des naissances masculines est .52.

On peut tester l'adéquation de la distribution du nombre de garçons à la loi binomiale de paramètres 12 et .52 en utilisant un test de type chi-deux.

Pour les trois autres problèmes, l'hypothèse nulle et l'alternative sont composées.

Définition 7.4 Test d'homogénéité. Soit $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$ deux échantillons de n points de Ω.

Les deux échantillons ont-ils été engendrés par n tirages indépendants selon une même loi (inconnue) ?

Nombre de garçons/12	Nombre de familles
0	3
1	24
2	104
3	286
4	670
5	1033
6	1343
7	1112
8	829
9	478
10	181
11	45
12	7

EXEMPLE 7.5 La table `Titanic` est une TABLE DE CONTINGENCES, c'est-à-dire un tableau à entrées multiples (4 ici : `Age, Sex, Class, Survived`) compilé à partir des archives du célèbre voyage. Pour chaque personne à bord (2201) , on a noté son `Age` (`Child/Adult`), sa `Class` (`1st,2nd,3rd,Crew`), son `Sex` (`Male, Female`), son destin (`Survived` : `No, Yes`). L'archive est un `data.frame` où les quatre variables sont catégorielles (ce sont des `factors`). La table de contingence indique pour chaque combinaison des modalités des variables l'effectif correspondant.

```
Titanic[c("3rd"), c("Male","Female"), "Child", ]
Survived
Sex      No Yes
Male    35  13
Female 17  14
```

Un effort de modélisation (discutable) nous conduit à concevoir le destin des personnes à bord comme la réalisation de tirages pile/face indépendants (c'est naïf compte tenu de la capacité insuffisante des canaux de sauvetage).

On peut définir une quantité de sous-populations, selon la `Class`, l'`Age`, etc, et se demander si ces différentes sous-populations ont subi le même aléa. Nous sommes conduits à mener un test d'homogénéité.

Le problème du test d'indépendance est le plus commun des tests d'hypothèse composites.

DÉFINITION 7.6 (Test d'INDÉPENDANCE.) Soit $((x_1, y_1), \dots, (x_n, y_n))$ un échantillon de n points de $\Omega \times \Omega'$.

L'échantillon a-t-il été engendré par n tirages indépendants selon une « loi produit » $P \otimes P'$ sur $\Omega \times \Omega'$?

Le problème de symétrie apparait plus rarement en pratique.

DÉFINITION 7.7 (Test de SYMÉTRIE). Soit $((x_1, y_1), \dots, (x_n, y_n))$ un échantillon de n points de $\Omega \times \Omega$.

L'échantillon a-t-il été engendré par n tirages indépendants selon une « loi » P sur $\Omega \times \Omega$, telle que $P\{A \times B\} = P\{B \times A\}$ pour tous $A, B \subseteq \Omega$?

7.4 TEST DU χ^2 D'ADÉQUATION

Comme dans le reste du chapitre, Ω est un ensemble fini de taille $p+1$ (identifié à $\{0, \dots, p\}$), l'hypothèse nulle est définie par $(\mathbf{p}_\theta(\alpha))_{\alpha \leq p}$. La paramétrisation naturelle est fournie par $\theta \in]0,1[^p$ qui vérifie $\sum_{\alpha=1}^{p} \theta[\alpha] < 1$.

On pourrait de fait aborder le problème du test d'adéquation à l'hypothèse simple définie par $(\theta[\alpha])_{\alpha>0}$ avec les outils développés en général pour les modèles exponentiels, notamment avec les régions de confiance inspirées par le phénomène de Wilks (voir Section 5.5).

Nous allons reprendre l'approche suivie par Pearson au début du vingtième siècle.

Définition 7.8 (statistique du χ^2 de pearson) Pour $\alpha \in \{0, \dots, p\}$, on note $N^n(\alpha)$ le nombre d'occurrences de la modalité α dans l'échantillon de taille n. La statistique de Pearson est égale à

$$\sum_{\alpha=0}^{p} \frac{(N^n(\alpha) - n\mathbf{p}_\theta(\alpha))^2}{n\mathbf{p}_\theta(\alpha)} = \left\| \left(\frac{N^n(\alpha) - n\mathbf{p}_\theta(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}} \right)_{\alpha \leq p} \right\|^2 .$$

On note cette statistique $C_n(\theta)$.

Si on note $O_\alpha := N^n(\alpha)$ (O comme *observed*), et $E_\alpha := n\mathbf{p}_\theta(\alpha)$ (E comme *expected*), cette statistique se présente comme le carré de la norme du vecteur aléatoire

$$\left(\frac{O_\alpha - E_\alpha}{\sqrt{E_\alpha}} \right)_{\alpha \in \Omega} ,$$

une forme que nous reverrons à propos des χ^2 d'adéquation à des hypothèses composites.

Cette statistique s'interprète aussi comme une *divergence d'information*. Nous avons déjà mentionné

$$\chi^2(P \mid Q) := \int \frac{\mathrm{d}Q}{\mathrm{d}\nu} \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1 \right)^2 \mathrm{d}\nu = \int \frac{\mathrm{d}P}{\mathrm{d}\nu} \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1 \right) \mathrm{d}\nu .$$

Si on note P_n la loi empirique définie par l'échantillon et P la loi définie par θ, la statistique de Pearson s'écrit aussi

$$n\chi^2 \left(P_n \mid P \right) .$$

Remarque 7.9 Vérifier que

$$D(P\|Q) \leq \chi^2(P \mid Q) .$$

Montrer que pour certaines couples de lois de probabilité, la divergence du χ^2 peut étre infinie alors que l'entropie relative est finie.

Rappelons

$$\frac{1}{n} \operatorname{cov}(N^n, N^n) = \operatorname{diag}(\mathbf{p}_\theta) - \begin{pmatrix} \mathbf{p}_\theta(0) \\ \vdots \\ \mathbf{p}_\theta(p) \end{pmatrix} \times \left(\mathbf{p}_\theta(0), \dots, \mathbf{p}_\theta(p) \right) = \Gamma(\theta) .$$

Observons aussi que le vecteur aléatoire N^n est une somme de vecteurs aléatoires indépendants et identiquement distribués : notons X_i la variable aléatoire à valeur dans $\{0, \dots, p\}$ qui vaut α si le i^{eme} élément de l'échantillon vaut α, pour chaque modalité $\alpha \in \{1, \dots, p\}$

$$N^n(\alpha) = \sum_{i=1}^{n} \mathbb{I}_{X_i = \alpha} .$$

Le théorème central limite vectoriel nous indique alors que sous l'hypothèse nulle

$$\left(\frac{N^n(\alpha) - n\mathbf{p}_\theta(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}} \right)_{\alpha \leq p} \rightsquigarrow \mathcal{N}(0, \Gamma(\theta))$$

où $\Gamma(\theta)$ est définie dans la section d'introduction.

Comme la matrice $\Gamma(\theta)$ est la matrice de projection sur le sous-espace orthogonal à la droite engendrée par $(\sqrt{\mathbf{p}_\theta(\alpha)})_{0 \leq \alpha \leq p}$, on peut conclure avec le théorème de Cochran que sous l'hypothèse nulle, la loi limite de la statistique de Pearson est χ^2_p.

Ceci nous fournit une méthode de test de niveau asymptotique garanti : pour atteindre le niveau $\eta \in]0, 1[$, on compare la statistique de Pearson $C_n(\theta)$ au quantile $q_{p,1-\eta}$ d'ordre $1 - \eta$ de la loi χ^2_p. Si $C_n(\theta) > q_{p,1-\eta}$, on rejette l'hypothèse nulle,

Exemple 7.10 (retour sur les données geissler) On effectue un test d'adéquation à une hypothèse simple : la probabilité d'observer une famille comportant α garçons est

$$\mathbf{p}_\theta(\alpha) = \binom{12}{\alpha}\theta^\alpha(1-\theta)^{12-\alpha}$$

pour $\alpha \in \{0,\dots,12\}$. On note O_α le nombre de familles comportant α garçons parmi 12 recensées par le docteur Geissler. On note $E_\alpha = n\mathbf{p}_\theta(\alpha)$ l'espérance du nombre de familles comportant α garçons parmi 12 dans un échantillon i.i.d. de $n = 6115$ observations.

La statistique de Pearson est

$$\sum_{\alpha=0}^{12}\frac{(O_\alpha - E_\alpha)^2}{E_\alpha}\,.$$

```
chisq.test(saxony[,2],p=dbinom(c(0:12),12,prob=.52))

Chi-squared test for given probabilities

data:  saxony[, 2]
X-squared = 110.53, df = 12, p-value < 2.2e-16

Warning message:
In chisq.test(saxony[, 2], p = dbinom(c(0:12), 12, prob = 0.52)):
Chi-squared approximation may be incorrect
```

Nous reviendrons plus loin sur l'avertissement.

La comparaison de la statistique de Pearson `X-squared` aux quantiles de la loi χ^2_{12} (chi-deux à `df`$= 12$ degrés de liberté) conduit à rejeter l'hypothèse nulle (nombre de garçons dans une famille de 12 enfants distribué selon une binomiale de paramètres 12 et .52) si on a choisi un niveau supérieur à 2.2×10^{-16}.

L'usage pratique des tests passe par la compréhension de la notion de p-valeur.

Définition 7.11 (p-value, degré de signification atteint). Soit une famille de tests binaires où l'hypothèse nulle est simple et définie par la loi P. Soient les tests définis en comparant une statistique S à un seuil et en rejetant l'hypothèse null lorsque ce seuil est dépassé (le seuil définit le niveau). La p-valeur est définie par $\overline{F}(S) = 1 - F(S)$ où F est la fonction de répartition de la loi de la statistique S sous l'hypothèse nulle.

Remarque 7.12 Lorsqu'on utilise une fonction de test dans un environment de calcul statistique, le niveau n'est en général pas décidé d'avance. Le fait de rapporter systématiquement la valeur de la statistique de test et la p-valeur associée, permet de reculer la décision. Lorsqu'on veut tester au niveau η, il n'est meme pas nécéssaire de comparer la statistique de test au quantile d'ordre $1-\eta$ de la loi de la statistique sous l'hypthèse nulle, il suffit de comparer la p-valeur à η. Si cette p-valeur est plus petite que η, on rejette l'hypothèse nulle. Cela garantit un test de niveau η.

Proposition 7.13 *Sous une hypothèse nulle simple, si la loi de la statistique S est diffuse, la loi de la p-valeur est la loi uniforme sur* $[0,1]$.

preuve. La proposition est le corollaire immédiat de résultat classique Si $X \sim F$ avec F continue, alors $F(X)$ est uniformément distribuée sur $[0,1]$. □

7.5 puissance du test du χ^2

Dans la mesure où l'alternative comprend toutes les lois sur Ω, l'étude de la puissance du test du χ^2 d'adéquation peut sembler un défi. L'étude asymptotique la rend relativement simple.

Alternative fixe

Si on choisit $\theta' \neq \theta$, pour un $\alpha \in \{0, \ldots, p\}$ au moins $\mathbf{p}_\theta(\alpha) \neq \mathbf{p}_{\theta'}(\alpha)$. Sous la loi définie par $\mathbf{p}_{\theta'}$, $\frac{N^n(\alpha)}{n}$ converge presque sûrement vers $\mathbf{p}_{\theta'}(\alpha)$ et donc

$$\left(\frac{N^n(\alpha) - n\mathbf{p}_\theta(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}}\right)^2 = n\left(\frac{N^n(\alpha)/n - \mathbf{p}_\theta(\alpha)}{\sqrt{\mathbf{p}_\theta(\alpha)}}\right)^2 \xrightarrow{P} +\infty$$

Quelque soit le niveau fixé, quelque soit le point de l'alternative, la probabilité d'erreur de seconde espèce tend vers 1.

Cette observation est rassurante mais pas très informative.

Suite d'alternatives contigue

Pour apprécier les performances de la statistique du χ^2 d'adéquation, il est bon de considérer non pas une alternative fixe, mais une suite d'alternatives qui se rapprochent de l'hypothèse nulle. Dans la suite $h \in \mathbb{R}^{p+1}$ vérifie $\sum_{\alpha=0}^p h[\alpha] = 0$. Pour n assez grand

$$\mathbf{p}_{\theta_n} := \mathbf{p}_\theta + \frac{h}{\sqrt{n}}$$

définit bien une probabilité P_{θ_n} sur $\{0, \ldots, p\}$. La divergence du χ^2 entre P_{θ_n} et P_θ se calcule simplement :

$$\chi^2(P_{\theta_n} \mid P_\theta) = \frac{1}{n}\sum_{\alpha=0}^p \frac{h[\alpha]^2}{\theta[\alpha]} = \frac{1}{n}h^t I(\theta)h\,.$$

Dans la dernière expression nous abusons des notations : $I(\theta)$ est une matrice de dimension p par p, h est introduit comme un vecteur indicé entre 0 et p. Il faut comprentre

$$h^t I(\theta)h = \sum_{1\leq\alpha,\beta\leq p} I(\theta)[\alpha,\beta]h(\alpha)h(\beta)\,.$$

Elle permet de quantifier la vitesse à laquelle P_{θ_n} s'approche de P_θ dans la direction définie par h. On peut aussi calculer un équivalent de la distance de Hellinger entre P_{θ_n} et P_θ :

$$\lim_{n\to\infty} nH(P_{\theta_n}, P_\theta)^2 = \frac{1}{4}\lim_n n\chi^2(P_{\theta_n} \mid P_\theta)\,.$$

La limite des lois images de $P_{\theta_n}^{\otimes n}$ par les statistiques de Pearson $n\chi^2(P_{\theta_n} \mid P_\theta)$ se détermine.

PROPOSITION 7.14 *La limite des lois de la statistique de Pearson $C_n(\theta)$ sous $(P_{\theta_n}^{\otimes n})_n$ est la loi du χ^2 à p degrés de liberté décentrée, de paramètre de décentrage*

$$h^t I(\theta)h = \sum_{\alpha=0}^p h_\alpha^2/\theta_\alpha = n\chi^2(P_{\theta_n} \mid P_\theta)\,.$$

Cette loi est notée $\chi_p^2(h^t I(\theta)h)$.

PREUVE. Comme

$$\left(\frac{N^n(\alpha) - n\mathbf{p}_\theta(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}}\right)_\alpha = \left(\frac{N^n(\alpha) - n\mathbf{p}_{\theta_n}(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}}\right)_\alpha + \left(\frac{h(\alpha)}{\sqrt{\mathbf{p}_\theta(\alpha)}}\right)_\alpha\,,$$

pour établir le résultat, il suffit d'établir que sous la suite $(P_{\theta_n}^{\otimes n})$,

$$\left(\frac{N^n(\alpha) - n\mathbf{p}_{\theta_n}(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}}\right)_\alpha \rightsquigarrow \mathcal{N}(0, \Gamma(\theta))\,.$$

Le mécanisme de Cramer-Wold (Théorème C.6 en appendice) et le lemme de Slutsky nous indiquent qu'il suffit de vérifier cette convergence pour les formes linéaires appliquées aux vecteurs aléatoires

$$\frac{N^n - n\mathbf{p}_{\theta_n}}{\sqrt{n\mathbf{p}_{\theta_n}}} := \left(\frac{N^n(\alpha) - n\mathbf{p}_{\theta_n}(\alpha)}{\sqrt{n\mathbf{p}_{\theta_n}(\alpha)}}\right)_\alpha .$$

Soit $\lambda \in \mathbb{R}^p$, on note $\sigma_n^2 := \lambda^t \Gamma(\theta_n)\lambda$ et $\sigma^2 := \lambda^t \Gamma(\theta)\lambda$, soit $\lim_n \sigma_n^2 = \sigma$.

On définit

$$S_n := \left\langle \lambda, \frac{N^n - n\mathbf{p}_{\theta_n}}{\sqrt{n\mathbf{p}_{\theta_n}}} \right\rangle .$$

Pour chaque n, S_n est une somme de variables aléatoires indépendantes identiquement distribuées, centrées. La variance de S_n est σ_n^2. La distribution des variables aléatoires dont la somme centrée et normalisée donne S_n dépend de n. On dit qu'on a affaire à un *tableau triangulaire*. Le théorème Central Limite de Lindeberg-Feller (Théorème C.7) est justement valable pour les tableaux triangulaires. On vérifie facilement que les conditions sont remplies dans le cas des vecteurs multinomiaux.

□

Lors de l'étude du test de Fisher en régression linéaire gaussienne, nous avons vu que la loi du χ^2 décentrée domine stochastiquement la loi du χ^2 centrée. Nous pouvons donc conclure que sous la suite P_{θ_n} le test du χ^2 de niveau α est sous la suite d'alternatives contigue $(\theta_n)_n$, de puissance asymptotique supérieure ou égale à α (on dit que le test est *sans biais*).

On peut regarder cette étude de la puissance asymptotique du test du χ^2 sur une suite l'alternatives contigue dans la perspective du théorème 6.13. Ce théorème nous indique que pour tout seuil $\tau > 0$

$$\max\left(P_\theta^{\otimes n}\{C(n, \mathbf{p}_\theta) \geq \tau\}, P_{\theta_n}^{\otimes n}\{C(n, \mathbf{p}_\theta) < \tau\}\right) \geq \frac{1}{4} - \frac{1}{2} n H(P_\theta, P_{\theta_n})^2 .$$

Si F_p (resp. $F_{p,\delta}$) désigne la fonction de répartition de la loi χ_p^2 (resp. $\chi_p^2(\delta)$), en choisissant

$$\delta = \sum_{\alpha=0}^{p} h(\alpha)^2 / \mathbf{p}_\theta(\alpha) = h^t I(\theta) h$$

(on choisit bien sûr h de façon à avoir $\delta < 2$), la limite du membre gauche est

$$\max(\eta, F_{p,\delta}(F_p^{\leftarrow}(1-\eta)))$$

alors que celle du membre droit est $\frac{1}{4} - \frac{\delta}{8}$. Pour $\delta = 1$, $p = 12$, le minimum du membre gauche et atteint autour de $\eta \approx .45$ et il est proche de .45. L'écart avec la minoration (1/8) ne permet pas de déclarer que le test du χ^2 est asymptotiquement optimal (la minoration basée sur la distance de Hellinger).

Ce calcul suscite une question : peut-on concevoir un test d'adéquation à P_θ qui soit asymptotiquement de niveau η et pour toute suite $(\theta_n)_n$ d'alternatives contigue vérifiant $n\chi^2(P_{\theta_n} \mid P_\theta) \to \delta$, de niveau asymptotique supérieur à $F_{p,\delta}(F_p^{\leftarrow}(1-\eta))$?

7.6 HYPOTHÈSE NULLE COMPOSITE

Modifications de la statistique de Pearson

Notons Θ_0 un ouvert de $\mathbb{R}^r$ (avec $r < p$), on supposera dans la suite qu'il existe une fonction deux fois (continuement) différentiable de Θ_0 dans l'ensemble des fonctions de masse de probabilité sur $\{0, \ldots, p\}$. On notera encore $\mathbf{p}_\theta$ la (fonction de masse de) probabilité associée à $\theta \in \Theta_0$.

On peut généraliser le problème du test d'ajustement à la situation suivante : si l'échantillon $x_1, \ldots, x_n$ a été tiré selon une loi P inconnue sur $\{0, \ldots, p\}$, tester

H_0 : $\exists \theta \in \Theta_0$ tel que la loi P est définie par $\mathbf{p}_\theta$
H_1 : $\nexists \theta \in \Theta_0, P$ est définie par $\mathbf{p}_\theta$.

On pourra vérifier que les trois situations données en introduction (homogénéité, indépendance, symétrie) rentrent dans ce cadre. Si on considère Θ_0 comme un sous-modèle et qu'on dispose d'un (d'une suite d') estimateur(s) $\widehat{\theta}$ pour ce modèle, on peut aussi procéder de la façon suivante.

Définition 7.15 Si on note $N^n(\alpha)$ le nombre d'occurrences de $\alpha \in \{0,\dots,p\}$ dans l'échantillon ($N_i^n \equiv \sum_{t=1}^n \mathbb{1}_{x_t=i}$), le test du χ^2 d'adéquation à $\Theta_0 \subset \mathbb{R}^r$ consiste à comparer la statistique de Pearson modifiée :

$$C(n, \mathbf{p}_{\widehat{\theta}}) \equiv \sum_{\alpha=0}^{p} \left(\frac{N^n(\alpha) - n\mathbf{p}_{\widehat{\theta}}(\alpha)}{\sqrt{n\mathbf{p}_{\widehat{\theta}}(\alpha)}} \right)^2$$

où $\widehat{\theta}$ est un estimateur pour le modèle Θ_0, à un seuil $\tau > 0$, à rejeter si $C(n, \Theta_0)$ est supérieur à ce seuil.

Exemple 7.16 (retour sur les données geissler) Le nombre moyen de garçons dans les familles de l'échantillon 0.519 est à peine différent de celui qu'on apprend dans les livres (.52), l'écart est en tous cas plus petit que l'écart-type de la moyenne empirique de 73380 épreuves de Bernoulli de paramètre .52.

Si on soumet les données `saxony` au test d'adéquation du χ^2 calculé par la fonction `chisq.test`, on obtient le résultat suivant.

```
chisq.test(saxony[,2],p=dbinom(c(0:12),12,prob=hattheta))

Chi-squared test for given probabilities

data:  saxony[, 2]
X-squared = 110.5, df = 12, p-value < 2.2e-16

Warning message:
In chisq.test(saxony[, 2], p = dbinom(c(0:12), 12, prob = hattheta)):
Chi-squared approximation may be incorrect
```

La statistique `X-squared` est bien la statistique de Pearson modifiée, puisque le paramètre `p` est défini à partir de $\widehat{\theta}_n \approx .519$ calculé sur les données. Cette statistique n'est pas comparée au quantile de la loi χ^2_{11} mais au quantile de la loi χ^2_{12}. La fonction `chisq.test` ne peut tenir compte du fait que les paramètres sont estimés.

Le message d'avertissement est lié au fait que pour $\alpha \leq 1$ et $\alpha \geq 11$, les effectifs attendus E_α sont inférieurs à 5, dans ce cas $(N^n_\alpha - n\mathbf{p}_{\widehat{\theta}_n}(\alpha))/\sqrt{n\mathbf{p}_{\widehat{\theta}_n}}$ n'est pas approximativement normale et le raisonnement asymptotique qui justifie la comparaison de la statistique de Pearson aux quantiles de la loi χ^2_{11} n'est pas tenable.

On regroupe les familles comportant 0 et 1 garçons, ainsi que les familles comportant 11 et 12 garçons. On compare à nouveau les fréquences observées et les probabilités attendues en utilisant la loi binomiale de paramètres 12 et .5192...

```
chisq.test(Saxonygrp,p=dbingrp)

Chi-squared test for given probabilities

data:  Saxonygrp
X-squared = 105.8463, df = 10, p-value < 2.2e-16
```

Si l'hypothèse binomiale est correcte, asymptotiquement, cette statistique se comporte selon une loi χ^2_9. La probabilité qu'une variable χ^2_9 distribuée atteigne la valeur 105.8 est inférieure à la précision de la machine.

Si on a choisi un niveau supérieur à 10^{-16}, on est conduit à rejeter l'hypothèse nulle (dans une même famille, les sexes des enfants sont indépendamment et identiquement distribués selon une loi de Bernoulli dont la probabilité est un propriété de l'espèce).

Loi limite de la statistique de Pearson modifiée

Dans cette section, nous caractérisons la loi limite de $C(n, \mathbf{p}_{\widehat{\theta}})$ sous l'hypothèse nulle lorsque Θ_0 est un modèle « régulier » et $\widehat{\theta}$ un estimateur du maximum de vraisemblance. Pour établir que cette loi limite est une loi du χ^2 dont le nombre de degrés de liberté est égal à la différence entre p et la « dimension » de Θ_0 (soit r si Θ_0 est un ouvert de $\mathbb{R}^r$), nous allons recourir à quelques outils probabilistes rappelés ici.

Hypothèses techniques : modèles multinomiaux réguliers
On suppose que :

i) l'espace des paramètres $\Theta \subset \mathbb{R}^r$ est ouvert,
ii) la fonction de $\Theta_0 \to]0,1[^{p+1}$, $\theta \mapsto \mathbf{p}_\theta$ est deux fois continument différentiable en θ.
iii) la matrice d'information de Fisher est encore définie par :

$$I(\theta) \equiv \mathbb{E}_\theta \left[\nabla \log \mathbf{p}_\theta \times \nabla^t \log \mathbf{p}_\theta \right] = \left[\mathbb{E}_\theta \left[\frac{\partial_i p_\theta}{p_\theta} \frac{\partial_{i'} p_\theta}{p_\theta} \right] \right]_{i,i' \leq r} = \sum_{\alpha=0}^{p} \frac{\nabla \mathbf{p}_\theta(\alpha)}{\sqrt{\mathbf{p}_\theta(\alpha)}} \frac{\nabla \mathbf{p}_\theta(\alpha)^t}{\sqrt{\mathbf{p}_\theta(\alpha)}}.$$

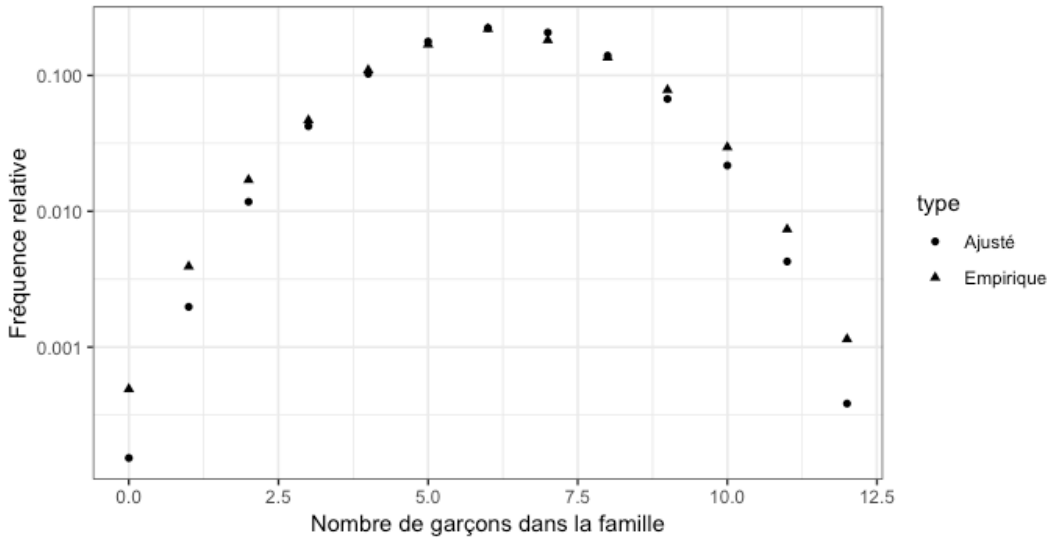

FIG. 7.1 : Les triangles représentent les proportions observées dans les données du docteur Geissler, les disques représentent les proportions attendues dans un échantillon de réalisations indépendantes de la loi binomiale de paramètres 12 et .519 (ajusté). On note que les familles équilibrées sont sous-représentées dans l'échantillon du Dr. Geissler. Cette visualisation complète l'interprétation de la statistique du χ^2. L'hypothèse d'une distribution binomiale du nombre de garçons dans la progéniture est rejetée. On pourrait tester d'autre hypothèses. L'une des plus simples postule que nous avons affaire à un *mélange* de deux types de famille. Dans chaque type de famille, le nombre de garçons est distribué selon une binomiale, et le paramètre de succès de la binomiale dépend du type de la famille. Ce modèle est défini par trois paramètres : les paramètres de succès des deux binomiales et les proportions du mélange. Ce n'est pas un modèle exponentiel.

Elle est supposée inversible, donc de rang r, en tout $\theta \in \Theta_0$.

iv) La fonction $\theta \mapsto I(\theta)$ est supposée continue.

On peut définir une matrice $B(\theta)$ de dimensions $r \times (p+1)$ et factoriser la matrice d'information $I(\theta)$:

$$B(\theta) := \left(\frac{\partial_j \mathbf{p}_\theta(\alpha)}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \right)_{\substack{1 \le j \le r \\ 0 \le \alpha \le p}} \quad \text{et } I(\theta) := B(\theta) \times B(\theta)^t. \tag{7.1}$$

La colonne de $B(\theta)$ indexée par α est $\frac{1}{\sqrt{\mathbf{p}_\theta(\alpha)}} \nabla \mathbf{p}_\theta(\alpha)$. On utilisera plus tard le fait que

$$B(\theta) \times \begin{pmatrix} \sqrt{\mathbf{p}_\theta}(0) \\ \vdots \\ \sqrt{\mathbf{p}_\theta}(p) \end{pmatrix} = \left(\sum_{\alpha=0}^{p} \partial_j \mathbf{p}_\theta(\alpha) \right)_{1 \le j \le r} = 0\,.$$

REMARQUE 7.17 On a aussi :

$$I(\theta) = \left(-\mathbb{E}_\theta \left[\partial_{j,j'} \log \mathbf{p}_\theta \right] \right)_{j,j' \le r} = -\mathbb{E}\left[\nabla^2 \log \mathbf{p}_\theta \right] .$$

Les coefficients de la matrice d'information peuvent en effet s'écrire

$$-\sum_{\alpha=0}^{p} \frac{\partial_{j,j'} \mathbf{p}_\theta\{\alpha\} \times \mathbf{p}_\theta\{\alpha\} - \partial_j \mathbf{p}_\theta\{\alpha\} \times \partial_{j'} \mathbf{p}_\theta\{\alpha\}}{\mathbf{p}_\theta\{\alpha\}} \qquad \text{pour } 1 \le j, j' \le r.$$

La possibilité d'écrire l'information de Fisher comme l'espérance du Hessien de la log-vraisemblance est une propriété de certains modèles statistiques, pas de tous, elle est notamment vérifiée dans les modèles exponentiels. L'information de Fisher est définie comme la covariance sous P_θ du vecteur score évalué en θ.

Rappelons que la matrice $(p+1)\times(p+1)$, $\Gamma(\theta)$ est définie par

$$\Gamma(\theta) := \mathrm{Id}_{p+1} - \sqrt{\mathbf{p}_\theta}\sqrt{\mathbf{p}_\theta}^t\,. \tag{7.2}$$

C'est la matrice de la projection orthogonale sur le sous-espace orthogonal à la droite engendrée par le vecteur unitaire $\sqrt{\mathbf{p}_\theta}$.

Nous notons $\mathrm{diag}\left(\frac{1}{\sqrt{n\mathbf{p}_\theta}}\right)$ la matrice diagonale dont les coefficients diagonaux sont les $\frac{1}{\sqrt{n\mathbf{p}_\theta(\alpha)}}$ pour $\alpha\in\{0,\dots p\}$.

On note

$$Z^n := \left(\frac{N^n(\alpha)-n\mathbf{p}_\theta\{\alpha\}}{\sqrt{n\mathbf{p}_\theta\{\alpha\}}}\right)_{\alpha\le p} = \mathrm{diag}\left(\tfrac{1}{\sqrt{n\mathbf{p}_\theta}}\right)\times(N^n - n\mathbf{p}_\theta)\,.$$

Dans la suite $\ell_n(\theta')$ désigne la log-vraisemblance de l'échantillon en $\theta'\in\Theta_0$

$$\ell_n(\theta') \equiv \sum_{l=1}^n \log \mathbf{p}_{\theta'}\{x_l\}\,,$$

pour chaque $j\le r$,

$$\partial_j\ell_n(\theta') = \sum_{l=1}^n \frac{\partial_j\mathbf{p}_{\theta'}\{x_l\}}{\mathbf{p}_{\theta'}\{x_l\}}\,.$$

La fonction score est définie comme le gradient de la log-vraisemblance : $\nabla\ell_n(\theta') = (\partial_j\ell_n(\theta'))_{j\le r}$, c'est un vecteur colonne.

On définit la matrice symétrique $A(\theta)$ par

$$A(\theta) := \begin{pmatrix} \Gamma(\theta) & \vdots & B(\theta)^tI^{-1}(\theta) \\ \dots & \dots & \dots \\ I^{-1}(\theta)B(\theta) & \vdots & I^{-1}(\theta)\end{pmatrix}$$

où $\Gamma(\theta)$ est défini par (7.2) et $B(\theta)$ et $I(\theta)$ sont définis par (7.1). On note la factorisation

$$A(\theta) = \begin{pmatrix} \mathrm{Id}_{p+1} \\ \dots \\ I^{-1}(\theta)B(\theta)\end{pmatrix}\times\Gamma(\theta)\times\begin{pmatrix}\mathrm{Id}_{p+1} & \vdots & B(\theta)^tI^{-1}(\theta)\end{pmatrix}$$

qui se vérifie en n'oubliant pas que $\Gamma(\theta)B(\theta)^tI^{-1}(\theta) = B(\theta)^tI^{-1}(\theta)$. En effet

$$\begin{aligned}\Gamma(\theta)B(\theta)^t &= B(\theta)^t - \begin{pmatrix}\sqrt{\mathbf{p}_\theta(0)}\\ \vdots \\ \sqrt{\mathbf{p}_\theta(p)}\end{pmatrix}\times\left(\sqrt{\mathbf{p}_\theta(0)},\dots,\sqrt{\mathbf{p}_\theta(p)}\right)\times B(\theta)^t \\ &= B(\theta)^t\,.\end{aligned}$$

Le résultat essentiel sur le comportement asymptotique de la statistique de Pearson (étendue) est donné dans le théorème suivant.

THÉORÈME 7.18 *Si $\Theta_0\subset\mathbb{R}^r$ définit un modèle multinomial régulier et si $\hat\theta$ est un estimateur (consistant) au maximum de vraisemblance alors pour tout $\theta\in\Theta_0$, sous $\mathbf{p}_\theta$*

$$C(n,\mathbf{p}_{\hat\theta}) \rightsquigarrow X \quad où \quad X\sim\chi^2_{p-r}.$$

La preuve du Théorème s'appuie sur le lemme suivant qui implique au passage la normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle Θ_0.

LEMME 7.19 *Sous les conditions du théorème 7.18, pour tout $\theta \in \Theta_0$, sous $\mathbf{p}_\theta$,*

$$I(\theta)\sqrt{n}(\widehat{\theta}_n - \theta) - \frac{1}{\sqrt{n}}\nabla\ell_n(\theta) \xrightarrow{\mathbf{p}_\theta} 0. \tag{7.3}$$

REMARQUE 7.20 Ce couplage asymptotique entre $I(\theta)^{-1}\frac{1}{\sqrt{n}}\nabla\ell_n(\theta)$ et $\sqrt{n}(\widehat{\theta}_n - \theta)$ n'est pas une propriété spécifique des modèles multinomiaux réguliers mais des modèles dits réguliers en général. Dans le cadre des modèles exponentiels, cette convergence en probabilité a déjà été mise à profit pour étudier les méthodes d'estimation « à un pas ».

Ce couplage peut être établi dans des conditions beaucoup moins exigeantes que celles énoncées ici.

LEMME 7.21 *Sous les conditions du théorème 7.18, pour tout $\theta \in \Theta_0$, sous $\mathbf{p}_\theta$, la suite de vecteurs aléatoires de dimension $p+1+r$*

$$X^n := \begin{pmatrix} Z^n \\ \cdots \\ \sqrt{n}\left(\hat{\theta}_n - \theta\right) \end{pmatrix}$$

converge en loi vers un vecteur gaussien centré de matrice de covariance $A(\theta)$.

La démonstration du Lemme 7.19 commence par vérifier que le score (le gradient de la log-vraisemblance $\ell_n(.)$) en θ est une transformation linéaire de Z^n.

PREUVE.(LEMME 7.19) Le théorème central limite vectoriel nous assure que la suite des vecteurs (Z_n) converge en loi vers $\mathcal{N}(0, \Gamma(\theta))$.

On vérifie d'abord que la fonction score est une transformation linéaire de Z_n.

$$\begin{aligned}
\frac{1}{\sqrt{n}}\nabla\ell_n(\theta) &= \frac{1}{\sqrt{n}}\sum_{\alpha=0}^{p} N^n(\alpha)\frac{\nabla\mathbf{p}_\theta\{\alpha\}}{\mathbf{p}_\theta\{\alpha\}} \\
&\qquad \text{comme } 0 = \sum_{\alpha=0}^{p}\nabla\,\mathbf{p}_\theta\{\alpha\}, \\
&= \sum_{\alpha=0}^{p}\frac{1}{\sqrt{n}}\frac{(N^n(\alpha) - n\mathbf{p}_\theta\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \times \frac{\nabla\mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \\
&= B(\theta) \times Z^n.
\end{aligned}$$

Passons au « couplage » (7.3). Par définition du maximum de vraisemblance $\nabla\ell_n(\hat{\theta}) = 0$, soit $0 = \partial_i\ell_n(\hat{\theta})$ pour $i \leq r$. On utilise la différentiabilité de la log-vraisemblance et un développement autour de θ par la formule de Taylor avec reste intégral. Pour tout $i \leq r$,

$$0 = \partial_i\ell_n(\widehat{\theta}_n) = \partial_i\ell_n(\theta) + \sum_{j=1}^{r}(\widehat{\theta}_n[j] - \theta[j])\int_0^1 \partial_{j,i}\ell_n(\theta + u(\widehat{\theta}_n - \theta))\mathrm{d}u$$

soit en divisant les 2 membres par $1/\sqrt{n}$:

$$0 = \sum_{\alpha=0}^{p}\frac{\partial_i\mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}\frac{(N^n(\alpha) - n\mathbf{p}_\theta\{\alpha\})}{\sqrt{n\mathbf{p}_\theta\{\alpha\}}} + \sum_{j=1}^{r}\sqrt{n}(\widehat{\theta}_n[j] - \theta[j])\frac{1}{n}\int_0^1 \partial_{j,i}\ell_n(\theta + u(\widehat{\theta}_n - \theta))\mathrm{d}u.$$

Vérifions maintenant que la matrice aléatoire de terme général

$$\left(\frac{1}{n}\int_0^1 \partial_{i,j}\ell_n(\theta + u(\widehat{\theta}_n - \theta))\mathrm{d}u\right)$$

converge en probabilité vers $-I(\theta)$ dont le terme général est aussi la limite en probabilité de

$$\frac{1}{n}\partial_{i,j}\ell_n(\theta) = \frac{1}{n}\int_0^1 \partial_{i,j}\ell_n(\theta)\mathrm{d}u.$$

Nous voulons établir

$$\frac{1}{n}\int_0^1 \left[\partial_{i,j}\ell_n(\theta + u(\widehat{\theta}_n - \theta)) - \partial_{i,j}\ell_n(\theta)\right]\mathrm{d}u \xrightarrow{P} 0.$$

Notons $w(\delta) \equiv \sup_{\theta':\|\theta-\theta'\|\leq\delta}\sup_{i,j\leq r}\sup_\alpha |\partial_{i,j}\log \mathbf{p}_{\theta'}\{\alpha\} - \partial_{i,j}\log \mathbf{p}_\theta\{\alpha\}|$. Cette quantité est une fonction continue de δ et tend vers 0 lorsque δ tend vers 0. On a par ailleurs

$$\left|\frac{1}{n}\int_0^1 \left[\partial_{i,j}\ell_n(\theta + u(\widehat{\theta}_n - \theta)) - \partial_{i,j}\ell_n(\theta)\right]\mathrm{d}u\right| \leq w\left(\|\theta - \widehat{\theta}_n\|\right).$$

L'hypothèse de consistance de l'estimateur $\widehat{\theta}_n$ et l'hypothèse de continuité de la matrice d'information de Fisher permettent alors d'établir la convergence en probabilité recherchée.

Par ailleurs la loi des grands nombres permet de conclure que

$$\frac{1}{n}\partial_{i,j}\ell_n(\theta) \xrightarrow{P} \mathbb{E}_\theta\left[\partial_{i,j}\log \mathbf{p}_\theta\right].$$

□

PREUVE. (LEMME 7.21) D'après le lemme 7.19, la limite en loi de la suite $(X^n)_{n\in\mathbb{N}}$ coïncide avec la limite en loi des images des vecteurs Z^n par la transformation linéaire définie par

$$\begin{pmatrix} \mathrm{Id}_{p+1} \\ \cdots \\ I^{-1}(\theta)B(\theta) \end{pmatrix}.$$

La limite est un vecteur gaussien centré. La covariance de cette limite est donnée par

$$\begin{pmatrix} \mathrm{Id}_{p+1} \\ \cdots \\ I^{-1}(\theta)B(\theta) \end{pmatrix} \times \Gamma(\theta) \times \begin{pmatrix} \mathrm{Id}_{p+1} & \vdots & B(\theta)^t I^{-1}(\theta) \end{pmatrix} = A(\theta).$$

□

Nous pouvons passer à la démonstration du théorème.

PREUVE.(THÉORÈME 7.18) En θ, la différentielle de l'application

$$\theta' \mapsto \left(\frac{\mathbf{p}_{\theta'}\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}\right)_{\alpha\leq p}$$

vaut exactement $B(\theta)$. En invoquant à nouveau le principe du Delta et le lemme 7.21, on en conclut que

$$\sqrt{n}\left(\frac{\frac{N^n(\alpha)}{n} - \mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}, \frac{(\mathbf{p}_\theta\{\alpha\} - \mathbf{p}_{\widehat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}\right)_{\alpha\leq p}$$

converge en loi vers un vecteur Gaussien de matrice de covariance

$$\begin{pmatrix} \Gamma(\theta) & \vdots & B^t(\theta)I^{-1}(\theta)B(\theta) \\ \cdots & & \cdots \\ B(\theta)^t I^{-1}(\theta)B(\theta) & \vdots & B(\theta)^t I^{-1}(\theta)B(\theta) \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{Id}_k & \vdots & B^t(\theta) \end{pmatrix} \begin{pmatrix} \Gamma(\theta) & \vdots & B(\theta)^t I^{-1}(\theta) \\ \cdots & & \cdots \\ I^{-1}(\theta)B(\theta) & \vdots & I^{-1}(\theta) \end{pmatrix} \begin{pmatrix} \mathrm{Id}_k \\ \cdots \\ B(\theta) \end{pmatrix}.$$

On déduit que la suite de vecteurs aléatoires $(Y^n)_{n\in\mathbb{N}}$

$$Y^n := \sqrt{n}\left(\frac{\frac{N^n(\alpha)}{n} - \mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} - \frac{(\mathbf{p}_\theta\{\alpha\} - \mathbf{p}_{\widehat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}\right)_{\alpha\leq p}$$

converge en loi vers un vecteur Gaussien de matrice de covariance

$$\Gamma(\theta) - B(\theta)^t I^{-1}(\theta)B(\theta)$$
$$= \mathrm{Id} - \begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_k} \end{pmatrix} \begin{pmatrix} \sqrt{p_1} & \sqrt{p_2} & \dots & \sqrt{p_k} \end{pmatrix} - B(\theta)^t \left(B(\theta)\times B(\theta)^t\right)^{-1} B(\theta).$$

La matrice symétrique

$$D(\theta) \equiv B(\theta)^t \left(B(\theta)\times B(\theta)^t\right)^{-1} B(\theta) = B(\theta)^t I^{-1}(\theta)B(\theta)$$

est idempotente. C'est une matrice de projection orthogonale de rang r.
De plus, le vecteur $\begin{pmatrix} \sqrt{p_\theta\{1\}} & \sqrt{p_\theta\{2\}} & \dots & \sqrt{p_\theta\{k\}} \end{pmatrix}^t$ appartient à son noyau.

Sous $\mathbf{p}_\theta$, les limites en loi de

$$C(n, \mathbf{p}_{\hat{\theta}}) \equiv \sum_{\alpha=0}^{p} \left(\frac{\sqrt{n}(N^n(\alpha)/n - \mathbf{p}_{\widehat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_{\widehat{\theta}_n}\{\alpha\}}}\right)^2$$

et de

$$\sum_{\alpha=0}^{p} \left(\frac{\sqrt{n}(N^n(\alpha)/n - \mathbf{p}_{\widehat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}\right)^2$$

sont identiques. La dernière quantité s'écrit

$$\sum_{\alpha=0}^{p} \left(\frac{\sqrt{n}(N^n(\alpha)/n - \mathbf{p}_\theta\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} - \frac{\sqrt{n}(\mathbf{p}_\theta\{\alpha\} - \mathbf{p}_{\widehat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}\right)^2,$$

c'est le carré de la norme euclidienne de Y^n. Du principe de l'image continue, on déduit que la suite $(\|Y^n\|_2^2)_n$ converge en loi vers la loi du carré de la norme du vecteur Gaussien de matrice de covariance $\Gamma(\theta) - D(\theta)$. D'après le Théorème de Cochran, le carré de cette norme suit la loi d'une somme pondérées de variables distribuées selon la loi χ_1^2. Les valeurs propres de $\Gamma(\theta) - D(\theta)$ sont 1 (multiplicité $p - r$) et 0 (multiplicité $r + 1$).

□

7.7 LE TEST D'INDÉPENDANCE

On dispose d'un échantillon de couples d'éléments de $\Omega' \times \Omega''$ (deux ensembles finis de cardinalités $k'+1$ et $k''+1$) : $(\omega'_1, \omega''_1), \dots, (\omega'_n, \omega''_n)$. Cet échantillon a été obtenu en réalisant n tirages indépendants selon une loi P inconnue sur $\Omega' \times \Omega''$. L'hypothèse d'indépendance s'écrit : « la loi jointe est le produit de ses deux marginales » .

L'ensemble des lois qui sont le produit de leur marginales est un petit sous-ensemble de l'ensemble des lois sur $\Omega' \times \Omega''$. Ce sous-ensemble peut être paramétrisé par un sous-ensemble de dimension $k'+k''$ alors que pour décrire toutes les lois sur $\Omega' \times \Omega''$, on doit utiliser un ensemble de dimension $(k'+1)\times(k''+1)-1$.

On vérifie simplement que les conditions qui précèdent l'énoncé du Théorème 7.18 sont remplies. Dans le cas du test d'indépendance, l'espace Θ_0 est formé par le produit d'un ouvert de $\mathbb{R}^{k'}$ par un ouvert de $\mathbb{R}^{k''}$. Si on note $\theta := (\theta', \theta'')$ avec $\theta' \in \mathbb{R}^{k'}, \theta'' \in \mathbb{R}^{k''}$, un élément de Θ_0 on a

$$\mathbf{p}_\theta\{\langle\alpha', \alpha''\rangle\} = \begin{cases} \theta'[\alpha'] \times \theta''[\alpha''] & \text{si} \quad 0 < \alpha' \leq k' \text{ et } 0 < \alpha'' \leq k'' \\ \theta'[\alpha'] \times \left(1 - \sum_{\beta''=1}^{k''} \theta''[\beta'']\right) & \text{si} \quad 0 < \alpha' \leq k' \text{ et } \alpha'' = 0 \\ \left(1 - \sum_{\beta'=1}^{k'} \theta'[\beta']\right) \times \theta''[\alpha''] & \text{si} \quad \alpha' = 0 \text{ et } 0 < \alpha'' \leq k'' \\ \left(1 - \sum_{\beta'=1}^{k'} \theta'[\beta']\right) \times \left(1 - \sum_{\beta''=1}^{k''} \theta''[\beta'']\right) & \text{si } \alpha' = 0 \text{ et } \alpha'' = 0. \end{cases}$$

L'estimateur au maximum de vraisemblance est simple. On note $N_{\alpha',\alpha''}$ le nombre d'occurrences du couple $(\alpha', \alpha'') \in \Omega \times \Omega'$ dans l'échantillon, $N_{\alpha',\cdot} = \sum_{\alpha'' \leq k''} N_{\alpha',\alpha''}$, et $N_{\cdot,\alpha''} = \sum_{\alpha' \leq k'} N_{\alpha',\alpha''}$. Le maximum de vraisemblance dans le sous-modèle Θ_0 est atteint en $(\hat{\theta}', \hat{\theta}'')$, où $\mathbf{p}_{\hat{\theta}'}[\alpha'] = N_{\alpha',\cdot}/n$ et $\mathbf{p}_{\hat{\theta}''}[\alpha''] = N_{\cdot,\alpha''}/n$. La statistique de Pearson modifiée s'écrit alors

$$\sum_{\alpha'=0}^{k'} \sum_{\alpha''=0}^{k''} \frac{(N_{\alpha',\alpha''} - N_{\alpha',\cdot} N_{\cdot,\alpha''}/n)^2}{N_{\alpha',\cdot} N_{\cdot,\alpha''}/n} .$$

Le théorème 7.18 implique que sous l'hypothèse nulle, la statistique du χ^2 d'indépendance converge en distribution vers une loi du χ^2 à $((k'+1)(k''+1)-1) - (k'+k'') = k' \times k''$ degrés de liberté.

7.8 REMARQUES BIBLIOGRAPHIQUES

L'optimalité du test d'adéquation du χ^2 contre une collection d'alternatives locales est établie dans [1, Chapitre 14, Theorem 14.3.2].

A. VAN DER VAART [2] aborde le problème des tests du type χ^2 pour des hypothèses composites plus générales. Il établit une version beaucoup plus générale du Lemme 7.19.

Références

[1] E. L. LEHMANN et J. P. ROMANO. **Testing statistical hypotheses**. Third. Springer Texts in Statistics. Springer, New York, 2005, p. xiv+784.

[2] A. VAN DER VAART. **Asymptotic statistics**. Cambridge University Press, 1998.

8.1 UN PROBLÈME

Nous travaillerons durant tout ce cours avec une loi P sur $\mathbb{R}$. Nous ne supposerons en général rien de spécial à propos de P si ce n'est l'absolue continuité par rapport à la mesure de Lebesgue (soit l'existence d'une densité f). Nous noterons F la fonction de répartition de P, (pour tous $a < b$: $\int_{[a,b]} f(x)\mathrm{d}(x) = F(b) - F(a) = P[a,b]$).

Dans ce type de situation, on dispose d'un « résumé » de l'échantillon $S_n = (X_1, X_2, \ldots, X_n)$, l'échantillon trié en ordre croissant : $(X_{(1)}, \ldots, X_{(n)})$.

Comme dans le cas du test d'adéquation du χ^2, on veut aborder le problème de décision suivant. Soit f une densité sur $\mathbb{R}$ et $(x_1, \ldots, x_n)$ un échantillon de n points de $\mathbb{R}$.

H_0 : L'échantillon $(x_1, \ldots, x_n)$ a été engendré par n tirages indépendants selon la loi P de densité f ?
H_1 : L'échantillon $(x_1, \ldots, x_n)$ a été engendré par n tirages indépendants selon une loi différente de P ?

REMARQUE 8.1 Dans ce problème, l'hypothèse nulle est simple, l'alternative est composite et même extrêmement riche.

EXEMPLE 8.2 C'est par exemple le problème que cherchera à résoudre une association de consommateurs à laquelle un constructeur affirme que les ampoules « dur-à-cuir » ont une durée de vie distribuée exponentiellement avec une paramètre .001 (La probabilité que l'ampoule fonctionne plus de x jours est $\exp(-0.001 \times x)$). L'association va collecter des informations sur la durée de vie des ampoules « dur-à-cuir » : elle va rassembler un échantillon de durées de vie observées $(x_1, \ldots, x_n)$ et se demander si cet échantillon peut raisonnablement être considéré comme la réalisation de n variables exponentiellement distribuées (avec le paramètre .001).

Il faut construire un test qui réponde au cahier des charges suivant :

i) La statistique du test doit être facile à calculer (comme la statistique de Pearson).
ii) La loi de la statistique de test ne doit pas dépendre de la loi P qui définit le problème d'ajustement. Ceci permettra de calibrer le test pour atteindre un niveau donné.
iii) Le test doit être consistant : si l'hypothèse nulle n'est pas vérifiée et si la taille de l'échantillon tend vers l'infini, la probabilité de rejeter doit tendre vers 1.

8.2 LE PRINCIPE DU TEST DE KOLMOGOROV ET SMIRNOV

Le test de Kolmogorov-Smirnov (par la suite) résume à peine les données : il oublie simplement l'ordre dans lequel elles ont été collectées. Les données $x_1, \ldots, x_n$ définissent une loi sur $\mathbb{R}$ appelée loi empirique, notée P_n :

$$P_n\{[a,b]\} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[a,b]}(x_i) \text{pour tout intervalle } [a,b].$$

Cette loi P_n est une loi discrète, dont la fonction de répartition (appelée fonction de répartition empirique) est notée F_n

$$F_n(t) = P_n\{(-\infty, t]\} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(x_i).$$

C'est une fonction en escalier, croissante, qui saute de $1/n$ en chaque point de l'échantillon (voir Figure 8.2).

REMARQUE 8.3 Si on considère le modèle formé par les lois sur $\mathbb{R}$ dominées par la mesure de Lebesgue, les statistiques d'ordre, ou de manière équivalente, la mesure empirique, forment une statistique suffisante (ou exhaustive). Conditionnellement à la mesure empirique la loi de l'échantillon est la loi uniforme sur l'ensemble des permutations des statistiques d'ordre.

DÉFINITION 8.4 La statistique de Kolmogorov-Smirnov D_n est définie par

$$D_n \equiv \sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| .$$

Le test de Kolmogorov-Smirnov rejette l'hypothèse nulle si D_n est trop grande.

Remarque 8.5 Ici la fonction de répartitin F est supposée connue. C'est l'adéquation par rapport à F que l'on teste. La statistique de Kolmogorov-Smirnov est bien une statistique, calculable à partir des données (F_n) et des connaissances préalables (F).

Nous allons maintenant étudier quelques aspects de la statistique D_n :

i) Sous l'hypothèse nulle (c'est à dire si les $X_i \sim_{i.i.d} P$), la loi de D_n est *libre*, elle ne dépend pas de P (pourvu que P ait une densité), elle ne dépend que de n.
ii) La calcul de D_n se réduit à un tri de l'échantillon, à l'évaluation de F sur les points de l'échantillon et à la recherche d'un maximum dans un tableau de $2n+1$ valeurs.
iii) Si l'hypothèse nulle n'est pas vérifiée, par exemple si les $X_i \sim_{i.i.d} P'$ avec $P' \neq P$. alors pour tout $M > 0$
$$P' \{D_n > M\} \to_n 1.$$
Le test est donc consistant.
iv) Pour terminer nous donnerons quelques résultats qui suggèreront (mais ne montreront pas) que sous l'hypothèse nulle,
$$D_n \rightsquigarrow D$$
où D est une loi « bien connue » (la loi du supremum du pont brownien réfléchi). Cette convergence en loi, qui peut être vue comme une conséquence d'un théorème central limite généralisé, permet une calibration « asymptotique » du test : si $d_{1-\alpha}$ désigne le $1-\alpha$ quantile de la loi de D alors le test qui rejette l'hypothèse nulle si $D_n > d_{1-\alpha}$ est asymptotiquement de niveau α (voir Figure 8.2).

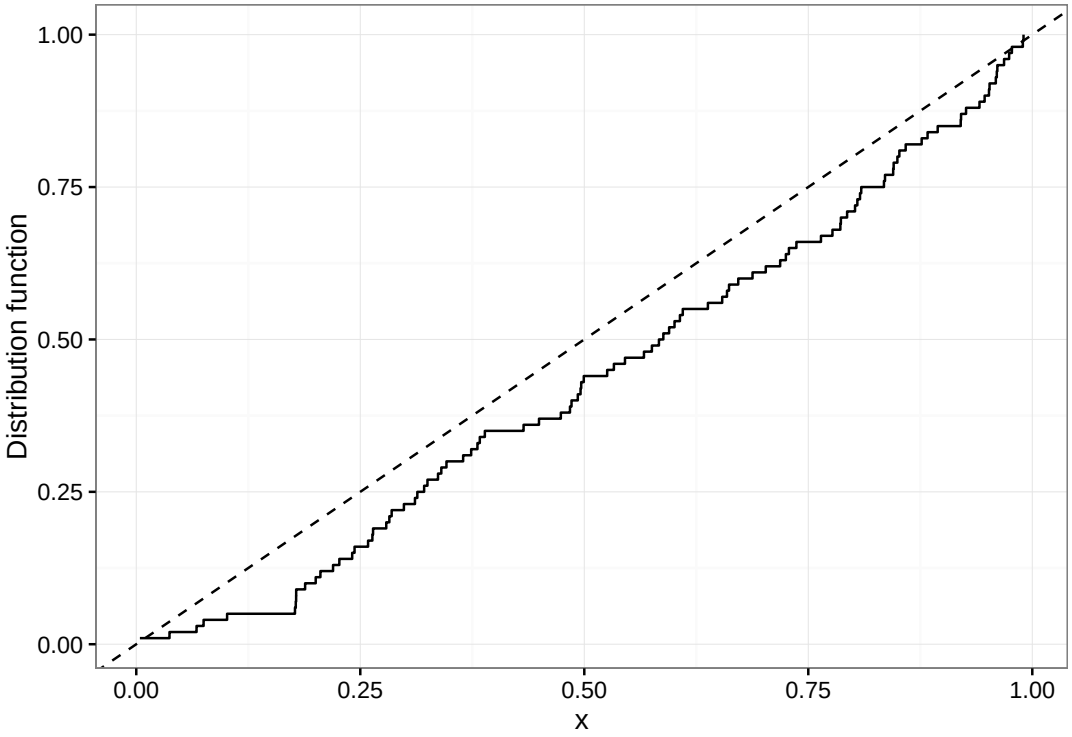

Fig. 8.1 : Fonction de répartition empirique F_n d'un échantillon de 100 points collectés indépendamment selon des tirages uniformes sur $[0,1]$. La courbe en tirets est la fonction de répartition de la loi uniforme.

8.3 la transformation quantile

Dans cette section nous allons formuler quelques observations très importantes sur le comportement de la variable aléatoire $F(X)$ lorsque $X \sim P$ de fonction de répartition F. Ces observations seront importantes pour nous, mais elles sont aussi très fécondes dans le domaine de la simulation de phénomènes

aléatoires, elles permettent d'engendrer des variables aléatoires de fonction de répartition donnée F à partir de variables aléatoires distribuées uniforméments sur $[0,1]$.

Une fonction de répartition F est une fonction positive, qui prend ses valeurs entre 0 et 1, est croissante (au sens large, c'est-à-dire non-décroissante), continue à droite, et possède une limite à gauche en tout point. Si la fonction de répartition correspond à une loi de probabilité qui ne charge aucun point (diffuse), elle est continue.

Dans tous les cas, on définit la *fonction quantile* F^{-1} comme une pseudo-inverse de F.

DÉFINITION 8.6 La fonction quantile F^{-1} d'une variable aléatoire X distribuée selon P (de fonction de répartition F) est définie par

$$F^{-1}(p) := \inf\{x : P\{X \leq x\} \geq p\} = \inf\{x : F(x) \geq p\}\,.$$

La fonction quantile est croissante et continue à droite elle aussi. La proposition suivante résume quelques propriétés simples de la fonction quantile.

PROPOSITION 8.7 *Si X a pour fonction de répartition F et pour fonction quantile F^{-1}, alors on a les propriétés suivantes, pour $p \in]0,1[$:*

i) $p \leq F(x)$ *si et seulement si* $F^{-1}(p) \leq x$.
ii) $F \circ F^{-1}(p) \geq p$.
iii) $F^{-1} \circ F(x) \leq x$.
iv) Si F admet une densité $F \circ F^{-1}(p) = p$

PREUVE. D'après la définition de F^{-1} si $F(x) \geq p$ alors $F^{-1}(p) \leq x$.

Maintenant pour établir la réciproque, il suffit de vérifier que $F \circ F^{-1}(p) \geq p$.

En effet, si $x \geq F^{-1}(p)$, comme F est croissante $F(x) \geq F \circ F^{-1}(p)$. Si $y = F^{-1}(p)$, par définition de $y = F^{-1}(p)$, il existe une suite décroissante $(z_n)_{n\in\mathbb{N}}$ qui converge vers y telle que $F(z_n) \geq p$. Mais comme F est continue à droite $\lim_n F(z_n) = F(\lim_n z_n) = F(y)$. Donc $F(y) \geq p$.

Remarquons que nous venons de prouver à la fois i) et ii).

iii) est une conséquence immédiate de i). Notons $p \equiv F(x)$. Donc $p \leq F(x)$, ce qui est équivalent d'après 1) à $F^{-1}(p) \leq x$, soit $F^{-1} \circ F(x) \leq x$.

iv) Si F admet une densité f, F est continue (même *absolument continue*). Pour toute valeur p dans $]0,1[$, il existe au moins une valeur x tel que $p = F(x)$ (Théorème des valeurs intermédiaires). Soit y l'infimum des valeurs x telles que $F(x) = p$, évidemment $y = F^{-1}(p)$. D'après 1) $F(y) \geq p$. Maintenant si $(z_n)_{n\in\mathbb{N}}$ est une suite strictement croissante convergente vers y, on a pour tout n, $F(z_n) < p$, et par continuité à gauche, $F(y) = F(\lim_n z_n) = \lim_n F(z_n) \leq p$. Donc $F(y) = p$, soit $F \circ F^{-1}(p) = p$. □

A partir des propriétés de la fonction quantile, il est aisé de déduire le résultat suivant.

COROLLAIRE 8.8 *Si X est une variable aléatoire de loi P dont la fonction de répartition F est continue, alors les variables aléatoires $F(X)$ et $1 - F(X)$ sont uniformément réparties sur $[0,1]$.*

Dans le contexte pour lequel nous voulons construire des tests, on peut donc affirmer que sous l'hypothèse nulle, l'échantillon $F(X_1), \ldots, F(X_n)$ est distribué de la même façon que $U_1, \ldots, U_n$ où les $U_i \sim$ Uniforme$[0,1]$.

On peut aussi en conclure que pour simuler un tirage selon la loi de fonction de répartition F, il suffit de disposer d'un générateur de nombres aléatoires uniformément répartis sur $[0,1]$ et de leur appliquer la transformation F^{-1}.

Nous allons maintenant vérifier que D_n est facile à calculer. Cette vérification, combinée aux observations de la section précédente nous permettra de conclure que sous l'hypothèse nulle, la loi de D_n est bien libre.

Nous adopterons la convention que $x_{(1)}, \ldots, x_{(n)}$ désigne la version triée en ordre croissant de $(x_1, \ldots, x_n)$, les variables aléatoires $X_{(i)}$ sont appelées les statistiques d'ordre de l'échantillon.

LEMME 8.9 *Le supremum dans la définition de D_n peut être calculé efficacement à partir de l'échantillon trié en ordre croissant $(x_{(1)}, \ldots, x_{(n)})$:*

$$D_n = \sqrt{n} \max_{0 \leq i < n} \max \left(\left| \frac{i}{n} - F(x_{(i)}) \right|, \left| \frac{i}{n} - F(x_{(i+1)}) \right| \right),$$

avec la convention $x_{(0)} = -\infty$.

PREUVE. Considérons un réel t compris entre $x_{(i)}$ et $x_{(i+1)}$. On a

$$F_n(t) = \frac{i}{n} = F_n\left(x_{(i)}\right) < F_n(x_{(i+1)}) = \frac{i+1}{n},$$

et

$$F(x_{(i)}) \leq F(t) \leq F(x_{(i+1)}).$$

On a donc

$$F_n(x_{(i)}) - F\left(x_{(i+1)}\right) \leq F_n(t) - F(t) \leq F_n\left(x_{(i)}\right) - F\left(x_{(i)}\right)$$

et

$$F(x_{(i)}) - F_n\left(x_{(i)}\right) \leq F(t) - F_n(t) \leq F\left(x_{(i+1)}\right) - F_n\left(x_{(i)}\right)$$

ce qui nous conduit à

$$|F(t) - F_n(t)| \leq \max \left(\left| \frac{i}{n} - F(x_{(i)}) \right|, \left| \frac{i}{n} - F(x_{(i+1)}) \right| \right)$$

pour tout $t \in [x_{(i)}, x_{(i+1)})$.

On peut par ailleurs vérifier avec le même genre d'arguments que pour tout $t < x_{(1)}$

$$|F(t) - F_n(t)| \leq F(x_{(1)}).$$

□

REMARQUE 8.10 Cette caractérisation de D_n nous assure au passage que D_n est bien une variable aléatoire. Cela n'allait pas de soi avec la définition de départ qui considérait un supremum sur un ensemble non-dénombrable.

Une conséquence importante du Lemme précédent est résumée dans le théorème suivant :

THÉORÈME 8.11 *Si la fonction de répartition F est absolument continue alors la statistique de Kolmogorov-Smirnov D_n est distribuée comme*

$$\sqrt{n} \max_{i \leq n} \max \left(\left| \frac{i}{n} - U_{(i)} \right|, \left| \frac{i}{n} - U_{(i+1)} \right| \right)$$

ou $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$ est le réarrangement croissant d'un échantillon i.i.d selon la loi uniforme sur $[0, 1]$.

Si $t_{n,\alpha}$ désigne le $1 - \alpha$ quantile de la loi de D_n, le test $T_{n,\alpha}$ de Kolmogorov-Smirnov est défini par

$$T_{n,\alpha}(S_n) = \begin{cases} 1 & \text{si } D_n \geq t_{n,\alpha} \\ 0 & \text{sinon .} \end{cases}$$

Devant un test de cette forme, on peut aborder le problème de calibration (détermination du seuil $t_{n,\alpha}$ en fonction du niveau de confiance recherché) de deux façons :

i) développer un algorithme de calcul de $t_{n,\alpha}$. Ceci est aujourd'hui envisageable en utilisant les possibilités de simulation des ordinateurs.

ii) montrer que la suite D_n converge en loi vers une variable aléatoire non-dégénérée et tabuler une fois pour toute la loi limite. Ceci permet de construire de suites de tests de niveau asymptotique donné. Cette démarche (la seule raisonnable dans les années 1930) a soulevé des questions très intéressantes en théorie des probabilités. Elle a marquée le début de la théorie des « processus empiriques » , qui sous-tend la théorie de la reconnaissance des formes ainsi que la théorie statistique de l'apprentissage.

Nous n'avons pas les moyens techniques et conceptuels pour traiter le point 2 ci-dessus. Mais nous allons tout de même vérifier que sous l'hypothèse nulle, D_n ne grandit pas trop vite, en fait certainement pas plus vite que $\sqrt{n}$, car $D_n/\sqrt{n}$ tend en probabilité vers 0.

8.4 UNE LOI DES GRANDS NOMBRES « FONCTIONNELLE » : LE THÉORÈME DE GLIVENKO-CANTELLI

Convenons de la définition suivante

$$Z_n \equiv \frac{D_n}{\sqrt{n}} = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| .$$

THÉORÈME 8.12 [THÉORÈME CLASSIQUE DE GLIVENKO-CANTELLI] *Pour toute probabilité P sur $\mathbb{R}$, de fonction de répartition F continue, la suite $(Z_n)_{n\in\mathbb{N}}$ converge en probabilité vers 0 :*

$$\forall \varepsilon > 0, \quad \lim_n P\left\{\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| > \varepsilon\right\} = 0.$$

REMARQUE 8.13 Pour chaque valeur particulière de x, on peut écrire

$$F_n(t) - F(t) = \frac{1}{n}\sum_{i\leq n} \left(\mathbf{1}_{x_i\leq t} - \mathbb{E}\left[\mathbf{1}_{X\leq t}\right]\right).$$

La différence $F_n(t) - F(t)$ est donc une somme de variables de Bernoulli indépendantes, normalisée par $1/n$. La loi des grands nombres habituelle nous indique alors

$$\forall \varepsilon > 0, \forall t \in \mathbb{R}, \quad P\left\{|F_n(t) - F(t)| > \varepsilon\right\} \to_n 0.$$

Le théorème de Glivenko-Cantelli constitue une loi des grands nombres *uniforme.*

PREUVE. Soit ϵ une réel strictement positif. Pour $k \in \mathbb{N}$, $1 \leq k \leq \lfloor \frac{1}{\epsilon} \rfloor$, on définit

$$\mathsf{y}_k^\epsilon \equiv F^{-1}(k\epsilon).$$

Etant donné un échantillon $x_1, \ldots, x_n$, et la fonction de répartition empirique F_n associée, si $\mathsf{y}_k^\epsilon \leq x \leq \mathsf{y}_{k+1}^\epsilon$, alors

$$\begin{array}{rcl} F(\mathsf{y}_k^\epsilon) & \leq & F(x) \leq F(\mathsf{y}_{k+1}^\epsilon) \\ F_n(\mathsf{y}_k^\epsilon) & \leq & F_n(x) \leq F_n(\mathsf{y}_{k+1}^\epsilon). \end{array}$$

Donc

$$F(\mathsf{y}_k^\epsilon) - F_n(\mathsf{y}_{k+1}^\epsilon) \leq F(x) - F_n(x) \leq F(\mathsf{y}_{k+1}^\epsilon) - F_n(\mathsf{y}_k^\epsilon)$$

ce qui entraine

$$|F(x) - F_n(x)| \leq \max\left(\left|F(\mathsf{y}_k^\epsilon) - F_n(\mathsf{y}_k^\epsilon)\right|, \left|F(\mathsf{y}_{k+1}^\epsilon) - F_n(\mathsf{y}_{k+1}^\epsilon)\right|\right) + \epsilon$$

soit

$$\sup_{\mathsf{y}_1^\epsilon \leq x \leq \mathsf{y}_{\lfloor 1/\epsilon\rfloor}^\epsilon} |F(x) - F_n(x)| \leq \max_{1\leq k\leq \lfloor 1/\epsilon\rfloor} \max\left(|F(\mathsf{y}_k^\epsilon) - F_n(\mathsf{y}_k^\epsilon)|\right) + \epsilon.$$

Par ailleurs, pour $x \leq \mathsf{y}_1^\epsilon$,

$$|F(x) - F_n(x)| \leq |F_n(\mathsf{y}_1^\epsilon) - F(\mathsf{y}_1^\epsilon)| + \epsilon$$

et de même pour $x \geq \mathsf{y}_{\lfloor 1/\epsilon \rfloor}^\epsilon$

$$|F(x) - F_n(x)| \leq \left|F_n(\mathsf{y}_{\lfloor 1/\epsilon \rfloor}^\epsilon) - F(\mathsf{y}_{\lfloor 1/\epsilon \rfloor}^\epsilon)\right| + \epsilon.$$

En récapitulant

$$\frac{D_n}{\sqrt{n}} \leq \max_{1 \leq k \leq \lfloor 1/\epsilon \rfloor} \max\left(|F(\mathsf{y}_k^\epsilon) - F_n(\mathsf{y}_k^\epsilon)|\right) + \epsilon.$$

On peut maintenant observer que $F(\mathsf{y}_k^\epsilon) - F_n(\mathsf{y}_k^\epsilon)$ est un somme normalisée de variables aléatoires centrées d'espérance nulle et de variance $k\epsilon(1-k\epsilon)$. Donc d'après l'inégalité de Bienaymée-Chebyshev

$$P^n\left\{|F(\mathsf{y}_k^\epsilon) - F_n(\mathsf{y}_k^\epsilon)| \geq u\right\} \leq \frac{k\epsilon(1-k\epsilon)}{nu^2} \leq \frac{1}{4nu^2}.$$

Et

$$P^n\left\{\max_{k \leq \lfloor 1/\epsilon \rfloor} |F(\mathsf{y}_k^\epsilon) - F_n(\mathsf{y}_k^\epsilon)| \geq u\right\} \leq \frac{\lfloor 1/\epsilon \rfloor}{4nu^2}.$$

En choisissant $u = \epsilon$, on peut en déduire que, lorque n tend vers l'infini :

$$\lim_n P^n\left\{\left|\frac{D_n}{\sqrt{n}}\right| \geq 2\epsilon\right\} = 0.$$

Comme on peut choisir ϵ arbitrairement petit, la convergence en probabilité est démontrée.

□

Remarque 8.14 Cette démonstration du théorème de Glivenko-Cantelli ne s'appuie que sur deux choses : la loi des grands nombres pour des sommes de variables de Bernoulli, et sur la possibilité d'encadrer chaque indicatrice de demi-droite, par deux indicatrices assez proches. Cette dernière idée est souvent utilisée sous une forme généralisée en théorie des processus empiriques : c'est la possibilité de couvrir la classe de fonctions intéressante $\mathcal{F}$ à l'aide d'un nombre fini d'intervalles ou braquets (*brackets*).

Consistence du test de Kolmogorov-Smirnov

Le théorème de Glivenko-Cantelli nous donne une intuition sur le comportement de la statistique de Kolmogorov-Smirnov. Si on veut se convaincre

Théorème 8.15 *Si $(X_1, \ldots, X_n)$ est collecté par des tirages i.i.d. selon une loi P' de fonction de répartition $F' \neq F$, alors $D_n = \sqrt{n}\sup_x |F_n(x) - F(x)|$ converge en P'-probabilité vers l'infini.*

preuve. Il existe $\epsilon > 0$, tel qu'il existe x vérifiant $F'(x) - F(x) \geq \epsilon > 0$ ou $F(x) - F'(x) \geq \epsilon > 0$. Sans perdre en généralité supposons que nous sommes dans la première situation. Comme

$$\sqrt{n}\sup_t |F_n(t) - F(t)| \geq \sqrt{n}\left(F_n(x) - F(x)\right) = \sqrt{n}\left(F_n(x) - F'(x)\right) + \sqrt{n}\left(F'(x) - F(x)\right),$$

on a donc

$$D_n \geq \sqrt{n}\left(F_n(x) - F'(x)\right) + \sqrt{n}\epsilon.$$

La loi des grands nombres, indique que sous P', $|F'(x) - F_n(x)| \to^{P'} 0$. Donc

$$P'\left\{F_n(x) - F'(x) < -\epsilon/2\right\} \to_n 0,$$

soit pour tout $M > 0$

$$P'\left\{\sqrt{n}\sup_t |F_n(t) - F(t)| > M\right\} \to 1.$$

□

Ceci implique que quelque soit le choix d'un seuil $\tau > 0$, pour le test qui rejette l'hypothèse nulle lorsque $D_n \geq \tau$, la probabilité de rejeter lorsque l'hypothèse nulle n'est pas vérifiée (sous une alternative fixe) tend vers 1 lorsque la taille de l'échantillon tend vers l'infini.

Nous ne sommes pas en mesure d'établir que sous l'hypothèse nulle la suite des lois de $(D_n)_n$ converge étroitement (les simulations présentées sur la figure 8.2 suggèrent que D_n est bien normalisée). C'est pourtant le cas.

THÉORÈME 8.16 *Sous l'hypothèse nulle, la statistique de Kolmogorov-Smirnov converge en loi. Pour $x > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{D_n \geq x\right\} = 2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2x^2}\,.$$

Nous allons établir des inégalités de déviation pour D_n sous l'hypothèse nulle. Ces inégalités sont non-asymptotiques (valables pour tout n) et elles montrent que la suite des lois de $(D_n)_n$ est uniformément tendue. Ceci implique que la suite des lois de $(D_n)_n$ est relativement compacte pour la topologie de la convergence étroite.

L'inégalité de DVORETZKY, KIEFER et WOLFOWITZ [1] et MASSART [4] s'énonce ainsi :

THÉORÈME 8.17 *Pour tout entier $n > 1$, tout $\epsilon > 0$,*

$$\mathbb{P}\left\{D_n \geq \epsilon\right\} \leq 2\mathrm{e}^{-2\epsilon^2}\,.$$

Si on choisit $x = F^{\leftarrow}(1/2)$, $nF_n(x)$ est distribué comme une binomiale de paramètres n et $1/2$. L'inégalité de Hoeffding nous indique alors

$$\mathbb{P}\left\{\sqrt{n}\,|F_n(x) - F(x)| \geq \epsilon\right\} \leq 2\mathrm{e}^{-2\epsilon^2}\,.$$

Le théorème 8.17 établit donc une inégalité de déviation de type Hoeffding assez remarquable pour le supremum. Il s'agit d'un résultat difficile.

Ce résultat difficile ne peut être amélioré. Une version affaiblie peut être enseignée :

PROPOSITION 8.18 *Pour tout entier $n > 0$, tout $\epsilon > 0$*

$$\mathbb{P}\left\{D_n \geq \epsilon\right\} \leq 4\mathrm{e}^{-\frac{\epsilon^2}{8}}\,.$$

La preuve de la Proposition 8.18 est l'occasion d'introduire la technique dite de *symétrisation*, d'explorer les premières étapes de la preuve des *inégalités de Vapnik-Chervonenkis*, et d'appliquer en statistique des outils élégants issus de la théorie des marches aléatoires (*principe de réflexion*).

La technique de symétrisation est une technique générale, simple et puissante de la théorie des processus empiriques. Elle permet de majorer élégamment des espérances de suprema de processus empiriques.

Pour formuler le Lemme de symétrisation, nous utiliserons des variables de Rademacher, autrement des signes aléatoires : une variable de Rademacher vaut 1 ou -1 avec probabilité $1/2$.

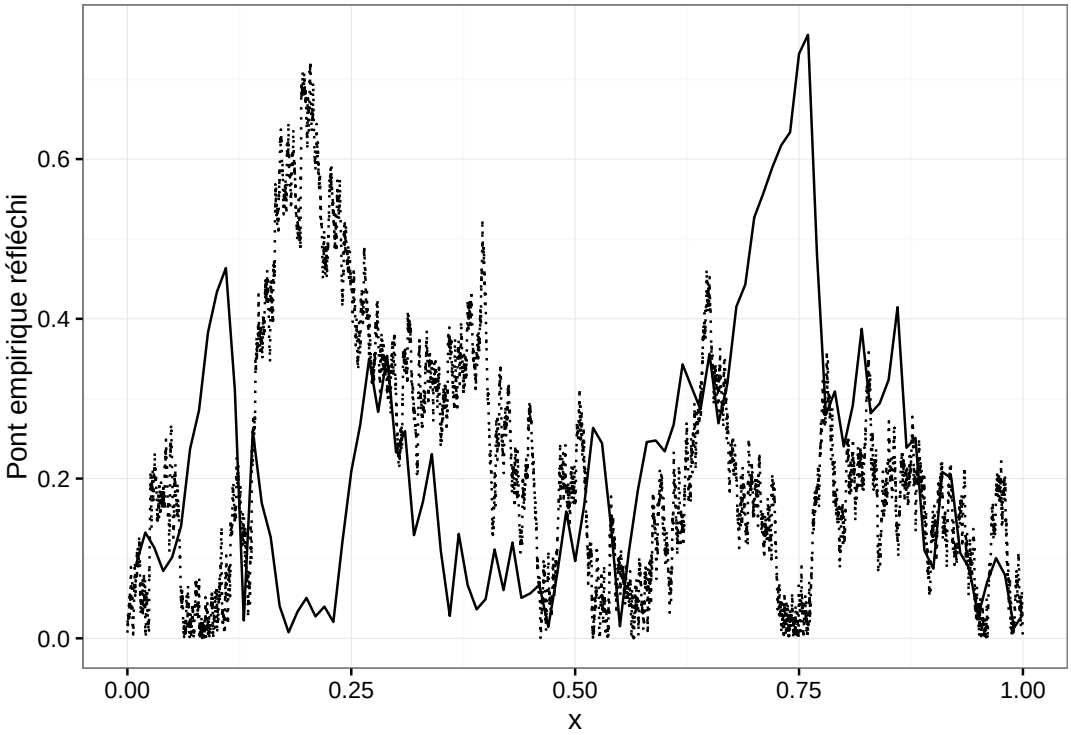

FIG. 8.2 : Graphe de la fonction $s \mapsto \sqrt{n}|F_n(s) - F(s)|$ pour un échantillon de taille 100 de la loi uniforme. Le processus illustré est souvent appelé le pont empirique. La limite en loi du pont empirique est la loi du pont brownien (mouvement brownien conditionné à valoir 0 au temps 1.) La ligne pointillée est calculée à partir d'un échantillon de taille 10000.

LEMME 8.19 *Soit $X_1, \ldots, X_n$ une suite de vecteurs aléatoires indépendants indexés par $\mathcal{T}$. Soient $\epsilon_1, \ldots, \epsilon_n$ une suite de variables de Rademacher indépendantes entre elles et indépendantes des $X_1, \ldots X_n$. Si $\Psi : \mathbb{R} \to \mathbb{R}$ est convexe et croissante, alors*

$$\mathbb{E}\left[\Psi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} X_{i,t} - \mathbb{E}X_{i,t}\right|\right)\right] \leq \mathbb{E}\left[\Psi\left(2\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\epsilon_i X_{i,t}\right|\right)\right] .$$

REMARQUE 8.20 Si on choisit $\Psi(x) = x$, on obtient l'inégalité

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} X_{i,t} - \mathbb{E}X_{i,t}\right|\right] \leq 2\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\epsilon_i X_{i,t}\right|\right] .$$

Cette inégalité de symétrisation permet de se concentrer sur une quantité souvent plus simple à étudier

$$\mathbb{E}\left[\Psi\left(2\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\epsilon_i X_{i,t}\right|\right) \mid X_1, \ldots, X_n\right]$$

où l'espérance conditionnelle n'est que l'espérance calculée par rapport aux variables de Rademacher (c'est une simple somme).

PREUVE.(Lemme de symétrisation) Dans la suite les Y_i sont i.i.d. uniformément sur $[0, 1]$. On introduit par souci de commodité

$$Z = \sqrt{n}D_n = \sup_{s\in[0,1]}\left|\sum_{i=1}^{n} X_{i,s} - \mathbb{E}X_{i,s}\right|$$

où $X_{i,s} = 1$ si $Y_i \leq s$ et $X_{i,s} = 0$ sinon. Dans la suite pour $i \leq n$, Y_i' est une copie indépendante de Y_i, et $X_{i,s}' = \mathbb{I}_{Y_i' \leq s}$. On peut réécrire Z en

$$Z = \sup_{s\in[0,1]} \left| \sum_{i=1}^{n} X_{i,s} - \mathbb{E}X_{i,s}' \right| .$$

Les ϵ_i sont des variables de Rademacher ($\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$) indépendantes des Y_i, Y_i'. On utilisera le fait que $\epsilon_i(X_i - X_i') \sim (X_i - X_i')$ car la variable aléatoire fonctionnelle $X_i - X_i'$ est *symétrique* : elle a même loi que son opposé. Dans la suite les espérances sont prises par rapport à $X_1, \ldots, X_n, X_1', \ldots, X_n', \epsilon_1, \ldots, \epsilon_n$.

$$\begin{aligned}
\mathbb{E}\left[\Psi\left(Z\right)\right] &= \mathbb{E}\left[\sup_{s\in[0,1]} \Psi\left(\left|\sum_{i=1}^{n} X_{i,s} - \mathbb{E}X_{i,s}'\right|\right)\right] \\
&\leq \mathbb{E}\left[\sup_{s\in[0,1]} \Psi\left(\left|\sum_{i=1}^{n} X_{i,s} - X_{i,s}'\right|\right)\right] && \text{(Inégalité de Jensen)} \\
&= \mathbb{E}\left[\sup_{s\in[0,1]} \Psi\left(\left|\sum_{i=1}^{n} \epsilon_i(X_{i,s} - X_{i,s}')\right|\right)\right] && (X_i - X_i' \text{ est symétrique}) \\
&\leq \mathbb{E}\left[\sup_{s\in[0,1]} \frac{1}{2}\left(\Psi\left(2\left|\sum_{i=1}^{n} \epsilon_i X_{i,s}\right|\right) + \Psi\left(2\left|-\sum_{i=1}^{n} \epsilon_i X_{i,s}'\right|\right)\right)\right] && \text{(Inégalité de Jensen)} \\
&\leq \mathbb{E}\left[\sup_{s\in[0,1]} \Psi\left(2\left|\sum_{i=1}^{n} \epsilon_i X_{i,s}\right|\right)\right] .
\end{aligned}$$

□

En choisissant $\Psi(x) = \mathrm{e}^{\lambda x}$ pour $\lambda > 0$, nous venons d'établir une *inégalité de symétrisation* qui sera l'élément clé de la preuve de la proposition théorème 8.18 (version facile du théorème 8.17 qui n'est pas fondée sur un argument de symétrisation) :

$$\mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{\lambda\left|\sum_{i=1}^{n} X_{i,s} - \mathbb{E}X_{i,s}'\right|}\right] \leq \mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{2\lambda\left|\sum_{i=1}^{n} \epsilon_i X_{i,s}\right|}\right] .$$

Remarque 8.21 Cette inégalité est valable pour des *suprema de processus empiriques* beaucoup plus généraux. Si les $X_{i,s}$ étaient des indicatrices d'ensembles plus compliqués que des demi-droites, elle resterait valable. De fait cette inégalité peut constituer la première étape d'une preuve des inégalités de Vapnik-Chervonenkis (voir Chapitre 9).

Nous allons maintenant exploiter le fait que nous avons affaire à des indicatrices de demi-droites.

preuve.(Proposition 8.18) Presque sûrement, les Y_i sont deux à deux distincts. Pour une réalisation des ϵ_i et des Y_i, on a

$$\begin{aligned}
\sup_{s\in[0,1]} \mathrm{e}^{2\lambda\left|\sum_{i=1}^{n} \epsilon_i X_{i,s}\right|} &= \sup_{s\in[0,1]} \mathrm{e}^{2\lambda\left|\sum_{i=1}^{n} \epsilon_i \mathbb{I}_{Y_i \leq s}\right|} \\
&= \max_{k\leq n} \mathrm{e}^{2\lambda\left|\sum_{i=1}^{n} \epsilon_i \mathbb{I}_{Y_i \leq Y_{(k)}}\right|} \\
&= \max_{k\leq n} \mathrm{e}^{2\lambda\left|\sum_{i=1}^{k} \epsilon_i\right|} \qquad \text{(en loi).}
\end{aligned}$$

Donc

$$\mathbb{E}\left[\sup_{s\in[0,1]} \mathrm{e}^{2\lambda\left|\sum_{i=1}^{n} \epsilon_i X_{i,s}\right|} \mid Y_1, \ldots, Y_n\right] = \mathbb{E}\left[\max_{k\leq n} \mathrm{e}^{2\lambda\left|\sum_{i\leq k} \epsilon_i\right|}\right] . \tag{8.1}$$

Le membre droit est un moment exponentiel de la marche aléatoire symétrique réfléchie. Comme pour toute variable aléatoire positive, l'espérance peut s'écrire comme l'intégrale des probabilités de déviation.

$$\mathbb{E}\left[\max_{k\leq n} \mathrm{e}^{2\lambda\left|\sum_{i\leq k} \epsilon_i\right|}\right] = \int_0^\infty \mathbb{P}\left\{\max_{k\leq n} \mathrm{e}^{2\lambda\left|\sum_{i\leq k} \epsilon_i\right|} \geq a\right\} \mathrm{d}a\,.$$

On note

$$A_k = \left\{ \mathrm{e}^{2\lambda|\sum_{i\leq k}\epsilon_i|} \geq a, \text{ et } \mathrm{e}^{2\lambda|\sum_{i\leq j}\epsilon_i|} < a \text{ pour } j<k \right\}.$$

Les A_k forment une famille d'événements deux à deux disjoints. D'une part,

$$\left\{ \max_{k\leq n} \mathrm{e}^{2\lambda|\sum_{i\leq k}\epsilon_i|} \geq a \right\} = \bigcup_{k\leq n} A_k$$

et d'autre part, pour chaque $k \in \{1,\ldots,n\}$,

$$\mathbb{P}\left\{ \mathrm{e}^{2\lambda|\sum_{i\leq n}\epsilon_i|} \geq a \mid A_k \right\} \quad \geq \quad \frac{1}{2},$$

en effet, avec probabilité supérieure ou égale à $1/2$, $\sum_{k<i\leq n}\epsilon_i$ est du même signe que $\sum_{i\leq k}\epsilon_i$ (nous utilisons ici le *principe de réflexion de Désiré André*). On peut déduire une inégalité maximale :

$$\begin{aligned}
\mathbb{P}\left\{ \max_{k\leq n} \mathrm{e}^{2\lambda|\sum_{i\leq k}\epsilon_i|} \geq a \right\} &\leq 2\sum_{k\leq n} \mathbb{P}\{A_k\}\mathbb{P} \times \left\{ \mathrm{e}^{2\lambda|\sum_{i\leq n}\epsilon_i|} \geq a \mid A_k \right\} \\
&= 2\mathbb{P}\left\{ \mathrm{e}^{2\lambda|\sum_{i\leq n}\epsilon_i|} \geq a \right\}
\end{aligned}$$

D'où

$$\begin{aligned}
\mathbb{E}\left[\max_{k\leq n} \mathrm{e}^{2\lambda|\sum_{i\leq k}\epsilon_i|} \right] &\leq 2\mathbb{E}\left[\mathrm{e}^{2\lambda|\sum_{i\leq n}\epsilon_i|} \right] \\
&\leq 2\mathbb{E}\left[\mathrm{e}^{2\lambda\sum_{i\leq n}\epsilon_i} + \mathrm{e}^{-2\lambda\sum_{i\leq n}\epsilon_i} \right] \\
&\leq 4\mathbb{E}\left[\mathrm{e}^{2\lambda\sum_{i\leq n}\epsilon_i} \right] && \text{(symétrie)} \\
&\leq 4\,\mathrm{e}^{n\frac{(2\lambda)^2}{2}} && \text{(Lemme de Hoeffding).}
\end{aligned}$$

On aboutit à

$$\mathbb{E}\left[\mathrm{e}^{\lambda Z} \right] \leq 4\mathrm{e}^{2n\lambda^2}$$

et la proposition 8.18 s'obtient en choisissant $\lambda = \epsilon/(4\sqrt{n})$ et invoquant le Lemme de Markov.

□

8.6 ADÉQUATION À UNE FAMILLE DE LOIS

De même qu'on peut utiliser les statistiques du χ^2 pour tester l'ajustement à une famille de lois, on peut utiliser les statistiques de Kolmogorov-Smirnov pour tester l'adéquation à une famille de lois. Prenons le cas de la famille des lois exponentielles dont les densités sont de la forme $1/\sigma \exp(-x/\sigma)$, pour $x>0, \sigma>0$.

On peut chercher à modifier la statistique de Kolmogorov-Smirnov, en substituant à la fonction de répartition inconnue, une fonction de répartition estimée, par exemple via la méthode du maximum de vraisemblance.

$$\sqrt{n}\sup_{x\in\mathbb{R}} \left| F_n(x) - \left(1 - \mathrm{e}^{-x/\widehat{\sigma}}\right) \right|$$

où $\widehat{\sigma} = \overline{X}_n$. On peut vérifier que sous l'hypothèse nulle, la loi de cette statistique ne dépend pas du paramètre σ (la loi de $(X_i/\overline{X}_n)$ ne dépend pas de σ, il n'y a là rien de spécifique à la famille des lois exponentielles. Cette construction fonctionne pour les familles définies par un paramètre d'échelle). La statistique de test devient

$$\sqrt{n}\sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\frac{X_i}{\widehat{\sigma}}\leq x} - \left(1-\mathrm{e}^{-x}\right) \right|.$$

Elle n'est pas distribuée de la même façon que la statistique de Kolmogorov-Smirnov calculée sous l'hypothèse nulle sur un n-échantillon. La distribution sous l'hypothèse nulle dépend de la famille de lois considérée.

Le lemme suivant nous fournit une autre voie.

LEMME 8.22 *Soient $Y_1, \ldots, Y_n$ des variables aléatoires positives i.i.d. On note $S_n := \sum_{i=1}^n Y_i$. La loi commune des Y_i est exponentielle si et seulement si conditionnellement à S_n, $Y_1, \ldots, Y_{n-1}$ est distribuée comme les écarts d'un $n-1$-échantillon de la loi uniforme sur $[0, S_n]$.*

PREUVE. Une suite de calculs astucieux mais élémentaires montre que la densité de la loi commune des Y_i est une solution mesurable de l'équation fonctionnelle de Cauchy

$$f(x+y) = f(x)f(y) \text{ pour } x, y \geq 0\,.$$

Les seules solutions mesurables de cette équation sont de la forme $f(x) = b\exp(ax)$ pour $b > 0, a \in \mathbb{R}$.

La densité de la loi commune des Y_i est une solution mesurable de l'équation fonctionnelle de Cauchy. □

On en tire le corollaire.

COROLLAIRE 8.23 *Soient $Y_1, \ldots, Y_n$ des variables aléatoires positives i.i.d., On note $S_n := \sum_{i=1}^n Y_i$. La loi commune des Y_i est exponentielle si et seulement si la suite $(Z_j)_{j<n}$ définie par $Z_j = \sum_{k=1}^j Y_j/S_n$ est distribuée comme les statistiques d'ordre d'un $n-1$-échantillon de la loi uniforme sur $[0,1]$.*

Tester l'adéquation à la famille des lois exponentielles revient donc à calculer la statistique de Kolmogorov-Smirnov d'ajustement à la loi uniforme sur la suite $Z_1, \ldots, Z_{n-1}$ et à comparer cette statistique aux quantiles de la loi de la statistique de Kolmogorov-Smirnov pour les échantillons de taille $n-1$.

8.7 REMARQUES BIBLIOGRAPHIQUES

La symétrisation, consiste (entre autres choses) à réduire l'étude d'une suite de variables aléatoires vectorielles centrées $(X_i)_i$ à celle d'une suite de variables alétoires $(X'_i)_i$ *symétriques*. Une variable aléatoire X est dite symétrique si X et $-X$ ont même loi. Ledoux et Talagrand [8] utilisent ce type de techniques pour caractériser la convergence presque sûre de sommes de vecteurs aléatoires indépendants (en particulier pour les vecteurs aléatoires à valeur dans des espaces de Banach). Ils renvoient aux travaux de Paul Lévy. L'utilisation des arguments de symétrisation est centrale dans les travaux de Vapnik et Chervonenkis [5, 6, 9], sous la forme présentée ici, elle a été popularisée par Giné et Zinn [2].

En statistique, la symétrisation est utilisée sans être toujours mentionnée. En apprentissage, les moyennes de Rademacher, conditionnelles ou non, sont une variante du *bootstrap* à poids qui repose sur la symétrisation [3]. Les tests de permutation peuvent aussi être considérés comme des techniques de symétrisation.

La résolution de l'équation fonctionnelle de Cauchy est étudiée dans [7].

Références

[1] A. DVORETZKY, J. KIEFER et J. WOLFOWITZ. « Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator ». In : *Annals of Mathematical Statistics* 33 (1956), p. 642-669.

[2] E. GINÉ et J. ZINN. « Some limit theorems for empirical processes ». In : *The Annals of Probability* 12 (1984), p. 929-989.

[3] V. KOLTCHINSKII. « Local Rademacher complexities and oracle inequalities in risk minimization ». In : *The Annals of Statistics* 36 (2006), p. 00-00.

[4] P. MASSART. « The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality ». In : *The Annals of Probability* 18 (1990), p. 1269-1283.

[5] V. Vapnik et A. Chervonenkis. « Necessary and sufficient conditions for the uniform convergence of means to their expectations ». In : *Theory of Probability and its Applications* 26 (1981), p. 821-832.

[6] V. Vapnik et A. Chervonenkis. « On the uniform convergence of relative frequencies of events to their probabilities ». In : *Theory of Probability and its Applications* 16 (1971), p. 264-280.

[7] N. Bingham, C. Goldie et J. Teugels. **Regular variation**. Cambridge University Press, 1987.

[8] M. Ledoux et M. Talagrand. **Probability in Banach Space**. New York : Springer-Verlag, 1991.

[9] V. Vapnik. **Estimation of Dependencies Based on Empirical Data**. New York : Springer-Verlag, 1982.

9.1 MOTIVATIONS

Le problème de *classification supervisée* est un problème de *prédiction*. C'est le problème qui se pose à un herboriste qui a ramassé dans fleurs dans la nature et qui essaye de déterminer pour chaque fleur l'espèce à laquelle elle appartient. Un être humain comparera la fleur aux spécimens décrit dans une flore, il utilisera de fait une procédure de classification arborescente : la classification des espèces est (était) un arbre, chaque critère de comparaison proposé par la flore indique une branche à suivre. Les critères utilisés dans les flores sont souvent qualitatifs.

On peut chercher et on a cherché depuis longtemps à automatiser les procédures de classification supervisée. L'une des premières incursions de la statistique dans le domaine de la classification supervisée fut l'étude de 150 specimens d'Iris par R. Fisher dans les années 30 (voir wikipedia). Le jeu de données (échantillon en statistique, ensemble d'entrainement dans le vocabulaire du *machine learning*), comporte 50 représentants de chacune des trois espèces : *Iris setosa*, *Iris versicolor*, *Iris virginica*. Pour chaque individu/observation (élément de l'échantillon), on dispose de 4 mesures (longueur, largeur des pétales et des sépales) et d'une étiquette l'espèce.

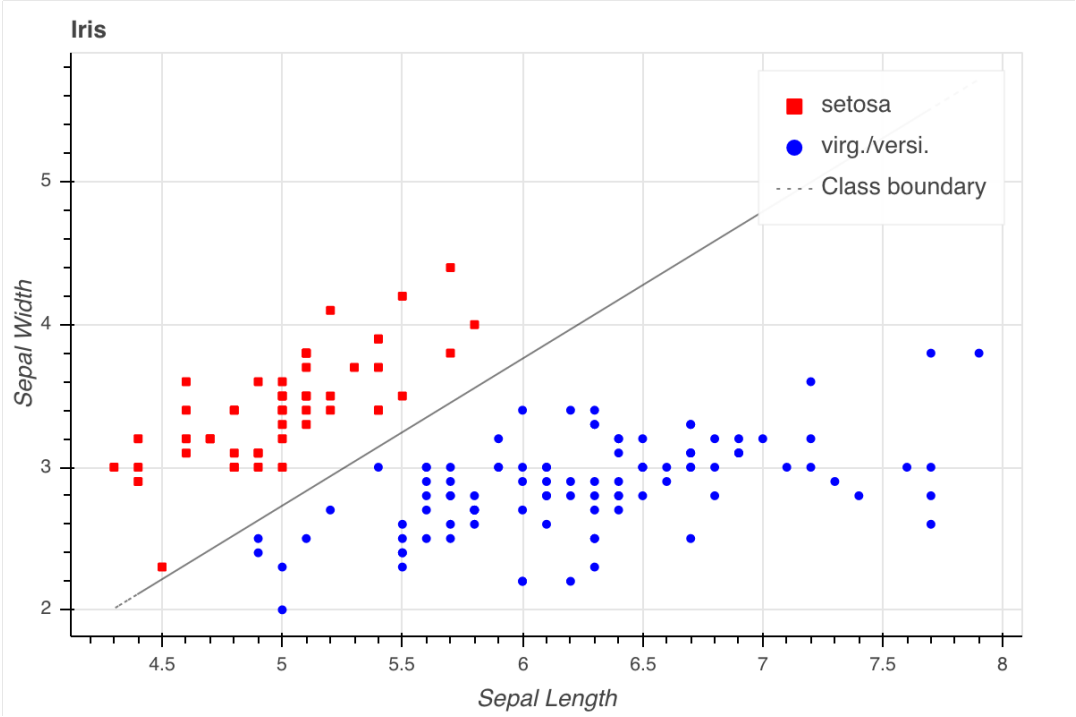

FIG. 9.1 : Classification binaire des Iris de Fisher selon la longueur et la largeur des sépales. La droite a été ajustée par régression logistique.

L'objectif de la classification binaire supervisée est de construire à partir de cet échantillon une fonction de $\mathbb{R}^4$ dans $\{-1, 1\}$ (associons 1 à *setosa*). L'espoir est que cette fonction (appelée *classifieur*) sera capable de classer correctement une nouvelle observation.

9.2 CLASSIFIEUR BAYÉSIEN, EXCÈS DE RISQUE

Pour formaliser le problème de classification supervisée, nous introduisons un cadre probabiliste : (X, Y) est un couple aléatoire à valeur dans $\mathcal{X} \times \mathcal{Y}$ où $\mathcal{Y} = \{-1, 1\}$. La loi P de (X, Y) est fixée, inconnue et nous ne souhaitons pas faire d'hypothèses à son sujet. Un tirage selon P modélise le tirage d'une observation X et de sa classe Y.

Nous supposerons ici que $\mathcal{X} = \mathbb{R}^D$ pour un D fini (on rencontre parfois des cadres plus sophistiqués où $\mathcal{X}$ est un espace de Hilbert, voire un Banach). L'espace $\mathcal{X}$ est muni de sa tribu habituelle.

Un *classifieur* (binaire) est une fonction mesurable de $\mathcal{X}$ dans $\{-1,1\}$. Un classifieur est utilisé pour prédire l'étiquette d'une observation.

Le risque du classifieur f est le probabilité de mal prédire Y à l'aide de f :

$$L(f) = \mathbb{P}\{f(X) \neq Y\} = \mathbb{E}[\mathbb{I}_{f(X)\neq Y}]\,.$$

Comme la loi P de (X,Y) n'est pas connue, nous serons amenés à approcher le risque par une quantité empirique, calculable à partir de l'échantillon (aussi appelé ensemble d'apprentissage) : le risque empirique du classifieur f,

$$L_n(f) = P_n\{f(X) \neq Y\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{f(X_i)\neq Y_i}\,.$$

Parmi les classifieurs, il en existe un qui minimise le risque. Pour définir ce classifieur, appelé classifieur bayésien, on introduit la fonction de régression, qui est l'espérance conditionnelle de Y sachant X

$$\eta(X) = \mathbb{E}[Y \mid X]\,.$$

La quantité $\eta(X)$ définit la loi conditionnelle de Y sachant X :

$$P(Y = 1 \mid X = x) = \frac{1+\eta(X)}{2}\,.$$

PROPOSITION 9.1 *Le classifieur bayésien est défini par*

$$f^*(X) = sign(\eta(X))\,.$$

avec la convention $sign(x) = 1$ *si* $x \geq 0$, -1 *sinon.*

$$\mathbb{E}[\mathbb{I}_{f^*(X)\neq Y} \mid X] = \frac{1-|\eta(X)|}{2}$$

Le classifieur bayésien minimise L. Pour tout classifieur f :

$$L(f) - L(f^*) = \mathbb{E}\left[\mathbb{I}_{f(X)f^*(X)<0}|\eta(X)|\right]\,.$$

PREUVE. L'excès de risque s'écrit :

$$\begin{aligned}
L(f) - L(f^*) &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{f(X)\neq Y} - \mathbb{I}_{f^*(X)\neq Y} \mid X\right]\right] \\
&= \mathbb{E}\left[\mathbb{I}_{f(X)\neq f^*(X)}\mathbb{E}\left[\mathbb{I}_{f^*(X)=Y} - \mathbb{I}_{f^*(X)\neq Y} \mid X\right]\right] \\
&= \mathbb{E}\left[\mathbb{I}_{f(X)\neq f^*(X)}\left(\frac{1+|\eta(X)|}{2} - \frac{1-|\eta(X)|}{2}\right)\right] \\
&= \mathbb{E}\left[\mathbb{I}_{f(X)\neq f^*(X)}|\eta(X)|\right]\,.
\end{aligned}$$

□

En apprentissage machine, l'objectif est de proposer une fonction f qui prédise Y presque aussi bien que le classifieur de Bayes. On se donne une classe de fonctions (mesurables) $\mathcal{F}$ de $\mathcal{X}$ dans $\{-1,1\}$. L'objectif est de minimiser le risque L dans $\mathcal{F}$. Pour alléger notations et calcul on suppose qu'il existe $\overline{f} \in \mathcal{F}$ tel que

$$L(\overline{f}) = \min_{f\in\mathcal{F}} L(f)\,.$$

Ce minimisant n'a pas de raison d'être unique (en général). Comme L n'est pas calculable (la loi de (X,Y) n'est pas connue), on adopte le *principe de substitution*, on minimise L_n. Le minimiseur du risque empirique est défini par

$$\widehat{f}_n = \arg\min L_n(f) \qquad \text{avec } f \in \mathcal{F}\,.$$

On se pose des questions

i) Que peut-on dire de $L(\widehat{f}_n) - L_n(\widehat{f}_n)$? Il s'agit de préciser

$$\mathbb{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] \geq 0\,.$$

ii) Que peut-on dire de l'*excès de risque* $L(\widehat{f}_n) - L(\overline{f})$ au delà du fait que cette variable aléatoire est positive ?

En se concentrant a priori sur la méthode de minimisation du risque empirique, nous adoptons une démarche de nature comparable à celle qui consiste à étudier la maximisation de la vraisemblance dans les modèles dominés. On ramène le problème de l'inférence statistique à un problème d'optimisation.

Techniquement, la situation est très différente de la maximisation de la vraisemblance dans les modèles exponentiels. Nous ne supposons pas grand chose sur le mécanisme de génération des données (sinon qu'elles sont i.i.d. selon P). La minimisation du risque de classification empirique n'est pas (en général) un problème d'optimisation convexe. Le minimisant du risque empirique n'est en général pas connu sous une forme close (même lorsqu'il est défini). En apprentissage machine, on ne cherche pas à s'appuyer sur des statistiques suffisantes fini-dimensionnelles (comme quand on travaille sur des modèles exponentiels), la seule statistique suffisante utilisée est la mesure empirique définie par l'échantillon (qui associe un point $1/n$ à chaque élément de l'échantillon).

Cette mesure empirique et l'ensemble de classifieurs $\mathcal{F}$ définissent un *processus empirique* indexé par $\mathcal{F}$:

$$(L_n(f) - L(f))_{f\in\mathcal{F}} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}_{f(X_i)\neq Y_i} - \mathbb{E}\mathbb{I}_{f(X_i)\neq Y_i}\right)\right)_{f\in\mathcal{F}}\,.$$

Pour étudier l'excès de risque et les quantités apparentées, nous allons nous concentrer sur le supremum du processus empirique.

L'étude de la minimisation du risque empirique passe en effet par le contrôle d'un processus empirique indexé par $\mathcal{F}$:

$$\begin{aligned} L(\widehat{f}_n) - L_n(\widehat{f}) &\leq \sup_{f\in\mathcal{F}} L(f) - L_n(f) \\ &\leq \sup_{f\in\mathcal{F}} |L(f) - L_n(f)| \end{aligned}$$

et

$$\begin{aligned} L(\widehat{f}_n) - L(\overline{f}) &\leq L(\widehat{f}_n) - L_n(\widehat{f}_n) - (L(\overline{f}) - L_n(\overline{f}) \\ &\leq \sup_{f\in\mathcal{F}} \left(L(f) - L_n(f) - (L(\overline{f}) - L_n(\overline{f})\right) \\ &\leq 2\sup_{f\in\mathcal{F}} |L(f) - L_n(f)|\,. \end{aligned}$$

Contrôler le supremum du processus empirique va nous obliger à développer de nouvelles techniques d'analyse substantiellement différentes des méthodes asymptotiques utilisées pour étudier le maximum de vraisemblance dans les modèles paramétriques.

9.3 CLASSIFIEURS LINÉAIRES

Pour illustrer le problème général, nous utiliserons un cas concret : les classifieurs linéaires.

Un classifieur linéaire définit un demi-espace :

$$f_w(x) = \text{sign}(\langle w, x\rangle)$$

On pourrait adopter la notation un peu plus générale

$$f_{w,\tau}(x) = \text{sign}(\langle w, x\rangle - \tau)$$

avec $w \in \mathbb{R}^p$, $\tau \in \mathbb{R}$. Ce dernier cadre n'est pas plus général que le précédent. Il suffit d'ajouter une coordonnée constante égale à 1 à chaque exemple, pour constater que tout algorithme d'apprentissage capable de résoudre le premier problème est capable de résoudre le second.

Lorsque nous chercherons à contrôler le supremum du processus empirique indexé par l'ensemble de classifieurs $\mathcal{F}$, nous utiliserons une inégalité de Bonferroni (*union bound*), nous serons amenés à manipuler le cardinal de

$$\{(\text{sign}(f(x_i)), \dots, \text{sign}(f(x_n)) : f \in \mathcal{F}\}$$

pour tout $(x_1, \dots, x_n) \in \mathcal{X}^n$. Les notions développées par Vapnik, Chervonenkis et quelques autres s'avéreront très utiles.

DÉFINITION 9.2 (COEFFICIENTS DE PULVÉRISATION ET DIMENSION DE VAPNIK-CHERVONENKIS) Soit $\mathcal{X}$ un ensemble, $\mathcal{A}$ un ensemble de parties de $\mathcal{X}$, et $\{x_1, \dots, x_n\} \subseteq \mathcal{X}$, on appelle *trace* de $\mathcal{A}$ sur $\{x_1, \dots, x_n\} \subseteq \mathcal{X}$

$$\text{Tr}_{\mathcal{A}}(x_1, \dots, x_n) = \Big\{ \{x_1, \dots, x_n\} \cap A : A \in \mathcal{A} \Big\}$$

On note

$$N_{\mathcal{A}}(x_1, \dots, x_n) = \Big| \text{Tr}_{\mathcal{A}}(x_1, \dots, x_n) \Big| .$$

Si $N_{\mathcal{A}}(x_1, \dots, x_n) = 2^n$, on dit que $(x_1, \dots, x_n)$ est *pulvérisé* par $\mathcal{A}$.

Les *coefficients de pulvérisation* de $\mathcal{A}$ sont définis par

$$\$(\mathcal{A}, n) = \max_{x_1, \dots, x_n} N_{\mathcal{A}}(x_1, \dots, x_n) .$$

La *dimension de Vapnik-Chervonenkis* de $\mathcal{A}$ est la cardinalité maximale des ensembles pulvérisés par $\mathcal{A}$:

$$\text{VC}(\mathcal{A}) = \max \big\{ n : \$(\mathcal{A}, n) = 2^n \big\} .$$

REMARQUE 9.3 La dimension de Vapnik-Chervonenkis peut être infinie. Il suffit que considérer $\mathcal{X}$ infini $\mathcal{A}$ l'ensemble de toutes les parties finies de $\mathcal{X}$.

L'ensemble des parties de $\mathcal{X}$ comportant chacune au plus v éléments de $\mathcal{X}$ possède une dimension de Vapnik-Chervonenkis exactement égale à v.

LEMME 9.4 *(Vapnik-Chervonenkis, Sauer, Shelah) Si* $\text{VC}(\mathcal{A}) = v < \infty$*, alors pour tout* n

$$\$(\mathcal{A}, n) \leq \sum_{i=0}^{v} \binom{n}{i} .$$

Dans la preuve, nous nous ramenons à l'étude d'un sous-ensemble du n-cube discret $\{0,1\}^n$. On définit un ordre partiel sur $\{0,1\}^n$, $u \preceq y$ si pour tout $i \leq n$, $u_i \leq v_i$. Un sous-ensemble E de $\{0,1\}^n$ est dit *monotone* ou *héréditaire*, si pour tout $v \in E$, pour tout $u \preceq v$, on a $u \in E$.

PREUVE. Si $n \leq v$, il n'y a rien à démontrer. Dans la preuve $n > v$, et $x_1, \dots, x_n$ est une suite de n éléments distincts de $\mathcal{X}$. On définit une bijection de

$$\text{Tr}_{\mathcal{A}}(x_1, \dots, x_n) = \{A \cap \{x_1, \dots, x_n\} : A \in \mathcal{A}\}$$

vers un sous-ensemble du n-cube discret $\{01, \}^n$:

$$E = \{w : w \in \{0,1\}^n, \exists A \in \mathcal{A}; w_i = \mathbb{I}_A(x_i)\} .$$

On note que $\mathcal{A}$ pulvérise $x_{i_1}, \dots, x_{i_k}$ si et seulement si la projection de E sur les dimensions $i_1, \dots, i_k$ est le k-cube discret (on dit alors par abus que E pulvérise $i_1, \dots, i_k$).

Pour $w \in \{0,1\}^n$, on note $w^{(i)}$ le vecteur $w_1, \dots, w_{i-1}, 0, w_{i+1}, \dots, w_n$ (on a toujours $w^{(i)} \preceq w$).

On définit l'opération de décalage selon la dimension $i \leq n$, T_i, pour $F \subseteq \{0,1\}^n$

$$T_i(F) = \{T_i(w, F) : w \in F\}$$

avec pour tout $w \in F$

$$T_i(w, F) = \begin{cases} w^{(i)} & \text{si } w^{(i)} \notin \mathcal{F} \\ w & w\text{sinon} \end{cases}$$

On note

i) $|T_i(F)| = |F|$
ii) $|T_i(F)|$ et $|F|$ pulvérisent les mêmes ensembles d'indices.
iii) $T_i(F) = F$, si pour tout $w \in F$, $w^{(i)} \in F$.

On cherche ensuite à appliquer les transformations T_i aux transformées de E, on s'arrête lorsqu'aucune transformation ne modifie le sous-ensemble courant F.

```
F, monotone, i  = E, False, 1   # initialisations
while not monotone :
  while i < n+1 :
    if (T[i](F) != F) :
      F = T[i](F)
      break
    else :
      i = i+1
  if i == n+1 :   #
    monotone = True
```

L'algorithme termine car lors d'une itération de la boucle `while` externe, soit une des transformations modifie F et on fait diminuer d'au moins 1, la somme des poids de Hamming des éléments de F, soit on constate que F est monotone. La somme des poids de Hamming des éléments de F est une quantité positive.

A chaque itération de la boucle `while` externe, on préserve le cardinal de F et sa VC-dimension.

Notons $\widetilde{E}$ la valeur de F lorsqu'on sort de la boucle `while` externe.

Si un sous-ensemble monotone de $\{0,1\}^n$ contient un élément avec d coordonnées non-nulles alors, il pulvérise ces d coordonnées. Comme la dimension de Vapnik-Chervonenkis n'augmente pas avec les décalages, on conclut que $\widetilde{E}$ ne contient pas d'élément comportant plus de v coordonnées non-nulles, soit

$$|E| = |\widetilde{E}| \leq \sum_{i=0}^{v} \binom{n}{i}.$$

□

Les coefficients de pulvérisation des classes de Vapnik-Chervonenkis sont souvent majorés à l'aide du résultat technique suivant. Cette idée est précisée par le pseudo-code `python` qui suit.

PROPOSITION 9.5 *Pour $n \geq v$,*

$$\sum_{i=0}^{v} \binom{n}{i} \leq \left(\frac{\mathrm{e}n}{v}\right)^v.$$

PREUVE. On a

$$\begin{aligned}
\sum_{i=0}^{v} \binom{n}{i} \left(\frac{v}{n}\right)^v &\leq \sum_{i=0}^{v} \binom{n}{i} \left(\frac{v}{n}\right)^i \\
&\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{v}{n}\right)^i \\
&= \left(1 + \frac{v}{n}\right)^n \\
&\leq \mathrm{e}^v
\end{aligned}$$

où la première inégalité vient de $v/n \leq 1$, □

Pour se figurer un cas concret et utile de classe de Vapnik, il suffit de penser aux demi-espaces.

PROPOSITION 9.6 *La dimension de Vapnik-Chervonenkis des demi-espaces de $\mathbb{R}^d$ est inférieure où égale à $d+1$.*

La proposition 9.6 est un corollaire de la proposition suivante.

PROPOSITION 9.7 *Soit $\phi_1, \ldots, \phi_k$ une collection de fonctions mesurables de $\mathcal{X}$ dans $\mathbb{R}$. On note r la dimension de l'espace vectoriel engendré $\mathcal{G}$ par $\phi_1, \ldots, \phi_k$. On considère l'ensemble des classifieurs $\mathcal{F} = \{\text{sign} \circ g : g \in \mathcal{G}\}$.*

$$\text{VC}(\mathcal{F}) \leq r = \dim(\mathcal{G})$$

PREUVE.(Proposition 9.7)

On procède par contradiction. Supposons $\{x_1, x_2, \ldots, x_n\}$ pulvérisée par $\mathcal{F}$ avec $n \geq r + 1$. Cet ensemble pulvérisé définit une fonction linéaire de $\mathcal{G}$ dans $\mathbb{R}^n$:

$$L(g) = (g(x_1), \ldots, g(x_n))^t .$$

L'ensemble $L(\mathcal{G})$ est un sous-espace vectoriel de $\mathbb{R}^n$ de dimension inférieure ou égale à $r < n$. Il existe donc un vecteur $u \in \mathbb{R}^n \setminus \{0\}$ orthogonal à $L(\mathcal{G})$, soit

$$\sum_{i=1}^{n} u_i g(x_i) = 0 \qquad \forall g \in \mathcal{G} .$$

Comme on a supposé $\{x_1, x_2, \ldots, x_n\}$ pulvérisée par $\mathcal{F}$, il existe $g \in \mathcal{G}$ tel que $g(x_1), g(x_2), \ldots, g(x_n)$ sont tous de même signe. Les coefficients u_i ne peuvent être tous de même signe. On peut écrire pour tout $g \in \mathcal{G}$

$$-\sum_{i:u_i<0} u_i g(x_i) = \sum_{i:u_i \geq 0} u_i g(x_i) .$$

Maintenant, l'hypothèse de pulvérisation implique qu'il existe $g \in \mathcal{G}$ tel que $g(x_i)u_i \geq 0$ pour tout $i \leq n$. □

9.5 INÉGALITÉS DE VAPNIK-CHERVONENKIS

THÉORÈME 9.8 *(INÉGALITÉS DE VAPNIK-CHERVONENKIS) Soient $X_1, \ldots, X_n$ i.i.d. sur un espace probabilisé et $\mathcal{F}$ une classe d'indicatrices (mesurables) sur l'espace des valeurs des X_i, alors pour tout n, tout $\eta > 0$,*

$$\mathbb{E}\left[\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right|\right] \leq \sqrt{\frac{2\log(2\mathbb{S}(\mathcal{F}, n))}{n}}$$

et, pour $\alpha > 0$,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right| \geq \alpha\right\} \leq \mathbb{S}(\mathcal{F}, n)\mathrm{e}^{-\frac{n\alpha^2}{16}} .$$

REMARQUE 9.9 Les bornes du théorème ne sont pas optimales. Elles peuvent être améliorées en utilisant des arguments plus fins que les bornes de Bonferroni, les arguments de chainage.

Si la classe d'indicatrices est une classe de Vapnik-Chervonenkis, on obtient des majorants faciles à exploiter (mais conservateurs). En exploitant le lemme 9.7 et la proposition 9.5, on déduit du théorème **??** le corollaire suivant.

COROLLAIRE 9.10 *Si $\mathcal{F}$ est une classe de Vapnik-Chervonenkis de dimension v, pour $n \geq v$,*

$$\mathbb{E}\left[\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right|\right] \leq \sqrt{8\frac{v}{n}\log\frac{2en}{v}}$$

et pour $\alpha > 0$

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right| \geq \alpha\right\} \leq 2\left(\frac{en}{v}\right)^{v} \mathrm{e}^{-\frac{n\alpha^2}{8}}.$$

La preuve du théorème 9.8 repose sur un principe simple : la symétrisation (déjà utilisée dans le cours 8) et sur l'inégalité de Hoeffding (lemme 1.13). La symétrisation mettra en avant la quantité suivante.

DÉFINITION 9.11 (COMPLEXITÉ DE RADEMACHER) Soit $\mathcal{F}$ une collection de fonctions mesurables de $\mathcal{X}$ dans $\mathbb{R}$, $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes à valeur dans $\mathcal{X}$, et $\epsilon_1, \ldots, \epsilon_n$ une suite de variables de Rademacher indépendantes et indépendante de $X_1, \ldots, X_n$, on appelle *complexité de Rademacher* de $\mathcal{F}$, la quantité suivante :

$$R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right]$$

La complexité de Rademacher et encore plus la complexité conditionnelle de Rademacher :

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right| \mid X_1, \ldots, X_n\right]$$

sont des outils commodes pour majorer l'espérance du supremum d'un processus empirique.

Rappelons maintenant le lemme de symétrisation, déjà utilisé pour établir des inégalités de déviation pour la statistique de Kolmogorov-Smirnov.

LEMME 9.12 *Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes à valeur dans $\mathcal{X}$, $\mathcal{F}$ une classe de fonctions mesurables de $\mathcal{V}$ dans $\mathbb{R}$. Soient $\epsilon_1, \ldots, \epsilon_n$ une suite de variables de Rademacher indépendantes entre elles et indépendantes des $X_1, \ldots X_n$. Si $\Psi : \mathbb{R} \to \mathbb{R}$ est convexe et croissante, alors*

$$\mathbb{E}\left[\Psi\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right|\right)\right] \leq \mathbb{E}\left[\Psi\left(2\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right)\right].$$

En choisissant $\Psi = \mathrm{Id}$, on obtient

COROLLAIRE 9.13

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right|\right] \leq 2R_n(\mathcal{F}).$$

Pour majorer $R_n(\mathcal{F})$, nous utiliserons l'inégalité maximale suivante.

LEMME 9.14 *Si $Z_1, \dots, Z_n$ sont des variables aléatoires (pas nécessairement indépendantes) telles qu'il existe une fonction convexe Υ vérifiant pour $\lambda \geq 0$,*

$$\log \mathbb{E}\mathrm{e}^{\lambda Z_i} \leq \Upsilon(\lambda) \qquad \textit{pour tout } i \leq n$$

alors

$$\mathbb{E}[\max_{i \leq n} Z_i] \leq \inf_{\lambda \geq 0} \frac{\log n + \Upsilon(\lambda)}{\lambda} .$$

PREUVE.(Lemme 9.14) Pour tout $\lambda \geq 0$, en commençant par invoquer l'inégalité de Jensen,

$$\begin{aligned} \mathrm{e}^{\mathbb{E}[\lambda \max_i Z_i]} &\leq \mathbb{E}\left[\max_{i \leq n} \mathrm{e}^{\lambda Z_i}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{e}^{\lambda Z_i}\right] \\ &\leq n\mathrm{e}^{\Upsilon(\lambda)} \end{aligned}$$

On peut les logarithmes des deux cotés de l'inégalité, et on optimise en λ pour obtenir le lemme. □

Nous pouvons maintenant établir le théorème.

PREUVE.(Théorème 9.8) Pour obtenir la première inégalité du théorème, on peut partir du lemme de symétrisation avec $\Psi = \mathrm{Id}$, on obtient d'abord le corollaire 9.13

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right|\right] &\leq 2R_n(\mathcal{F}) \\ &= 2\mathbb{E}\left[\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \epsilon_i f(X_i)\right| \mid X_1, \dots, X_n\right]\right] . \end{aligned}$$

L'égalité repose sur le théorème de Fubini.

Considérons maintenant une réalisation des X_i ($X_i = x_i$ pour $i \leq n$), le supremum en $\mathcal{F}$ n'est qu'un maximum calculé sur une collection finie de vecteurs à coefficients dans $\{-1, 1\}$, ces vecteurs correspondent à la trace de $\mathcal{F}$ dans $x_1, \dots, x_n$: $(f(x_1), \dots, f(x_n))$ est à valeur dans $\{0, 1\}^n$. À cause de la valeur absolue, il faut aussi prendre en compte le vecteur opposé.

$$\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \epsilon_i f(x_i)\right| = \max_{A \in \mathrm{Tr}_{\mathcal{F}}(x_1, \dots, x_n)} \max\left(\sum_{i=1}^{n} \epsilon_i \mathbb{I}_A(x_i), -\sum_{i=1}^{n} \epsilon_i \mathbb{I}_A(x_i)\right) .$$

Pour chaque $f \in \mathcal{F}$, en invoquant le lemme de Hoeffding (Lemme **??**)

$$\log \mathbb{E}\mathrm{e}^{\lambda \sum_{i=1}^{n} \epsilon_i f(x_i)} \leq \mathrm{e}^{n\lambda^2/2} .$$

On combine avec le lemme 9.14, et on optimise en λ pour obtenir :

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \epsilon_i f(X_i)\right| \mid X_1, \dots, X_n\right] &\leq \inf_{\lambda > 0} \frac{\log(2\mathbb{S}(\mathcal{F}, n)) + \frac{n\lambda^2}{2}}{\lambda} \\ &\leq \sqrt{2n \log(2\mathbb{S}(\mathcal{F}, n))} . \end{aligned}$$

Pour obtenir les inégalités exponentielles, on part de nouveau du lemme de symétrisation 9.12 mais avec $\Psi(x) = \mathrm{e}^{\lambda x}$, $\lambda > 0$. On raisonne encore conditionnellement à $X_1, \dots, X_n$. À nouveau, on peut

oublier la structure de $\mathcal{F}$, c'est $\mathrm{Tr}_{\mathcal{F}}(x_1,\dots,x_n)$ qui importe

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(2\lambda\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\epsilon_i f(X_i)\right|\right)\mid X_1,\dots,X_n\right]\\
&=\mathbb{E}\left[\max_{A\in\mathrm{Tr}_{\mathcal{F}}(x_1,\dots,x_n)}\exp\left(2\lambda\left|\sum_{i=1}^n\epsilon_i\mathbb{I}_A(x_i)\right|\right)\right]\\
&=\mathbb{E}\left[\sum_{A\in\mathrm{Tr}_{\mathcal{F}}(x_1,\dots,x_n)}\exp\left(2\lambda\left|\sum_{i=1}^n\epsilon_i\mathbb{I}_A(x_i)\right|\right)\right]\\
&\le 2\$(\mathcal{F},n)\times \mathrm{e}^{2n\lambda^2},
\end{aligned}$$

à la dernière étape, nous avons encore invoqué l'inégalité de Hoeffding.

Pour obtenir les probabilités de queue, on utilise l'inégalité de Markov et on optimise en λ.

$$\begin{aligned}
\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^n f(X_i)-\mathbb{E}f(X_i)\right|\ge\alpha\right\}&\le\inf_{\lambda>0}\mathrm{e}^{-n\lambda\alpha}\mathbb{E}\left[\exp\left(\lambda\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\epsilon_i f(X_i)\right|\right)\right]\\
&\le 2\mathbb{E}\big[2\$(\mathcal{F},n)\big]\inf_{\lambda>0}\mathrm{e}^{-n(\lambda\alpha-2\lambda^2)}\\
&=2\mathbb{E}\big[2\$(\mathcal{F},n)\big]\,\mathrm{e}^{-n\alpha^2/8}.
\end{aligned}$$

□

9.6 BORNES DE RISQUE POUR LA MINIMISATION DU RISQUE EMPIRIQUE

Nous allons maintenant mettre à profit les inégalités de Vapnik-Chervonenkis pour établir des bornes de risque qui seront valables quelque soit la loi de (X,Y), dès que l'ensemble des classifieurs $\mathcal{F}$ possède une dimension de Vapnik-Chervonenkis v finie et dès que $v\ge n$.

Analyse globale

On considère la classe de parties $\mathcal{F}_L$, $\mathcal{X}\times\{-1,1\}$ définie par

$$\mathcal{F}_L=\Big\{\{(x,y):(x,y)\in\mathcal{X}\times\{-1,1\},f(x)\ne y\},\qquad f\in\mathcal{F}\Big\}.$$

Si $\mathcal{F}$ est une classe de Vapnik-Chervonenkis de dimension v, il en est de même pour $\mathcal{F}_L$. On peut invoquer les inégalités de Vapnik-Chervonenkis à propos de $\mathcal{F}_L$.

THÉORÈME 9.15 *Si $\mathcal{F}$ est un ensemble de classifieurs qui forme une classe de Vapnik-Chervonenkis de dimension v, et $n\ge v$,*

$$0\le\mathbb{E}\left[L(\widehat{f}_n)-L_n(\widehat{f}_n)\right]\le\sqrt{2\frac{v}{n}\log\frac{2en}{v}}.$$

$$0\le\mathbb{E}\left[L(\widehat{f}_n)-L(\overline{f})\right]\le 2\sqrt{8\frac{v}{n}\log\frac{en}{v}}.$$

Pour $\alpha>0$,

$$\mathbb{P}\left\{\left|L(\widehat{f}_n)-L_n(\widehat{f}_n)\right|\ge\alpha\right\}\le\left(\frac{en}{v}\right)^v\mathrm{e}^{-\frac{n\alpha^2}{16}}.$$

PREUVE. Il faut vérifier la première inégalité. En effet pour chaque f fixé $L_n(f)$ est un estimateur sans biais de $L(f)$. Mais le minimisant du risque empirique $\widehat{f}_n$ est lui meme une quantité aléatoire.

$$\begin{aligned}\mathbb{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] &= \mathbb{E}\left[L(\widehat{f}_n) - L(\overline{f})\right] + \mathbb{E}\left[L(\overline{f}) - L_n(\widehat{f}_n)\right] \\ &\geq \mathbb{E}\left[L(\overline{f}) - L_n(\widehat{f}_n)\right] \\ &= \mathbb{E}\left[L_n(\overline{f}) - L_n(\widehat{f}_n)\right] \\ &\geq 0\,,\end{aligned}$$

où la dernière égalité provient du fait que pour chaque f, $L_n(f)$ est un estimateur sans biais de $L(f)$.

Pour obtenir les majorants des espérances, il suffit de majorer

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}_L}\left|\sum_{i=1}^{n} g(X_i,Y_i) - \mathbb{E}g(X_i,Y_i)\right|\right]\,.$$

On observe que pour tout $\{(x_i,y_i),\ldots,(x_n,y_n)\}$),

$$|\mathrm{Tr}_{\mathcal{G}}\{(x_i,y_i),\ldots,(x_n,y_n)\}| \leq |\mathrm{Tr}_{\mathcal{F}}\{x_i,\ldots,x_n\}|\,.$$

Les bornes de risque se déduisent alors du Corollaire 9.10. □

Raffinement, analyse localisée

Les bornes de risque présentées ici constituent un résultat préliminaire. Elles sont pessimistes. Pour obtenir des bornes de risques fines, dont l'ordre de grandeur coïncide avec les bornes inférieures, il est nécessaire d'analyser finement

$$L(\widehat{f}_n) - L_n(\widehat{f}_n) - (L(\overline{f}) - L_n(\overline{f})\,.$$

Cela permet, sous des conditions favorables, notamment lorsque on dispose d'un minorant presque sur de $|\eta(X)|$ d'établir que l'excès de risque décroit comme $1/n$ plutôt que comme $1/\sqrt{n}$.

9.7 CONVEXIFICATION DU RISQUE, RISQUE LOGISTIQUE

Cette analyse statistique de la minimisation du risque de classification empirique n'a pas tenu compte des coûts de calcul. Dans la plupart des situations d'intérêt pratique, minimiser l'erreur de classification empirique est un problème difficile (NP-difficile). Cela a longtemps été une source de malentendus entre praticiens et théoriciens. Les premiers objectant que les bornes de risque conservatrices annoncées par les seconds n'avaient rien à voir avec l'apprentissage machine dans la vraie vie.

Pour les classifieurs linéaires, la minimisation du risque de classification empirique est une méthode justifiée du point de vue statistique. Mais cette méthode n'est pas praticable du point de vue calculatoire.

Ce constat a conduit les praticiens à changer d'objectif. Plutôt que de minimiser à cout prohibitif un risque empirique de classification, ils ont choisis de minimiser des risques empiriques plus faciles à traiter en machine. Cela les a conduits à travailler sur des risques convexes. Parmi ces risques convexes, on trouve le *risque logistique* dérivé de la perte logistique définis pour les fonctions g de $\mathcal{X}$ dans $\mathbb{R}$ par :

$$\log\left(1+\mathrm{e}^{-yg(x)}\right) \qquad \text{pour } x\in\mathcal{X}, y\in\{-1,1\}\,.$$

La perte logistique de g est reliée à la perte de classification de $\mathrm{sign}\circ g$ par :

$$\mathbb{I}_{\mathrm{sign}(g(x))y<0} \leq \log_2\left(1+\mathrm{e}^{-yg(x)}\right)$$

Pour les classifieurs linéaires, définis par $w\in\mathbb{R}^d, b\in\mathbb{R}$, la perte logistique

$$\log\left(1+\mathrm{e}^{-y(\langle w,x\rangle - b)}\right)$$

est convexe en w, b.

Dans la suite, on notera $\phi(z) = \log(1 + \mathrm{e}^{-z})$. On appellera ϕ la fonction de perte logistique.

Le risque logistique du classifieur linéaire défini par w est défini par :

$$L^{\log}(w) = \mathbb{E}\left[\log\left(1 + \mathrm{e}^{-Y\langle w, X\rangle}\right)\right] = \mathbb{E}\left[\phi(Y\langle w, X\rangle)\right]$$

Le risque empirique correspondant est

$$L_n^{\log}(w) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-Y_i\langle w, X_i\rangle}\right) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y\langle w, X_i\rangle)\,.$$

Peut-on justifier le changement de la définition du risque empirique à minimiser ? Peut-on contrôler le risque d'un classifieur construit par minimisation du risque logistique empirique ?

Nous allons ici montrer que l'excès de risque de classification de $\mathrm{sign} \circ g$ est lié à l'excès de risque logistique de g.

Nous allons maintenant nous intéresser à l'analogue du classifieur bayésien pour la régression logistique, c'est à dire chercher à déterminer s'il existe une fonction mesurable de X, qui minimise le risque logistique et si oui déterminer sa forme.

Pour une fonction mesurable $t : \mathcal{X} \to \mathbb{R}$, on cherche à minimiser

$$\mathbb{E}[\phi(Y\,t(X))] = \mathbb{E}\left[\mathbb{E}[\phi(Y\,t(X)) \mid X]\right]\,.$$

Pour une valeur de $t \in \mathbb{R}$, le risque logistique conditionnel est

$$\mathbb{E}[\phi(Y\,t) \mid X] = \frac{1 + \eta(X)}{2}\phi(t) + \frac{1 - \eta(X)}{2}\phi(-t)\,.$$

On note au passage que

$$\mathbb{E}[\phi(Yt) \mid X] = \mathbb{E}[\phi(tY) \mid \eta],$$

On écrira par la suite $t(\eta)$ plutôt que $t(X)$. Si $|\eta| \neq 1$, l'infimum est atteint en

$$t^*(\eta) = \log\frac{1 + \eta}{1 - \eta}\,.$$

Sans surprise, il ne dépend que de $\eta(X)$ (comme le classifieur bayésien), et son risque logistique conditionnel optimal est

$$H(\eta) = -\frac{1 + \eta}{2}\log\frac{1 + \eta}{2} - \frac{1 - \eta}{2}\log\frac{1 - \eta}{2}$$

c'est l'*entropie de Shannon* d'une variable de Bernoulli de paramètre $\frac{1+\eta}{2}$.

Ces observations peuvent être résumées à l'aide de la définition et du lemme suivants.

Définition 9.16 Une fonction de perte ℓ est *calibrée pour la classification* si (et seulement si), la fonction de X qui minimise la perte conditionnelle

$$t(x) = t(\eta) = \arg\min_{t \in \mathbb{R}} \mathbb{E}[\ell(tY) \mid X = x]$$

est du même signe que le classifieur bayésien, c'est-à-dire si elle vérifie

$$\mathrm{sign}(t(X)) = \mathrm{sign}(\eta(X))\,.$$

Lemme 9.17 *La fonction de perte logistique est calibrée pour la classification,*

$$t(x) = \arg\min_{t \in \mathbb{R}} \mathbb{E}[\phi(tY) \mid X = x] = \log\frac{1 + \eta(x)}{1 - \eta(x)}$$

et la perte logistique optimale conditionnellement à X ne dépend que de $\eta = \eta(X)$:

$$H(\eta) = -\frac{1 + \eta}{2}\log\frac{1 + \eta}{2} - \frac{1 - \eta}{2}\log\frac{1 - \eta}{2}$$

*c'est l'*entropie de Shannon *d'une variable de Bernoulli de paramètre $\frac{1+\eta}{2}$.*

Si le risque logistique $L^{\log}$ était calculable, sa minimisation nous livrerait le classifieur bayésien. Mais le risque logistique n'est pas plus calculable que le risque de classification. En pratique, on doit se contenter de chercher à minimiser le risque logistique empirique dans une classe de fonctions. Pour nous il s'agira des fonctions linéaires. Si on veut analyser la minimisation du risque logistique empirique, il faut, comme pour le risque de classification, contrôler l'excès de risque (logistique), c'est un problème de processus empirique, plus facile que celui du contrôle de l'excès de risque de classification en raison de la convexité de la fonction de perte, mais plus difficile en raison du fait que la perte n'est plus bornée. Mais contrôler l'excès de risque logistique ne suffit pas, il faut aussi relier l'excès de risque logistique à l'excès de risque de classification.

C'est à cette dernière tache que nous allons nous consacrer.

Il est utile d'introduire quelques notions qui nous permettrons de facturer l'excès de risque de classification en terme d'excès de risque logistique.

$$H^-(\eta) = \inf_{t:\eta t \leq 0} \mathbb{E}[\phi(tY) \mid \eta] \,.$$

Comme

$$\mathbb{E}[\phi(tY) \mid \eta] = \frac{1+\eta}{2}\phi(t) + \frac{1-\eta}{2}\phi(-t)$$

déterminer H^- revient à minimiser une fonction convexe sur le domaine $(-\infty, 0]$ si $\eta > 0$, $[0, \infty)$ si $\eta < 0$. La dérivée du cout est croissante. En 0, elle vaut $\eta\phi'(0)$. Comme $\phi'(0) < 0$, si $\eta > 0$, la dérivée du cout est négative sur $(-\infty, 0]$, si $\eta < 0$, la dérivée du cout est positive sur $[0, \infty)$. Dans tous les cas, le minimum sous-contrainte est atteint en $t = 0$ et vaut $\phi(0)$.

On définit G par

$$G(\eta) = H^-(\eta) - H(\eta) \,.$$

C'est l'excès de risque logistique conditionnel minimal lorsque le minimisant du risque logistique conditionnel conduit à une prédiction erronée du signe de η. Les calcul menés précédemment nous conduisent à

$$G(\eta) = D\left(\mathcal{B}(\frac{1+\eta}{2}) \| \mathcal{B}(\frac{1}{2})\right)$$

G est l'entropie relative entre deux lois de Bernoulli. On note au passage que G est convexe en η.

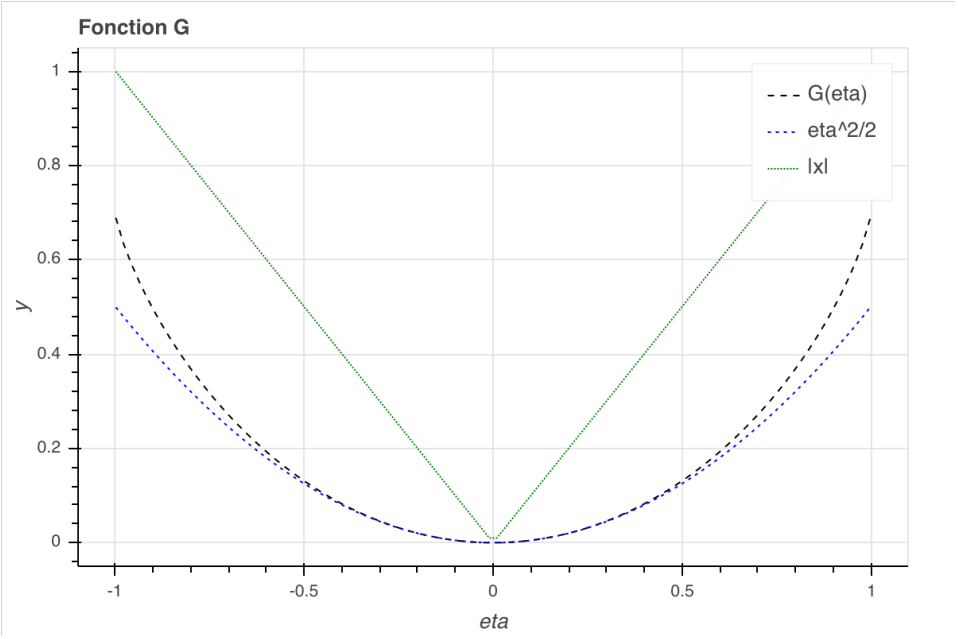

FIG. 9.2 : Fonction $G(\eta)$ comparée aux fonctions $\eta^2/2$ et $|\eta|$.

Insistons sur le fait que $G(\eta)$ représente l'excès de risque logistique conditionnel minimal d'un classifieur qui n'est pas du même signe que le classifieur bayésien. L'excès de risque de classification correspondant est $|\eta|$.

Nous sommes maintenant en mesure d'établir une relation entre excès de risque de classification et excès de risque logistique

THÉORÈME 9.18 *Si f désigne une fonction de $\mathcal{X}$ dans $\mathbb{R}$, et $\operatorname{sign} \circ f$ le classifieur associé, alors l'excès de risque de classification de $\operatorname{sign} \circ f$ est relié à l'excès de risque logistique de*

$$G(L(\operatorname{sign} \circ f) - L(f^*)) \leq L^{\log}(f) - \inf_f L^{\log}(f)\,.$$

PREUVE. En invoquant l'inégalité de Jensen, la convexité de G et le fait que $G(0) = 0$,

$$\begin{aligned}
G\left(L(\operatorname{sign} \circ f) - L(f^*)\right) &= G\left(\mathbb{E}\left[\mathbb{I}_{\operatorname{sign}(f(X))\eta(X)<0}|\eta(X)|\right]\right) \\
&\leq \mathbb{E}\left[\mathbb{I}_{\operatorname{sign}(f(X))\eta(X)<0} G(|\eta(X)|)\right] \\
&\leq \mathbb{E}\left[\mathbb{I}_{\operatorname{sign}(f(X))\eta(X)<0}\left(H^-(\eta(X)) - H(\eta(X))\right)\right] \\
&\leq \mathbb{E}\left[\mathbb{I}_{\operatorname{sign}(f(X))\eta(X)<0}\left(\mathbb{E}[\phi(Yf(X)) \mid X] - H(\eta(X))\right]\right. \\
&\leq \mathbb{E}\left[\phi(Yf(X)) - H(\eta(X))\right] \\
&= L^{\log}(f) - \inf_f L^{\log}(f)\,.
\end{aligned}$$

□

REMARQUE 9.19 AUTRE MOTIVATION DE LA MINIMISATION DU RISQUE LOGISTIQUE. La régression logistique peut être introduite comme problème de classification soulevé dans le cadre des modèles paramétriques réguliers. Supposons que nous ayons affaire à un modèle exponentiel sous forme canonique. Cela signifie que nous avons une mesure σ-finie sur $(\mathcal{X}, \mathcal{F})$, une fonction mesurable $T : \mathcal{X} \to \mathbb{R}^d$ et Θ° est l'intérieur (non-vide) de l'ensemble convexe

$$\Theta = \left\{\beta : \beta \in \mathbb{R}^d,\ ; Z(\beta) = \int \exp(\langle \beta, T(x)\rangle)\nu(\mathrm{d}x) < \infty\right\}\,.$$

Le modèle exponentiel est défini par un ensemble paramétrique de densités sur $\mathcal{X} : x \mapsto p(x \mid \beta) = \exp(\langle \beta, T(x)\rangle - \log Z(\beta)), \beta \in \Theta^\circ$. Considérons maintenant les distributions de probabilités sur $\mathcal{X} \times \{-1, 1\}$ paramétrées par des triplets $(\pi, \beta^+, \beta^-) \in (0, 1) \times \Theta^\circ \times \Theta^\circ$, avec des densités conditionnelles

$$\begin{aligned}
p(X = x \mid Y = 1) &= p(x \mid \beta^+) \\
p(X = x \mid Y = -1) &= p(x \mid \beta^-) \\
p(Y = 1) &= \pi\,.
\end{aligned}$$

Pour une telle densité jointe sur $\mathcal{X} \times \{-1, 1\}$, le logarithme du rapport des cotes satisfait

$$\log \frac{p(Y = 1 \mid X = x)}{p(Y = -1 \mid X = x)} = \langle \beta^+ - \beta^-, T(x)\rangle - \log\left(\frac{1-\pi}{\pi}\frac{Z(\beta^+)}{Z(\beta^-)}\right)\,.$$

Cette dépendance linéaire du logarithme du rapport des cotes par rapport au vecteur $T(x)$ est la caractéristique de la régression logistique. Dans ce réglage, le classificateur Bayes est linéaire par rapport à T :

$$g^*(x) = \begin{cases} 1 & \text{si } \langle \beta^+ - \beta^-, T(x)\rangle > \log\left(\frac{1-\pi}{\pi}\frac{Z(\beta^+)}{Z(\beta^-)}Z(\beta^)\right) \\ -1 & \text{sinon.} \end{cases}$$

Le classificateur Bayes est linéaire et dépend de $\beta^+, \beta^-, \beta^-, \pi$ à $\beta + -\beta^-$ et l'intercept $\log\left(\frac{1-\pi}{\pi}\frac{Z(\beta^+)}{Z(\beta^-)}\right)$. Étant donné un échantillon de formation $(x_i, y_i)_{i\leq n}$, la régression logistique aborde le problème de l'estimation de $\beta^+ - \beta^-$. Le but ultime n'est peut-être pas la prédiction des étiquettes à partir d'exemples, mais l'estimation même des coefficients avec des valeurs possibles (applications à l'inférence causale [5, 7, 9]). L'inférence au maximum de vraisemblance des paramètres du classificateur de Bayes exige seulement de maximiser la probabilité conditionnelle de les étiquettes $y_1, \ldots, y_n$ étant donné les exemples $x_1, \ldots, x_n$. Quelques lignes de calculs révèlent que cela revient à minimiser le risque logistique empirique

$$\sum_{i=1}^n \log\left(1 + \mathrm{e}^{-y_i\langle \beta, T(x_i)\rangle + b}\right)$$

par rapport $\beta \in \mathbb{R}^d, b \in \mathbb{R}$. Ce problème de minimisation s'avère être un problème de minimisation « lisse » et « fortement convexe ». Ce problème de minimisation peut être résolu efficacement par des méthodes du second ordre (la méthode de Newton-Raphson se résume à résoudre une séquence de problèmes de moindres carrés pondérés).

Avant d'etre utilisée en apprentissage pour convexifier le risque de classification, la régression logistique a été utilisée en statistique classique pour estimer les (des) paramètres d'un modèle censé engendrer les données.

9.8 REMARQUES BIBLIOGRAPHIQUES

La classification (binaire) est une (toute) petite partie des problèmes qui forment le *machine learning* et son versant théorique la théorie statistique de l'apprentissage. Pour un panorama des problèmes et des méthodes, on peut se rapporter à [8, 12].

Les livres [13] et [**mohri**] fournissent une introduction à la théorie de l'apprentissage du point de vue de l'informatique, c'est-à-dire du point de vue de la théorie de la complexité. La théorie calculatoire de l'apprentissage est décrite par son fondateur dans le petit volume [14].

La théorie statistique de l'apprentissage est vigoureusement expliquée dans [15]. L'éventail des méthodes classiques au milieu des années 90 est présenté de façon rigoureuse mais accessible dans [6]. L'étude fine de l'excès de risque en classification a été initiée par [3]. On trouvera un survol de la théorie de la classification en 2005 dans [1] avec notamment une introduction aux analyses localisées. [2, 4, 10, 11] présentent en détail une analyse fine de l'excès de risque en classification.

L'article [16] donne une présentation générale de la convexification du risque en classification. La bibliographie de cet article donne des références précises pour les différents risques de substitution envisagés.

Références

[1] S. Boucheron, O. Bousquet et G. Lugosi. « Theory of classification : a survey of some recent advances ». In : *ESAIM Probability and Statistics* 9 (2005), p. 323-375.

[2] V. Koltchinskii. « Local Rademacher complexities and oracle inequalities in risk minimization ». In : *The Annals of Statistics* 36 (2006), p. 00-00.

[3] E. Mammen et A. B. Tsybakov. « Smooth discrimination analysis ». In : *Annals of Statistics* 27.6 (1999), p. 1808-1829. ISSN : 0090-5364. MR : MR1765618(2001i:62074).

[4] P. Massart et E. Nedelec. « Risk bounds for classification ». In : *Annals of Statistics* 34.5 (2006), p. 2326.

[5] D. R. Cox. **The analysis of binary data**. Methuen & Co., Ltd., London, 1970, p. viii+142. MR : 0282453.

[6] L. Devroye, L. Györfi et G. Lugosi. **A Probabilistic Theory of Pattern Recognition**. Springer-Verlag, New York, 1996.

[7] D. Freedman et al. **Statistical models and causal inference : a dialogue with the social sciences**. Cambridge University Press, 2009.

[8] T. Hastie, R. Tibshirani et J. Friedman. **The Elements of Statistical Learning**. Springer Series in Statistics. New York : Springer Verlag, 2001.

[9] R. H. Keogh et D. R. Cox. **Case-control studies**. T. 4. Cambridge University Press, 2014.

[10] V. Koltchinskii. **Oracle inequalities in empirical risk minimization and sparse recovery problems. Lectures from the 38th Probability Summer School, Saint-Flour**. T. 2033. Lecture Notes in Mathematics. Springer, 2008.

[11] P. Massart. **Concentration inequalities and model selection**. T. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Berlin : Springer, 2007, p. xiv+337. ISBN : 978-3-540-48497-4 ; 3-540-48497-3. MR : 2319879(2010a:62008).

[12] K. P. Murphy. **Machine learning - a probabilistic perspective**. Adaptive computation and machine learning series. MIT Press, 2012. ISBN : 0262018020.

[13] S. Shalev-Shwartz et S. Ben-David. **Understanding machine learning. From theory to algorithms.** English. Cambridge : Cambridge University Press, 2014, p. xvi + 397. ISBN : 978-1-107-05713-5/hbk ; 978-1-107-29801-9/ebook.

[14] L. Valiant. **Probably Approximately Correct : NatureÕs Algorithms for Learning and Prospering in a Complex World**. Basic Books (AZ), 2013.

[15] V. Vapnik. **Statistical Learning Theory**. New York : John Wiley, 1998.

[16] P. Bartlett, M. Jordan et J. McAuliffe. « Convexity, classification, and risk bounds. » Berkeley. 2003.

10.1 LE PROGRAMME DE FISHER

Les modèles réguliers sont les modèles qui, sans être nécessairement des modèles exponentiels, possèdent des propriétés mises en évidence pour les modèles exponentiels :

i) ce sont des modèles paramétriques où l'espace des paramètres Θ et un ouvert de $\mathbb{R}^d$;
ii) ils sont identifiables ;
iii) ils sont dominés ;
iv) le maximum de vraisemblance est défini (mais pas nécessairement unique) avec une probabilité qui tend vers 1 lorsque la taille de l'échantillon tend vers l'infini ;
v) une fonction score peut être définie sur Θ (dans les modèles exponentiels c'est le gradient de la log vraisemblance) , sous P_θ, la fonction score en θ est centrée et sa matrice de covariance appelée information de Fisher. Celle-ci est inversible.

La première partie du programme de Fisher peut se résumer (caricaturer) de la façon suivante :

i) sous P_θ, la différence $\left(1/\sqrt{n}I(\theta)^{-1}\dot{\ell}_n(\theta) - \sqrt{n}(\widehat{\theta}_n - \theta)\right)$ converge en probabilité vers 0 ;
ii) sous P_θ, $\sqrt{n}\big(\widehat{\theta}_n - \theta\big) \rightsquigarrow \mathcal{N}\big(0, I(\theta)^{-1}\big)$.

La définition des modèles réguliers a muri pendant les décennies qui ont suivi la seconde guerre mondiale. La théorie de Le Cam a légitimé cette partie du programme de Fisher. Elle a construit une notion de modèle régulier capable de rendre compte de tous les modèles possédant ces bonnes propriétés des modèles exponentiels et montré que les propriétés mentionnées plus haut sont vérifiées. Cette conception des modèles réguliers s'appuie sur la définition suivante.

DÉFINITION 10.1 (DIFFÉRENTIABILITÉ EN MOYENNE QUADRATIQUE) Une modèle dominé $(P_\theta)_{\theta\in\Theta}$ (Θ ouvert de $\mathbb{R}^d$), avec fonction de vraisemblance $p(\theta, x), \theta \in \Theta, x \in \mathcal{X}$ par rapport à la mesure dominante ν est dit *différentiable en moyenne quadratique* si il existe un fonction $\dot{p} : \Theta \times \mathcal{X} \to \mathbb{R}^d$ (avec $\|\dot{p}(\theta, \cdot)\| \in L_2(P_\theta)$ pour tout $\theta \in \Theta$) telle que

$$\int_{\mathcal{X}} \left(\sqrt{p(\theta', x)} - \sqrt{p(\theta, x)} - \frac{\sqrt{p(\theta, x)}}{2}(\theta' - \theta)^t \dot{p}(\theta, x) \right)^2 \mathrm{d}\nu(x) = o\left(\|\theta - \theta'\|^2\right) .$$

au voisinage de θ. La fonction $\dot{p}(\theta, \cdot)$ est appelée fonction score. L'information de Fisher du modèle en θ est

$$\mathbb{E}_\theta \left[\dot{p}(\theta, X)\dot{p}(\theta, X)^t\right] .$$

On peut vérifier que les modèles exponentiels en forme canonique sont différentiables en moyenne quadratique et que la fonction score habituelle (gradient de la log-vraisemblance) est compatible avec la définition qui précède. Cette définition et la théorie développée autour permettent de traiter des situations intéressantes.

EXEMPLE 10.2 Le modèle défini par les translations de la loi exponentielle bilatère (les densité sur $\mathbb{R}$ de la forme $f_\theta(x) := \frac{1}{2}\exp(-|x - \theta|)$ pour $\theta \in \mathbb{R}$) fournit un exemple de modèle régulier sans être exponentiel. Dans un échantillon, toute médiane empirique maximise la vraisemblance. L'information de Fisher est constante, et quelque soit le choix de la médiane empirique comme maximisant de la vraisemblance, $\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0, 1)$.

Fisher a aussi formulé la conjecture suivante : si $(T_n)_n$ est une suite d'estimateurs asymptotiquement normaux tel que pour tout $\theta \in \Theta$, $\sqrt{n}(T_n - \theta) \rightsquigarrow \mathcal{N}(0, v(\theta))$, où $v(\cdot)$ est une fonction de Θ dans l'ensemble des matrices semi-définies positives, alors pour tout θ, $v(\theta) \succeq I(\theta)^{-1}$. Cette conjecture revient à postuler que dans ces modèles réguliers, la méthode du maximum de vraisemblance est, asymptotiquement au moins, uniformément la meilleure.

Cette exigence d'optimalité uniforme s'est avérée intenable. Dès 1950, Hodges a construit un contre exemple.

Exemple 10.3 On se place dans une modèle unidimensionnel $\Theta \subseteq \mathbb{R}$, et on dispose d'une suite d'estimateurs $(T_n)_n$ de θ telle que pour tout $\theta \in \Theta$, sous $P_\theta^{\mathbb{N}}$

$$\sqrt{n}(T_n - \theta) \rightsquigarrow \mathcal{N}\left(0, \sigma^2(\theta)\right)$$

et aussi qu'il existe $\theta_0, \delta > 0$ tels que

$$\sup_n \mathbb{E}_{\theta_0}\left[\left(\sqrt{n}(T_n - \theta_0)\right)^{2+\delta}\right] \leq \infty\,.$$

Hodge étudie alors la suite d'estimateurs $(\widetilde{T}_n)_n$ définie par

$$\widetilde{T}_n := \begin{cases} T_n, & \text{si } |T_n - \theta_0| > n^{-1/4} \\ 0 & \text{sinon.} \end{cases}$$

On vérifie alors que pour $\theta \neq \theta_0$, sous $P_\theta^{\mathbb{N}}$

$$\sqrt{n}\left(\widetilde{T}_n - \theta\right) \rightsquigarrow \mathcal{N}(0, \sigma^2(\theta))$$

alors que sous $P_{\theta_0}^{\mathbb{N}}$,

$$\sqrt{n}\left(\widetilde{T}_n - \theta_0\right) \rightsquigarrow \delta_0\,.$$

Autrement dit, $(\widetilde{T}_n)_n$ est au moins aussi bonne que T_n partout et bien meilleure en θ_0.

Cette construction peut être appliquée lorsque T_n est un estimateur au maximum de vraisemblance. On construit ainsi un estimateur *super-efficace* en θ_0. Elle conduit à réfuter la seconde conjecture de Fisher. Cette construction peut sembler pathologique à plusieurs titres :

i) Elle n'améliore l'estimateur initial qu'en un point ;
ii) Pour une valeur fixée de n, au voisinage de θ_0, le risque de $\widetilde{T}_n$ est supérieur à celui de T_n. L'amélioration uniforme n'est qu'asymptotique.

Nous allons voir dans la section suivante qu'on peut néanmoins travailler le contre-exemple de Hodge pour développer une idée féconde : les estimateurs « par contraction » qui s'avéreront meilleurs que les estimateurs au maximum de vraisemblance dans un sens non-asymptotique.

10.2 le phénomène de stein

On se place dans le modèle de la localisation Gaussienne : $Y \sim \mathcal{N}(\theta, \sigma^2 \mathrm{Id})$ avec $\theta \in \mathbb{R}^d$ le paramètre à estimer (σ^2 est supposé connu). L'information de Fisher est constante $I(\theta) = \frac{1}{\sigma^2}\mathrm{Id}$. On s'intéresse au cas où $d \geq 3$.

On définit l'estimateur par contraction (appelé estimateur de James-Stein) :

$$\widehat{\theta} := \left(1 - \sigma^2 \frac{d-2}{\|Y\|^2}\right) Y\,.$$

Comme $\|Y\|^2$ tend à être plus grand que $\sigma^2(d-2)$, cet estimateur tend à contracter l'estimateur au maximum de vraisemblance qui n'est autre que Y.

Théorème 10.4 *Pour tout $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}_\theta\left[\|\widehat{\theta} - \theta\|^2\right] \leq \mathbb{E}_\theta\left[\|Y - \theta\|^2\right] = d\sigma^2\,.$$

Pour $\theta = 0$,

$$\mathbb{E}_\theta\left[\|\widehat{\theta} - \theta\|^2\right] = 2\sigma^2\,.$$

Pour $\theta \neq 0$,

$$\lim_{\sigma \to 0} \frac{\mathbb{E}_\theta\left[\|\widehat{\theta} - \theta\|^2\right]}{\mathbb{E}_\theta\left[\|Y - \theta\|^2\right]} = 1\,.$$

Les deux ingrédients de la preuve de ce théorème sont des résultats gaussiens. Le premier est l'identité de Stein pour les gaussiennes (ce type d'identité peut être développé pour les modèles exponentiels en général).

On rappelle qu'une fonction $g : \mathbb{R} \to \mathbb{R}$ est absolument continue sur $[a, b]$ si et seulement si il existe une fonction g' intégrable de $[a, b]$ dans $\mathbb{R}$ telle que $g(x) - g(a) = \int_{[a,x]} g'(y)\mathrm{d}y$ (la fonction de répartition d'une loi absolument continue par rapport à la mesure de Lebesgue est une fonction absolument continue).

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$, est dite presque partout absolument continue en chaque coordonnée si et seulement si pour tout i, pour presque tout $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, la fonction

$$x \mapsto f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$$

est absolument continue. On note en abrégé $\partial_i f$ une fonction qui vérifie

$$\begin{aligned} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) - & f(x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n) \\ & = \int_a^x \partial_i f(x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_n)\mathrm{d}u \end{aligned}$$

(elle est implicitement paramétrée par $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$).

En dimension 1, l'identité de Stein caractérise les gaussiennes, voir lemmes 2.1 et 2.3. La proposition suivante généralise l'identité de Stein aux vecteurs gaussiens à matrice de covariance diagonale.

PROPOSITION 10.5 (IDENTITÉ DE STEIN) *Soit $Y \sim \mathcal{N}(\theta, \sigma^2 \mathrm{Id}_d)$ avec $\theta \in \mathbb{R}^d$. Soit $f : \mathbb{R}^d \to \mathbb{R}$, presque partout absolument continue en chaque coordonnée, vérifiant pour chaque $i \leq d$*

$$\mathbb{E}\left[|\partial_i f|\right] := \mathbb{E}\left[|\partial_i f(Y)|\right] < \infty\,.$$

Alors pour chaque $i \leq d$,

$$\mathbb{E}\left[(\theta_i - Y_i) f(Y)\right] = -\sigma^2 \mathbb{E}\left[\partial_i f\right]\,.$$

PREUVE. L'identité de Stein univariée (voir Lemmes 2.1 et 2.3) s'écrit

$$\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)] \qquad f \text{ absolument continue et } X \sim \mathcal{N}(0,1)\,.$$

Si $Y \sim \mathcal{N}(\theta, \sigma^2)$, $Y \sim \theta + \sigma X$, avec $X \sim \mathcal{N}(\theta, \sigma^2)$,

$$\begin{aligned} \mathbb{E}[(Y-\theta)f(Y)] &= \mathbb{E}[\sigma X f(\sigma X + \theta)] \\ &= \sigma \mathbb{E}[\sigma f'(\sigma X + \theta)] \\ &= \sigma^2 \mathbb{E}[f'(Y)]\,. \end{aligned}$$

Soit $i \leq d$, et f presque partout absolument continue en chaque coordonnée. Sans perdre en généralité, on suppose que $f(\theta) = 0$. On note $\mathcal{F}^{(i)}$ la tribu engendrée par $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$:

$$\mathbb{E}\left[(\theta_i - Y_i) f(Y)\right] = \mathbb{E}\left[\mathbb{E}\left[(\theta_i - Y_i) f(Y) \mid \mathcal{F}^{(i)}\right]\right]\,.$$

L'expression

$$\mathbb{E}\left[(\theta_i - Y_i) f(Y) \mid \mathcal{F}^{(i)}\right]$$

peut se calculer en intégrant selon la $i^{\text{ème}}$ par

$$\int (\theta_i - z) f(y_1, \ldots, y_{i-1}, z, y_{i+1}, \ldots, y_d) \phi\left(\frac{z - \theta_i}{\sigma}\right) \mathrm{d}z\,.$$

En invoquant l'identité de Stein univariée, et l'absolue continuité presque partout en chaque coordonnée de f pour presque tout $y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_d$, on obtient (p.p.)

$$\begin{aligned}\mathbb{E}\left[(\theta_i - Y_i) f(Y) \mid \mathcal{F}^{(i)}\right] &= \int (\theta_i - z) f(y_1, \dots, y_{i-1}, z, y_{i+1}, \dots, y_d) \phi\left(\frac{z-\theta_i}{\sigma}\right) \mathrm{d}z \\ &= -\sigma^2 \int \partial_i f(y_1, \dots, y_{i-1}, z, y_{i+1}, \dots, y_d) \phi\left(\frac{z-\theta_i}{\sigma}\right) \mathrm{d}z \\ &= -\sigma^2 \mathbb{E}[\partial_i f \mid \mathcal{F}^{(i)}] \,.\end{aligned}$$

On obtient le résultat désiré en prenant l'espérance. □

Le second ingrédient repose sur les propriétés des normes de vecteurs gaussiens.

PROPOSITION 10.6 (NORMES DE VECTEURS GAUSSIENS) *Soit $Y \sim \mathcal{N}(\theta, \sigma^2 \mathrm{Id}_d)$ avec $\theta \in \mathbb{R}^d$, alors si $d \geq 3$,*

$$\mathbb{E}\left[\frac{1}{\|Y\|^2}\right] \leq \frac{1}{\sigma^2(d-2)} \,.$$

PREUVE. Dans un premier temps, on se souvient que l'on peut construire un espace probabilisé avec deux vecteurs gaussiens $Y \sim \mathcal{N}(\theta, \sigma^2 \mathrm{Id}_d)$ et $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_d)$ tel que $\|Z\|^2 \leq \|Y\|^2$. C'est une manifestation du fait qu'une loi du χ^2 centrée est toujours stochastiquement dominée par une loi du χ^2 décentrée (voir Théorème 2.36).

Il suffit donc de vérifier l'inégalité dans le cas où $\theta = 0$. Mais on sait alors que $\|Y\|^2$ est distribuée selon une loi Gamma de paramètre de forme $d/2$ et de paramètre d'échelle 2. On peut calculer explicitement l'espérance. Pour $d \geq 3$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\|Y\|^2}\right] &= \int_0^\infty \frac{1}{2x} \frac{(x/2)^{d/2-1}}{\Gamma(d/2)} \mathrm{e}^{-x/2} \mathrm{d}x \\ &= \frac{\Gamma((d-2)/2)}{2\Gamma(d/2)} \int_0^\infty \frac{(x/2)^{(d-2)/2-1}}{2\Gamma((d-2)/2)} \mathrm{e}^{-x/2} \mathrm{d}x \\ &= \frac{\Gamma((d-2)/2)}{2\Gamma(d/2)} \\ &= \frac{1}{d-2} \,.\end{aligned}$$

□

PREUVE.(théorème 10.4) On s'intéresse à des estimateurs de la forme $\widehat{\theta} := g(Y)Y$ où $g : \mathbb{R}^d \to \mathbb{R}$. On attend de la fonction g qu'elle contracte l'estimateur au maximum de vraisemblance Y. On part du développement

$$\mathbb{E}\left[\|\widehat{\theta} - \theta\|^2\right] = \underbrace{\mathbb{E}\left[\|Y-\theta\|^2\right]}_{(a)} + 2\underbrace{\mathbb{E}\left[\langle \theta - Y, (1-g(Y))Y\rangle\right]}_{(b)} + \underbrace{\mathbb{E}\left[(1-g(Y))^2 \|Y\|^2\right]}_{(c)} \,.$$

On a immédiatement

$$(a) = \sigma^2 d \,.$$

Pour traiter (b), on utilise l'identité de Stein, pour chaque $i \leq d$,

$$\mathbb{E}\left[(\theta_i - Y_i)(1-g(Y))Y_i\right] = -\sigma^2 \mathbb{E}\left[(1-g(Y)) - Y_i \partial_i g\right] ,$$

en sommant sur i,

$$(b) = -2\sigma^2 \mathbb{E}\left[d(1-g(Y)) - \langle Y, \nabla g\rangle\right] \,.$$

Soit

$$(b) + (c) = \mathbb{E}\left[-2d\sigma^2(1-g(Y)) + 2\sigma^2\langle Y, \nabla g\rangle + \|Y\|^2(1-g(Y))^2\right] \,.$$

Nous allons maintenant chercher à minimiser $(b)+(c)$ en nous concentrant sur des fonctions g de la forme $g(y) = 1 - \kappa/\|y\|^2$, soit

$$\partial_i g(y) = \frac{2\kappa y_i}{\|y\|^4}\,.$$

$$(b)+(c) = \mathbb{E}\left[-\frac{2d\kappa\sigma^2}{\|Y\|^2} + 2\frac{2\kappa\sigma^2}{\|Y\|^2} + \frac{\kappa^2}{\|Y\|^2}\right] = (\kappa^2 - 2\kappa(d-2)\sigma^2)\mathbb{E}\left[\frac{1}{\|Y\|^2}\right]\,.$$

Cette expression est minimisée en choisissant $\kappa = (d-2)\sigma^2$, elle vaut alors

$$(b)+(c) = -\mathbb{E}\left[\frac{\sigma^4(d-2)^2}{\|Y\|^2}\right]$$

Cette quantité est toujours négative ou nulle. La première inégalité du théorème est donc établie.

Lorsque $\theta = 0$, on peut invoquer la proposition 10.6 pour établir

$$(b)+(c) = -\sigma^2(d-2)\,,$$

qui permet de vérifier la deuxième assertion du théorème.

Pour conclure, lorsque $\theta \neq 0$, $\|Y\|^2 \sim (\|\theta\| + \sigma Z_1)^2 + \sum_{i=2}^d \sigma^2 Z_i^2$ où $Z_1, \ldots, Z_d$ sont i.i.d. $\mathcal{N}(0,1)$. Lorsque $\sigma \to 0$, $\frac{\sigma^2(d-2)^2}{\|Y\|^2}$ converge presque sûrement vers 0. En exploitant la domination stochastique, on peut construire une variable aléatoire $Z' \sim \mathcal{N}(0,1)$ qui vit sur le même espace probabilisé que $Z_1, \ldots, Z_d$ telle que

$$\frac{\sigma^2(d-2)^2}{(\|\theta\| + \sigma Z_1)^2 + \sum_{i=1}^d \sigma^2 Z_i^2} \leq \frac{(d-2)^2}{(Z_1')^2 + \sum_{i=1}^{d-1} Z_i^2}$$

et invoquer le théorème de convergence dominée (le membre droit est intégrable) pour conclure

$$\lim_{\sigma\to 0} \mathbb{E}\left[\frac{\sigma^2(d-2)^2}{\|Y\|^2}\right] < \infty$$

ce qui suffit à établir la dernière affirmation du théorème. □

10.3 PERTES, RISQUES, RISQUE MINIMAX

Plus encore que le contre-exemple de Hodges, le phénomène de Stein montre ce qui ne va pas dans la seconde partie du programme de Fisher. C'est la volonté de montrer qu'une méthode est meilleure en tout point de Θ que n'importe quelle autre. Pour éviter cet écueil, la théorie de la décision propose deux perspectives qu'on oppose parfois, mais qui se complètent très bien.

i) Perspective *minimax* : le statisticien cherche une assurance contre les situations les plus défavorables.
ii) Perspective *bayésienne* : le statisticien suppose que la loi qui engendre les observations est elle-même choisie au hasard selon une loi appelée *a priori*. Le statisticien vise alors à optimiser la *performance moyenne* de ses estimateurs, régions de confiance et tests.

L'opposition entre ces deux points de vue occulte le fait que les constructions bayésiennes sont très utiles pour établir des bornes inférieures même en perspective minimax (ou fréquentiste) ; les estimateurs bayésiens peuvent être étudiés selon une perspective fréquentiste.

Pour poser les problèmes, il est utile d'introduire les notions de *fonction de pertes* et de *risque*. Nous nous plaçons dans le contexte d'une expérience statistique $(\Omega, \mathcal{F}, \{P_\theta : \theta \in \Theta\}, \mathcal{X}, \mathcal{G}, X)$. Généralement, nous considérerons des expériences échantillonnées, éventuellement une seule fois. Nous nous posons le problème d'estimer une fonction de θ, $g(\theta)$ où $g : \Theta \to E$. Dans la suite T désigne un estimateur de $g(\theta)$, c'est à dire une fonction des observations $X_1, \ldots, X_n$ dans E. Pour évaluer les performances de T, nous utilisons une *fonction de perte* L, c'est à dire une fonction de $E \times E$ dans $\mathbb{R}_+$ qui vérifie $L(z,z) = 0, \forall z \in E$, et nous définissons le *risque de T en θ* par

$$R(T,\theta) := \mathbb{E}_\theta[L(T(X), g(\theta))]$$

EXEMPLE 10.7 Lorsque $\Theta \subseteq \mathbb{R}^d$, $g(\theta) = \theta$, la perte L_2 est $L(z, z') := \|z - z'\|^2$, et le risque s'appelle lui aussi risque L_2 ou risque quadratique. On peut aussi s'intéresser à la perte absolue, $L(z, z') := \|z - z'\|_1 = \sum_{i=1}^d |z(i) - z'(i)|$.

En apprentissage/classification (voir chapitre 9), la situation est plus complexe. Les observations prennent leur valeur dans $\mathcal{Z} \times \mathcal{Y}$ avec ($\mathcal{Y} := \{-1, 1\}$). On dispose d'une collection de *classifieurs*, c'est à dire de fonctions de $\mathcal{Z} \to \mathcal{Y}$, l'erreur de classification de f en (z, y) est $\frac{1}{2}|f(z) - y|$ que l'on peut aussi écrire $(1 - f(z)y)/2$ (car $f(z), y \in \{-1, 1\}$). Le risque (ou erreur) du classifieur f est

$$\mathbb{E}\left[\frac{1}{2}|f(Z) - Y|\right].$$

Si f est choisie dans une collection de classifieurs à partir d'un échantillon de données (l'ensemble d'apprentissage) $((Z_1, Y_1), \ldots, (Z_n, Y_n))$, la qualité du classifieur choisi $\widehat{f}_n$ est évaluée par son erreur. Le risque d'une méthode d'apprentissage sous une loi jointe sur $\mathcal{Z} \times \mathcal{Y}$:

$$\mathbb{E}\left[\mathbb{E}\left[\frac{1}{2}|\widehat{f}_n(Z) - Y|\right]\right]$$

où l'espérance interne porte sur (Z, Y) et l'espérance externe sur $((Z_1, Y_1), \ldots, (Z_n, Y_n))$. Notons qu'en apprentissage/classification, on ne cherche pas à estimer un paramètre, mais à effectuer la moins mauvaise classification.

Le *risque maximal d'un estimateur* T contre Θ est défini par

$$R(T, \Theta) := \sup_{\theta \in \Theta} R(T, \theta)$$

Le *risque minimax* sur Θ $\overline{R}(\Theta)$ est défini par

$$\overline{R}(\Theta) := \inf_T R(T, \Theta) = \inf_T \sup_{\theta \in \Theta} R(T, \theta).$$

L'existence d'estimateurs T réalisant le risque minimax ne va pas de soi. La construction d'estimateurs minimax lorsqu'il en existe n'est pas non plus immédiate. Enfin la vérification du caractère minimax d'un estimateur n'est pas toujours une tâche aisée.

Par exemple, le fait que l'estimateur au maximum de vraisemblance ne soit pas *admissible* dans le modèle de localisation gaussienne ne préjuge pas du fait qu'il ne soit pas minimax. Nous n'avons pas déterminé le risque minimax dans ce modèle.

10.4 RISQUE BAYÉSIEN

On suppose Θ muni d'une tribu $\mathcal{A}$ qui en fait un espace probabilisable. Cet espace probabilisable est muni d'une mesure que nous appellerons *loi a priori* lorsque la mesure est une loi de probabilité. Nous noterons cette mesure Π. Lorsque Θ est un ouvert de $\mathbb{R}^d$ et que la mesure a priori est absolument continue par rapport à la mesure de Lebesgue, nous noterons sa densité π.

On suppose dans la suite que pour tout événement B de la tribu $\mathcal{F}$ sur $\mathcal{X}$, $\theta \mapsto P_\theta(B)$ est $\mathcal{A}$-mesurable. Techniquement, la famille (P_θ) définit un *noyau de probabilité* de $(\Theta, \mathcal{A})$ vers $(\mathcal{X}, \mathcal{F})$ (et plus généralement de $(\Theta, \mathcal{A})$ vers $\mathcal{X}^n$ muni de la tribu produit).

La donnée de la loi a priori Π, des lois $(P_\theta), \theta \in \Theta$ (et des lois produit) définit une loi jointe sur $\Theta \times \mathcal{X}^n$. La loi marginale de $X_1, \ldots, X_n$ sera appelée *loi de mélange* définie par l'a priori Π et notée P :

$$P(A) := \int_\Theta \int_{\mathcal{X}^n} \mathbb{I}_{(x_1, \ldots, x_n) \in A} \mathrm{d}P_\theta(x_1) \ldots \mathrm{d}P_\theta(x_n) \mathrm{d}\Pi(\theta).$$

Comme Θ est un ouvert de $\mathbb{R}^d$ (espace métrique complet séparable) et comme $\mathcal{X}$ est aussi supposé être un ouvert d'un espace $\mathbb{R}^k$, l'existence de probabilités conditionnelles ne pose pas de problèmes.

La loi conditionnelle de θ sachant $X_1 = x_1, \ldots, X_n = x_n$ est appelée *loi a posteriori* $\Pi(\cdot \mid x_1, \ldots, x_n)$.

Pour toute statistique T, pour tout $x \in \mathcal{X}$, la fonction de $\Theta \to \mathbb{R}$ qui associe à une statistique $T(x)$, sa perte en x contre $\theta : \theta \mapsto L(T(x), \theta)$, est $\mathcal{A}$-mesurable. On peut alors définir le risque intégré contre l'a priori Π

$$\int_\Theta R(T, \theta) \Pi(\mathrm{d}\theta)$$

et lorsque Π est une loi de probabilité

$$\int_\Theta R(T, \theta) \Pi(\mathrm{d}\theta) = \mathbb{E}_\Pi \left[\mathbb{E}_{P_\theta}\left[L(T, \theta)\right]\right].$$

Le *risque maximin* est défini par

$$\underline{R}(\Theta) := \sup_{\Pi} \inf_{T} \int_{\Theta} R(T,\theta)\Pi(\mathrm{d}\theta)$$

où le supremum à gauche est pris sur les lois de probabilité définies sur Π.

PROPOSITION 10.8 *Le* risque maximin $\underline{R}(\Theta)$ *est toujours inférieur ou égal au risque minimax :*

$$\underline{R}(\Theta) = \sup_{\Pi} \inf_{T} \int_{\Theta} R(T,\theta)\Pi(\mathrm{d}\theta) \leq \inf_{T} \sup_{\theta\in\Theta} R(T,\theta) = \overline{R}(\Theta)\,.$$

PREUVE. Pour une loi a priori Π et un estimateur T quelconques

$$\int_{\Theta} R(T,\theta)\Pi(\mathrm{d}\theta) \leq \sup_{\theta\in\Theta} R(T,\theta)\,,$$

donc, pour toute loi a priori Π, le risque bayésien est inférieur ou égal au risque minimax :

$$\inf_{T} \int_{\Theta} R(T,\theta)\Pi(\mathrm{d}\theta) \leq \inf_{T} \sup_{\theta\in\Theta} R(T,\theta)\,.$$

Si maintenant, on maximise le risque bayésien, celui ci reste inférieur au risque minimax. □

Pour chaque problème de statistique, on peut se demander s'il y a égalité entre risque maximin et risque minimax.

10.5 LIENS ENTRE ESTIMATEURS BAYÉSIENS ET MINIMAX

THÉORÈME 10.9 *Soit T un estimateur de $g(\theta)$ tel que $\sup_{\theta\in\Theta} R(T,\theta) = r < \infty$. Soit $\Theta' \subseteq \Theta$ vérifiant*

$$\forall\theta \in \Theta', \qquad R(T,\theta) = r\,.$$

Si Θ' est non-vide et si T est bayésien pour une loi a priori Π telle que $\Pi\{\Theta'\} = 1$, alors T est minimax et cette loi a priori est maximin.

PREUVE. Supposons l'existence d'un estimateur T' tel que

$$\sup_{\theta\in\Theta} R(T',\theta) = r' < r\,.$$

Alors

$$\int_{\Theta} R(T',\theta)\mathrm{d}\Pi(\theta) \leq r' < r = \int_{\Theta} R(T,\theta)\mathrm{d}\Pi(\theta)\,.$$

Ceci contredit l'hypothèse du caractère bayésien de l'estimateur T relativement à l'a priori Π. On a donc

$$\forall T', \qquad r \leq \sup_{\theta\in\Theta'} R(T',\theta) \leq \sup_{\theta\in\Theta} R(T',\theta)\,.$$

La quantité r est donc le risque minimax qui coïncide avec le risque bayésien sous l'a priori Π. La loi a priori Π est donc maximin. □

Si un estimateur est de risque constant sur Θ et s'il est bayésien pour une loi a priori alors il est minimax.

On peut se demander si tout estimateur minimax est aussi un estimateur bayésien.

THÉORÈME 10.10 *Soit T, tel que $r := \sup_{\theta\in\Theta} R(T,\theta) < \infty$ est atteint sur Θ. S'il existe une suite $(\Pi_k)_{k\in\mathbb{N}}$ de lois a priori à support inclus dans Θ telle que*

$$\lim_{k\to\infty} \inf_{T'} \int_\Theta R(T',\theta)\mathrm{d}\Pi_k(\theta) = r$$

alors T est un estimateur minimax.

PREUVE. Pour tout k, on note

$$r_k := \inf_{T'} \int_\Theta R(T',\theta)\mathrm{d}\Pi_k(\theta)\,,$$

et $\underline{r}$ le risque maximin. Par définition $\limsup_k r_k \leq \underline{r}$. Et $\underline{r}$ est inférieure ou égale au risque minimax, qui est inférieur ou égal à r. Si $\limsup_k r_k = r$, le risque maximin et le risque minimax coïncident. Et ils sont égaux à r. □

On peut se demander si un estimateur, par exemple un estimateur au maximum de vraisemblance, peut être Bayésien pour une loi a priori à déterminer.

10.6 RAPPELS SUR LES LOIS CONDITIONNELLES

Le conditionnement joue un role essentiel en statistique bayésienne. En effet, toutes les constructions bayésiennes sont définies à partir de la loi conditionnelle du paramètre sachant les observations. Cette loi est appelée *loi a psoteriori*. Pour travailler, nous aurons besoin des notions suivantes.

DÉFINITION 10.11 (PROBABILITÉ CONDITIONNELLE RÉGULIÈRE) Une fonction $P(\cdot \mid \mathcal{G})(\cdot)$ de $\mathcal{F}\times\Omega$ dans $[0,1]$ est appelée probabilité conditionnelle régulière sachant $\mathcal{G}$ si et seulement si

i) pour tout $B\in\mathcal{F}$, $P(B \mid \mathcal{G})(\omega) = \mathbb{E}[\mathbb{I}_B \mid \mathcal{G}](\omega)$ P-p.s.
ii) $P(\cdot \mid \mathcal{G})(\omega)$ est P-p.s. une loi de probabilité sur $(\Omega,\mathcal{F})$.

En statistique bayésiennne, la notion dont nous avons le plus besoin est celle de probabilité conditionnelle d'une variable aléatoire Y sachant une autre X. Elle peut être construite à partir de la définition 10.11.

DÉFINITION 10.12 (PROBABILITÉ CONDITIONNELLE DE Y SACHANT X) Soit $(\Omega,\mathcal{F},P)$ un espace probabilisé sur lequel vivent deux variables aléatoires X,Y. Soit $\mathcal{G}$ la sous-tribu engendrée par X, soit $(\mathcal{Y},\mathcal{H})$ l'espace probabilisable image de $(\Omega,\mathcal{F})$ par Y. Une fonction $P(\cdot \mid X)(\cdot)$ de $\mathcal{H}\times X(\Omega)$ définie par

$$P(H \mid X)(X(\omega)) = P(Y^{-1}(H) \mid \mathcal{G})(\omega) \qquad \text{pour } H\in\mathcal{H}, \omega\in\Omega\,,$$

est une probabilité conditionnelle de Y sachant X. On notera $P(\cdot \mid x)$ la fonction $H \mapsto P(H \mid X)(x)$ qui est presque sûrement une loi de probabilité sur $(\mathcal{Y},\mathcal{H})$.

Les probabilités conditionnelles permettent de calculer des espérances et des espérances conditionnelles par intégration itérée.

PROPOSITION 10.13 *Si (X,Y) sont des variables aléatoires sur $(\Omega,\mathcal{F},P)$ à valeurs dans $\mathcal{X},\mathcal{Y}$ et f une fonction de $\mathcal{X}\times\mathcal{Y}$ dans $\mathbb{R}$. S'il existe une probabilité conditionnelle régulière de Y sachant X, alors*

$$\mathbb{E}\left[f(X,Y) \mid X\right](\omega) = \int_{\mathcal{Y}} f(x,y)P(\mathrm{d}y \mid x) \qquad \textit{pour } x = X(\omega)$$

et

$$\mathbb{E}[f(X,Y)] = \int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)P(\mathrm{d}y \mid x)P_X(\mathrm{d}x)$$

où $P_X = P\circ X^{-1}$.

Comme nous considérons toujours $\Theta \subseteq \mathbb{R}^d$ et $\mathcal{X} \subseteq \mathbb{R}^k$, l'existence de probabilités conditionnelles régulières est garantie par le théorème B.64.

10.7 CONSTRUCTION D'ESTIMATEURS BAYÉSIENS

La construction des estimateurs bayésiens repose sur la notion de loi a posteriori.

La loi a priori Π sur $(\Theta, \mathcal{A})$ et la famille de lois $(P_\theta)_{\theta\in\Theta}$ sur $(\mathcal{X}, \mathcal{F})$ définissent une loi jointe $\mathbb{P}$ sur l'espace produit $(\Theta \times \mathcal{X}, \sigma(\mathcal{A} \times \mathcal{F}))$ si pour tout $E \in \mathcal{F}$,

$$\theta \mapsto P_\theta(E)$$

est $\mathcal{A}$-mesurable. Cette loi jointe $\mathbb{P}$ est complètement définie par :

$$\mathbb{P}(A \times E) = \int_\Theta \mathbb{I}_A(\theta) \times P_\theta(E)\, \Pi(\mathrm{d}\theta)$$

pour $A \in \mathcal{A}, E \in \mathcal{F}$.

La loi jointe définit une loi marginale $\mathbb{Q}$ sur $(\mathcal{X}, \mathcal{F})$:

$$\mathbb{Q}(E) = \int_\Theta P_\theta(E)\, \Pi(\mathrm{d}\theta)\,.$$

Toute famille de lois $(P(\cdot \mid x))_{x\in\mathcal{X}}$ sur $(\Theta, \mathcal{A})$ telle que pour tout $A \in \mathcal{A}$, $x \mapsto P(A \mid x)$ est $\mathcal{F}$ mesurable et

$$\mathbb{P}(A \times E) = \int_\mathcal{X} \mathbb{I}_E(x) \times P(A \mid x)\, \mathbb{Q}(\mathrm{d}x)$$

pour tous $A \in \mathcal{A}, E \in \mathcal{F}$, est appelée famille de *lois a posteriori* pour l'a priori Π. La loi $P(\cdot \mid x)$ est a appelée loi a posteriori sachant x.

L'existence de lois a posteriori est garantie par des résultats généraux sur l'existence des probabilités conditionnelles régulières décrits plus hauts (Théorème B.64).

PROPOSITION 10.14 (EXISTENCE DE LA LOI A POSTERIORI) *Si Θ peut être muni d'une structure d'espace métrique complet séparable et si la loi a priori Π est définie sur les boréliens, alors il existe une loi a posteriori notée génériquement $\Pi(\cdot \mid X)$.*

Dans le cas (dont nous ne sortons pas) où le modèle $(P_\theta)_{\theta\in\Theta}$ est dominé et où l'a priori possède une densité par rapport à la mesure de Lebesgue, la loi a posteriori possède une densité par rapport à la mesure de Lebesgue.

La mesure dominante (notée ν) sera en général sous-entendue. Il pourra s'agir de la mesure de Lebesgue sur $\mathbb{R}^d \supseteq \Theta$.

PROPOSITION 10.15 (DENSITÉ A POSTERIORI) *Si la mesure a priori Π est une loi de probabilité de densité π par rapport à une mesure de référence, si le modèle est dominé par une mesure σ-additive ν, et admet une fonction de vraisemblance $p(x \mid \theta)$ alors la mesure a posteriori étant donné $X_1, \ldots, X_n$ est p.s. absolument continue par rapport à la loi a priori et sa densité en θ est proportionnelle à la vraisemblance en $\theta, X_1, \ldots, X_n$*

$$\pi\left(\theta \mid X_1, \ldots, X_n\right) \propto \pi(\theta) \times \prod_{i=1}^{n} p(X_i \mid \theta)\,.$$

PREUVE. On traite le cas d'une observation ($n = 1$).

On définit

$$p_X(x) = \int_\Theta p(x \mid \theta)\pi(\theta)\mathrm{d}\theta\,,$$

c'est une densité de la loi marginale de X par rapport à ν sous la loi de mélange définie par l'a priori Π. Son existence ne pose pas de problème (Théorème de Tonnelli-Fubini), mais son calcul peut être délicat.

L'existence d'une loi a posteriori est garantie par le Théorème B.64. Il suffit de vérifier que P_X presque sûrement, pour tout $A \in \mathcal{F}$

$$\mathbb{E}\left[\mathbb{I}_A \mid X\right](\omega) = \int_\Theta \mathbb{I}_A(\theta)\Pi(\mathrm{d}\theta \mid x) \qquad \text{avec } x = X(\omega)$$

est égal à

$$\int_\Theta \mathbb{I}_A(\omega)\frac{p(\theta \mid x)\pi(\theta)}{p_X(x)}\mathrm{d}\theta\,.$$

Sans perdre en généralité, on peut se restreindre à l'ensemble $\{x : p_X(x) > 0\}$.

Pour tout $A \subseteq \Theta$ mesurable,

$$\begin{aligned}\mathbb{E}\mathbb{I}_{\theta\in A} &= \int_\Theta \int_{\mathcal{X}} \mathbb{I}_{p_X(x)>0}\,\mathbb{I}_{\theta\in A}\,p(x \mid \theta)\pi(\theta)\,\nu(\mathrm{d}x)\mathrm{d}\theta \\ &= \int_{\mathcal{X}} p_X(x)\mathbb{I}_{p_X(x)>0}\left(\int_\Theta \mathbb{I}_{\theta\in A}\,\frac{p(x \mid \theta)\pi(\theta)}{P_X(x)}\,\mathrm{d}\theta\right)\mathbb{I}_{p_X(x)>0}\;\nu(\mathrm{d}x)\end{aligned}$$

ce qui établit que $\frac{\pi(\theta)p(x|\theta)}{p_X(x)}$ est une densité conditionnelle de la loi a posteriori de θ sachant l'observation $X = x$. □

La densité de la loi a posteriori est donc en général connue à une constante près. Dans des modèles sophistiqués, le calcul de cette constante de normalisation peut être non-trivial. C'est une des premières motivations du développement des statistiques computationelles et de la popularité des méthodes MCMC (Monte-Carlo Markov chains).

Lorsqu'on travaille sur des familles exponentielles en forme canonique, il est possible de choisir une loi a priori *conjuguée* au modèle. Le calcul des densités a posteriori s'en trouve grandement facilité.

EXEMPLE 10.16 (CONJUGAISON POISSON-GAMMA) $\Theta =]0, \infty)$, P_θ est la loi de Poisson d'espérance θ. La fonction de vraisemblance en l'observation k est

$$\mathrm{e}^{-\theta}\frac{\theta^k}{k!}$$

(la dominante est la mesure de comptage sur $\mathbb{N}$). La loi a priori est une loi Gamma de paramètres de forme $p > 0$ et d'intensité $\lambda > 0$ (l'intensité est l'inverse du paramètre d'échelle), de densité :

$$\mathbb{I}_{\theta>0}\lambda\frac{(\lambda\theta)^{p-1}}{\Gamma(p)}\mathrm{e}^{-\lambda\theta}\,.$$

La densité de la loi a posteriori en θ est proportionnelle au produit de la vraisemblance par la densité de la loi a priori, soit si l'observation est k

$$\propto \mathrm{e}^{-\theta}\frac{\theta^k}{k!} \times \mathbb{I}_{\theta>0}\lambda\frac{(\lambda\theta)^{p-1}}{\Gamma(p)}\mathrm{e}^{-\lambda\theta} \propto \mathbb{I}_{\theta>0}(\lambda+1)\frac{((\lambda+1)\theta)^{p+k-1}}{\Gamma(p+k)}\mathrm{e}^{-(\lambda+1)\theta}\,.$$

On reconnait la densité d'une loi Gamma de paramètres de forme $p+k$ et d'intensité $\lambda+1$, ce qui nous dispense du calcul de la constante de normalisation.

Si au lieu d'une seule observation, $X_1 = k$, on disposait d'un échantillon $X_1, \ldots, X_n$, en notant $S_n := \sum_{i=1}^n X_i$, on reconnaitrait que la loi a posteriori $\Pi(\cdot \mid X_1, \ldots, X_n)$ est la loi Gamma de paramètres de forme $p + S_n$ et d'intensité $\lambda + n$.

Remarquons que la loi a posteriori est caractérisée par la statistique suffisante S_n.

Adoptons maintenant le point de vue fréquentiste sur les méthodes bayésiennes. On fixe θ, on échantillonne selon P_θ et on étudie la suite des lois a posteriori. La loi des grands nombres indique que S_n/n tend presque sûrement vers θ. La suite des espérances des lois a posteriori de $\left(\frac{p+S_n}{\lambda+n}\right)_n$ converge donc presque sûrement vers θ elle aussi. La suites des variances des loi a posteriori $\left(\frac{p+S_n}{(\lambda+n)^2}\right)_n$ converge presque sûrement vers 0. Presque sûrement, la suite des lois a posteriori converge étroitement vers la masse de Dirac en θ. On observe donc ce qu'il faut appeler un résultat de consistance presque sûre de la suite des lois a posteriori.

L'étude fréquentiste systématique des suites (aléatoires) de lois a posteriori est un sujet très actif en statistique non-paramétrique (lorsqu'il n'est plus possible de représenter les lois qui forment le modèle par un paramètre fini-dimensionnel). Cette étude porte sur des suites de mesures aléatoires. Le choix du cadre (topologie placée sur l'espace des mesures, tribu) devient alors un problème non trivial.

PROPOSITION 10.17 *Si S_n est une statistique suffisante pour le modèle échantillonné $\{P_\theta^{\otimes n}, \theta \in \Theta\}$, alors la densité de la loi a posteriori est une fonction de S_n.*

PREUVE. Si S_n est une statistique suffisante alors pour tout θ, la densité produit $\prod_{i=1}^n p(x_i \mid \theta)$ se factorise en

$$\prod_{i=1}^n p(x_i \mid \theta) = q(s \mid \theta) \times h(x_1, \ldots, x_n, s) \qquad \text{pour } s = S_n(x_1, \ldots, x_n)$$

où $q(\cdot \mid \cdot)$ et h sont des fonctions mesurables. La densité de la loi a posteriori est donc proportionnelle à

$$\pi(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta) q(s \mid \theta) h(x_1, \ldots, x_n, s) \qquad \text{pour } s = S_n(x_1, \ldots, x_n)\,.$$

En tant que fonction de θ elle est proportionnelle à $\pi(\theta)q(s \mid \theta)$. □

La notion de loi a posteriori est centrale dans la construction d'estimateurs bayésiens.

PROPOSITION 10.18 *Un estimateur T est bayésien (réalise le risque bayésien sous l'a priori Π) si et seulement si sous la loi de mélange P-presque sûrement, il minimise la perte sous la loi a posteriori $\Pi(. \mid x_1, \ldots, x_n)$:*

$$T(x_1, \ldots, x_n) = \mathrm{argmin}_t \int_\Theta L(t, g(\theta'))\mathrm{d}\Pi(\theta' \mid x_1, \ldots, x_n)\,.$$

PREUVE. Soit T un estimateur, son risque moyen peut s'écrire comme l'espérance de la perte a posteriori sous la loi de mélange sur les échantillons définie par la loi a priori, c'est une illustration du théorème de Fubini.

$$\int_\Theta \mathbb{E}_\theta \left[L(T(X_1, \ldots, X_n), g(\theta))\right] \mathrm{d}\Pi(\theta)$$
$$= \int_{\mathcal{X}^n} \int_\Theta L(T(x_1, \ldots, x_n), g(\theta))\mathrm{d}\Pi(\theta \mid x_1, \ldots, x_n)\mathrm{d}P(x_1, \ldots, x_n)\,.$$

Si un estimateur minimise l'intégrande $\int_\Theta L(T(x_1, \ldots, x_n), g(\theta))\mathrm{d}\Pi(\theta \mid x_1, \ldots, x_n)$, il minimise le risque bayésien. □

Pour le risque quadratique, l'estimateur bayésien prend une forme particulièrement simple.

COROLLAIRE 10.19 *Si la fonction de perte est quadratique (carré d'une distance L_2), l'estimateur bayésien est la moyenne a posteriori.*

PREUVE. Si $T(X_1, \ldots, X_n) = \int_\Theta g(\theta)\mathrm{d}\Pi(\theta \mid X_1, \ldots, X_n)$ P-p.s., alors pour tout autre estimateur T',

$$\int_\Theta \left(T'(x_1, \ldots, x_n) - g(\theta)\right)^2 \mathrm{d}\Pi(\theta \mid x_1, \ldots, x_n) \geq \int_\Theta \left(T(x_1, \ldots, x_n) - g(\theta)\right)^2 \mathrm{d}\Pi(\theta \mid x_1, \ldots, x_n)$$

car l'espérance minimise l'écart quadratique. Donc en intégrant par rapport à la loi de mélange,

$$\int_{\mathcal{X}^n}\int_{\Theta}\left(T'(x_1,\dots,x_n)-g(\theta)\right)^2 \mathrm{d}\Pi(\theta \mid x_1,\dots,x_n)\,\mathrm{d}P(x_1,\dots,x_n) \geq$$
$$\int_{\mathcal{X}^n}\int_{\Theta}\left(T(x_1,\dots,x_n)-g(\theta)\right)^2 \mathrm{d}\Pi(\theta \mid x_1,\dots,x_n)\,\mathrm{d}P(x_1,\dots,x_n)\,.$$

□

PROPOSITION 10.20 *Pour le risque quadratique, s'il n'est pas trivial, un estimateur Bayésien n'est jamais sans biais.*

PREUVE. Soit T un estimateur bayésien de $g(\theta)$, d'après le corollaire 10.19, T est l'espérance de $g(\theta')$ lorsque θ' est distribuée sous la loi a posteriori :

$$T(X_1,\dots,X_n)=\int_{\Theta} g(\theta')\mathrm{d}\Pi(\theta' \mid X_1,\dots,X_n)\,.$$

Si T est sans biais on a aussi

$$g(\theta)=\mathbb{E}_\theta[T(X_1,\dots,X_n)]\,.$$

Dans la suite du calcul, on abrège $T(X_1,\dots,X_n)$ en T. On note P la loi de mélange sur $\mathcal{X}^n$ définie par Π et $(P_\theta)_{\theta\in\Theta}$.

Sous ces hypothèses,

$$\begin{aligned}
&\int_{\Theta} R(T,\theta)\mathrm{d}\Pi(\theta)\\
&= \int_{\Theta}\left[\int_{\mathcal{X}^n}(T(x_1,\dots,x_n)-g(\theta))^2\prod_{i=1}^n p(x_i\mid\theta)\mathrm{d}\nu(x_1)\dots\mathrm{d}\nu(x_n)\right]\mathrm{d}\Pi(\theta)\\
&= \int_{\Theta}\left[\int_{\mathcal{X}^n}(T^2-g(\theta)^2)\prod_{i=1}^n p(x_i\mid\theta)\mathrm{d}\nu(x_1)\dots\mathrm{d}\nu(x_n)\right]\mathrm{d}\Pi(\theta)\\
&\quad \text{car } T \text{ est sans biais}\\
&= \int_{\mathcal{X}^n}\int_{\Theta}(T^2-g(\theta)^2)\mathrm{d}\Pi(\theta \mid x_1,\dots,x_n)\mathrm{d}P(x_1,\dots,x_n)\\
&\quad \text{par désintégration}\\
&= \int_{\mathcal{X}^n}\int_{\Theta}\left(\left(\int_{\Theta} g(\theta')\mathrm{d}\Pi(\theta' \mid x_1,\dots,x_n)\right)^2-g(\theta)^2\right)\mathrm{d}\Pi(\theta \mid x_1,\dots,x_n)\mathrm{d}P(x_1,\dots,x_n)\\
&\quad T \text{ est Bayésien, donc moyenne sous loi a posteriori}\\
&\leq \int_{\mathcal{X}^n}\int_{\Theta}\left(\left(\int_{\Theta} g(\theta')^2\mathrm{d}\Pi(\theta' \mid x_1,\dots,x_n)\right)-g(\theta)^2\right)\mathrm{d}\Pi(\theta \mid x_1,\dots,x_n)\mathrm{d}P(x_1,\dots,x_n)\\
&\quad \text{inégalité de Jensen}\\
&= \int_{\mathcal{X}^n}\int_{\Theta} g(\theta')^2\mathrm{d}\Pi(\theta' \mid x_1,\dots,x_n)\mathrm{d}P(x_1,\dots,x_n)\\
&\quad -\int_{\mathcal{X}^n}\int_{\Theta} g(\theta)^2\mathrm{d}\Pi(\theta \mid x_1,\dots,x_n)\mathrm{d}P(x_1,\dots,x_n)\\
&= 0\,,
\end{aligned}$$

autrement dit, sur le support de Π, $R(T,\theta)=0$, ce qui revient à écrire $T(X_1,\dots,X_n)=g(\theta)$ P-presque sûrement. Ceci n'est possible que si la loi a posteriori et donc la loi a priori ne chargent qu'une seule valeur de θ. Cela trivialise l'estimateur. □

REMARQUE 10.21 Dans le problème de la localisation gaussienne en dimension d, on peut se demander si l'estimateur au maximum de vraisemblance ou l'estimateur de James-Stein sont des estimateurs bayésiens. Notons que l'un est sans biais mais pas l'autre. On peut déjà écarter l'estimateur au maximum de vraisemblance.

La démarche minimax conduit parfois à des estimateurs dont l'intérêt peut être questionné.

Considérons le modèle binomial, avec $\Theta =]0,1[$ où P_θ désigne la loi de Bernoulli de probabilité de succès θ. On choisit la fonction de perte quadratique.

On choisit comme loi a priori Beta(p,q) (ou Dirichlet) de densité

$$\mathbb{I}_{0<x<1}\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\times x^{p-1}(1-x)^{q-1}\,.$$

On peut profiter de la conjugaison entre lois de Bernoulli (et lois binomiales) et lois Beta pour calculer les lois a posteriori. Si $n_1 := \sum_{i=1}^n X_i = n\overline{X}_n$ et $n_0 = n - n_1$; la loi a posteriori est une loi Beta $(p+n_1, q+n_0)$. L'estimateur bayésien est alors

$$\widehat{\theta}_n := \frac{p+n_1}{p+q+n} = \frac{\overline{X}_n + \frac{p}{n}}{1+\frac{p+q}{n}}$$

On peut rechercher parmi les lois Beta celle qui forme l'a priori le plus défavorable, celle qui maximise le risque Bayesien (et réalise donc le risque maximin pour le modèle). Nous allons nous contenter de vérifier que l'a priori maximin est la loi Beta$(\sqrt{n}/2, \sqrt{n}/2)$ (c'est une loi d'espérance $1/2$ et de variance $\frac{1}{4(\sqrt{n}+1)}$). L'estimateur Bayésien est pour cet a priori

$$\widehat{\theta}_n := \frac{\overline{X}_n + \frac{1}{2\sqrt{n}}}{1+\frac{1}{\sqrt{n}}}\,,$$

son biais et sa variance en θ étant respectivement

$$\frac{1}{\sqrt{n}}\frac{\frac{1}{2}-\theta}{1+\frac{1}{\sqrt{n}}} \qquad \text{et} \qquad \frac{\theta(1-\theta)}{n\left(1+\frac{1}{\sqrt{n}}\right)^2}\,.$$

Le risque quadratique en θ (somme de la variance et du carré du biais) est alors

$$\frac{1}{4(\sqrt{n}+1)^2}\,.$$

Le risque quadratique de l'estimateur bayésien ne dépend pas de l'estimande ! D'après le théorème 10.9, l'estimateur bayésien est minimax.

Nous avons par ce calcul élémentaire calculé le risque minimax pour le modèle binomial. Au passage nous constatons que l'estimateur au maximum de vraisemblance n'est pas minimax, puisque son risque maximal est $1/(4n)$. En revanche, on peut dire que l'estimateur au maximum de vraisemblance est asymptotiquement minimax, puisque son risque maximal est asymptotiquement équivalent au risque minimax.

Comme les résultats de super-efficacité, ce résultat ne doit pas être pris trop au sérieux. En effet, l'estimateur minimax améliore un peu l'estimateur au maximum de vraisemblance en $\theta = 1/2$, mais il est moins bon en tout θ tel que

$$\left|\theta - \frac{1}{2}\right| \geq \frac{1}{2n^{1/4}}\frac{\sqrt{1+\frac{1}{\sqrt{n}}}}{1+\frac{1}{\sqrt{n}}}\,.$$

Toujours dans le modèle binomial, on peut envisager d'autres risques que le risque quadratique. Par exemple,

$$L(T,\theta) = \frac{(T-\theta)^2}{\theta(1-\theta)}\,,$$

en effet rien n'oblige la fonction de perte à être une fonction de la différence entre l'estimande et l'estimateur. Pour cette fonction de perte, le risque de l'estimateur au maximum de vraisemblance est constant et égal à $1/n$. Toujours d'après le théorème 10.9, c'est donc un estimateur minimax (et bayésien pour l'a priori le plus défavorable). Le fait que cet estimateur soit sans biais ne contredit pas la proposition 10.20.

10.9 INÉGALITÉ DE VAN TREES

Nous avons déjà vu quelques techniques de minoration du risque. Pour les tests entre hypothèses simples, on peut partir de la distance en variation, mais aussi de la distance de Hellinger ou de l'entropie relative entre les deux hypothèses. Pour l'estimation ponctuelle, on peut réduire un problème de test à un problème d'estimation. C'est en fait une méthode très générale. Quand on s'intéresse au risque quadratique d'estimateurs sans biais, sous des conditions de régularité du modèle, l'inégalité de Cramer-Rao fournit une borne inférieure. Se contenter des estimateurs sans biais n'est pas raisonnable. Les inégalités de van Trees permettent de minorer le risque quadratique pour des estimateurs généraux. Comme les techniques utilisées ailleurs (Assouad, Fano, Le Cam), elles partent d'un jeu bayésien.

Le modèle $(P_\theta)_{\theta\in\Theta}$ avec $\Theta \subseteq \mathbb{R}$ est supposé dominé par ν. Dans la suite $\Theta =]a, b[$ avec a, b éventuellement infinis. La vraisemblance en θ, x est notée $p(x \mid \theta)$. La fonction $\theta \mapsto p(x \mid \theta)$ est supposée mesurable.

Pour tout θ, on suppose que pour tout θ, la fonction score $\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x \mid \theta)$ est dans $L_2(P_\theta)$, et qu'elle est centrée

$$\int_{\mathcal{X}} \frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x \mid \theta) p(x \mid \theta) \mathrm{d}\nu(x) = 0\,.$$

On note $I(\theta)$ la variance de la fonction score

$$I(\theta) := \int_{\mathcal{X}} \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x \mid \theta)\right)^2 p(x \mid \theta) \mathrm{d}\nu(x)\,.$$

On suppose que c'est une fonction continue de θ (en fait le centrage de la fonction score peut être prouvé à partir de la seule continuité de $\theta \mapsto I(\theta)$.

i) La loi a priori Π est absolument continue (de densité π) et l'information de Fisher pour le modèle de localisation défini par $\Pi(\cdot - \theta), \theta \in \Theta$ est finie :

$$I(\Pi) := \int_\Theta \frac{\pi'(\theta)^2}{\pi(\theta)} \mathrm{d}\theta < \infty\,.$$

ii) Les fonctions $\theta \mapsto \theta\pi(\theta)p(x \mid \theta)$ et $\theta \mapsto \pi(\theta)p(x \mid \theta)$ sont supposées tendre vers 0 assez vite lorsque θ tend vers un point fini de $\partial\Theta = \{a, b\}$, pour que pour ν-presque tout x,

a)

$$\int_{]a,b[} \frac{\mathrm{d}\left(p(x \mid \theta)\pi(\theta)\right)}{\mathrm{d}\theta} \mathrm{d}\theta = \left[p(x \mid \theta)\pi(x)\right]_a^b = 0$$

b)

$$\int_{]a,b[} \theta \frac{\mathrm{d}\left(p(x \mid \theta)\pi(\theta)\right)}{\mathrm{d}\theta} \mathrm{d}\theta = \left[\theta p(x \mid \theta)\pi(\theta)\right]_a^b - \int_{]a,b[} p(x \mid \theta)\pi(\theta)\mathrm{d}\theta = -\int_{]a,b[} p(x \mid \theta)\mathrm{d}\Pi(\theta)$$

THÉORÈME 10.22 (INÉGALITÉ DE VAN TREES) *Sous les hypothèses énoncées ci-dessus, si $T = T(X)$ est une statistique,*

$$\int_\Theta \mathbb{E}_\theta\left[(T-\theta)^2\right] \mathrm{d}\Pi(\theta) \geq \frac{1}{\int_\Theta I(\theta)\mathrm{d}\Pi(\theta) + I(\Pi)}\,.$$

PREUVE. L'étape cruciale de la preuve consiste à vérifier

$$\int_\Theta \int_{\mathcal{X}} \left[(T(x)-\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\left(p(x \mid \theta) \times \pi(\theta)\right)\right] \mathrm{d}\nu(x)\mathrm{d}\theta = 1\,. \tag{10.1}$$

La vérification revient à invoquer les hypothèses du théorème à l'aide du théorème de Fubini. D'une part, on a

$$\begin{aligned}
&\int_\Theta \int_{\mathcal{X}} T(x) \frac{\mathrm{d}}{\mathrm{d}\theta} \left(p(x \mid \theta) \times \pi(\theta)\right) \mathrm{d}\nu(x)\mathrm{d}\theta \\
&\quad = \int_{\mathcal{X}} T(x) \int_\Theta \frac{\mathrm{d}}{\mathrm{d}\theta} \left(p(x \mid \theta) \times \pi(\theta)\right) \mathrm{d}\theta \mathrm{d}\nu(x) \\
&\quad = \int_{\mathcal{X}} T(x) \left[p(x \mid \theta) \times \pi(\theta)\right]_a^b \mathrm{d}\nu(x) \\
&\quad = \int_{\mathcal{X}} 0 \, \mathrm{d}\nu(x) \\
&\quad = 0 \, .
\end{aligned}$$

Et, on a aussi

$$\begin{aligned}
&\int_\Theta \int_{\mathcal{X}} \left[\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \left(p(x \mid \theta) \times \pi(\theta)\right)\right] \mathrm{d}\nu(x)\mathrm{d}\theta \\
&\quad = \int_{\mathcal{X}} \int_\Theta \theta \frac{\mathrm{d}}{\mathrm{d}\theta} \left[(p(x \mid \theta) \times \pi(\theta))\right] \mathrm{d}\theta \mathrm{d}\nu(x) \\
&\quad = \int_{\mathcal{X}} \left(\left[\theta p(x \mid \theta) \times \pi(\theta)\right]_a^b - \int_\Theta p(x \mid \theta) \times \pi(\theta) \mathrm{d}\theta \right) \mathrm{d}\nu(x) \\
&\quad = -\int_\Theta \int_{\mathcal{X}} p(x \mid \theta) \mathrm{d}\nu(x) \pi(\theta) \mathrm{d}\theta \\
&\quad = -1 \, .
\end{aligned}$$

A partir de (10.1), on peut réécrire

$$\begin{aligned}
1 &= \int_\Theta \int_{\mathcal{X}} \left[(T(x) - \theta) \frac{\mathrm{d}}{\mathrm{d}\theta} \log \left(p(x \mid \theta) \times \pi(\theta)\right)\right] p(x \mid \theta) \times \pi(\theta) \mathrm{d}\nu(x) \mathrm{d}\theta \\
&\leq \left(\int_\Theta \int_{\mathcal{X}} \left[(T(x) - \theta)^2\right] p(x \mid \theta) \times \pi(\theta) \mathrm{d}\nu(x) \mathrm{d}\theta \right)^{1/2} \\
&\quad \times \left(\int_\Theta \int_{\mathcal{X}} \left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log \left(p(x \mid \theta) \times \pi(\theta)\right) \right)^2 \right] p(x \mid \theta) \times \pi(\theta) \mathrm{d}\nu(x) \mathrm{d}\theta \right)^{1/2}
\end{aligned}$$

où l'inégalité est une application de l'inégalité de Cauchy-Schwarz.

$$\begin{aligned}
&\int_\Theta \int_{\mathcal{X}} \left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log \left(p(x \mid \theta) \times \pi(\theta)\right) \right)^2 \right] p(x \mid \theta) \times \pi(\theta) \mathrm{d}\nu(x) \mathrm{d}\theta \\
&\quad = \int_\Theta \int_{\mathcal{X}} \left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x \mid \theta) \right)^2 + \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log \pi(\theta) \right)^2 \right. \\
&\qquad\qquad \left. + 2 \frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x \mid \theta) \times \frac{\mathrm{d}}{\mathrm{d}\theta} \log \pi(\theta) \right] p(x \mid \theta) \times \pi(\theta) \mathrm{d}\nu(x) \mathrm{d}\theta \\
&\quad = \int_\Theta I(\theta) \pi(\theta) \mathrm{d}\theta + I(\Pi) \\
&\qquad + 2 \int_\Theta \left(\int_{\mathcal{X}} \frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x \mid \theta) p(x \mid \theta) \mathrm{d}\nu(x) \right) \pi(\theta) \mathrm{d}\theta \\
&\quad = \int_\Theta I(\theta) \pi(\theta) \mathrm{d}\theta + I(\Pi) \, ,
\end{aligned}$$

où la dernière égalité est une conséquence du centrage des fonctions scores dans les modèles réguliers. □

En posant les hypethèses pertinantes d'intégrabilité sur g, on prouve de la même façon.

Corollaire 10.23

$$\int_\Theta \mathbb{E}_\theta\left[(T(X_1,\dots,X_n)-g(\theta))^2\right]\mathrm{d}\Pi(\theta) \geq \frac{\left(\int_\Theta g'(\theta)\mathrm{d}\Pi(\theta)\right)^2}{n\int_\Theta I(\theta)\mathrm{d}\Pi(\theta)+I(\Pi)}\,.$$

On se contente d'écrire la preuve pour $n=1$ (1 observation).

preuve. Il suffit de vérifier que pour ν-presque tout x,

$$\begin{aligned}\int_\Theta g(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\left(p(x\mid\theta)\times\pi(\theta)\right)\mathrm{d}\theta &\\ = \left[g(\theta)\left(p(x\mid\theta)\times\pi(\theta)\right)\right]_a^b - \int_\Theta g'(\theta)p(x\mid\theta)\pi(\theta)\mathrm{d}\theta &\\ = -\int_\Theta g'(\theta)p(x\mid\theta)\pi(\theta)\mathrm{d}\theta\,.&\end{aligned}$$

pour remplacer (10.1) par

$$\int_\Theta\int_{\mathcal{X}}\left[(T(x)-g(\theta))\frac{\mathrm{d}}{\mathrm{d}\theta}\left(p(x\mid\theta)\times\pi(\theta)\right)\right]\mathrm{d}\nu(x)\mathrm{d}\theta = \int_\Theta g'(\theta)\mathrm{d}\Pi(\theta)\,.$$

□

Exemple 10.24 On considère le modèle de translation gaussienne $\mathcal{N}(\theta,1)$ avec $\theta\in\Theta:=]0,\infty)$. On veut estimer $g(\theta):=\theta^\alpha$ pour $\alpha\in[0,1[$.

Un estimateur intuitif pour θ^α est $(\overline{X}_n)_+^\alpha$. Il est consistant en tout $\theta\in\Theta$. Si $\theta>0$, la méthode delta nous indique que

$$\sqrt{n}\left((\overline{X}_n)_+^\alpha-\theta^\alpha\right)\rightsquigarrow\mathcal{N}\left(0,\alpha^2\theta^{2\alpha-2}\right)\,.$$

On doit se demander comment se comporte le risque quadratique en $\theta=0$ et comment se comporte le risque minimax à n fixé. En $\theta=0$, $\left((\overline{X}_n)_+^\alpha-\theta^\alpha\right)=(\overline{X}_n)_+^\alpha$ vaut 0 avec probabilité 1/2 et avec probabilité 1/2 est distribuée comme $n^{-\alpha/2}$ multiplié par la puissance α de la valeur absolue d'une gaussienne standard. Le risque quadratique décroit comme $n^{-\alpha}$ plutôt que comme $1/n$.

Les inégalités de Van Trees permettent de minorer le risque minimax.

Dans ce modèle $I(\theta)=1$ pour tout $\theta\in\Theta$. Si on choisit une loi a priori qui vérifie les conditions du théorème 10.22, pour tout estimateur T

$$\sup_{\theta\geq0}\mathbb{E}_\theta\left[(T(X_1,\dots,X_n)-\theta^\alpha)^2\right]\geq\int_\Theta\mathbb{E}_\theta\left[(T(X_1,\dots,X_n)-\theta^\alpha)^2\right]\pi(\theta)\mathrm{d}\theta\geq\frac{\left(\int_\Theta\alpha\theta^{\alpha-1}\pi(\theta)\right)^2}{nI(\theta)+I(\pi)}\,.$$

Si la densité π satisfait les conditions de l'inégalité de van Trees, alors les densités π_a obtenues par changement d'échelle $\pi_a(\cdot):=\frac{1}{a}\pi(\frac{\cdot}{a})$ satisfont aussi les conditions de l'inégalité. L'information de Fisher de π_a est facilement obtenue à partir de $I(\pi)$:

$$I(\pi_a)=\frac{1}{a^2}I(\pi)\,.$$

Le membre droit de l'inégalité de van Trees obtenue à partir de l'a priori π_a est

$$\frac{a^{2(\alpha-1)}\left(\int_\Theta\alpha\theta^{\alpha-1}\pi(\theta)\right)^2}{n+I(\pi)/a^2}\,.$$

On peut optimiser le facteur d'échelle en choisissant

$$a=\sqrt{\frac{\alpha I(\pi)}{n(1-\alpha)}}\,.$$

On obtient finalement

$$\sup_{\theta\geq0}\mathbb{E}_\theta\left[(T(X_1,\dots,X_n)-\theta^\alpha)^2\right]\geq\frac{(I(\pi))^{\alpha-1}\alpha^\alpha(1-\alpha)^{1-\alpha}}{n^\alpha}\left(\int_\Theta\alpha\theta^{\alpha-1}\pi(\theta)\right)^2\,.$$

10.10 REMARQUES BIBLIOGRAPHIQUES

Le livre de Lehmann et Casella [7] est la source principale d'inspiration du cours. L'estimateur de James-Stein est aussi exposé dans [6, Chapitre 3] [9].

L'étude systématique des modèles différentiables en moyenne quadratique (théorie de Hajek-Le Cam) est au coeur de [2, 10]. Elle est aussi illustrée sur les tests dans [8] et explorée en profondeur dans [6].

[1] décrit la méthode de Stein qui utilise l'identité de Stein comme point de départ pour démontrer des théorèmes centraux limites avec vitesse de convergence.

Les résultats de la section 10.5 sont attribués à Lehman et Hodges, voir [6, Chapitre 3].

La section sur les probabilités conditionnelles est inspirée du chapitre 10 de [3]. On y trouve toutes les preuves.

L'étude fréquentiste des performances des méthodes bayésiennes fait l'objet d'une étude intensive depuis une vingtaine d'années. On trouve dans [5] et surtout dans [4] un exposé de ces progrès.

Références

[1] N. Ross. « Fundamentals of Stein's method ». In : *ArXiv e-prints* (sept. 2011).

[2] A. van der Vaart. « The statistical work of Lucien Le Cam ». In : *Annals of Statistics* 30.3 (2002), p. 631-682.

[3] R. M. Dudley. **Real analysis and probability**. T. 74. Cambridge Studies in Advanced Mathematics. Cambridge : Cambridge University Press, 2002, p. x+555. MR : MR1932358(2003h:60001).

[4] S. Ghosal et A. van der Vaart. **Fundamentals of nonparametric Bayesian inference**. T. 44. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2017, p. xxiv+646. ISBN : 978-0-521-87826-5. MR : 3587782.

[5] E. Giné et R. Nickl. **Mathematical Foundations of Infinite-Dimensional Statistical Models**. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015. ISBN : 9781107043169.

[6] I. Ibragimov et R. Khasminskii. **Statistical Estimation : Asymptotic Theory**. New York : Springer-Verlag, 1981.

[7] E. L. Lehmann et G. Casella. **Theory of point estimation**. Second. Springer Texts in Statistics. Springer-Verlag, New York, 1998, p. xxvi+589.

[8] E. L. Lehmann et J. P. Romano. **Testing statistical hypotheses**. Third. Springer Texts in Statistics. Springer, New York, 2005, p. xiv+784.

[9] A. Tsybakov. **Introduction à l'estimation non-paramétrique**. T. 41. Mathématiques & Applications. Berlin : Springer-Verlag, 2004, p. x+175. MR : MR2013911(2005a:62007).

[10] A. van der Vaart. **Asymptotic statistics**. Cambridge University Press, 1998.

11.1 INTRODUCTION

L'objectif de ce chapitre est d'introduire certains outils utilisés par les praticiens de l'inférence bayésienne.

Le paradigme bayésien peut sembler plus complexe à mettre en œuvre que les méthodes dites fréquentistes ou « classiques », et notamment que le maximum de vraisemblance ; pour autant, les méthodes bayésiennes sont très populaires dans de nombreux domaines d'application. On peut citer plusieurs explications :

i) la mesure de l'incertitude est aisée et explicite, puisqu'on obtient une distribution pour le paramètre à estimer θ (la distribution a posteriori) et jamais une simple estimation ponctuelle ;
ii) les méthodes bayésiennes peuvent être plus aisées à appliquer sur des modèles complexes ;
iii) le paradigme bayésien est facile à utiliser pour des modèles ou expériences imbriqués : la loi a posteriori de l'expérience 1 peut être utilisée comme loi a priori de l'expérience 2 ;
iv) l'étude de fonctions complexes des paramètres est facilitée (par exemple : lois marginales, lois conditionnelles...) ;
v) l'étude de plusieurs modèles concurrents est plus aisée ;
vi) certains mettent en avant des raisons philosophiques ;
vii) les estimateurs bayésiens et les régions de crédibilité associées peuvent être plus intuitifs que les notions d'intervalle de confiance ou de p-valeur, qui sont mal comprises des non-mathématiciens ;
viii) le cadre mathématique est parfois plus simple à utiliser.

Pour autant, il existe bien entendu un grand nombre de situations où l'estimateur du maximum de vraisemblance (EMV) est utilisé en pratique, notamment pour des modèles suffisamment simples pour que l'EMV soit calculable et avec des données de taille suffisamment grande pour que les propriétés asymptotiques de l'EMV entrent en action.

Dans la suite, on considère un modèle statistique dominé avec un paramètre $\theta \in \Theta \subseteq \mathbb{R}^k$ inconnu, des observations y et une fonction de vraisemblance $L(\theta; y) = p(y|\theta)$. On note $\pi(\theta)$ la (densité de la) loi a priori et $\pi(\theta|y) \propto \pi(\theta)L(\theta; y)$ la (densité de la) loi a posteriori. On suppose que toutes les lois considérées ont une densité par rapport à la mesure de Lebesgue ou de comptage. Pour un événement A, on note $P(A)$ et $P(A|y)$ sa probabilité respectivement a priori et a posteriori ; on utilise la notation équivalente pour l'espérance.

Il existe plusieurs estimateurs ponctuels possibles. Les plus usités sont :

i) l'espérance a posteriori $\mathbb{E}[\theta|y]$ (qui s'impose lorsqu'on cherche à minimiser un risque quadratique, voir chapitre précédent) ;
ii) la médiane a posteriori ;
iii) le maximum a posteriori (MAP) $\arg\max_\theta \pi(\theta|y)$;

mais il serait dommage de se cantonner à un estimateur ponctuel : l'intérêt du paradigme bayésien réside dans la distribution a posteriori prise dans son entier.

Définissons tout d'abord les pendants bayésiens des notions d'ensemble de confiance et de test de rapport de vraisemblance.

DÉFINITION 11.1 (ENSEMBLE DE CRÉDIBILITÉ) Soit $\Pi(\cdot \mid y)$ la loi a posteriori sachant la donnée $y \in \mathcal{Y}$. Soit $\alpha \in]0, 1[$. Un ensemble C_y est un α-ensemble de crédibilité si

$$\Pi\left[\theta \in C_y | y\right] \geq 1 - \alpha.$$

Dans cette définition, remarquons que θ est aléatoire, distribuée selon la loi a posteriori, alors que C_y est une fonction habituellement déterministe des observations y (on n'exclut pas la possibilité d'une randomisation comme dans la construction de Neyman-Pearson) : les rôles sont inversés par rapport à la définition d'un ensemble de confiance.

DÉFINITION 11.2 (HPD) Un α-ensemble de crédibilité est dit HPD (*highest posterior density*) si on peut écrire

$$\{\theta : \pi(\theta|y) > k_\alpha\} \subset C_y \subset \{\theta : \pi(\theta|y) \geq k_\alpha\}$$

où k_α est le plus grand seuil tel que C_y soit un α-ensemble de crédibilité.

Un ensemble HPD est de volume minimal parmi les α-ensembles de crédibilité. Par exemple, si θ est de dimension 1 et que C_y est connexe, alors C_y est l'intervalle de crédibilité le plus court.

11.2 CHOIX DE LA LOI A PRIORI

Le choix de la loi a priori est une des problématiques principales de toute analyse bayésienne appliquée. La subjectivité du scientifique est rendue explicite par le choix de la loi a priori. Cela présente un intérêt philosophique, puisque l'analyse en est plus transparente, mais ce choix doit pouvoir être défendu. Parmi les méthodes pour choisir une loi a priori, citons :

i) la loi « non-informative » ou a priori de Jeffreys : $\pi(\theta) \propto \sqrt{I(\theta)}$ où $I(\theta)$ est l'information de Fisher du modèle ;
ii) une loi a priori fournie par un expert du domaine d'application : selon les cas, l'expert peut soit fournir une loi, soit au moins les premiers moments (espérance, variance) ce qui permet de choisir les paramètres d'une loi conjuguée ;
iii) une loi a priori qui provient d'une expérience précédente : en cas d'expériences successives, la loi a posteriori de l'expérience k peut servir de loi a priori pour l'expérience $k+1$;
iv) dans tous les cas, et encore plus quand aucun choix de la loi a priori ne s'impose, il faut vérifier l'influence de la loi a priori sur la loi a posteriori : on effectue l'analyse avec plusieurs lois a priori différentes, et on vérifie que les loi a posteriori concordent.

EXEMPLE 11.3 Considérons le modèle simple suivant : $Y_1, \ldots, Y_n$ sont iid de loi de Poisson $\mathcal{P}(\lambda)$, avec la loi a priori conjuguée $\lambda \sim \Gamma(a, b)$. La loi a posteriori est alors $\lambda|Y \sim \Gamma(a + \sum y_i, b + n)$. La Figure 11.3 montre l'influence de la loi a priori pour différentes valeurs de n.

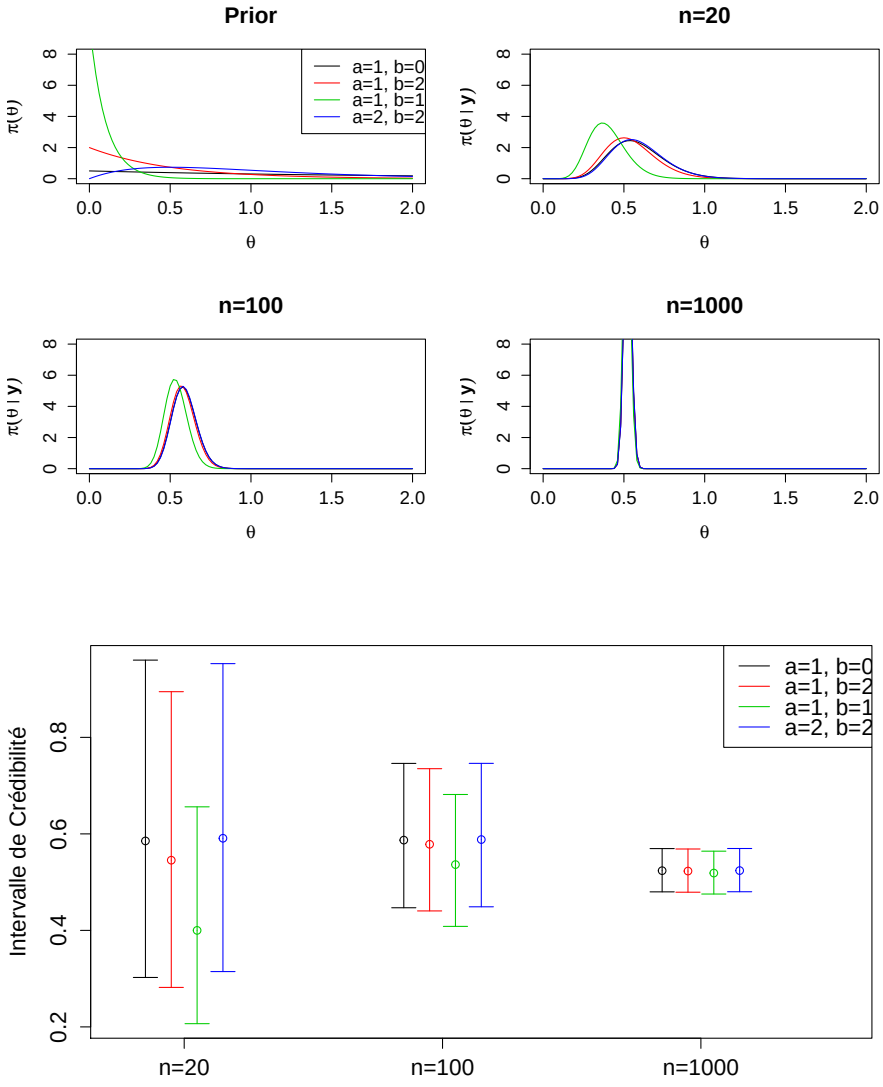

FIG. 11.1 : Pour un modèle de Poisson $\mathcal{P}(\lambda)$ et différentes lois a priori $\lambda \sim \Gamma(a, b)$, on observe que les lois a posteriori correspondantes (haut) se rapprochent au fur et à mesure que la taille d'échantillon n augmente, et notamment les intervalles de crédibilité (bas). Pour $n = 1000$, l'influence de la loi a priori est négligeable.

Avec $(\Omega, \mathcal{F})$ un espace probabilisable, si $T : \Omega \to \mathbb{R}^p$ est $\mathcal{F}$-mesurable, et ν une mesure σ-finie sur $(\Omega, \mathcal{F})$, le modèle exponentiel défini par ν et T est paramétré par l'intérieur de

$$\Theta := \left\{ \theta \in \mathbb{R}^p : \qquad Z(\theta) := \int_\Omega \exp\left(\langle \theta, T(\omega) \rangle\right) \nu(\mathrm{d}\omega) < \infty \right\}$$

supposé non vide. Chaque loi P_θ est définie par sa densité par rapport à ν :

$$p(\omega \mid \theta) = \exp\left(\langle \theta, T(\omega) \rangle\right) / Z(\theta) \,.$$

On peut définir une loi a priori sur Θ par sa densité par rapport à la mesure de Lebesgue

$$\pi(\theta) \propto \exp\left(\langle \theta, t \rangle - t_0 \log Z(\theta)\right)$$

avec $t \in \mathbb{R}^p, t_0 \in \mathbb{R}$ tels que

$$\int_{\Theta^\circ} \exp\left(\langle \theta, t \rangle - t_0 \log Z(\theta)\right) \mathrm{d}\theta < \infty \,.$$

Cette loi a priori est paramétrée par $(t_0, t) \in \mathbb{R}^{p+1}$.

La densité de la loi a posteriori définie par les données $D_n = \omega_1, \ldots, \omega_n$ est proportionnelle à

$$\exp\left(\left\langle \theta, t + \sum_{i=1}^n T(\omega_i) \right\rangle - (n + t_0) \log Z(\theta) \right) ,$$

la densité $\pi(\cdot \mid \omega_1, \ldots, \omega_n)$ est de la même forme que la densité a priori mais les paramètres sont modifiés par les données :

$$\left(t_0 + n, t + \sum_{i=1}^n T(\omega_i) \right) .$$

Dans les modèles exponentiels (en forme canonique ici) il est donc très facile de définir des lois a priori issues de familles conjuguées au modèle. Le calcul de la loi a posteriori s'en trouve grandement simplifié.

En choisissant une loi a priori conjuguée dans un modèle exponentiel, il est relativement simple de mener à bien une *analyse fréquentiste* de l'inférence bayésienne, c'est-à-dire d'étudier l'évolution de la loi a posteriori.

Supposons pour le reste de cette section, que les données sont i.i.d. selon P_{θ_0} avec $\theta_0 \in \Theta^\circ$. La suite des maxima a posteriori est donnée par la suite des solutions de

$$t + \sum_{i=1}^n T(\omega_i) - (t_0 + n) \nabla \log Z(\theta) = 0$$

soit, avec probabilité qui tend vers 1 lorsque $n \to \infty$,

$$\widetilde{\theta}_n := \nabla \log Z^{-1} \left(\frac{t + \sum_{i=1}^n T(\omega_i)}{t_0 + n} \right) .$$

Lorsque n tend vers l'infini, $\widetilde{\theta}_n$ tend à s'approcher de l'EMV : $\widehat{\theta}_n := \nabla \log Z^{-1} \left(\frac{\sum_{i=1}^n T(\omega_i)}{n} \right)$.

On peut se demander à quoi ressemble la loi a posteriori recentrée autour de $\widetilde{\theta}_n$ et renormalisée par $\sqrt{n}$, soit la loi de densité

$$h \mapsto \frac{1}{\sqrt{n}} \, \pi \left(\widehat{\theta}_n + \frac{h}{\sqrt{n}} \mid D_n \right)$$

La densité de la loi a posteriori recentrée autour de $\widetilde{\theta}_n$ et renormalisée par $\sqrt{n}$ est proportionnelle à

$$\exp\left(-\frac{t_0 + n}{2n} h^t \nabla^2 \log Z\left(\widetilde{\theta}_n\right) h + \frac{(t_0 + n)\|h\|^2}{n} R_{\widetilde{\theta}_n}(h/\sqrt{n}) \right)$$

où $R_{\widetilde{\theta}_n}(x) = o(\|x\|)$.

La régularité de la fonction $\theta \mapsto \log Z(\theta)$ et la convergence presque sûre de $\widetilde{\theta}_n$ vers θ_0 nous indiquent que $\nabla^2 \log Z\left(\widetilde{\theta}_n\right)$ convergence presque sûrement vers $\nabla^2 \log Z(\theta_0) = I(\theta_0)$. Un peu de travail supplémentaire nous permettrait de vérifier que presque sûrement

$$R_{\widetilde{\theta}_n}(h/\sqrt{n}) \longrightarrow 0\,.$$

Cela revient à établir que presque sûrement, la densité a posteriori recentrée et renormalisée converge simplement vers la densité de la loi gaussienne centrée et de covariance $I^{-1}(\theta)$. La convergence simple des densité implique la convergence en distribution, et même la convergence au sens de la distance en variation (Lemme 12.2 dit de Scheffé).

Ce résultat peut se démontrer rigoureusement pour les modèles différentiables en moyenne quadratique, c'est le théorème dit de Bernstein-von Mises. Celui-ci peut aussi s'énoncer ainsi : dans un modèle différentiable en moyenne quadratique, sous des conditions bénignes sur la loi a priori, la suite des lois a posteriori recentrées autour du maximum a posteriori et renormalisées converge presque sûrement en distribution vers la gaussienne centrée de covariance égale à l'inverse de la matrice d'information de Fisher.

Le théorème de Bernstein-von Mises nous montre au passage que les ensembles de crédibilité définis par les ensembles de niveau de la densité a posteriori sont asymptotiquement des régions de confiance. Ils approchent les ellipsoïdes de confiance introduits à partir de la normalité asymptotique de $\sqrt{n}I(\widehat{\theta}_n)^{1/2}(\widehat{\theta}_n - \theta_0)$.

11.4 FACTEUR DE BAYES

Considérons maintenant la question du choix de modèle.

On considère deux modèles M_0 et M_1, qu'on souhaite comparer. Le modèle M_i dépend d'un paramètre θ_i, pour lequel on a une loi a priori $\pi_i(\theta_i)$ et une fonction de vraisemblance L_i. Enfin, on choisit la probabilité qu'on attribue a priori à chacun des deux modèles (par exemple 1/2 et 1/2). A priori, le rapport des chances des deux modèles est $\Pi[M_0]/\Pi[M_1]$; a posteriori, ce rapport des chances est $\Pi[M_0|y]/\Pi[M_1|y]$.

DÉFINITION 11.4 (FACTEUR DE BAYES)
On appelle facteur de Bayes la quantité

$$B^{\pi}_{01}(y) = \frac{\Pi[M_0|y]}{\Pi[M_1|y]} \Big/ \frac{\Pi[M_0]}{\Pi[M_1]}\,.$$

On utilise le facteur de Bayes pour mesurer comment les données ont fait évoluer le rapport des chances. Intuitivement, si $B^{\pi}_{01}(y)$ est grand, alors les données y favorisent M_0 ; si $B^{\pi}_{01}(y)$ est petit, les données favorisent M_1. Le facteur de Bayes mesure à quel point les données ont fait évoluer notre croyance, entre le ratio des probabilités a priori et le ratio a posteriori.

Par le théorème de Bayes, on a

$$\Pi[M_i|y] = \frac{P[y|M_i]\Pi[M_i]}{P[y]}$$

et on peut récrire le facteur de Bayes :

$$B^{\pi}_{01} = \frac{m_0(y)}{m_1(y)} = \frac{\int \pi_0(\theta_0)L_0(\theta_0;y)\,\mathrm{d}\theta_0}{\int \pi_1(\theta_1)L_1(\theta_1;y)\,\mathrm{d}\theta_1}$$

où

$$m_i(y) = \mathbb{E}_{\pi_i}\left[L_i(\theta_i;y)\right] = \int \pi_i(\theta_i)L_i(\theta_i;y)\,d\theta_i$$

n'est autre que la constante de normalisation de la formule $\pi_i(\theta_i|y) \propto \pi_i(\theta_i)L_i(\theta_i;y)$.

REMARQUE 11.5

i) alors que dans le cadre classique du test d'hypothèses, on accorde une place privilégiée à l'hypothèse nulle, ici les deux hypothèses jouent des rôles symétriques ;

ii) cette dernière expression ressemble au ratio de vraisemblances utilisé dans le cadre classique, dans lequel on aurait remplacé arg max par une intégrale. Cette modification permet d'avoir une pénalisation naturelle de la taille du modèle, et d'éviter ainsi les problèmes de surapprentissage.

De même qu'il existe des échelles pour interpréter les p-valeurs, on cite souvent l'échelle de Jeffreys pour interpréter le facteur de Bayes. Si $\log_{10} B^{\pi}_{01}(y) > 0$, alors les données favorisent le modèle M_0 ; plus précisément :

i) entre 0 et $\frac{1}{2}$, l'évidence est *faible* ;
ii) entre $\frac{1}{2}$ et 1, elle est *substantielle* ;
iii) entre 1 et 2, elle est *forte*
iv) au-dessus de 2, elle est *décisive*

et symétriquement en faveur du modèle M_1 pour les valeurs négatives.

Exemple 11.6 On reprend l'exemple Poisson-Gamma exposé ci-dessus : on dispose d'observations $Y_1, \dots, Y_n$, mais également de covariables $Z_1, \dots, Z_n$ avec $Z_i \in \{1, 2\}$ et on souhaite choisir entre les deux modèles suivants :

$$\mathcal{M}_0 : \quad Y_i \sim \mathcal{P}(\lambda) \qquad \lambda \sim \Gamma(a, b)$$

$$\mathcal{M}_1 : \quad Y_i | Z_i = k \sim \mathcal{P}(\lambda_k) \qquad \lambda_1, \lambda_2 \sim \Gamma(a, b)$$

autrement dit, on cherche à savoir si les (Y_i) proviennent tous de la même distribution, ou s'ils proviennent de deux distributions différentes selon la valeur des (Z_i).

Pour choisir entre ces deux modèles, on calcule les vraisemblances marginales (avec les notations $n_1 = \sum \mathbb{I}_{Z_i=1}$, $S_1 = \sum y_i \mathbb{I}_{Z_i=1}$, et les équivalents pour n_2 et S_2) :

$$m_0(y) = E_{\pi_0}\left[L_0(y|\theta_0\right] = \int_0^\infty \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} e^{-n\lambda} \lambda^{\sum y_i} \prod \frac{1}{y_i!} \, d\lambda = \frac{b^a}{\Gamma(a)} \prod \frac{1}{y_i!} \frac{\Gamma(a + \sum y_i)}{(b+n)^{a+\sum y_i}};$$

$$m_1(y) = \frac{b^{2a}}{\Gamma(a)^2} \frac{\Gamma(a+S_1)}{(b+n_1)^{a+S_1}} \frac{\Gamma(a+S_2)}{(b+n_2)^{a+S_2}} \prod \frac{1}{y_i!}$$

d'où le facteur de Bayes

$$B_{01}(y) = \frac{\Gamma(a}{b^a} \frac{\Gamma(a+S_1+S_2}{\Gamma(a+S_1)\Gamma(a+S_2)} \frac{(b+n_1)^{a+S_1}(b+n_2)^{a+S_2}}{(b+n)^{a+S_1+S_2}}$$

11.5 MONTE-CARLO

Beaucoup de quantités d'intérêt peuvent s'écrire sous la forme

$$I = \mathbb{E}_p[h(\theta)] = \int h(\theta) p(\theta) \, \mathrm{d}\theta.$$

Par exemple, l'espérance a posteriori correspond à $h(\theta) = \theta$ et p la loi a posteriori ; la vraisemblance marginale correspond à $h(\theta) = L(\theta; y)$ et p la loi a priori...

Dans les exemples simples considérés ci-dessus, ces espérances peuvent être calculables sous forme analytique, mais cela n'est bien sûr pas le cas en général. On va alors chercher un estimateur de I. L'estimateur le plus simple est celui de Monte-Carlo : on simule $\theta_1, \dots, \theta_T \sim p$ et on pose

$$\hat{I}_T^{\mathrm{MC}} = \frac{1}{T} \sum_{t=1}^{T} h(\theta_t).$$

Notons que $\hat{I}_T^{\mathrm{MC}}$ est un estimateur sans biais de I, et que sous des conditions très générales la Loi des Grands Nombres garantit qu'il est convergent. Enfin, on a

$$\mathrm{var}(\hat{I}_T^{\mathrm{MC}}) = \frac{1}{T} \mathrm{var}_p(h(\theta)).$$

Cette approche simple est rarement optimale : d'une part $\mathrm{var}_p(h(\theta))$ peut être élevée, d'autre part (et c'est plus gênant) il peut être difficile ou impossible de générer les θ_t selon p. Il est alors plus efficace de procéder par *échantillonnage d'importance* (*importance sampling*) : on peut récrire, pour une densité γ bien choisie,

$$I = \int \frac{h(\theta)p(\theta)}{\gamma(\theta)} \gamma(\theta) \, \mathrm{d}\theta = \mathbb{E}_\gamma \left[\frac{h(\theta)p(\theta)}{\gamma(\theta)} \right]$$

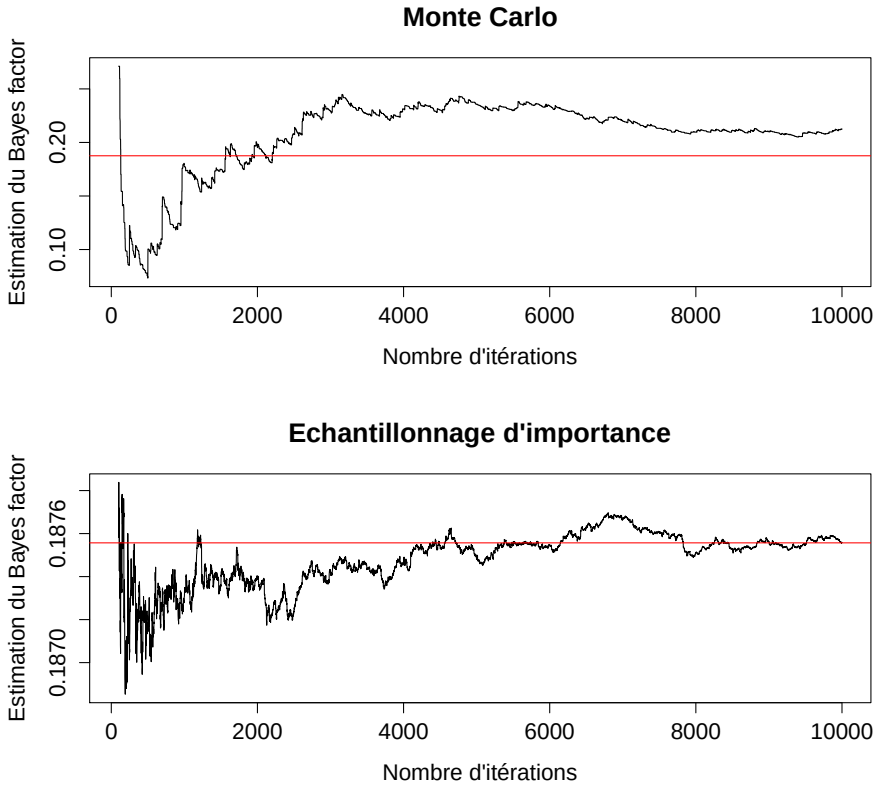

FIG. 11.2 : Estimation d'un Bayes factor par Monte-Carlo et par Échantillonage d'importance (échantillonnage préférentiel). En rouge, la valeur analytique. Notez que l'échelle des ordonnées n'est pas la même sur les deux figures.

et on reprend l'idée précédente : on simule $\theta'_1, \ldots, \theta'_T \sim \gamma$ et on pose

$$\hat{I}_T^{\text{IS}} = \frac{1}{T} \sum_{t=1}^{T} \frac{h(\theta_t) p(\theta_t)}{\gamma(\theta_t)}.$$

Comment choisir γ ? Il faut bien sûr que $\gamma(\theta)$ soit non nulle dès que $p(\theta)$ est non nulle, et que γ soit facile à simuler. De plus, comme

$$\text{var}(\hat{I}_T^{\text{IS}}) = \frac{1}{T} \text{var}_\gamma \left(\frac{h(\theta) p(\theta)}{\gamma(\theta)} \right)$$

on va chercher à minimiser cette variance, ce qui revient à prendre $\gamma \approx hp$ à une constante multiplicative près.

EXEMPLE 11.7 Reprenons le calcul du *Bayes factor* de la section précédente. Pour nos données ($n = 577$), le calcul analytique montre que le *Bayes factor* vaut environ 0.1876 ; on va chercher à l'estimer par Monte-Carlo et par échantillonnage d'importance. Pour la loi d'importance γ, on choisit la loi gaussienne de paramètres l'espérance et la variance a posteriori : il s'agit bien d'une loi aisée à simuler, et on s'attend à ce quelle amène une faible variance.

La figure 11.7 montre l'évolution des estimateurs $\hat{I}_T^{\text{MC}}$ et $\hat{I}_T^{\text{IS}}$, en fonction du nombre d'itérations T ; la ligne horizontale rouge correspond à la valeur analytique. Prenez garde à l'axe des ordonnées : en une centaine d'itérations, l'échantillonnage d'importance donne une estimation avec une erreur inférieure à 10^{-3}, précision que le Monte-Carlo n'a toujours pas atteinte en 10 000 itérations.

11.6 ALGORITHME DE METROPOLIS-HASTINGS

Lorsque les méthodes de Monte-Carlo et d'échantillonnage d'importance ne peuvent pas être appliquées (par exemple parce que la densité p dépend d'une constante de normalisation inconnue), on peut

avoir recours au Markov Chain Monte Carlo (MCMC). L'idée est de créer une chaîne de Markov (Z_t) de loi stationnaire p. Pour peu qu'on puisse garantir la convergence en loi $Z_t \xrightarrow[t\to\infty]{\mathcal{L}} p$, on aura l'estimateur asymptotiquement sans biais

$$\hat{I}_T^{\text{MCMC}} = \frac{1}{T}\sum_{t=1}^{T} h(Z_t).$$

Il existe de nombreuses manières de construire une telle chaîne de Markov (Z_t) ; nous nous contentons ici de décrire l'algorithme de Metropolis-Hastings, qui est l'un des plus génériques.

DÉFINITION 11.8 (BILAN DÉTAILLÉ/DETAILED BALANCE) Soit p une mesure de probabilité sur $\mathcal{X}$ et (Z_t) une chaîne de Markov à valeurs dans $\mathcal{X}$, de fonction de transition $K(x \to x')$. On dit que (Z_t) vérifie le *detailed balance* pour p si

$$\forall x, x' \in \mathcal{X},\ p(x)K(x \to x') = p(x')K(x' \to x).$$

On peut aussi dire que (Z_t) est *réversible*.

THÉORÈME 11.9 *Si la chaîne de Markov (Z_t) vérifie le bilan détaillé pour la loi p, alors p est une loi stationnaire pour (Z_t).*

PREUVE. On donne la preuve dans le cas où $\mathcal{X}$ est discret ; elle s'adapte aisément.

Supposons que $Z_t \sim p$. Alors

$$P[Z_{t+1} = i] = \sum_j P[Z_t = j]K(j \to i) = \sum_j p(j)K(j \to i) = \sum_j p(i)K(i \to j) = p(i)$$

et on a donc $Z_{t+1} \sim p$. □

DÉFINITION 11.10 (ALGORITHME DE METROPOLIS-HASTINGS) Soit p une mesure de probabilité sur $\mathcal{X}$. On se donne un noyau de proposition $q(\cdot|\cdot)$ sur $\mathcal{X}$, c'est à dire que $\forall x \in \mathcal{X}$, $q(\cdot|x)$ est une densité de probabilité sur $\mathcal{X}$. On se donne une valeur initiale Z_0 arbitraire. On appelle algorithme de Metropolis-Hastings l'algorithme dont l'itération $t+1$ est donnée par :

i) Tirer $Y_{t+1} \sim q(\cdot|z_t)$;
ii) Poser la probabilité d'acceptation

$$\alpha_t = \min\left(1, \frac{p(y_{t+1})}{p(z_t)}\frac{q(z_t|y_{t+1})}{q(y_{t+1}|z_t)}\right)$$

iii) Poser

$$Z_{t+1} = \begin{cases} Y_{t+1} & \text{avec probabilité } \alpha_t \\ Z_t & \text{avec probabilité } 1-\alpha_t \end{cases}$$

THÉORÈME 11.11 *La chaîne de Markov définie par l'algorithme de Metropolis-Hastings vérifie le bilan détaillé pour p.*

PREUVE. Notons K la fonction de transition de la chaîne (Z_t). Soient $x, x' \in \mathcal{X}$. Étudions la transition $x \to x'$, *i.e.* $Z_t = x$ et $Y_{t+1} = x'$. Par symétrie, on peut supposer $\alpha_t \leq 1$ (ce qui signifie que si on avait proposé la transition $x' \to x$, la probabilité d'acceptation aurait été 1). On a

$$\frac{K(x \to x')}{K(x' \to x)} = \frac{q(x'|x)\alpha_t}{q(x|x')} = \frac{p(x')}{p(x)}$$

et le bilan détaillé est bien vérifié. □

Pour peu qu'on ait choisi q de façon à ce que (Z_t) soit irréductible, apériodique et positive récurrente, ce qui est aisé en pratique, le théorème ergodique garantit que l'estimateur $\hat{I}_T^{\text{MCMC}}$ converge en probabilité. Étudions sa variance :

$$\operatorname{var}\left(\hat{I}_T^{\text{MCMC}}\right) = \frac{1}{T}\operatorname{var}(h(Z)) + \frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{t-1}\operatorname{cov}\left(h(Z_t), h(Z_{t'})\right).$$

Comme Z_t converge en loi, on a à l'asymptote $\operatorname{cov}\left(h(Z_t), h(Z_{t'})\right) = \operatorname{cov}\left(h(Z_{t+k}), h(Z_{t'+k})\right)$ pour tout k et chaque terme apparaît donc environ T fois dans la double somme, d'où

$$\begin{aligned}
\operatorname{var}\left(\hat{I}_T^{\text{MCMC}}\right) &\approx \frac{1}{T}\operatorname{var}(h(Z)) + \frac{2}{T}\sum_{k=1}^{T}\operatorname{cov}\left(h(Z_t), h(Z_{t+k})\right) \text{ pour un } t \text{ arbitraire suffisamment grand} \\
&\approx \operatorname{var}(h(Z))\left(\frac{1}{T} + \frac{2}{T}\sum \rho_k\right) \text{ où } \rho_k = \operatorname{cor}(h(Z_t), h(Z_{t+k})) \\
&\approx \frac{1}{M}\operatorname{var}(h(Z))
\end{aligned}$$

où $M = \frac{T}{1+2\sum \rho_k}$ est la taille d'échantillon effective (effective sample size, ESS) : notre estimateur MCMC après T itérations $\hat{I}_T^{\text{MCMC}}$ est de même qualité qu'un estimateur Monte-Carlo $\hat{I}_M^{\text{MC}}$ après M itérations (avec $M \leq T$).

Ceci nous donne une heuristique pour choisir une bonne proposition q : on cherche à minimiser l'autocorrélation de la chaîne (Z_t), c'est-à-dire la corrélation entre Z_t et Z_{t+k}, pour tout k.

Prenons l'exemple simple où $q(\cdot|x) = \mathcal{N}(x, \sigma^2)$ et où cherche à optimiser en σ. Si $\sigma \to 0$, alors on propose une valeur Y_{t+1} très proche de Z_t (et qui sera acceptée avec une probabilité α_t proche de 1), donc l'auto-corrélation sera proche de 1. Si $\sigma \to \infty$, alors on propose une valeur Y_{t+1} très éloignée de Z_t, qui ne sera généralement pas acceptée ($\alpha_t \approx 0$) : on a une forte probabilité que $Z_{t+1} = Z_t$, donc l'auto-corrélation sera encore proche de 1. Il existe donc une valeur optimale finie de σ qui minimise l'auto-corrélation. Dans un cas simple, on peut montrer que l'optimum est atteint si $E[\alpha_t] = 0.234$; l'expérience montre que des valeurs proches de 0.234 fonctionnent également bien dans un grand nombre de cas plus complexes.

11.7 REMARQUES BIBLIOGRAPHIQUES

L'ouvrage [4] constitue un plaidoyer (vibrant) pour les méthodes bayésiennes.

La version générale du théorème de Bernstein-von Mises (valable pour les modèles différentiables en moyenne quadratique) est due à Le Cam. On en trouve une démonstration dans [5].

En statistique non-paramétrique, l'étude fréquentiste des méthodes bayésiennes est un sujet très actif [1].

Les méthodes de Monte-Carlo jouent un rôle central en probabilités numériques [3]. Les propriétés de mélange des chaines de Markov sont étudiées dans [2].

Références

[1] S. Ghosal et A. van der Vaart. **Fundamentals of nonparametric Bayesian inference**. T. 44. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2017, p. xxiv+646. ISBN : 978-0-521-87826-5. MR : 3587782.

[2] D. A. Levin, Y. Peres et E. L. Wilmer. **Markov chains and mixing times**. With a chapter by James G. Propp and David B. Wilson. American Mathematical Society, Providence, RI, 2009, p. xviii+371. ISBN : 978-0-8218-4739-8. MR : 2466937.

[3] J. S. Liu. **Monte Carlo strategies in scientific computing**. Springer Series in Statistics. Springer-Verlag, New York, 2001, p. xvi+343. ISBN : 0-387-95230-6. MR : 1842342.

[4] C. Robert. **The Bayesian choice**. Second. Springer Texts in Statistics. From decision-theoretic foundations to computational implementation, Translated and revised from the French original by the author. New York : Springer-Verlag, 2001, p. xxiv+604. ISBN : 0-387-95231-4. MR : MR1835885(2002c:62040).

[5] A. van der Vaart. **Asymptotic statistics**. Cambridge University Press, 1998.

12.1 PROBLÈME

En statistique dite paramétrique, on suppose que l'on dispose d'un bon modèle du phénomène étudié. Par exemple, lorsqu'on plaque un modèle linéaire gaussien sur une étude (voir par exemple notre exploration des données `whiteside`), on postule qu'une certaine distribution (celle du bruit) possède une forme (très) particulière (gaussienne). Ce genre de postulat peut être plus ou moins motivé : s'il y a de bonnes raisons de penser que le bruit est une somme de petites perturbations indépendamment distribuées, le TCL justifie l'hypothèse gaussienne. Il peut être aussi motivé de façon discutable : on postule un bruit gaussien parce que cela facilite les calculs !

Il est aussi des situations où on se refuse à postuler un modèle paramétrique particulier. Par exemple, en statistique du risque (environnement, assurance, ...), on cherche à étudier les caractéristiques des queues de distribution sous la seule hypothèse que les fonctions de répartition ou les fonctions quantiles possèdent des propriétés de *variation régulière* (par exemple $\lim_{t\to\infty} \overline{F}(tx)/\overline{F}(t)$ existe pour tout $x > 0$). Une partie des polémiques à propos de l'usage des mathématiques en finance portait sur l'usage des modèles gaussiens pour décrire l'évolution des cours d'une variété d'actifs. On souhaite souvent effectuer une modélisation minimaliste, par exemple supposer simplement que les données sont i.i.d. selon une loi qui admet une densité. Le problème est alors d'estimer cette densité. Il s'agit d'un problème beaucoup plus difficile que l'estimation de la fonction de répartition.

Le problème de l'estimation de densité (en dimension 1) est défini ainsi : étant donné un échantillon $X_1, \dots, X_n$ de tirages indépendants selon une loi (inconnue) de fonction de répartition F et de densité f (c'est à dire une fonction positive intégrable, d'intégrale égale à 1), estimer la fonction f. Nous noterons nos estimateurs $\widehat{f}_n$.

12.2 FONCTIONS DE PERTE

Pour apprécier la qualité d'un estimateur, il faut choisir une fonction de perte L. Comme les densités appartiennent à $L_1(\mathbb{R})$, il est naturel d'utiliser la distance L_1 pour quantifier la perte subie en estimant f par $\widehat{f}_n$:

$$L(\widehat{f}_n, f) := \|f - \widehat{f}_n\|_1 = \int_{\mathbb{R}} |f(x) - \widehat{f}_n(x)| \mathrm{d}x \,.$$

Ce n'est pas la seule façon de faire, il existe une littérature abondante où on s'intéresse à la perte quadratique :

$$\int_{\mathbb{R}} |f(x) - \widehat{f}_n(x)|^2 \mathrm{d}x \,.$$

On peut aussi s'intéresser à la distance de Hellinger, ou plutôt à son carré :

$$\int_{\mathbb{R}} \left| \sqrt{f(x)} - \sqrt{\widehat{f}_n(x)} \right|^2 \mathrm{d}x \,.$$

La perte L_1 possède un grand mérite. Comme c'est le double de la distance en variation, on réalise immédiatement qu'elle est invariante par transformation injective des données. En particulier, elle est insensible aux changements d'échelle et aux translation. Nous nous concentrerons sur cette perte.

LEMME 12.1 *Soient P, Q deux lois de probabilité sur un univers $(\Omega, \mathcal{F})$, absolument continues par rapport à une mesure σ-finie ν. Soient f et g des dérivées de Radon-Nikodym de P et Q par rapport à ν,*

$$\sup_{A\in\mathcal{F}} P(A) - Q(A) =: \mathrm{d}_{\mathrm{TV}}(P, Q) = \frac{1}{2} \int_\Omega |f(x) - g(x)| \, \mathrm{d}\nu(x) = \int_\Omega (f(x) - g(x))_+ \, \mathrm{d}\nu(x)$$

PREUVE. Comme $\int_\Omega f(x)\mathrm{d}\nu(x) = \int_\Omega g(x)\mathrm{d}\nu(x) = 1$, la seconde égalité est immédiate.

Si on définit A par $A := \{x : f(x) > g(x)\}$,

$$P(A) - Q(A) = \int_A f(x) - g(x) \mathrm{d}\nu(x) = \int_\Omega (f(x) - g(x))_+ \mathrm{d}\nu(x)\,.$$

On peut donc déjà conclure que

$$\mathrm{d}_{\mathrm{TV}}(P, Q) \geq \int_\Omega (f(x) - g(x))_+ \,\mathrm{d}\nu(x)\,.$$

Soit B une partie mesurable

$$\begin{aligned} P(B) - Q(B) &= \int_{A\cap B} f(x) - g(x) \mathrm{d}\nu(x) + \int_{A^c\cap B} f(x) - g(x) \mathrm{d}\nu(x) \\ &= \int_{A\cap B} (f(x) - g(x))_+ \mathrm{d}\nu(x) - \int_{A^c\cap B} (f(x) - g(x))_- \mathrm{d}\nu(x) \\ &\leq \int_A (f(x) - g(x))_+ \mathrm{d}\nu(x) \\ &= \int_\Omega (f(x) - g(x))_+ \mathrm{d}\nu(x)\,. \end{aligned}$$

Soit

$$\mathrm{d}_{\mathrm{TV}}(P, Q) \leq \int_\Omega (f(x) - g(x))_+ \,\mathrm{d}\nu(x)\,.$$

□

LEMME 12.2 (LEMME DE SCHEFFÉ) *Soient (P_n) une suite de lois absolument continues par rapport à une mesure σ-finie ν, et $(f_n)_n$ une suite de densités associées. Si la suite (f_n) converge simplement vers la (une) densité f d'une loi P, alors la suite (P_n) converge vers P au sens de la distance en variation (la suite $(f_n)_n$ converge vers f dans $L_1(\nu)$).*

PREUVE. D'après le lemme précédent, il suffit de vérifier que

$$\lim_{n\to\infty} \int_\Omega (f(x) - f_n(x))_+ \mathrm{d}\nu(x) = 0\,.$$

Mais d'une part $(f - f_n)_+ \leq f$ qui est intégrable et en tout x $\lim_n (f(x) - f_n(x))_+ = 0$. Il est donc possible de conclure à l'aide du théorème de convergence dominée. □

12.3 ESTIMATION PAR HISTOGRAMMES

La plus simple des méthodes d'estimation de densité, consiste à estimer la densité par une densité constante par morceaux. Un estimateur par histogramme est défini par une partition de $\mathbb{R}$ en intervalles. Les intervalles sont définis par $b_1 \leq \ldots \leq b_k$. Pour que la construction soit sans défauts, on se place dans le cas où aucun point de l'échantillon n'appartient à $(-\infty, b_1] \cup [b_k, \infty) \cup \{b_2, \ldots, b_{k-1}\}$. La densité estimée est nulle sur $(-\infty, b_1]$ et sur $[b_k, +\infty)$ sur $]b_j, b_{j+1})$,

$$\widehat{f}_n(x) := \frac{F_n(b_{j+1}) - F_n(b_j)}{b_{j+1} - b_j}\,.$$

Cette définition convient lorsqu'on peut supposer que la densité à estimer est supportée par $[b_1, b_k]$. Très souvent, on se concentre sur des histogrammes réguliers : $b_{j+1} - b_j = h$ pour $1 \leq j < k$.

Quoiqu'il en soit, si la partition est bien choisie avant la collecte des données, l'estimateur par histogramme est un *estimateur par projection* : la loi obtenue est la projection orthogonale de la mesure empirique sur le sous-espace fermé des lois à densité constante sur les classes de la partition.

Cette interprétation de l'estimation par histogramme comme estimateur par projection, explique que la théorie L_2 (avec perte quadratique) de l'estimation par histogramme soit accessible.

Les histogrammes sont une des techniques de base de la statistique descriptive. Dans cette spécialité aussi, ils trouvent leurs limites : l'allure de l'histogramme obtenu peut dépendre subtilement d'un décalage de la partition.

Nous allons introduire les méthodes de noyau par la méthode des fenêtres glissantes. Comme pour les histogrammes, la densité estimée est une fonction constante par morceaux. Mais dans le cas des fenêtres glissantes les morceaux dépendent des points de l'échantillon.

12.4 FENÊTRES GLISSANTES

Un résultat de consistance

Dans la suite, un noyau sera une fonction intégrable K ($\int_{\mathbb{R}} |K(x)|\mathrm{d}x < \infty$), on dira qu'il est positif si $K(x) \geq 0$ en tout x, et un noyau positif est une densité si $\int_{\mathbb{R}} K(x)\mathrm{d}x = 1$. Pour la largeur de bande $h > 0$, $K_h(x) := \frac{1}{h}K(\frac{x}{h})$. On a bien sûr $\int_{\mathbb{R}} K_h(x)\mathrm{d}x = \int_{\mathbb{R}} K(x)\mathrm{d}x$.

La méthode des fenêtres glissantes est définie à partir d'un *noyau*, appelé *noyau rectangulaire*

$$K(x) := \frac{1}{2}\mathbb{I}_{[-1,1]}(x)$$

qui n'est autre que la densité de la loi uniforme sur l'intervalle $[-1, 1]$. Pour chaque *largeur de bande* $h > 0$, on définit le noyau K_h (densité de la loi uniforme sur $[-h, h]$) par

$$K_h(x) := \frac{1}{2h}\mathbb{I}_{[-1,1]}(x/h)\,.$$

L'estimateur de densité défini par le noyau rectangulaire et la largeur de bande h est

$$\widehat{f}_n(x) := \frac{1}{2h}\left(F_n(x+h) - F_n(x-h)\right)\,.$$

Dans la suite, la largeur de bande notée h_n dépendra généralement de la taille de l'échantillon n et on conviendra de

$$\widehat{f}_n(x) := \frac{1}{2h_n}\left(F_n(x+h_n) - F_n(x-h_n)\right)\,.$$

Le théorème fondamental du calcul infinitésimal nous rappelle que si F est une fonction de répartition absolument continue (fonction de répartition d'une loi absolument continue), de densité f pour presque tout $x \in \mathbb{R}$,

$$\lim_{h\searrow 0} \frac{F(x+h) - F(x-h)}{2h} = f(x)$$

et

$$F(x+h) - F(x-h) = \int_{x-h}^{x+h} f(y)\mathrm{d}y\,.$$

Cela nous suffira par la suite. Si on veut étendre les résultats qui suivent à l'estimation par noyaux en dimension $d \geq 2$, ou en dimension 1, à des noyaux plus généraux, il est utile d'avoir en tête le théorème suivant.

THÉORÈME 12.3 (THÉORÈME DE DENSITÉ DE LEBESGUE) *Soit c un réel (strictement) positif. Soit $\mathcal{Q}$ une collection des boréliens de $\mathbb{R}^d$, tels que pour tout $Q \in \mathcal{Q}$*

$$\frac{\mathrm{Leb}(Q^*)}{\mathrm{Leb}(Q)} \leq c$$

où Q^* est le plus petit cube (boule L_∞) contenant Q. Soit $\mathcal{Q}_r$ le sous-ensemble des éléments de $\mathcal{Q}$, de mesure inférieure à $r > 0$.

Soit f une densité sur $\mathbb{R}^d$. Pour presque tout $x \in \mathbb{R}^d$,

$$\lim_{r\searrow 0} \sup_{Q\in\mathcal{Q}_r} \left|\frac{1}{\mathrm{Leb}(Q)}\int_{x+Q} f(y)\mathrm{d}y - f(x)\right| = 0\,.$$

L'ensemble des points où la convergence a lieu est appelé l'ensemble des points de Lebesgue de f.

Les boules euclidiennes, les boules $\ell_p, p \geq 1$ appartiennent à $\mathcal{Q}_r$ pour un choix de r qui peut dépendre de la dimension d et de p. L'ensemble de tous les rectangles de $\mathbb{R}^d$ n'est inclus dans aucun $\mathcal{Q}_r$ avec r fixé.

En dimension 1, si x est un point de Lebesgue de f, et K dénote le noyau rectangulaire,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} \int_{x-h}^{x+h} f(y) \mathrm{d}y = \lim_{h \to 0} (K_h * f)(x) \,. \tag{12.1}$$

PROPOSITION 12.4 *Soit f la densité d'une loi P sur $\mathbb{R}$, si x est un point de Lebesgue, et si la suite des largeurs de bande $(h_n)_n$ vérifie $\lim_n h_n = 0$ et $\lim_n nh_n = \infty$, alors*

$$\lim_{n \to \infty} \mathbb{E}_{P^{\otimes n}} \left[(\widehat{f}_n(x) - f(x))^2 \right] = 0 \,.$$

PREUVE. Le risque quadratique de $\widehat{f}_n(x)$ admet une décomposition biais/variance.

$$\mathbb{E}_{P^{\otimes n}} \left[(\widehat{f}_n(x) - f(x))^2 \right] = \mathrm{var} \left(\widehat{f}_n(x) \right) + \left(\mathbb{E} \left[\widehat{f}_n(x) \right] - f(x) \right)^2 .$$

La variance est traitée en notant que $\widehat{f}_n(x)$ est une moyenne de variables aléatoires indépendantes identiquement distribuées comme $K_{h_n}(x - X)$,

$$\mathrm{var} \left(\widehat{f}_n(x) \right) = \frac{1}{n} \mathrm{var} \left(K_{h_n}(x - X) \right) .$$

La variable aléatoire $K_{h_n}(x - X)$ est la multiplication par $1/(2h_n)$ d'une Bernoulli de paramètre $F(x + h_n) - F(x - h_n)$, d'où

$$\mathrm{var} \left(\widehat{f}_n(x) \right) = \frac{1}{2nh_n} \frac{F(x + h_n) - F(x - h_n)}{2h_n} \left(1 - F(x + h_n) - F(x - h_n) \right) .$$

Comme x est supposé être un point de Lebesgue, si $\lim_n h_n = 0$,

$$\mathrm{var} \left(\widehat{f}_n(x) \right) \sim \frac{f(x)}{2nh_n} \qquad \text{quand } n \to \infty \,.$$

Si de plus $\lim_n nh_n = \infty$, la variance tend vers 0.

Comme $\mathbb{E}\widehat{f}_n(x) = K_{h_n} * f(x)$, le biais s'écrit

$$\mathbb{E} \left[\widehat{f}_n(x) \right] - f(x) = K_{h_n} * f(x) - f(x) = \frac{F(x + h_n) - F(x - h_n)}{2h_n} - f(x)$$

qui tend vers 0 lorsque x est un point de Lebesgue et $\lim_n h_n = 0$. □

La méthode des fenêtres glissantes est universellement consistante.

THÉORÈME 12.5 *Soit f la densité d'une loi P sur $\mathbb{R}$, si la suite des largeurs de bande $(h_n)_n$ vérifie $\lim_n h_n = 0$ et $\lim_n nh_n = \infty$, alors*

$$\lim_{n \to \infty} \mathbb{E}_{P^{\otimes n}} \left[\int_{\mathbb{R}} \left| \widehat{f}_n(x) - f(x) \right| \mathrm{d}x \right] = 0 \,.$$

PREUVE.

$$\begin{aligned} \mathbb{E}_{P^{\otimes n}} \left[\int_{\mathbb{R}} \left| \widehat{f}_n(x) - f(x) \right| \mathrm{d}x \right] &= 2\mathbb{E}_{P^{\otimes n}} \left[\int_{\mathbb{R}} \left(f(x) - \widehat{f}_n(x) \right)_+ \mathrm{d}x \right] \\ &= 2 \int_{\mathbb{R}} \mathbb{E}_{P^{\otimes n}} \left[\left(f(x) - \widehat{f}_n(x) \right)_+ \right] \mathrm{d}x \,. \end{aligned}$$

En tout x,

$$\mathbb{E}_{P^{\otimes n}}\left[\left(f(x)-\widehat{f}_n(x)\right)_+\right] \leq f(x)\,.$$

Comme

$$\mathbb{E}_{P^{\otimes n}}\left[\left(f(x)-\widehat{f}_n(x)\right)_+\right] \leq \sqrt{\mathbb{E}_{P^{\otimes n}}\left[\left(f(x)-\widehat{f}_n(x)\right)^2\right]}\,,$$

en tout point de Lebesgue et donc presque partout,

$$\lim_n \mathbb{E}_{P^{\otimes n}}\left[\left(f(x)-\widehat{f}_n(x)\right)_+\right] = 0\,.$$

Le théorème s'ensuit par convergence dominée. □

Remarque 12.6 Ce résultat de consistance universelle doit être tempéré. La décroissance du risque peut être arbitrairement lente. La preuve souligne le fait que nous ne maitrisons absolument pas la décroissance du biais. La manière dont celui-ci tend vers 0 dépend de la densité à estimer (qui est inconnue) et du choix des largeurs de bande. Le choix des largeurs de bande pour optimiser le compromis biais/variance reste un problème difficile.

12.5 noyaux et outils

Définition

La méthode des fenêtres glissantes est un exemple de la méthode des noyaux de Parzen et Rosenblatt. Dans la suite K est une fonction intégrable, pas toujours une densité de probabilité. Elle n'est pas nécessairement positive. La forme générale de l'estimateur à noyau est la suivante.

Définition 12.7 (estimateur à noyau) Soit K une fonction intégrable, $h>0$ une largeur de bande, l'estimateur à noyau défini par K et h est

$$\widehat{f}_n(x) := \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\left(\frac{x-X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^n K_h\left(x-X_i\right)\,.$$

Outre le noyau rectangulaire qui sous-tend la méthode des fenêtres glissantes, on peut citer quelques noyaux fréquemment utilisés :

i) gaussien, $K=\phi$.
ii) Epanechnikov, $K(x)=\frac{3}{4}(1-x^2)_+$
iii) Silverman, $K(x)=\frac{1}{2}\mathrm{e}^{-|u|/\sqrt{2}}\sin\left(\frac{|u|}{\sqrt{2}}+\frac{\pi}{4}\right)$.

Convolution

Un estimateur à noyau est le résultat de la convolution de la mesure empirique avec le noyau. Nous allons ici rappeler quelques propriétés de la convolution des mesures et des fonctions.

Définition 12.8 (convolution de deux fonctions) Si f et g sont deux fonctions intégrables, la convolution de f et g, notée $f*g$ est une fonction définie par

$$f*g(x) := \int_{\mathbb{R}} f(x-y)g(y)\mathrm{d}y = \int_{\mathbb{R}} f(y)g(x-y)\mathrm{d}y\,.$$

La fonction $f*g$ est elle-même intégrable.

Théorème 12.9 (inégalité de young L_1) *Soit f,g deux fonctions intégrables, alors $f*g$ est intégrable et*

$$\|f*g\|_1 \leq \|f\|_1 \times \|g\|_1\,.$$

PREUVE.

$$\begin{aligned}
\|f*g\|_1 &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} f(x-y)\times g(y)\mathrm{d}y \right| \mathrm{d}x \\
&\leq \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x-y)\times g(y)|\,\mathrm{d}y\mathrm{d}x \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x-y)|\,\mathrm{d}x \times |g(y)|\mathrm{d}y \\
&= \int_{\mathbb{R}} \|f\|_1 \times |g(y)|\mathrm{d}y \\
&= \|f\|_1 \times \|g\|_1 \,.
\end{aligned}$$

□

REMARQUE 12.10 Cette inégalité est un cas facile de l'inégalité de Young générale. Une version à peine plus difficile affirme que pour $p \geq 1$,

$$\|f*g\|_p \leq \|f\|_p \times \|g\|_1 \,.$$

La convolution est une opération régularisante.

THÉORÈME 12.11 *Si f est k fois dérivable et à support compact, et $g \in L_1$ alors $g*f$ est k fois dérivable.*

Voir [1, Proposition 4.20].

Les arguments de convolution permettent d'établir le théorème de densité suivant.

THÉORÈME 12.12 *L'ensemble des fonctions infiniment différentiables à support compact est dense dans L_1.*

Voir [1, Corollaire 4.23].

On peut convoluer un noyau et une loi de probabilité qui n'a pas nécessairement de densité (par exemple avec une loi empirique).

DÉFINITION 12.13 (CONVOLUTION D'UN NOYAU ET D'UNE LOI DE PROBABILITÉ) Soit F la fonction de répartition d'un loi de probabilité μ sur $\mathbb{R}$ et K un noyau, alors

$$K*\mu(x) := \int_{\mathbb{R}} K(x-y)\mathrm{d}F(y) \,.$$

Si K est une densité, $K*\mu$ est la densité de la loi de $X+Y$ où X, de densité K et Y, de loi μ sont indépendantes.

Si P_n désigne la loi empirique $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{X_i \in A}$, alors l'estimateur de densité défini par le noyau K et la largeur de bande h s'écrit

$$\widehat{f}_n(x) = P_n * K_h(x) \,.$$

12.6 CONSISTANCE UNIVERSELLE DES MÉTHODES DE NOYAU

Le résultat de consistance établi pour les fenêtres glissantes n'est pas un accident lié au noyau rectangulaire. Sous les mêmes conditions sur les largeurs de bande, on obtient la consistance universelle pour une large classe de noyaux.

Un estimateur à noyau est en général biaisé. Pour tout $x \in \mathbb{R}$:

$$\mathbb{E}\left[\widehat{f}_n(x)\right] = K_{h_n} * f(x) = \int_{\mathbb{R}} \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) f(y)\mathrm{d}y\,.$$

Si K est une densité et que h est petit, $K_{h_n} * f(x)$ est une moyenne de f autour de x. Si f est assez régulière/lisse, on s'attend à ce que cette moyenne soit proche de f.

Le théorème suivant nous dit que si la largeur de bande tend vers 0, sous des hypothèses minimales, le biais intégré tend vers 0.

THÉORÈME 12.14 (LIMITES DE CONVOLUÉS) *Si le noyau positif $K \in L_1$ avec $\int_{\mathbb{R}} |K(x)|\mathrm{d}x = 1$ et $K_h(x) := \frac{1}{h}K(\frac{x}{h})$, alors pour tout $f \in L_1$,*

$$\lim_{h\to 0} \int_{\mathbb{R}} |f * K_h(x) - f(x)|\,\mathrm{d}x = 0\,.$$

Pour établir ce théorème, nous partirons des deux propositions suivantes :

PROPOSITION 12.15 *Si le noyau positif $K \in L_1$ avec $\int_{\mathbb{R}} |K(x)|\mathrm{d}x = 1$ et $K_h(x) := \frac{1}{h}K(\frac{x}{h})$, alors pour toute fonction g lipschtzienne et à support borné*

$$\lim_{h\to 0} \int_{\mathbb{R}} |g * K_h(x) - g(x)|\,\mathrm{d}x = 0\,.$$

PROPOSITION 12.16 *Les fonctions lipschtziennes et à support borné sont denses dans L_1.*

La proposition 12.16 est un corollaire du théorème 12.12.

PREUVE.(Proposition 12.15) Dans la suite g est C-lipschtzienne à support compact inclus dans $[-M, M]$.

On définit une noyau tronqué L_h par

$$L_h = K_h \times \mathbb{I}_{[-rh,rh]}$$

pour $r \in \mathbb{R}^+$ que l'on pourra régler selon nos besoins.

Le biais peut se majorer par une somme

$$\begin{aligned}
&\int_{\mathbb{R}} |g * K_h(x) - g(x)|\,\mathrm{d}x \\
&\quad\leq \underbrace{\int_{\mathbb{R}} |g * (K_h - L_h)(x)|\,\mathrm{d}x}_{(a)} + \\
&\qquad \underbrace{\int_{\mathbb{R}} \left|g * L_h(x) - g(x)\int_{\mathbb{R}} L_h(y)\mathrm{d}y\right|\mathrm{d}x}_{(b)} + \\
&\qquad \underbrace{\int_{\mathbb{R}} \left|g(x)\int_{\mathbb{R}} (K_h(y) - L_h(y))\mathrm{d}y\right|\mathrm{d}x}_{(c)}\,.
\end{aligned}$$

D'après l'inégalité de Young,

$$(a) \quad \leq \quad \|g\|_1 \times \|K_h - L_h\|_1 \ ,$$

alors que

$$\begin{aligned}(c) \quad &\leq \quad \int_{\mathbb{R}} |g(x)| \int_{\mathbb{R}} |K_h(y) - L_h(y)| \, \mathrm{d}y \mathrm{d}x \\ &\leq \quad \int_{\mathbb{R}} |g(x)| \, \mathrm{d}x \times \int_{\mathbb{R}} |K_h(y) - L_h(y)| \, \mathrm{d}y \\ &= \quad \|g\|_1 \times \|K_h - L_h\|_1 \ .\end{aligned}$$

Comme

$$\begin{aligned}\|K_h - L_h\|_1 \quad &\leq \quad \int_{\mathbb{R}} |K(x)| \mathbb{I}_{|x|>r} \mathrm{d}(x) \, , \\ (a) + (c) \leq 2 \, \|g\|_1 \times &\int_{\mathbb{R}} |K(x)| \mathbb{I}_{|x|>r} \mathrm{d}(x) \, .\end{aligned}$$

Le membre droit peut être rendu arbitrairement petit en choisissant r assez grand.

Pour majorer (b), nous allons utiliser le fait que g est à support compact et que le noyau L_h est aussi à support compact.

$$\begin{aligned}(b) \quad &= \quad \int_{\mathbb{R}} \left| \int_{\mathbb{R}} (g(x-y) - g(x)) L_h(y) \mathrm{d}(y) \right| \mathrm{d}x \\ &\leq \quad \int_{\mathbb{R}} \int_{-M-rh}^{M+rh} |(g(x-y) - g(x))| \mathrm{d}x L_h(y) \mathrm{d}(y) \\ &\leq \quad \int_{-rh}^{rh} \int_{-M-rh}^{M+rh} C|y| \mathrm{d}x |L_h(y)| \mathrm{d}(y) \\ &\leq \quad 2(M+rh)Crh \int_{-r}^{r} |L(y)| \mathrm{d}(y) \\ &\leq \quad 2(M+rh)Crh \int_{\mathbb{R}} |K(y)| \mathrm{d}(y) \, .\end{aligned}$$

En combinant les majorations de $(a), (b)$ et (c),

$$\limsup_{h \to 0} \int_{\mathbb{R}} |g * K_h(x) - g(x)| \, \mathrm{d}x \quad \leq \quad 2 \, \|g\|_1 \times \int_{\mathbb{R}} |K(x)| \mathbb{I}_{|x|>r} \mathrm{d}(x) \, .$$

En faisant tendre r vers ∞, on conclut

$$\limsup_{h \to 0} \int_{\mathbb{R}} |g * K_h(x) - g(x)| \, \mathrm{d}x = 0 \, .$$

□

L'étude de l'approximation des fonctions intégrables devient alors facile.

PREUVE.(Théorème 12.14) Dans la preuve f est intégrable, et g est Lipschtzienne à support compact.

$$\int_{\mathbb{R}} |f * K_h(x) - f(x)| \, \mathrm{d}x \quad \leq \quad \underbrace{\int_{\mathbb{R}} |(f-g) * K_h(x)| \, \mathrm{d}x}_{(a)} + \underbrace{\int_{\mathbb{R}} |g * K_h(x) - g(x)| \, \mathrm{d}x}_{(b)} + \underbrace{\int_{\mathbb{R}} |g(x) - f(x)| \, \mathrm{d}x}_{(c)} \ .$$

L'inégalité de Young nous permet de majorer (a) :

$$(a) \leq \|f - g\|_1 \times \|K_h\|_1$$

et donc $(a) + (c)$ par $(1 + \|K\|_1)\|g - f\|_1$.

Si maintenant on suppose g lipschtzienne à support borné, on peut déduire de la proposition 12.15, que

$$\limsup_{h \to 0} \int_{\mathbb{R}} |f * K_h(x) - f(x)| \, \mathrm{d}x \leq (1 + \|K\|_1)\|g - f\|_1 \, .$$

Comme l'ensemble des fonctions lipschtziennes à support borné est dense dans L_1, on peut rendre le membre droit arbitrairement proche de 0. □

REMARQUE 12.17 Dans le cas du noyau rectangulaire, il n'était pas nécessaire de procéder à une troncature.

Comme dans la plupart des cas, lorsqu'on étudie des méthodes de noyau, l'analyse du biais est la partie délicate.

Pour établir la consistance universelle de l'estimation par noyaux, nous aurons besoin du résultat suivant qui étend la limite (12.1) au cas de noyaux plus généraux.

DÉFINITION 12.18 Si K est un noyau positif, intégrable, vérifiant $\int K(x)\mathrm{d}x = 1$, la famille $(K_h)_{h>0}$ est dite approximation de l'identité s'il existe $A > 0$ tel que

$$K_h(x) \leq Ah^{-1} \qquad \text{et} \qquad K_h(x) \leq A\frac{h}{x^2} \qquad \forall h > 0, x \in \mathbb{R}\,,$$

Notons que les deux dernières hypothèses impliquent l'intégrabilité.

Un noyau positif, borné, normé, à support compact définit une approximation de l'identité.

THÉORÈME 12.19 *Si K est un noyau positif, intégrable, normé, et $(K_h)_{h>0}$ définit une approximation de l'identité alors pour $f \in L_1(\mathbb{R})$, pour tout x point de Lebesgue de f*

$$\lim_{h\to 0}(f * K_h)(x) = f(x)\,.$$

THÉORÈME 12.20 (CONSISTANCE UNIVERSELLE POUR LE RISQUE L_1) *Si le noyau $K \in L_1$ est un noyau positif, intégrable, normé, qui définit une approximation de l'identité et $\kappa := \int_{\mathbb{R}} K^2(x)\mathrm{d}x < \infty$, si $h_n \to 0$ et $nh_n \to \infty$, alors pour toute densité f sur $\mathbb{R}$, en définissant $\widehat{f}_n := P_n * K_{h_n}$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\int_{\mathbb{R}} \left|\widehat{f}_n(x) - f(x)\right| \mathrm{d}x\right] = 0\,.$$

L'argument reprend l'analyse déjà effectuée à propos de la méthode des fenêtres glissantes.

PREUVE.

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathbb{R}} \left|\widehat{f}_n(x) - f(x)\right| \mathrm{d}x\right] \\
&\quad\leq \quad \mathbb{E}\left[\int_{\mathbb{R}} \left|\widehat{f}_n(x) - f * K_{h_n}(x)\right| \mathrm{d}x\right] + \mathbb{E}\left[\int_{\mathbb{R}} |f(x) - f * K_{h_n}(x)|\, \mathrm{d}x\right] \\
&\quad= \quad \mathbb{E}\left[\int_{\mathbb{R}} \left|\widehat{f}_n(x) - f * K_{h_n}(x)\right| \mathrm{d}x\right] + \int_{\mathbb{R}} |f(x) - f * K_{h_n}(x)|\, \mathrm{d}x\,.
\end{aligned}$$

Le théorème 12.14 garantit que le second terme tend vers 0 quand $h_n \to 0$.

Nous allons maintenant nous attacher à prouver que le premier terme tend vers 0 quand $nh_n \to \infty$ et $h_n \to 0$.

$$
\begin{aligned}
&\mathbb{E}\left[\int_{\mathbb{R}} \left|\widehat{f}_n(x) - f * K_{h_n}(x)\right| \mathrm{d}x\right] \\
&= 2\int_{\mathbb{R}} \mathbb{E}\left[\left(f * K_{h_n}(x) - \widehat{f}_n(x)\right)_+\right] \mathrm{d}x \\
&\leq 2\int_{\mathbb{R}} \min\left(\mathbb{E}\left[\left(f * K_{h_n}(x) - \widehat{f}_n(x)\right)_+\right], f * K_{h_n}(x)\right) \mathrm{d}x \\
&\leq 2\int_{\mathbb{R}} \min\left(\sqrt{\mathbb{E}\left[\left(f * K_{h_n}(x) - \widehat{f}_n(x)\right)^2\right]}, f * K_{h_n}(x)\right) \mathrm{d}x \\
&= 2\int_{\mathbb{R}} \min\left(\sqrt{\mathrm{var}\left(\widehat{f}_n(x)\right)}, f * K_{h_n}(x)\right) \mathrm{d}x \\
&\leq 2\int_{\mathbb{R}} \min\left(\sqrt{\mathrm{var}\left(\widehat{f}_n(x)\right)}, f(x)\right) \mathrm{d}x + 2\int_{\mathbb{R}} |f(x) - f * K_{h_n}(x)| \, \mathrm{d}x \,.
\end{aligned}
$$

Le dernier terme tend vers 0 quand h_n tend vers 0.

Pour étudier le premier terme, comme l'intégrande est dominé par la fonction intégrable f, il suffit de vérifier que l'intégrande converge simplement vers 0.

$$
\begin{aligned}
\mathrm{var}\left(\widehat{f}_n(x)\right) &\leq \frac{1}{n}\int_{\mathbb{R}} K_{h_n}(y-x)^2 f(y)\mathrm{d}y \\
&\leq \frac{1}{nh_n}\int_{\mathbb{R}} K_{h_n}^2(y-x) f(y)\mathrm{d}y \,.
\end{aligned}
$$

Comme $\kappa = \int_{\mathbb{R}} K^2(y)\mathrm{d}y$, K^2/κ est normalisé, et on peut lui appliquer le théorème 12.19

$$
\lim_{h_n \searrow 0} \frac{1}{\kappa}\int_{\mathbb{R}} K_{h_n}^2(y-x) f(y)\mathrm{d}y = f(x) \,.
$$

et même

$$
\lim_{h_n \searrow 0} \int_{\mathbb{R}} \left|\frac{1}{\kappa}\int_{\mathbb{R}} K_{h_n}^2(y-x) f(y)\mathrm{d}y - f(x)\right| \mathrm{d}x = 0 \,.
$$

Si $nh_n \nearrow \infty$, on a donc tout point de Lebesgue x, $\lim_n \mathrm{var}(\widehat{f}_n(x)) = 0$. Ce qui suffit pour conclure. □

12.7 VITESSE DE CONVERGENCE

Le résultat de consistance universelle n'est malheureusement pas accompagné d'une garantie d'uniformité.

THÉORÈME 12.21 (BORNE INFÉRIEURE SUR LE RISQUE MINIMAX) *Il n'existe pas de vitesse de convergence uniforme pour l'estimation de densité sur* $\mathbb{R}$. *Pour tout noyau* K,

$$
\sup_{f:\|f\|_1=1, f\geq 0} \inf_{h>0} \mathbb{E}\left[\int_{\mathbb{R}} |K_h * P_n(x) - f(x)| \mathrm{d}x\right] = 2 \,.
$$

Voir [2] et [3].

Pour offrir des garanties uniformes sur le risque, il faut se concentrer sur des densités assez régulières et éventuellement à support compact.

Dans la suite,

$$
f^*(x) := \sup\{f(y) : |x-y| \leq 1\} \,.
$$

Les densités qui nous intéresseront appartiendront à l'ensemble W.

Définition 12.22 (classe de densité W)

$$W := \left\{ f : f, f' \text{ absolument continues et } \int |f''| < \infty \text{ et } \int \sqrt{f^*} \ < \infty \right\}$$

Remarque 12.23 L'hypothèse d'absolue continuité de f' exclut la densité de la loi uniforme sur un intervalle.

On note $s(f)$ la longueur minimale d'un intervalle contenant le support de la densité f. Soit I un intervalle contenant le support de f, d'après l'inégalité de Cauchy-Schwarz :

$$\int_{\mathbb{R}} \sqrt{f(x)} \mathrm{d}x \leq \sqrt{|I|}\,,$$

et donc

$$\int_{\mathbb{R}} \sqrt{f(x)} \mathrm{d}x \leq \sqrt{s(f)}\,.$$

Si on note $c(f) := \int_{\mathbb{R}} |f''(x)| \mathrm{d}x$, pour $c > 0$, on définit $W_c \subset W$ par

$$W_c := \left\{ f : f \in W, \quad s(f)^2 \times c(f) < c \right\}\,.$$

Deux paramètres importants permettent de quantifier les performances d'un noyau.

$$\mathrm{sd}(K) := \int_{\mathbb{R}} x^2 K(x) \mathrm{d}x \qquad R(K) := \int_{\mathbb{R}} K(x)^2 \mathrm{d}x\,.$$

Théorème 12.24 *Soit K un noyau symétrique, borné à support inclus dans $[-1,1]$ tel que $\int_{-1}^{1} K(y)\mathrm{d}y = 1$ avec $\mathrm{sd}(K) < \infty$ et $R(K) < \infty$. Il existe une constante $\kappa(K)$ (dépendant de K), telle que pour toute densité $f \in W_c$, pour n assez grand, il existe une largeur de bande $h_n > 0$ vérifiant*

$$\mathbb{E}\left[\int_{\mathbb{R}} \left| f(x) - \widehat{f}_n(x) \right| \mathrm{d}x \right] \leq \kappa(K) \frac{c^{1/5}}{n^{2/5}}\,.$$

On peut choisir $\kappa(K) \leq 2(4\,\mathrm{sd}(K) R(K)^2)^{1/5}$.

Dans la preuve du théorème 12.24, nous utiliserons la notion suivante. Le noyau associé permet de décrire le biais de l'estimateur à noyau à l'aide de la dérivée seconde de la densité à estimer.

Définition 12.25 (noyau associé) Etant donné un noyau intégrable K, le noyau L associé à K est défini par

$$L(x) = \begin{cases} \int_x^\infty (y-x) K(y) \mathrm{d}y & \text{pour } x \geq 0 \\ L(-x) & \text{pour } x < 0\,. \end{cases}$$

Exemple 12.26 Pour le noyau rectangulaire $\frac{1}{2}\mathbb{I}_{[-1,1]}$, le noyau associé est $\frac{1}{4}(1-|x|)_+^2$.

Proposition 12.27 *Si K est un noyau intégrable et L son noyau associé, alors*

i) L est pair (symétrique) ;
ii) $\int_{\mathbb{R}} L(x)\mathrm{d}x = \int_0^\infty x^2 K(x)\mathrm{d}x$;
iii) Si K est pair, $\int_{\mathbb{R}} L(x)\mathrm{d}x = \frac{1}{2}\int_{\mathbb{R}} x^2 K(x)\mathrm{d}x$;
iv) Si K est pair, $\int_{\mathbb{R}} |L(x)|\mathrm{d}x \leq \frac{1}{2}\int_{\mathbb{R}} x^2 |K(x)|\mathrm{d}x$;

PREUVE. La parité est inscrite dans la définition.

On a

$$\begin{aligned}\int_{\mathbb{R}} L(x)\mathrm{d}x &= 2\int_0^\infty \int_0^\infty \mathbb{I}_{y>x}(y-x)K(y)\mathrm{d}y\mathrm{d}x \qquad \text{parité de } L\\ &= \int_0^\infty \int_0^y 2(y-x)\mathrm{d}xK(y)\mathrm{d}y \qquad \text{Fubini}\\ &= \int_0^\infty y^2K(y)\mathrm{d}y\,.\end{aligned}$$

Si K est pair,

$$\int_0^\infty y^2K(y)\mathrm{d}y = \frac{1}{2}\int_{\mathbb{R}} y^2K(y)\mathrm{d}y\,,$$

ce qui établit la troisième assertion.

Si K est pair, en invoquant encore la parité de L, puis le théorème de Fubini,

$$\begin{aligned}&\int_{\mathbb{R}} |L(x)|\mathrm{d}x\\ &= 2\int_0^\infty \left|\int_0^\infty \mathbb{I}_{y>x}(y-x)K(y)\mathrm{d}y\right|\mathrm{d}x\\ &\leq 2\int_0^\infty \int_0^\infty |\mathbb{I}_{y>x}(y-x)K(y)|\,\mathrm{d}y\mathrm{d}x\\ &= \int_0^\infty \int_0^y 2(y-x)\mathrm{d}x|K(y)|\mathrm{d}y\\ &= \int_0^\infty y^2|K(y)|\mathrm{d}y\\ &= \frac{1}{2}\int_{\mathbb{R}} y^2|K(y)|\mathrm{d}y\,.\end{aligned}$$

Ceci établit la quatrième assertion. □

PREUVE.(Théorème 12.24)

CONTRÔLE DU BIAIS

En combinant l'hypothèse de régularité sur la densité f (deux fois différentiable) et l'hypothèse de parité du noyau, on peut obtenir des majorations de $\int_{\mathbb{R}} |f(x) - K_h * f(x)|\,\mathrm{d}x$.

Le noyau associé à K_h est noté $(L)_h$, il est lié à L_h par la relation suivante :

$$(L)_h = h^2 L_h\,.$$

En effet, pour $x > 0$

$$\begin{aligned}(L)_h(x) &= \int_x^\infty (y-x)\frac{1}{h}K(\frac{y}{h})\mathrm{d}y\\ &= \int_x^\infty \frac{y-x}{h}K(\frac{y}{h})\mathrm{d}y\\ &= h^2 \times \frac{1}{h}\int_{x/h}^\infty (y - x/h)K(y)\mathrm{d}y\\ &= h^2 \times \frac{1}{h}L\left(\frac{x}{h}\right)\\ &= h^2 L_h(x)\,.\end{aligned}$$

Comme on suppose K symétrique et $\int_{-1}^1 K(y)\mathrm{d}y = 1$,

$$\begin{aligned}K_h * f(x) - f(x) &= \int_{\mathbb{R}} \left(f(x-y) - f(x)\right) K_h(y)\mathrm{d}y\\ &= \int_{\mathbb{R}} \left(f(x+y) - f(x)\right) K_h(y)\mathrm{d}y\,.\end{aligned}$$

Comme f et f' sont supposées absolument continues, pour presque tout couple (x, y),

$$f(x+y) - f(x) = yf'(x) + \int_x^{x+y} (x+y-v)f''(v)\mathrm{d}v$$

Ponctuellement, l'écart entre $K_h * f$ et f s'exprime à l'aide de la convoluée de f'' et du noyau associé L_h.

Rappelons que le noyau K est supposé symétrique (ce qui entraine $\int_{\mathbb{R}} yK(y)\mathrm{d}y = 0$).

$$\begin{aligned}
&K_h * f(x) - f(x) \\
&= \int_{\mathbb{R}} \int_x^{x+y} (x+y-v)f''(v)\mathrm{d}v K_h(y)\mathrm{d}y \\
&= \int_x^{\infty} \int_{v-x}^{\infty} (x+y-v)f''(v)K_h(y)\mathrm{d}y\mathrm{d}v - \int_{-\infty}^{x} \int_{-\infty}^{v-x} (x+y-v)f''(v)K_h(y)\mathrm{d}y\mathrm{d}v \\
&= \int_x^{\infty} f''(v) \int_{v-x}^{\infty} (x+y-v)K_h(y)\mathrm{d}y\mathrm{d}v - \int_{-\infty}^{x} f''(v) \int_{-\infty}^{v-x} (x+y-v)K_h(y)\mathrm{d}y\mathrm{d}v \\
&= \int_x^{\infty} f''(v)(L)_h(v-x)\mathrm{d}v + \int_{-\infty}^{x} f''(v)(L)_h(v-x)\mathrm{d}v \\
&= \int_{\mathbb{R}} f''(v)(L)_h(x-v)\mathrm{d}v \\
&= h^2 f'' * L_h(x)\,.
\end{aligned}$$

En intégrant et en invoquant l'inégalité de Young et la proposition 12.27, on obtient une majoration du biais qui fait intervenir largeur de bande, régularité de la densité, et moments du noyau K :

$$\begin{aligned}
\int_{\mathbb{R}} |K_h * f(x) - f(x)|\,\mathrm{d}x &\leq h^2 \int_{\mathbb{R}} |f'' * L_h(x)|\,\mathrm{d}x \\
&\leq h^2 \int_{\mathbb{R}} |f''|\,\mathrm{d}x \times \int_{\mathbb{R}} |L_h(x)|\,\mathrm{d}x \\
&\leq \frac{h^2}{2} \int_{\mathbb{R}} |f''|\,\mathrm{d}x \times \int_{\mathbb{R}} x^2 |K(x)|\,\mathrm{d}x \\
&= \frac{h^2}{2} \times c(f) \times \mathrm{sd}(K)\,.
\end{aligned}$$

Contrôle des fluctuations.

En chaque x, $\widehat{f}_n(x) - f * K_h(x)$ s'écrit comme une moyenne de variables aléatoires indépendantes centrées.

En invoquant d'abord l'inégalité de Cauchy-Schwarz,

$$\begin{aligned}
\mathbb{E}\left[\left|\widehat{f}_n(x) - f * K_h(x)\right|\right] &\leq \left(\frac{\mathrm{var}\left(K_h(x-X)\right)}{n}\right)^{1/2} \\
&\leq \left(\frac{\mathbb{E}\left[K_h(x-X)^2\right]}{n}\right)^{1/2} \\
&= \left(\frac{K_h^2 * f(x)}{nh}\right)^{1/2}
\end{aligned}$$

avec $K_h^2(x) := \frac{1}{h} K^2\left(\frac{x}{h}\right)$.

Des hypothèses de support borné pour f et K, on déduit que le support de $K_h * f$ est inclus dans un intervalle J de longueur au plus $s(f) + 2h$. On peut mettre à profit l'inégalité de Cauchy-Schwarz :

$$\begin{aligned}
\int_{\mathbb{R}} \left(\frac{K_h^2 * f(x)}{nh}\right)^{1/2} \mathrm{d}x &= |J| \times \int_J \frac{1}{|J|} \left(\frac{K_h^2 * f(x)}{nh}\right)^{1/2} \mathrm{d}x \\
&\leq \sqrt{|J|} \times \left(\int_J \frac{K_h^2 * f(x)}{nh} \mathrm{d}x\right)^{1/2}.
\end{aligned}$$

En revenant à notre objectif initial,

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathbb{R}} \left|\widehat{f}_n(x) - f * K_h(x)\right| \mathrm{d}x\right] &= \int_{\mathbb{R}} \mathbb{E}\left[\left|\widehat{f}_n(x) - f * K_h(x)\right|\right] \mathrm{d}x \\
&\leq \int_{\mathbb{R}} \left(\frac{K_h^2 * f(x)}{nh}\right)^{1/2} \mathrm{d}x \\
&\leq \frac{\sqrt{s(f)+2h}}{\sqrt{nh}} \sqrt{\int_{\mathbb{R}} K_h^2 * f(x) \mathrm{d}x} \\
&\leq \frac{\sqrt{s(f)+2h}}{\sqrt{nh}} \sqrt{\|K_h^2\|_1 \times \|f\|_1} \\
&\leq \frac{\sqrt{s(f)+2h}}{\sqrt{nh}} \sqrt{R(K)}\,.
\end{aligned}$$

En combinant la majoration du biais et celle des fluctuations, on obtient :

$$\mathbb{E}\left[\int_{\mathbb{R}} \left|f(x) - \widehat{f}_n(x)\right| \mathrm{d}x\right] \leq \frac{\sqrt{s(f)+2h}}{\sqrt{nh}} \sqrt{R(K)} + \frac{h^2}{2} c(f) \times \mathrm{sd}(K)$$

On impose $2h \leq s(f)$. On peut optimiser $\frac{\sqrt{2s(f)}}{\sqrt{nh}} \sqrt{R(K)} + \frac{h^2}{2} c(f) \times \mathrm{sd}(K)$ en h. On se contente d'équilibrer les deux termes en choisissant

$$h = \left(\frac{8 s(f) R(K)}{n c(f)^2 \,\mathrm{sd}(K)^2}\right)^{1/5}$$

Pour n assez grand, la contrainte $2h \leq s(f)$ est respectée. On obtient

$$\mathbb{E}\left[\int_{\mathbb{R}} \left|f(x) - \widehat{f}_n(x)\right| \mathrm{d}x\right] \leq 2 \frac{(s(f)^2 c(f))^{1/5}}{n^{2/5}} (4 R(K)^2 \,\mathrm{sd}(K))^{1/5} \leq 2 \frac{c^{1/5}}{n^{2/5}} (4 R(K)^2 \,\mathrm{sd}(K))^{1/5}\,.$$

□

Remarque 12.28 Comme la borne de risque final ne mentionne que la constante c qui définit la classe W_c, on peut se demander s'il n'est pas possible de définir une largeur de bande qui conviendrait à toutes les densités de la classe W_c simultanément. Si c'est possible, notre analyse ne le dit pas. La connaissance de $s(f)$ et de $c(f)$ semble nécessaire au bon équilibrage du biais et de la variance.

Le théorème 12.24 contient plusieurs messages.

i) Si on fait des hypothèses de régularité sur la densité à estimer, on peut choisir un noyau et une largeur de bande qui permettent une majoration quantitative du risque.
ii) Pour coller à la régularité de la fonction à estimer, il semble utile de bien choisir le noyau. Pour tirer parti de régularités d'ordre s (f $s-1$ fois dérivables, et dérivée d'ordre $s-1$ absolument continue), il faudrait utiliser un noyau d'ordre s, c'est à dire un noyau vérifiant $\int x^p K(x) \mathrm{d}x = 0$ pour tout entier $1 \leq p < s$, et tel que $\int |x|^s |K(x)| \mathrm{d}x < \infty$.
iii) Pour optimiser la largeur de bande, il faut connaître des quantités qui dépendent de la densité à estimer, ou au moins disposer de majorants de ces quantités.
iv) La quantité $\widehat{f}_n$ n'est pas un estimateur. Le choix de h_n dépend de $s(f)$ et de $c(f)$ et pas seulement de c. Il dépend donc de paramètres de la fonction f qui ne sont pas connus du statisticiens. Le problème de l'ajustement de la taille de la fenêtre est un problème très délicat d'équilibre entre biais et variance.

Il y a donc un (assez) long chemin à parcourir pour transformer un résultat comme le théorème 12.24 en une méthode pratique et utile pour l'estimation de densité.

La sélection de largeur de bande, de noyau, l'usage d'autres méthodes d'estimation (par ondelettes, splines, etc) font encore l'objet d'une recherche intense.

On peut se faire une idée de la tâche concrète représentée par l'estimation de densité, et examinant les graphiques produits sur les données `faithful` (durées d'éruption du geyser `faithful` dans le parc de Yellowstone). La méthode de sélection de largeur de bande utilisée avec deux noyaux différents conduits à des estimées sensiblement différentes.

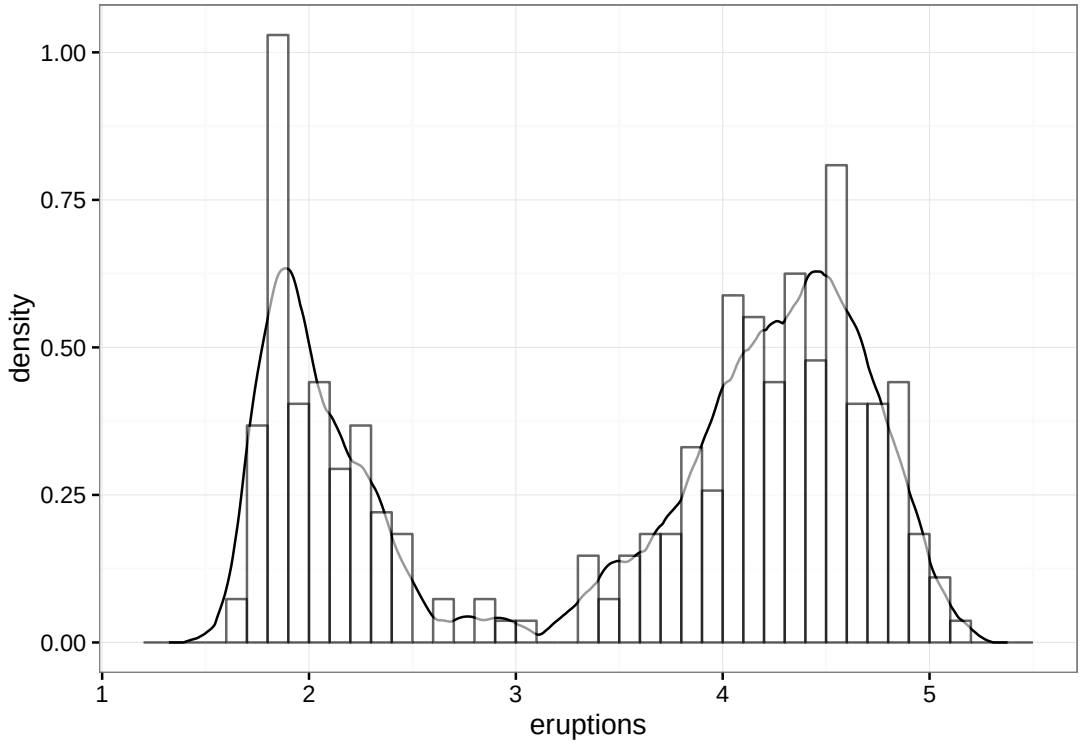

FIG. 12.1 : Estimation de densité sur les données `faithful` à l'aide de noyaux d'Epanechnikov en sélectionnant la largeur de bande avec la méthode de Sheather-Jones

```
data(faithful, package = "datasets")
geyser <- faithful
qplot(x=(density(x=geyser$eruptions,kernel="epanechnikov",bw="SJ",adjust=.66))$x,
      y=(density(x=geyser$eruptions,kernel="epanechnikov",bw="SJ",adjust=.66))$y,
      geom="line")+
  geom_histogram(data=geyser,mapping=aes(x=eruptions,y=..density..),
                                 colour="black",fill="white",binwidth=.1,alpha=.6) +
   xlab("eruptions") + ylab("density")
```

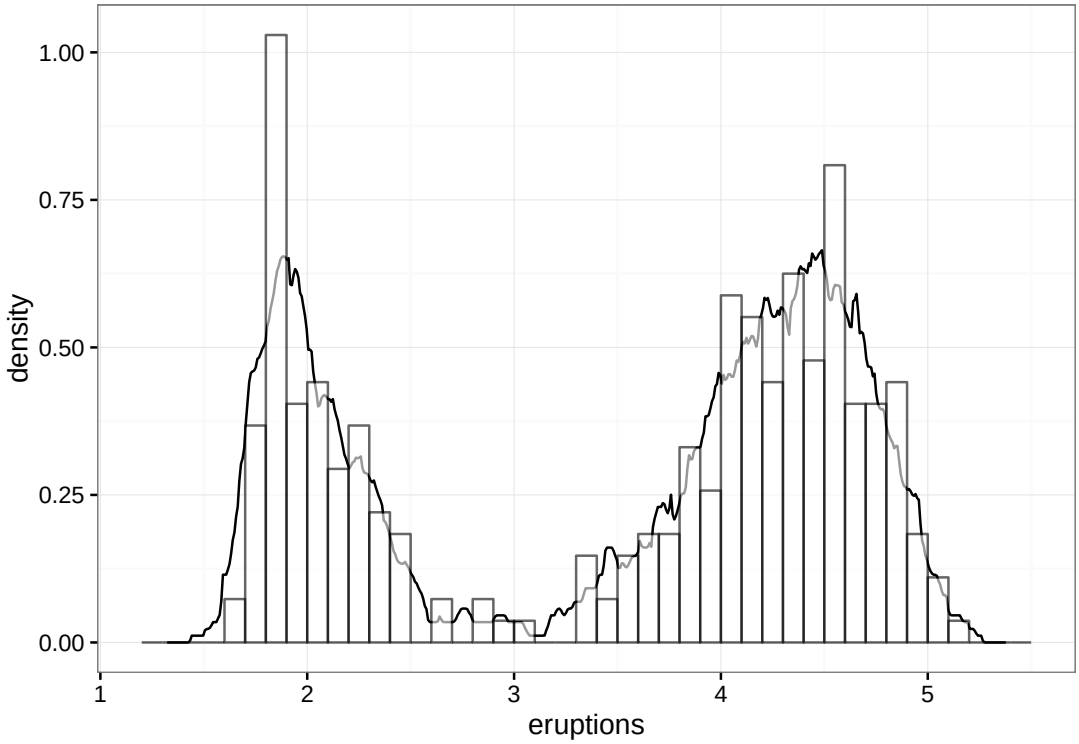

FIG. 12.2 : Estimation de densité sur les données `faithful` à l'aide de noyaux rectangulaires en sélectionnant la largeur de bande avec la méthode de Sheather-Jones

12.8 REMARQUES BIBLIOGRAPHIQUES

Le théorème de densité de Lebesgue et une multitude de résultats reliés, en particulier, le théorème 12.19 sont introduits dans le chapitre 3 de l'ouvrage [5].

Le point de vue L_1 sur l'estimation de densité a été popularisé par [2]. Une présentation concise se trouve dans [3].

BREZIS [1] présente les propriétés essentielles de la convolution (et beaucoup d'autres choses).

SILVERMAN [4] décrit la pratique de l'estimation de densité.

Références

[1] H. BREZIS. **Functional analysis, Sobolev spaces and partial differential equations**. Springer, New York, 2011.

[2] L. DEVROYE et L. GYÖRFI. **Nonparametric Density Estimation : The L_1 View**. New York : John Wiley, 1985.

[3] L. DEVROYE et G. LUGOSI. **Combinatorial Methods in Density Estimation**. Springer-Verlag, New York, 2000.

[4] B. W. SILVERMAN. **Density estimation for statistics and data analysis**. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[5] E. STEIN et R. SHAKARCHI. **Real analysis**. Princeton Lectures in Analysis III. Princeton University Press, 2005.

Dans cet appendice, on s'intéresse à des matrices à coefficients réels. On note $\mathcal{M}_{n,p}$ l'ensemble des matrices à n lignes et p colonnes.

A.1 FACTORISATION DE CHOLESKY

THÉORÈME A.1 (FACTORISATION DE CHOLESKY) *Soit* $\mathbf{K}$ *une matrice* $\in \mathcal{M}_{n,n}$ *symétrique, semi-définie positive. Il existe une matrice de permutation* $\mathbf{P} \in \mathcal{M}_{n,n}$ *et une matrice triangulaire inférieure* $\mathbf{L} \in \mathcal{M}_{n,n}$*, à coefficients diagonaux positifs telles que*

$$\mathbf{P} \times \mathbf{K} \times \mathbf{P}^t = \mathbf{L} \times \mathbf{L}^t .$$

Si $\mathbf{K}$ *est définie positive, on peut supposer que* $\mathbf{P}$ *est l'identité, la matrice* $\mathbf{L}$ *est unique.*

Nous traiterons simplement du cas défini positif. Nous allons vérifier une identité matricielle, qui est à la base de la factorisation de Cholesky.

PROPOSITION A.2 *Soit une matrice symétrique définie positive de* $\mathcal{M}_{n,n}$

$$\mathbf{K} = \left[\begin{array}{c:c} \mathbf{A} & \mathbf{B}^t \\ \hdashline \mathbf{B} & \mathbf{W} \end{array}\right]$$

où $\mathbf{A} \in \mathcal{M}_{k,k}$*,* $1 \leq k < n$*.*

Alors le complément de Schur *de* $\mathbf{A}$ *dans* $\mathbf{K}$*, défini par :*

$$\mathbf{W} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^t$$

est défini positif.

PREUVE.[Proposition A.2]

Le premier point à vérifier est le caractère défini positif du complément de Schur de $\mathbf{A}$ dans $\mathbf{K}$. Supposons qu'elle ne le soit pas. Soit $\vec{u}$ un vecteur non nul tel que

$$\vec{u}^t(\mathbf{W} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^t)\vec{u} \leq 0$$

alors

$$\vec{v} := \left[\begin{array}{c} -\mathbf{A}^{-1}\mathbf{B}^t\vec{u} \\ \hdashline \vec{u} \end{array}\right]$$

est un vecteur non-nul qui vérifie

$$0 \geq \vec{v}^t\mathbf{K}\vec{v} = \vec{u}^t(-\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^t + \mathbf{W})\vec{u}$$

ce qui contredit l'hypothèse de départ sur $\mathbf{K}$. □

PREUVE.[Theorem A.1]

Pour $n = 1$, l'existence d'une factorisation de Cholesky est triviale.

On peut ensuite raisonner par récurrence sur n. Supposons l'existence de la factorisation de Cholesky pour les matrices symétriques définies positives de dimension inférieure ou égale à n.

Soit une matrice $\mathbf{K}$ une symétrique définie positive de $\mathcal{M}_{n+1,n+1}$ décomposée en blocs

$$\mathbf{K} = \left[\begin{array}{c:c} \mathbf{A} & \mathbf{B}^t \\ \hdashline \mathbf{B} & \mathbf{W} \end{array}\right]$$

où $\mathbf{A} \in \mathcal{M}_{k,k}$, $1 \leq k \leq n$.

D'après l'hypothèse de récurrence et la proposition A.2, les deux sous-matrices $\mathbf{A}$ et $\mathbf{W} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^t$ admettent chacun une décomposition de Cholesky $\mathbf{A} = \mathbf{L}_1\mathbf{L}_1^t$, $\mathbf{W} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^t = \mathbf{L}_2\mathbf{L}_2^t$ où $\mathbf{L}_1, \mathbf{L}_2$ sont triangulaires inférieures à coefficients diagonaux positifs.

La factorisation de Cholesky de $\mathbf{K}$ s'écrit alors :

$$\mathbf{K} = \left[\begin{array}{c:c} \mathbf{L}_1 & 0 \\ \hdashline \mathbf{B}(\mathbf{L}_1^t)^{-1} & \mathbf{L}_2 \end{array}\right] \times \left[\begin{array}{c:c} \mathbf{L}_1^t & \mathbf{L}_1^{-1}\mathbf{B}^t \\ \hdashline 0 & \mathbf{L}_2^t \end{array}\right].$$

□

Remarque A.3 Nous n'avons pas vérifié l'unicité de la décomposition de Cholesky.

La décomposition de Cholesky est facile à calculer. L'algorithme simple requiert un nombre d'opérations élémentaires de l'ordre de n^3.

Si $\mathbf{A}$ est une matrice symétrique définie positive, son facteur de Cholesky n'est pas au sens strict une « racine carré » de $\mathbf{A}$. Une racine carré devrait vérifier

$$\mathbf{A} = \mathbf{A}^{1/2} \times \mathbf{A}^{1/2}.$$

La factorisation abordée dans la section suivante nous offrira une racine carré pour les matrices définies positives.

A.2 Décomposition en valeurs singulières et décomposition spectrale

La factorisation suivante appelée *décomposition en valeurs singulières* (SVD) joue un rôle primordial, en probabilités, statistiques, analyse numérique, étude des problèmes inverses, ...

Théorème A.4 (décomposition en valeurs singulières) *Si $\mathbf{Z} \in \mathcal{M}_{n,p}$ est de rang $r \leq \min(n,p)$, alors il existe trois matrices $\mathbf{U} \in \mathcal{M}_{n,n}, \mathbf{D} \in \mathcal{M}_{n,p}, \mathbf{V} \in \mathcal{M}_{p,p}$ telles que*

$$\mathbf{Z} = \mathbf{U} \times \mathbf{D} \times \mathbf{V}^t$$

$\mathbf{D}$ est $n \times p$, diagonale, positive, de rang r, $\mathbf{U}^t\mathbf{U} = \boldsymbol{I}_n$ et $\mathbf{V}^t\mathbf{V} = \boldsymbol{I}_p$.
Les coefficients diagonaux de $\mathbf{D}$ sont décroissants

- Les coefficients diagonaux de $\mathbf{D}$ sont les *valeurs singulières* de $\mathbf{Z}$;
- Les colonnes de $\mathbf{U}$ définissent les *vecteurs singuliers à droite* de $\mathbf{Z}$;
- Les colonnes de $\mathbf{V}$ définissent les *vecteurs singuliers à gauche* de $\mathbf{Z}$.

La norme d'opérateur $\|\cdot\|_{\text{op}}$ sur $\mathcal{M}_{n,p}$ est définie par

$$\|\mathbf{A}\|_{\text{op}} = \sup\left\{u^t\mathbf{A}v : u \in \mathbb{R}^n, v \in \mathbb{R}^p, \|u\| \leq 1, \|v\| \leq 1\right\}.$$

preuve. On note $\sigma_1 := \|\mathbf{Z}\|_{\text{op}}$. En dimension finie, les boules unité sont compactes, il existe donc $\widehat{u} \in \mathbb{R}^n$ et $\widehat{v} \in \mathbb{R}^p$ de normes 1, tels que $\sigma_1 = \widehat{u}^t\mathbf{Z}\widehat{v}$. Soit $\mathbf{A} \in \mathbb{M}_{n,n-1}$, $\mathbf{B} \in \mathcal{M}_{p,p-1}$ des matrices telles que $\left[\begin{array}{c|c} \widehat{u} & \mathbf{A} \end{array}\right]$ et $\left[\begin{array}{c|c} \widehat{v} & \mathbf{B} \end{array}\right]$ soient *orthognales* de dimensions $n \times n$ et $p \times p$

On décompose les matrices en blocs :

$$\left[\begin{array}{c} \widehat{u}^t \\ \hdashline \mathbf{A}^t \end{array}\right] \times \mathbf{Z} \times \left[\begin{array}{c|c} \widehat{v} & \mathbf{B} \end{array}\right] = \left[\begin{array}{c:c} \sigma_1 & w^t \\ \hdashline \mathbf{0} & \mathbf{A}^t\mathbf{Z}\mathbf{B} \end{array}\right] =: \mathbf{Y}.$$

En effet $\mathbf{A}^t\mathbf{Z}\ \widehat{v} = \mathbf{0}$, car $\mathbf{Z}\ \widehat{v}$ est colinéaire avec $\widehat{u}$ et les colonnes de $\mathbf{A}$ sont orthogonales à $\widehat{u}$.

On vérifie de plus que $w = 0$, car la multiplication par une matrice orthogonale ne change pas la norme d'opérateur : $\|\mathbf{Y}\|_{\text{op}} = \|\mathbf{Z}\|_{\text{op}} = \sigma_1$. Aussi,

$$\|\mathbf{Y}\|_{\text{op}} \geq \frac{\left\|\mathbf{Y}\cdot\begin{bmatrix}\sigma_1 \\ \hdashline w\end{bmatrix}\right\|}{\left\|\begin{bmatrix}\sigma_1 \\ \hdashline w\end{bmatrix}\right\|} \geq \frac{\sigma_1^2 + w^tw}{\sqrt{\sigma_1^2 + w^tw}} = \sqrt{\sigma_1^2 + w^tw}.$$

Pour avoir $\sigma_1 \geq \|\mathbf{Y}\|_{\text{op}}$, il est nécessaire que $w = \mathbf{0}$. La matrice $\mathbf{A}^t\mathbf{Z}\mathbf{B}$ (de rang $r-1$) satisfait donc l'hypothèse de récurrence ($\max(n-1, p-1) \leq m$).

Il existe des matrices orthogonales $\mathbf{U}'$ et $\mathbf{V}'$ telles que ${\mathbf{U}'}^t\mathbf{A}^t\mathbf{Z}\mathbf{B}\mathbf{V}'$ soit égal à une matrice diagonale non-négative $\mathbf{D}'$ à $r-1$ coefficients non nuls.

$$\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & {\mathbf{U}'}^t \end{bmatrix} \times \begin{bmatrix} \widehat{u}^t \\ \mathbf{A}^t \end{bmatrix} \times \mathbf{Z} \times \begin{bmatrix} \widehat{v} & \mathbf{B} \end{bmatrix} \times \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}' \end{bmatrix} = \begin{bmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}' \end{bmatrix}$$

Les matrices

$$\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & {\mathbf{U}'}^t \end{bmatrix} \times \begin{bmatrix} \widehat{u}^t \\ \mathbf{A}^t \end{bmatrix} \qquad \text{et} \qquad \begin{bmatrix} \widehat{v} & \mathbf{B} \end{bmatrix} \times \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}' \end{bmatrix}$$

sont orthogonales de dimensions $n \times n$ et $p \times p$. □

La décomposition décrite dans l'énoncé du théorème A.4 est la SVD *complète*. Si le rang$r$ de la matrice $\mathbf{Z}$ est inférieur à $\max(n, p)$ (par exemple lorsque $n \neq p$), la matrice $\mathbf{D}$ n'est pas inversible, ni même symétrique. On peut simplifier la SVD et obtenir la SVD *fine* en notant $\mathbf{U}_r$, $\mathbf{D}_r$ et $\mathbf{V}_r$ les matrices obtenues à partir de $\mathbf{U}, \mathbf{D}, \mathbf{V}$ en ne conservant que les r premières colonnes. On note alors que

- Les colonnes de $\mathbf{U}_r$ engendrent l'image de $\mathbf{Z}$;
- Le noyau/nullspace de $\mathbf{Z}$ est orthogonal au sous-espace engendré par les colonnes $\mathbf{V}_r$.

On a toujours

$$\mathbf{Z} = \mathbf{U}_r \times \mathbf{D}_r \times \mathbf{V}_r^t \,.$$

La décomposition spectrale des matrices symmetriques est liée à la SVD.

THÉORÈME A.5 (DÉCOMPOSITION SPECTRALE) *Pour toute matrice symétrique $\mathbf{K} \in \mathbb{R}^{d\times d}$, il existe une base orthogonale $e_1, e_2, \ldots, e_p$ of $\mathbb{R}^p$ et une suite décroissante de réels $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ tels que*

$$\mathbf{K} = \sum_{j=1}^{p} \lambda_j e_j {e_j}^t \,.$$

En notation matricielle,

$$\mathbf{K} = \mathbf{U} \times \mathbf{D} \times \mathbf{U}^t$$

où $\mathbf{U}$ est la matrice orthogonale dont les colonnes sont les vecteurs propres $(e_i)_{i\leq p}$ et $\mathbf{D}$ est la matrice diagonale définie par $\lambda_1 \geq \ldots \geq \lambda_n$.

REMARQUE A.6 Les λ_i sont appelées *valeurs propres* de K, les e_i sont appelés *vecteurs propres*. Le sous-espace engendré par les *vecteurs propres* associés à des valeurs propres identiques est appelé un *sous-espace propre*. La suite des valeurs propres est appelés le *spectre* de la matrice. $\mathbf{K}$ est symétrique non négative (resp. positive) ssi ses valeurs propres sont non-negatives (resp. positives) Pour une matrice symétrique, les valeurs propres et les sous-espaces propres sont définis de manière unique.

Pour une matrice symétrique $\mathbf{K}$, $\mathbf{K}\mathbf{K}^t = \mathbf{K}^t\mathbf{K}$. Eventuellement au signe près, les vecteurs singuliers à gauche peuvent être choisis comme vecteurs singuliers à droite, et les valeurs singulières sont les valeurs absolues des valeurs propres.

A.3 MEILLEURE APPROXIMATION DE RANG DONNÉ PAR RAPPORT AUX NORMES DE HILBERT-SCHMIDT ET D'OPÉRATEUR

La SVD permet de calculer des bonnes approximations d'une matrice en plusieurs sens. C'est ce qui explique le rôle de la SVD en statistique en grande dimension. Lorsqu'il est utile de *réduire la dimension*, la SVD fournit un guide. Ce guide est mis à profit dans des méthodes comme l'analyse en composantes principales (ACP).

L'espace $\mathcal{M}_{n,p}$ est un espace vectoriel de dimension $n \times p$. Il est muni d'un produit scalaire

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\text{HS}} = \text{Trace}\left(\mathbf{A} \times \mathbf{B}^t\right)$$

(HS renvoie à Hilbert-Schmidt). On fabrique une base orthonormée de cet espace à l'aide de bases de $\mathbb{R}^n$ et $\mathbb{R}^p$.

PROPOSITION A.7 *Soit $u_1, \ldots, u_n$ et $v_1, \ldots, v_p$ deux bases orthonormées de $\mathbb{R}^n$ et $\mathbb{R}^p$, alors $(u_i v_j^t)_{i \leq n, j \leq p}$ forment une base orthonormée de $\mathcal{M}_{n,p}$.*

PREUVE.

$$\begin{aligned} \langle u_i v_j^t, u_k v_\ell^t \rangle_{\text{HS}} &= \text{Trace}\left(u_i v_j^t v_\ell u_k^t\right) \\ &= \langle v_j, v_\ell \rangle \, \text{Trace}\left(u_i u_k^t\right) \\ &= \langle v_j, v_\ell \rangle \times \langle u_i, u_k \rangle \end{aligned}$$

□

Chaque SVD complète nous fournit donc une base orthonormée de $\mathcal{M}_{n,p}$.

COROLLAIRE A.8 *Les vecteurs singuliers gauche et droit dans la SVD complète d'une matrice $n \times p$ forment une base orthonormée de $\mathcal{M}_{n,p}$ doté de la norme Hilbert-Schmidt.*

Soit $k \leq r$ et

$$\mathbf{Z}_k := \sum_{i \leq k} \sigma_i u_i v_i^t = \mathbf{U}_k \times \mathbf{D}_k \times \mathbf{V}_k$$

où $\mathbf{U}_k, \mathbf{V}_k$ sont formées par les k premières colonnes de $\mathbf{U}, \mathbf{V}$, alors que $\mathbf{D}_k$ est formée des k premières lignes et colonnes de $\mathbf{D}$.

THÉORÈME A.9 *La meilleure approximation de rang k de $\mathbf{Z}$ au sens de la norme de Hilbert-Schmidt est $\mathbf{Z}_k$ et*

$$\|\mathbf{Z} - \mathbf{Z}_k\|_{\text{HS}}^2 = \sum_{j=k+1}^{p} \sigma_j^2 \, .$$

Le résultat suivant est (un peu) plus surprenant.

THÉORÈME A.10 (ECKHART-YOUNG) *La meilleure approximation de rang k de $\mathbf{Z}$ au sens de la norme d'opérateur de $\mathbf{Z}$ est $\mathbf{Z}_k$ et*

$$\|\mathbf{Z} - \mathbf{Z}_k\|_{\text{op}} = \sigma_{k+1} \, .$$

PREUVE.[Théorème A.9]
Supposons $n \geq p$. A partir de la SVD complète de $\mathbf{Z}$,

$$\mathbf{Z} = \sum_{i=1}^{p} \sigma_i u_i v_i^t \qquad \text{avec} \quad \sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0 \, .$$

Soit $\mathbf{W} = \sum_{i \leq n, j \leq p} \mu_{i,j} u_i v_j^t$, alors

$$\|\mathbf{Z}-\mathbf{W}\|_{\text{HS}}^2 = \sum_{i=1}^{p}(\sigma_i^2-\mu_{i,i})^2 + \sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{I}_{i\neq j}\mu_{i,j}^2$$

Tout minimisant devrait satisfaire $\mu_{i,j}=0$ pour $i\neq j$. La matrice $\sum_{i\leq p}\mu_{i,i}u_iv_i^t$ est de rang k ssi exactement k coefficients $\mu_{i,i}$ sont non nuls. La distance est minimisée en choisissant $\mu_{i,i}=\sigma_i$ pour $i\leq k$ et $\mu_{i,i}=0$ pour $k+1\leq i\leq p$. □

PREUVE.[Théorème A.10] La SVD de $\mathbf{Z}-\mathbf{Z}_k$ est

$$\mathbf{U}\times\text{diag}\begin{bmatrix}0\\ \vdots\\ 0\\ \sigma_{k+1}\\ \vdots\\ \sigma_p\end{bmatrix}\times\mathbf{V}^t.$$

La norme d'opérateur de la différence est égale à σ_{k+1}

Now, let $\mathbf{W}$ be of rank k. $\ker(\mathbf{W})$ has dimension $p-k$, is intersects the linear span of the first $k+1$ columns of $\mathbf{V}$. Let y be a unit vector in this intersection. By definition $\mathbf{W}y=0$.

$$\|\mathbf{Z}y\|^2 = \sum_{i=1}^{k+1}\sigma_i^2\langle v_i,y_i\rangle^2 \geq \sigma_{k+1}^2$$

which proves that $\|\mathbf{Z}-\mathbf{W}\|_2\geq\sigma_{k+1}$. □

A.4 PSEUDO-INVERSE

Dans l'étude de la régression multiple (qui étend l'étude des modèles linéaires gaussiens du chapitre 3), on s'intéresse à la minimisation de

$$\|y-\mathbf{A}x\|_2^2$$

lorsque l'inconnu $x\in\mathbb{R}^p$ alors que $y\in\mathbb{R}^n$ et $\mathbf{A}\in\mathcal{M}_{n,p}$. Dans le chapitre 3, nous avons supposé que $n\geq p$, et que $\mathbf{A}$ est de rang p. Ceci garantit l'unicité du minimisant. On dispose alors d'une forme explicite pour ce minimisant. On peut s'intéresser aux moindre carrés ordinaires sans supposer que $\mathbf{A}$ est de plein rang. Le problème de minimisation n'a plus de solution unique. On peut provilégier le minimisant dont la norme euclidienne est minimale. Cette solution se détermine à l'aide de la SVD fine. Elle met en avant une construction appelée pseudo-inverse.

DÉFINITION A.11 (PSEUDO-INVERSE) Soit $\mathbf{A}$ une matrice de $\mathcal{M}_{n,p}$, sa *pseudo-inverse de Moore-Penrose* est une matrice $\mathbf{A}^+$ de $\mathcal{M}_{p,n}$ qui vérifie :

i) $\mathbf{A}\times\mathbf{A}^+\times\mathbf{A}=\mathbf{A}$;
ii) $\mathbf{A}^+\times\mathbf{A}\times\mathbf{A}^+=\mathbf{A}^+$;
iii) $(\mathbf{A}\times\mathbf{A}^+)^t=\mathbf{A}\times\mathbf{A}^+$;
iv) $(\mathbf{A}^+\times\mathbf{A})^t=\mathbf{A}^+\times\mathbf{A}$.

THÉORÈME A.12 *Soit $\mathbf{A}$ une matrice de $\mathcal{M}_{n,p}$, sa pseudo-inverse de Moore-Penrose existe et est unique.*

PREUVE. L'existence se déduit de la décomposition en valeurs singulières fine. Soit

$$\mathbf{A}=\mathbf{U}\times\mathbf{D}\times\mathbf{V}^t$$

avec $\mathbf{U}\in\mathcal{M}_{n,r}$, $\mathbf{D}\in\mathcal{M}_{r,r}$ diagonale, strictement positive, et $\mathbf{V}\in\mathcal{M}_{p,r}$, $\mathbf{U}^t\times\mathbf{U}=\text{Id}_r$ et $\mathbf{V}^t\times\mathbf{V}=\text{Id}_r$. La matrice

$$\mathbf{A}^+=\mathbf{V}\times\mathbf{D}^{-1}\times\mathbf{U}^t$$

vérifie les propriétés de la pseudo-inverse de Moore-Penrose.

Pour vérifier l'unicité, notons que les quatre conditions de la définiton impliquent que pour toute pseudo-inverse $\mathbf{B}$, $\mathbf{AB}$ est la projection orthogonale sur l'espace engendré par les colonnes de $\mathbf{A}$, et $\mathbf{BA}$ est la projection orthogonale sur l'espace engendré par les colonnes de $\mathbf{B}$. □

La pseudo-inverse de Moore-Penrose coïncide avec l'inverse lorsque la matrice $\mathbf{A}$ est inversible. Dans le cas où $\mathbf{A}$ n'est pas inversible, la pseudo-inverse vérifie une partie des identités vérifiées par l'inverse classique.

PROPOSITION A.13

$$\begin{aligned}(\mathbf{A}^+)^+ &= \mathbf{A}\\ (\mathbf{A}^t)^+ &= (\mathbf{A}^+)^t\\ (\lambda\mathbf{A})^+ &= \frac{1}{\lambda}\mathbf{A}^+\\ (\mathbf{A}\times\mathbf{B})^+ &= \mathbf{B}^+\times\mathbf{A}^+\,.\end{aligned}$$

La pseudoinverse est intimement liée aux problèmes de moindre carrés.

PROPOSITION A.14 *Pour tout $y \in \mathbb{R}^n$, la solution de norme euclidienne minimale de $\arg\min \|y-\mathbf{X}\theta\|^2$ est donnée par $\mathbf{X}^+y$.*

PREUVE. Comme $\mathbf{XX}^+$ représente la projection orthogonale sur le sous-espace engendré par les colonnes de $\mathbf{X}$, $\widehat{\theta}=\mathbf{X}^+y$ minimise $\|y-\mathbf{X}\theta\|^2$. Il reste à vérifier qu'il s'agit de la solution de norme euclidienne minimale. Si la SVD fine de $\mathbf{X}$ est $\mathbf{U}\times\mathbf{D}\times\mathbf{V}$, tout minimisant de l'écart quadratique se décompose en $\widehat{\theta}+\eta$ où η est orthogonal au sous-espace engendré par les colonnes de $\mathbf{V}$ ($\widehat{\theta}$ appartient au sous-espace engendré par les colonnes de $\mathbf{V}$ et on le suppose non-nul). La norme euclidienne de $\widehat{\theta}+\eta$ est minimisée en choisissant $\eta=0$. □

REMARQUE A.15 Si $\mathbf{X}^t\times\mathbf{X}$ est inversible alors

$$\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t=\mathbf{X}^+\,.$$

On retrouve les résultats du Chapitre 3.

A.5 REMARQUES BIBLIOGRAPHIQUES

L'ouvrage R. HORN et C. JOHNSON. **Matrix analysis**. Cambridge University Press, 1990, fournit une présentation très accessible de l'analyse matricielle.

Le livre R. BHATIA. **Matrix analysis**. Springer-Verlag, 1997, contient une multitude de résultats pointus très utiles pour aborder les matrices aléatoires.

Le livre de G. H. GOLUB et C. F. VAN LOAN. **Matrix computations**. Third. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 1996, p. xxx+698, propose un lien assez systématique avec l'algorithmique. Le chapitre 8 décrit les algorithmes de calculs des valeurs propres.

L. ELDÉN. **Matrix methods in data mining and pattern recognition**. T. 4. Fundamentals of Algorithms. Society for Industrial et Applied Mathematics (SIAM), Philadelphia, PA, 2007, p. x+224. ISBN : 978-0-898716-26-9. MR : 2314399 décrit la régression selon les composantes principales (*principal component regression*) pour traiter les problèmes de moindre carrés lorsque le design n'est pas de plein rang ou est mal conditionné.

B.1 ESPÉRANCE CONDITIONNELLE

Conditionnement en statistique

Conditionnement par rapport à un événement

Dans le domaine du conditionnement, la notion la plus élémentaire est celle de probabilité conditionnelle par rapport à un événement de probabilité strictetment positive.

DÉFINITION B.1 Soit P une probabilité sur $(\Omega, \mathcal{F})$. Soit $A \in \mathcal{F}$ un événement tel que $P\{A\} \neq 0$. Soit B un autre événement ($B \in \mathcal{F}$), n appelle *probabilité de B sachant A* la quantité :

$$P\{B \mid A\} = \frac{P\{A \cap B\}}{P\{A\}}.$$

EXEMPLE B.2 Si X est une variable aléatoire Gaussienne standard sur $(\Omega, \mathcal{F})$, et A l'événement $\{\omega : X(\omega) \geq t\}$ pour $t \geq 0$, on peut conditionner par A et définir pour $P\{B \mid A\}$ pour $B = \{\omega : |X(\omega)| \geq 2t\}$. On aura

$$P\{B \mid A\} = \frac{P\{X \geq 2t\}}{P\{X \geq t\}}\,.$$

On peut vérifier la proposition suivante et réexaminant la définition de ce qu'est une loi de probabilités.

PROPOSITION B.3 *Soit P une probabilité sur $(\Omega, \mathcal{F})$. Soit $A \in \mathcal{F})$ tel que $P\{A\} \neq 0$, alors la probabilité sachant $P\{\cdot \mid A\}$ définit une probabilité sur $(\Omega, \mathcal{F})$.*

On peut étudier la loi des variables aléatoires sur $(\Omega, \mathcal{F})$ sous la loi $P\{\cdot \mid A\}$. On peut en particulier calculer l'espérance d'une variable aléatoire X sous la loi sachant A :

$$\mathbb{E}_{P\{\cdot|A\}}X = \frac{\mathbb{E}[\mathbb{I}_A\, X]}{P\{A\}}\,.$$

Cette espérance est souvent notée $\mathbb{E}[X \mid A]$, nous essaierons d'éviter cette notation.

EXEMPLE B.4 Dans la situation gaussienne décrite plus haut, on peut étudier la loi de X^2 sachant $A = \{\omega : X(\omega) \geq t\}$. Pour $t > 1$, on peut écrire

$$\begin{aligned}\mathbb{E}_{P\{\cdot|X\geq t\}}X^2 &= \frac{\int_t^\infty x^2\phi(x)\mathrm{d}x}{\int_t^\infty \phi(x)\mathrm{d}x}\\ &\leq \frac{t^2}{1-1/t} + 1\,.\end{aligned}$$

où la majoration est obtenue en utilisant plusieurs fois l'intégration par parties.

On observe que la loi de X sachant A n'est pas gaussienne, et que sous A, X tend à etre proche de t, surtout lorsque t est grand.

La probabilité sachant A permet d'étudier l'indépendance d'un événement par rapport à A.

PROPOSITION B.5 *Si A et $B \in \mathcal{F}$ vérifient $P\{A\} > 0$, alors A et B sont indépendants relativement à la probabilité P si et seulement si*

$$P\{B \mid A\} = P\{B\}.$$

FORMULE DE BAYES La formule de Bayes est aussi appelée « formule de probabilités des causes ». C'est une expression dont il faut se méfier. La causalité est affaire trop sérieuse pour etre enfermée dans une formule calculatoire.

PROPOSITION B.6 [FORMULE DE BAYES] *Soit P une probabilité sur $(\Omega, \mathcal{F})$, soit $(A_i)_{i\in\mathcal{I}\subseteq\mathbb{N}}$ une famille d'événements deux à deux disjoints, de probabilité non nulle telle que $\cup_{i\in\mathcal{I}} A_i = \Omega$ (les A_i forment un système complet d'événements), soit B un événement de probabilité non nulle, alors pour tout $i \in \mathcal{I}$*

$$P\{A_i \mid B\} = \frac{P\{A_i\} \times P\{B \mid A_i\}}{\sum_{j\in\mathcal{I}} P\{A_j\} \times P\{B \mid A_j\}}.$$

PREUVE. Par définition, $P\{A_i \mid B\} = P\{A_i \cap B\}/P\{B\} = P\{A_i\} \times P\{B \mid A_i\}/P\{B\}$. Par ailleurs

$$P\{B \cap (\cup_{j\in\mathcal{I}} A_j)\} = P\{\cup_{j\in\mathcal{I}}(B \cap A_j)\} = \sum_{j\in\mathcal{I}} P\{B \cap A_j\} = \sum_{j\in\mathcal{I}} P\{A_j\} \times P\{B \mid A_j\}.$$

□

REMARQUE B.7 Dans la proposition précédente, on appelle $P\{A_i\}$ la probabilité a priori de A_i et $P\{A_i \mid B\}$ la probabilité a posteriori.

Espérances conditionnelles par rapport à une tribu discrète

La notion d'espérance conditionnelle par rapport à une tribu discrète peut être introduite à partir de la notion élémentaire de probabilité sachant un événemement (de probabilité non nulle).

DÉFINITION B.8 Soit Ω un univers discret et $\mathcal{F}$ une tribu d'événements sur Ω et P une probabilité sur $(\Omega, \mathcal{F})$, soit $(A_i)_{i\in\mathcal{I}\subseteq\mathbb{N}}$ une famille d'événements deux à deux disjoints, de probabilité non nulle telle que $\cup_i A_i = \Omega$. On appelle $\mathcal{G}$ la tribu engendrée par la collection $(A_i)_{i\in\mathcal{I}}$.

Soit X une variable aléatoire de $(\Omega, \mathcal{F})$ vers $(\mathcal{X}, \mathcal{H})$, on appelle ESPÉRANCE CONDITIONNELLE DE X SACHANT $\mathcal{G}$ la variable aléatoire

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{i\in\mathcal{I}} \mathbb{E}_{P_{\{\cdot\mid A_i\}}}[X] \times \mathbf{1}_{A_i}.$$

REMARQUE B.9 Les quantités $\mathbb{E}_{P_{\{\cdot\mid A_i\}}}[X]$ sont des nombres alors que $\mathbb{E}[X \mid \mathcal{G}]$ est une fonction sur Ω. Il vaut mieux éviter de confondre ces objets de nature différente. Nous évitons la notation $\mathbb{E}[X \mid A_i]$ parce qu'elle peut prêter à confusion : s'agit-il de $\mathbb{E}_{P_{\{\cdot\mid A_i\}}}[X]$ ou de $\mathbb{E}[X \mid \sigma(A_i)]$ où $\sigma(A_i)$ est la tribu engendrée par A_i : $\{A_i, A_i^c, \Omega, \emptyset\}$.

PROPOSITION B.10 *Soit P une probabilité sur $(\Omega, \mathcal{F})$, soit $(A_i)_{i\in\mathcal{I}\subseteq\mathbb{N}}$ une famille d'événements deux à deux disjoints, de probabilité non nulle telle que $\cup_{i\in\mathcal{I}} A_i = \Omega$. On note $\mathcal{G} = \sigma\Big((A_i)_{i\in\mathcal{I}}\Big)$ la tribu engendrée par la collection $(A_i)_{i\in\mathcal{I}}$.*

La variable aléatoire X est supposée P intégrable.

L'espérance conditionnelle $\mathbb{E}[X \mid \mathcal{G}]$ est une variable aléatoire $\mathcal{G}$-mesurable, telle que pour toute variable aléatoire Y, elle aussi $\mathcal{G}$-mesurable

$$\mathbb{E}[YX] = \mathbb{E}[Y\mathbb{E}[X \mid \mathcal{G}]].$$

PREUVE. Il faut vérifier deux choses :

i) $\mathbb{E}[X \mid \mathcal{G}]$ vérifie bien la propriété (B.10).
ii) si Z vérifie (B.10), alors $Z = \mathbb{E}[X \mid \mathcal{G}]$.

Vérification de i.)

Si Y est $\mathcal{G}$-mesurable $Y = \sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i}$ pour une suite de réels $(\lambda_i)_{i\in\mathcal{I}}$. Alors

$$\begin{aligned}
\mathbb{E}[Y\mathbb{E}[X \mid \mathcal{G}]] &= \mathbb{E}\left[\left(\sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i}\right)\left(\sum_{j\in\mathcal{I}} \mathbf{1}_{A_j} \frac{\mathbb{E}[\mathbf{1}_{A_j} X]}{P\{A_j\}}\right)\right] \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i} \frac{\mathbb{E}[\mathbf{1}_{A_i} X]}{P\{A_i\}}\right)\Bigg] \\
&= \sum_{i\in\mathcal{I}} \lambda_i \mathbb{E}[\mathbf{1}_{A_i} X] \frac{\mathbb{E}[\mathbf{1}_{A_i}]}{P\{A_i\}} \quad \text{linéarité de l'espérance} \\
&= \sum_{i\in\mathcal{I}} \lambda_i \mathbb{E}[\mathbf{1}_{A_i} X] \\
&\quad \text{linéarité de l'espérance} \\
&= \mathbb{E}\left[\left(\sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i}\right) X\right] \\
&= \mathbb{E}[YX].
\end{aligned}$$

Vérification de ii.) Supposons que Z vérifie (B.10). Définissons Y par $Y = \mathbf{1}_{A_i}$, pour un indice $i \in \mathcal{I}$. Comme Z est $\mathcal{G}$-mesurable, il existe une suite de réels $(\mu_j)_{j\in\mathcal{I}}$, telle que $Z = \sum_{j\in\mathcal{I}} \mu_j \mathbf{1}_{A_j}$.Donc, en utilisant l'incompatibilité deux à deux des A_j :

$$\mathbb{E}[ZY] = \mathbb{E}\left[\sum_{j\in\mathcal{I}} \mu_j \mathbf{1}_{A_j} \mathbf{1}_{A_i}\right] = \mu_i P\{A_i\}$$

D'après (B.10), on a :

$$\mathbb{E}[ZY] = \mathbb{E}[XY] = \mathbb{E}[X\mathbf{1}_{A_i}].$$

On a finalement pour tout $i \in \mathcal{I}$, $\mu_i = \mathbb{E}[X\mathbf{1}_{A_i}]/P\{A_i\}$. Ce qui permet de conclure $Z = \mathbb{E}[X \mid \mathcal{G}]$. □

La prooposition suivante montre le role de l'espérance conditionnelle dans les problèmes de prédiction/approximation au sens de l'erreur quadratique.

PROPOSITION B.11 *Soit Y une variable aléatoire de carré intégrable sur $(\Omega, \mathcal{F}, P)$ et $\mathcal{G}$ une sous-tribu discrète de $\mathcal{F}$. L'espérance conditionnelle de Y par rapport à $\mathcal{G}$ minimise*

$$\mathbb{E}\left[\big(Y - Z\big)^2\right]$$

parmi les variables aléatoires Z $\mathcal{G}$-mesurables et de carré intégrable.

Rappelons qu'une une variable aléatoire $\mathcal{G}$-mesurable est une fonction constante sur chaque partie $A_i, i \in \mathcal{I}$.

PREUVE. Si Y est une variable aléatoire sur $(\Omega, \mathcal{F})$, et si on cherche à prédire au mieux Y à partir d'une variable aléatoire $\mathcal{G}$-mesurable, on cherche une suite de coefficients $(b_i)_{i\in\mathcal{I}}$ qui minimise :

$$\begin{aligned}
\mathbb{E}_P\left[\Big(Y - \sum_{i\in\mathcal{I}} b_i \mathbf{I}_{A_i}\Big)^2\right] &= \mathbb{E}_P\left[\Big(\sum_{i\in\mathcal{I}} (Y - b_i)\mathbf{I}_{A_i}\Big)^2\right] \\
&= \sum_{i\in\mathcal{I}} \mathbb{E}_P\left[(Y - b_i)^2 \mathbf{I}_{A_i}\right] \\
&= \sum_{i\in\mathcal{I}} P\{A_i\}\, \mathbb{E}_{P\{\cdot|A_i\}}\left[(Y - b_i)^2\right]
\end{aligned}$$

On voit alors que pour chaque i, b_i doit coïncider avec l'espérance de Y sous $P\{\cdot \mid A_i\}$. La meilleure prédiction de Y, au sens de l'erreur quadratique, parmi les fonctions $\mathcal{G}$-mesurables est l'espérance conditionnelle de Y par rapport à $\mathcal{G}$. □

Ce sont les propriétés dégagées par les propositions B.10 et B.11 qui serventt de définition à l'espérance conditionnelle par rapport à une sous-tribu quelconque.

Espérance conditionnelle par rapport à une tribu générale

Dans cette section et les suivantes, $(\Omega, \mathcal{F}, P)$ est un espace probabilisé, et $\mathcal{G}$ une sous-tribu (quelconque) de $\mathcal{F}$. Notre objectif est de définir une espérance conditionnelle par rapport à la sous-tribu $\mathcal{G}$. La définition générale de l'espérance conditionnelle par de la propriété décrite dans la proposition B.10.

Définition B.12 (Espérance conditionnelle.) Soient $X \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$ et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, alors une variable aléatoire Y est appelée version de l'espérance conditionnelle de X sachant $\mathcal{G}$ si et seulement si

i) Y est $\mathcal{G}$-mesurable.
ii) Pour tout événement B de $\mathcal{G}$
$$\mathbb{E}_P\left[\mathbf{I}_B X\right] = \mathbb{E}_P\left[\mathbf{I}_B Y\right] .$$

Sans préjuger de l'existence de versions de l'espérance conditionnelle de X, nous allons vérifier que s'existe différentes versions, elles ne diffèrent que sur un ensemble négligeable.

Proposition B.13 (Egalité presque-sûre des versions.) *Soient $X \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$ et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, alors si Y' et Y sont deux versions de l'espérance conditionnelle de X sachant $\mathcal{G}$:*
$$P\left\{Y = Y'\right\} = 1.$$

preuve. Comme Y et Y' sont $\mathcal{G}$-mesurables, l'événement
$$B = \{\omega \ : \ Y(\omega) > Y'(\omega)\}$$
appartient aussi à $\mathcal{G}$. Par ailleurs,
$$\begin{aligned} \mathbb{E}_P\left[\mathbf{I}_B\, X\right] &= \mathbb{E}_P\left[\mathbf{I}_B\, Y\right] \\ &= \mathbb{E}_P\left[\mathbf{I}_B\, Y'\right] . \end{aligned}$$
Donc
$$\mathbb{E}_P\left[\mathbf{I}_B(Y - Y')\right] = 0\,.$$
Comme la variable aléatoire $\mathbf{I}_B(Y - Y')$ est positive ou nulle, si son espérance est nulle, elle est nulle presque partout. Donc
$$P\{Y > Y'\} = 0\,.$$
En procédant de même pour l'événement $\{Y < Y'\}$, on termine la preuve du théorème. □

Sans préjuger de l'existence de versions de l'espérance conditionnelle de X, etablissons maintenant quelques propriétés des versions de l'espérance conditionnelle lorsqu'il en existe.

Proposition B.14 *Soient $X_1, X_2 \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, a_1, a_2 deux réels alors si Y_1 Y_2 et Z sont respectivement des versions de l'espérance conditionnelle de X_1, X_2 et de $a_1X_1 + a_2X_2$ sachant $\mathcal{G}$, on a*
$$P\{a_1Y_1 + a_2Y_2 = Z\} = 1\,.$$

PREUVE. Soit B l'événement de $\mathcal{G}$ défini par

$$\{a_1Y_1 + a_2Y_2 > Z\}\,.$$

On a immédiatement

$$\begin{aligned}\mathbb{E}[\mathbf{I}_B Z] &= \mathbb{E}[\mathbf{I}_B(a_1X_1 + a_2X_2)]\\ &= a_1\mathbb{E}[\mathbf{I}_B X_1] + a_2\mathbb{E}[\mathbf{I}_B X_2]\\ &= a_1\mathbb{E}[\mathbf{I}_B Y_1] + a_2\mathbb{E}[\mathbf{I}_B Y_2]\\ &= \mathbb{E}[\mathbf{I}_B(a_1Y_1 + a_2Y_2)]\,,\end{aligned}$$

et donc

$$\mathbb{E}[\mathbf{I}_B(Z - (a_1Y_1 + a_2Y_2))] = 0\,.$$

On en conclut comme dans la preuve précédente que $P\{B\} = 0$.

On achève la preuve en raisonnant de la même façon sur l'événement $\{a_1Y_1 + a_2Y_2 < Z\}$. □

PROPOSITION B.15 *Soit $X \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ une sous-tribu de $\mathcal{F}$. Si Z est une version de l'espérance conditionnelle de X sachant $\mathcal{G}$ et si X est positive P-p.s. alors*

$$P\{Z \geq 0\} = 1\,.$$

La démonstration reproduit l'argument qui prouve l'unicité presque sûre des versions l'espérance conditionnelle.

PREUVE. Pour $n \in \mathbb{N}$, notons B_n l'événement (appartenant à $\mathcal{G}$)

$$B_n = \left\{\mathbb{E}_P[X \mid \mathcal{G}] < -\frac{1}{n}\right\}.$$

Etablir la Proposition, il suffit de vérifier

$$P\{\cup_n B_n\} = 0.$$

Comme $P\{\cup_n B_n\} = \lim_n P\{B_n\}$, il suffit donc de vérifier $P\{B_n\} = 0$, pour tout n, $P\{B_n\} = 0$. Pour tout n, $\mathbb{E}_P[\mathbf{I}_{B_n} X] \geq 0$. Par ailleurs,

$$\begin{aligned}\mathbb{E}_P[\mathbf{I}_{B_n} X] &= \mathbb{E}_P[\mathbf{I}_{B_n}\mathbb{E}_P[X \mid \mathcal{G}]]\\ &\quad \text{propriété } \textbf{??} \text{ et } B_n \in \mathcal{G}\\ &\leq -\frac{P\{B_n\}}{n}\\ &\leq 0\\ &\quad \text{avec égalité si et seulement si } P\{B_n\} = 0.\end{aligned}$$

Donc $P\{B_n\} = 0$ pour tout n. □

On obtient alors presque immédiatement :

COROLLAIRE B.16 *Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires de $\mathfrak{L}_2(\Omega, \mathcal{F}, P)$ qui vérifie $X_{n+1} \geq X_n$ P-p.s. alors une suite de versions de l'espérance conditionnelle est croissante P-p.s.*

$$\forall n \in \mathbb{N}, \quad \mathbb{E}_P[X_{n+1} \mid \mathcal{F}] \geq \mathbb{E}_P[X_n \mid \mathcal{F}]\,.$$

Si on dispose de sous-tribus emboitées, les espérances conditionnelles vérifient les propriétés suivantes.

PROPOSITION B.17 *Si $(\Omega, \mathcal{F})$ est un espace, probabilisé, si $\mathcal{G}$ et $\mathcal{H}$ sont deux sous-tribus de $\mathcal{F}$ qui vérifient $\mathcal{G} \subseteq \mathcal{H}$, alors pour toute variable aléatoire X (sur $(\Omega, \mathcal{F})$), pour toute loi de probabilité sur $(\Omega, \mathcal{F})$:*

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}\right] \mid \mathcal{H}\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] \mid \mathcal{G}\right] = \left[\mathbb{E}\left[X \mid \mathcal{G}\right].$$

PROPOSITION B.18 *Soit P une probabilité sur $(\Omega, \mathcal{F})$, et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$.*

Soit X une variable aléatoire définie sur $(\Omega, \mathcal{F})$, de carré intégrable intégrable sous P. L'espérance conditionnelle $\mathbb{E}[X \mid \mathcal{G}]$ est une variable aléatoire $\mathcal{G}$-mesurable, de carré integrable, c'est la projection orthogonale de X sur l'espace des variables aléatoires $\mathcal{G}$-mesurables de carré intégrable.

Si Z est $\mathcal{G}$-mesurable, de carré intégrable, alors

$$\mathbb{E}\left[(X-Z)^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X|\mathcal{G}])^2\right] + \mathbb{E}\left[(Z - \mathbb{E}[X|\mathcal{G}])^2\right] .$$

EXISTENCE DE L'ESPÉRANCE CONDITIONNELLE, LE POINT DE VUE PRÉDICTIF

DÉFINITION B.19 Soit $(\Omega, \mathcal{F})$ un espace probabilisé, X une variable aléatoire, P une probabilité sur $(\Omega, \mathcal{F})$ pour laquelle X est de carré intégrable et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$. La VARIANCE CONDITIONNELLE de X sachant $\mathcal{G}$ est définie par

$$\mathrm{Var}_P\left[X \mid \mathcal{G}\right] = \mathbb{E}_P\left[\left(X - \mathbb{E}[X \mid \mathcal{G}]\right)^2 \mid \mathcal{G}\right].$$

La proposition suivante est un Théorème de Pythagore pour la variance.

PROPOSITION B.20 *Soit $(\Omega, \mathcal{F})$ un espace probabilisé, X une variable aléatoire, P une probabilité sur $(\Omega, \mathcal{F})$ pour laquelle X est de carré intégrable et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$.*

$$\mathrm{Var}_P[X] = \mathrm{Var}_P\left[\mathbb{E}_P\left[X \mid \mathcal{G}\right]\right] + \mathbb{E}_P\left[\mathrm{Var}_P\left[X \mid \mathcal{G}\right]\right].$$

PREUVE. Rappelons que $\mathbb{E}_P\left[\mathbb{E}_P\left[X \mid \mathcal{G}\right]\right] = \mathbb{E}_P\left[X\right]$.

Supposons que $\mathcal{G}$ est engendrée par les atomes $(A_i)_{i\in\mathcal{I}}$ et que ces atomes sont tous de probabilité non-nulle.

$$\begin{aligned}
\mathrm{Var}_P\left[X\right] &= \mathbb{E}_P\left[\left(X - \mathbb{E}_P\left[X\right]\right)^2\right] \\
&= \mathbb{E}_P\left[\left(X - \mathbb{E}_P\left[X \mid \mathcal{G}\right] + \mathbb{E}_P\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right]\right)^2\right] \\
&= \mathbb{E}_P\left[\left(X - \mathbb{E}_P\left[X \mid \mathcal{G}\right]\right)^2\right] \\
&\quad +2\mathbb{E}_P\left[\left(X - \mathbb{E}_P\left[X \mid \mathcal{G}\right]\right)\left(\mathbb{E}_P\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right]\right)\right] \\
&\quad +\mathbb{E}_P\left[\left(\mathbb{E}_P\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right]\right)^2\right] \\
&\quad \text{invoquer deux fois la proposition B.17} \\
&= \mathbb{E}_P\left[\mathbb{E}_P\left[\left(X - \mathbb{E}_P\left[X \mid \mathcal{G}\right]\right)^2 \mid \mathcal{G}\right]\right] \\
&\quad +2\mathbb{E}_P\left[\mathbb{E}_P\left[\left(X - \mathbb{E}_P\left[X \mid \mathcal{G}\right]\right) \mid \mathcal{G}\right]\left(\mathbb{E}_P\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right]\right)\right] \\
&\quad +\mathrm{Var}_P\left[\mathbb{E}_P\left[X \mid \mathcal{G}\right]\right] \\
&\quad \text{le second terme est nul} \\
&= \mathbb{E}_P\left[\mathrm{Var}_P\left[X \mid \mathcal{G}\right]\right] + \mathrm{Var}_P\left[\mathbb{E}_P\left[X \mid \mathcal{G}\right]\right].
\end{aligned}$$

□

Dans tout le texte, on parle d'un espace probabilisé $(\Omega, \mathcal{F}, P)$. On note $\mathfrak{L}_2(\Omega, \mathcal{F}, P)$ l'ensemble des variables aléatoires de carré intégrable sur $(\Omega, \mathcal{F}, P)$:

$$\mathfrak{L}_2(\Omega, \mathcal{F}, P) = \left\{ X : X \text{ Borel} - \text{ mesurable de}(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})), \mathbb{E}_P\left[X^2\right] < \infty \right\}.$$

Cet ensemble est un espace vectoriel muni d'une pseudo-norme

$$\|X\| = \left(\mathbb{E}_P\left[X^2\right]\right)^{1/2}.$$

On parle de pseudo-norme car deux variables aléatoires presque-sûrement égales sont confondues par cette pseudo-norme. Si on identifie les variables aléatoires presque-sûrement égales on obtient une véritable norme et l'espace $L_2(\Omega, \mathcal{F}, P)$ qui est un espace de Hilbert. Un élément de $L_2(\ldots)$ correspond à une classe d'équivalence d'éléments de $\mathfrak{L}_2(\ldots)$, chaque élément de classe d'équivalence est appelé version de la classe d'équivalence.

NOTATION B.21 On abrège souvent $L_2(\Omega, \mathcal{F}, P)$ (respectivement $\mathfrak{L}_2(\Omega, \mathcal{F}, P)$) en $L_2(P)$ (resp. en $\mathfrak{L}_2(P)$).

Considérons X et Y deux variables aléatoires, et supposons $X \in \mathfrak{L}_2(P)$, quelle est la meilleure prédiction de X à partir de Y ? Existe-t-il une fonction f de Y qui minimise

$$\mathbb{E}_P\left[(X - f(Y))^2\right]?$$

S'il existe une fonction de Y qui réalise ce minimum, est-elle unique ? Une variable aléatoire de la forme $f(Y)$ est une variable aléatoire qui appartient à la tribu engendrée par Y (notée $\sigma(Y)$). On peut reposer la question précédente en cherchant s'il existe une variable aléatoire $Z \in \mathfrak{L}_2(\Omega, \sigma(Y), P)$ qui minimise

$$\mathbb{E}_P\left[(X - Z)^2\right]?$$

Comme nous ne voulons pas recourir à l'artifice des lois conditionnelles de X sachant Y, il faut trouver un autre mécanisme de construction pour définir une espérance conditionnelle raisonnable. Ce mécanisme sera fourni par la structure hilbertienne de $L_2(\ldots)$

Sous-espaces fermés de $\mathfrak{L}_2(P)$. Rappelons quelques résultats de mesurabilité.

En combinant

LEMME B.22 *Soient $(\Omega, \mathcal{F})$ un espace probabilisable, et $(X_n)_n$ une suite de variables aléatoires de $(\Omega, \mathcal{F})$ dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Si la suite $(X_n)_n$ converge simplement vers une fonction X de Ω dans $\mathbb{R}$ alors X est une variable aléatoire.*

et

LEMME B.23 *Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$. Soit $(Z_n)_n$ une suite de variables aléatoires réelles $\mathcal{G}$-mesurables qui converge presque sûrement vers une variable aléatoire Z admet une version $\mathcal{G}$-mesurable.*

THÉORÈME B.24 *Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé, et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, l'ensemble des variables aléatoires (ayant une version) $\mathcal{G}$-mesurables de carré intégrable est un sous-espace fermé de $\mathfrak{L}_2(\Omega, \mathcal{F}, P)$, on le notera $\mathfrak{L}_2(\Omega, \mathcal{G}, P)$.*

REMARQUE B.25 La fermeture du sous-espace de $\mathfrak{L}_2(\Omega, \mathcal{G}, P)$ se démontre de la même façon que la complétude de $\mathfrak{L}_2(\Omega, \mathcal{F}, P)$, et que la complétude des espaces $\mathfrak{L}_p(\Omega, \mathcal{F}, P)$ pour $p \geq 1$. Le fait que P soit une mesure de probabilité n'est pas important. Les arguments fonctionneraient avec des mesures finies.

THÉORÈME B.26 [THÉORÈME DE PROJECTION] *Soit E un espace de Hilbert et F un sous-ensemble convexe fermé de F. Pour tout $x \in E$, il existe un unique $y \in F$, tel que*

$$\|x - y\| = \inf_{z \in F} \|x - z\|.$$

On appelle y la projection (orthogonale) de $x$sur F.

De plus, si F est un sous-espace vectoriel, x et y vérifient la relation de Pythagore

$$\|x\|^2 = \|y\|^2 + \|x - y\|^2.$$

Projection dans les espaces de Hilbert. Comme $L_2(\Omega, \mathcal{G}, P)$ est une partie convexe fermée de $L_2(\Omega, \mathcal{F}, P)$, l'existence et l'unicité de la projection sur une partie convexe fermée d'un espace de Hilbert nous donnent le corollaire suivant.

COROLLAIRE B.27 *Etant donné un élément X de $L_2(\Omega, \mathcal{F}, P)$ et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, il existe un unique élément de $L_2(\Omega, \mathcal{G}, P)$, qui minimise la distance à X.*

En Probabilités, on travaille sur l'espace $\mathfrak{L}_2(\Omega, \mathcal{F}, P)$ et ses sous-espaces. En utilisant le théorème de Projection, on obtient l'existence d'une espérance conditionnelle pour les variables de carré intégrable.

COROLLAIRE B.28 *Etant donnée $X \in \mathfrak{L}_2(\Omega, \mathcal{F}, P)$ et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, il existe un élément Y de $\mathfrak{L}_2(\Omega, \mathcal{G}, P)$, noté $\mathbb{E}_P[X \mid \mathcal{G}]$ et qui minimise*

$$\mathbb{E}_P\left[(X - Z)^2\right] \ pourZ \in \mathfrak{L}_2(\Omega, \mathcal{G}, P).$$

Tout autre élément de $\mathfrak{L}_2(\Omega, \mathcal{G}, P)$, qui minimise (B.28) est P-presque sûrement égal à Z.

On peut utiliser ce corollaire pour établir le théorème suivant, qui montre que pour les variables de carré intégrables, on peut construire des versions de l'espérance conditionnelle. Dans le cas de ces variables aléatoires de carré intégrable, l'espérance conditionnelle est une projection orthogonale.

THÉORÈME B.29 *. Soient $X \in \mathfrak{L}_2(\Omega, \mathcal{F}, P)$ et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, alors si Z est un représentant de la projection orthogonale de X sur $L_2(\Omega, \mathcal{F}, P)$*

$$\forall B \in \mathcal{G}, \quad \mathbb{E}_P[\mathbf{I}_B X] = \mathbb{E}_P[\mathbf{I}_B Z]$$

PREUVE. Soit Z un représentant de la projection orthogonale de X sur $L_2(\Omega, \mathcal{G}, P)$ et B un élément de $\mathcal{G}$.

Si on considère l'expression

$$\mathbb{E}_P[(X - Z)\mathbf{I}_B],$$

c'est le produit intérieur de $\mathbf{I}_B$ (de $\mathfrak{L}(\Omega, \mathcal{F}, P)$) et de $X - Z$ (qui appartient au sous espace orthogonal à $\mathfrak{L}(\Omega, \mathcal{F}, P)$). Cette quantité est donc nulle. □

Convention. Dans la suite $\mathbb{E}[X \mid \mathcal{G}]$ désignera une version de l'espérance conditionnelle de X sachant $\mathcal{G}$.

EXERCICE B.30 Vérifier que $\mathbb{E}[X \mid \mathcal{G}]$ est l'unique élément de $\mathfrak{L}_2(\Omega, \mathcal{G}, P)$ qui vérifie

$$\forall Z \in \mathfrak{L}_2(\Omega, \mathcal{G}, P), \quad \mathbb{E}_P[ZX] = \mathbb{E}_P[Z\mathbb{E}_P[X \mid \mathcal{G}]].$$

Espérance conditionnelle dans $\mathfrak{L}_1(P)$. Pour construire l'espérance conditionnelle d'une variable aléatoire, il n'est pas nécessaire que cette variable soit de carré intégrable, il suffit qu'elle soit intégrable. C'est la signification du théorème suivant.

THÉORÈME B.31 *Si Y est une variable aléatoire non-négative, ou si Y est intégrable ($Y \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$), alors il existe une variable aléatoire $\mathcal{G}$-mesurable, intégrable, notée $\mathbb{E}[Y \mid \mathcal{G}]$ appartenant à $\mathcal{L}_1(\Omega, \mathcal{F}, P)$ vérifiant*

$$\forall B \in \mathcal{G}, \mathbb{E}[\mathbf{I}_B Y] = \mathbb{E}[\mathbf{I}_B \mathbb{E}[Y \mid \mathcal{G}]].$$

*De plus, toute variable aléatoire $\mathcal{G}$-mesurable, intégrable Z qui satisfait aussi (**??**) est presque sûrement égale à $\mathbb{E}[Y \mid \mathcal{G}]$:*

$$P\{Z = \mathbb{E}[Y \mid \mathcal{G}]\} = 1.$$

EXERCICE B.32 Soit $\mathcal{G}'$ un π-système qui contient Ω et engendre $\mathcal{G}$. Si Z est une variable $\mathcal{G}$-mesurable, intégrable qui vérifie

$$\forall B \in \mathcal{G}', \mathbb{E}[\mathbf{I}_B Y] = \mathbb{E}[\mathbf{I}_B \mathbb{E}[Y \mid \mathcal{G}]]$$

alors $Z = \mathbb{E}[Y \mid \mathcal{G}]$.

Pour établir le théorème B.31, on utilise des arguments de passage à limite qui constituent la « mécanique habituelle » .

PROPOSITION B.33 *Si $(Y_n)_n$ est une suite croissante de variables aléatoires de carré intégrable positives telle que $Y_n \nearrow Y$ alors il existe une variable aléatoire $\mathcal{G}$-mesurable Z telle que*

$$\mathbb{E}_P[Y_n \mid \mathcal{G}] \nearrow Z \quad p.s.$$

PREUVE. Comme $(Y_n)_n$ est croissante, d'après la Proposition précédente $(\mathbb{E}_P[Y_n \mid \mathcal{G}])_n$ est une suite (p.s.) croissante de variables aléatoires $\mathcal{G}$-mesurables, elle admet une limite (finie ou non) $\mathcal{G}$-mesurable. □

PREUVE.(THÉORÈME B.31.) Sans perdre en généralité, on peut supposer $Y \geq 0$ (si ce n'est pas le cas, on pose $Y = Y_+ - Y_-$ où $Y_+ = |Y|\mathbf{I}_{Y \geq 0}$ et $Y_- = |Y|\mathbf{I}_{Y<0}$, on traite Y_+ et Y_- séparément en utilisant la linéarité de l'espérance).

Définissons

$$Y_n = Y\mathbf{I}_{|Y| \leq n}.$$

On a $Y_n \nearrow Y$ partout. La variable aléatoire Y_n est bornée donc de carré intégrable. Une variable aléatoire $\mathbb{E}_P[Y_n \mid \mathcal{G}]$ est donc bien définie pour chaque n.

La suite $\mathbb{E}_P[Y_n \mid \mathcal{G}]$ (de versions) est monotone P-p.s. Elle converge donc de manière monotone vers une variable aléatoire Z à valeurs dans $\mathbb{R}_+ \cup \{\infty\}$, $\mathcal{G}$-mesurable. Nous devons montrer que cette variable aléatoire Z est en fait dans $\mathcal{L}_1(\Omega, \mathcal{F}, P)$ et qu'elle satisfait la propriété **??**.

Par convergence monotone :

$$\mathbb{E}_P[Y] = \lim_n \mathbb{E}_P[Y_n] = \lim_n \mathbb{E}_P[\mathbb{E}_P[Y_n \mid \mathcal{G}]] = \mathbb{E}_P\left[\lim_n \mathbb{E}_P[Y_n \mid \mathcal{G}]\right] = \mathbb{E}_P[Z].$$

Soit $A \in \mathcal{G}$, on a par convergence monotone

$$\mathbb{E}_P[\mathbf{I}_A Y_n] \nearrow \mathbb{E}_P[\mathbf{I}_A Y]$$

et donc

$$\mathbb{E}_P\left[\mathbf{I}_A \mathbb{E}_P\left[Y_n \mid \mathcal{G}\right]\right] \nearrow \mathbb{E}_P\left[\mathbf{I}_A Y\right].$$

Par convergence monotone encore :

$$\mathbb{E}_P\left[\mathbf{I}_A \lim_n \mathbb{E}_P\left[Y_n \mid \mathcal{G}\right]\right] \nearrow \mathbb{E}_P\left[\mathbf{I}_A Z\right].$$

□

LA CONSTRUCTION DE L'ESPÉRANCE CONDITIONNELLE PAR LA DÉRIVÉE DE RADON-NYKODIM Tous les auteurs ne construisent pas l'espérance conditionnelle à partir de la notion de projection dans $L_2(\ldots)$. Beaucoup de livres classiques abordent directement la construction de l'espérance conditionnelle des variables aléatoires intégrables. Pour cela, ils s'appuient sur un résultat puissant de la théorie de la mesure (la version donnée ici n'est pas la plus générale possible) :

THÉORÈME B.34 THÉORÈME DE RADON-NIKODYM. *Soient P et Q deux mesures positives finies sur $(\Omega, \mathcal{F})$.Supposons P et absolument continue par rapport à Q ($P\{A\} > 0 \Rightarrow Q\{A\} > 0$), alors il existe une fonction f $\mathcal{F}$-mesurable (une variable aléatoire) telle que pour tout $A \in \mathcal{F}$*

$$\int_\Omega \mathbf{I}_A(\omega) f(\omega) \mathrm{d}P = \int_\Omega \mathbf{I}_A(\omega) \mathrm{d}Q.$$

REMARQUE B.35 La fonction f est appelée la dérivée de Radon-Nikodym de P par rapport à Q. Elle est notée $\frac{\mathrm{d}P}{\mathrm{d}Q}$. Elle est définie Q-presque sûrement.

EXERCICE B.36 Vérifier que s'il existe deux fonctions $\mathcal{F}$-mesurables f et f' qui vérifient les conditions du théorèmes, elles sont Q-presque sûrement égales.

Voici une démonstration du Théorème B.31 à partir du Théorème de Radon-Nikodym.

PREUVE. Soit X une variable aléatoire intégrable. Sans perdre en généralité, on suppose $X \geq 0$. On définit sur $\mathcal{G}$ la fonctionnelle :

$$A \mapsto \mathbb{E}_P\left[\mathbf{I}_A X\right].$$

Cette fonctionnelle définit une mesure positive Q sur $(\Omega, \mathcal{G})$ (linéarité de l'espérance, convergence dominée ...).

Cette mesure est absolument continue par rapport à la mesure de probabilité P. Le théorème de Radon-Nikodym nous indique alors qu'il existe une fonction $\mathcal{G}$-mesurable (une variable aléatoire $\mathcal{G}$-mesurable) f qui vérifie pour tout $A \in \mathcal{G}$:

$$\mathbb{E}_P\left[\mathbf{I}_A X\right] = \int_\Omega \mathbf{I}_A(\omega) f(\omega) \mathrm{d}P = \mathbb{E}_P\left[\mathbf{I}_A f\right].$$

Autrement dit f est une version de l'espérance conditionnelle de X sachant $\mathcal{G}$. □

Propriétés de l'espérance conditionnelle

AVERTISSEMENT B.37 Dans toute cette section, on se réfère à un espace probabilisé $(\Omega, \mathcal{F}, P)$, à une sous-tribu $\mathcal{G}$ de $\mathcal{F}$. Les variables aléatoires $(X_n)_n, (Y_n)_n, X, Y, Z$ sont supposées intégrables. L'expression p.s. signifie P-presque sûrement.

La propriété la plus simple, mais pas la moins utile est

PROPOSITION B.38 *Si $X \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$ alors*

$$\mathbb{E}_P\left[X\right] = \mathbb{E}_P\left[\mathbb{E}_P\left[X \mid \mathcal{G}\right]\right].$$

PROPOSITION B.39 *Si $X \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$ et X est $\mathcal{G}$-mesurable alors*

$$X = \mathbb{E}_P[X \mid \mathcal{G}] \quad P\text{-p.s.}$$

En passant par les fonctions étagées, on peut vérifier

PROPOSITION B.40 *Soit $X \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$ et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$, alors pour tout $Y \in \mathfrak{L}_1(\Omega, \mathcal{G}, P)$, tel que $\mathbb{E}_P[|XY|] < \infty$*

$$\mathbb{E}_P[XY] = \mathbb{E}_P[Y\mathbb{E}_P[X \mid \mathcal{G}]].$$

PROPOSITION B.41 *Si $X, Y \in \mathfrak{L}_1(\Omega, \mathcal{F}, P)$ et Y est $\mathcal{G}$-mesurable alors*

$$\mathbb{E}_P[XY \mid \mathcal{G}] = Y\mathbb{E}_P[X \mid \mathcal{G}] \quad P\text{-p.s.}$$

PREUVE. Comme $Y\mathbb{E}_P[X \mid \mathcal{G}]$ est bien $\mathcal{G}$-mesurable, il suffit de vérifier que pour tout $B \in \mathcal{G}$,

$$\mathbb{E}_P[\mathbf{I}_B XY] = \mathbb{E}_P[\mathbf{I}_B (Y\mathbb{E}_P[X \mid \mathcal{G}])].$$

Mais

$$\begin{aligned}
\mathbb{E}_P[\mathbf{I}_B XY] &= \mathbb{E}_P[(\mathbf{I}_B Y)X] \\
&\quad \text{comme } \mathbf{I}_B Y \in \mathfrak{L}_1(\Omega, \mathcal{G}, P) \text{d'après} \text{B.40} \\
&= \mathbb{E}_P[(\mathbf{I}_B Y)\mathbb{E}_P[X \mid \mathcal{G}]] \\
&= \mathbb{E}_P[\mathbf{I}_B (Y\mathbb{E}_P[X \mid \mathcal{G}])].
\end{aligned}$$

□

Opérations usuelles

OPÉRATIONS LINÉAIRES On peut étendre la Proposition B.14 au cas $\mathfrak{L}_1(\Omega, \mathcal{F}, P)$:

PROPOSITION B.42 *Soient X et Y deux variables aléatoires. Soit a_1, a_2 et b trois réels. Alors*

$$\mathbb{E}_P[a_1 X + a_2 Y + b \mid \mathcal{G}] = a_1\mathbb{E}_P[X \mid \mathcal{G}] + a_2\mathbb{E}_P[Y \mid \mathcal{G}] + b \quad p.s.$$

PREUVE. Il s'agit de montrer qu'il existe une version de $\mathbb{E}_P[a_1 X + a_2 Y + b \mid \mathcal{G}]$ qui est presque sûrement égale à une combinaison linéaire d'une version de $\mathbb{E}_P[X \mid \mathcal{G}]$ et d'une version de $\mathbb{E}_P[Y \mid \mathcal{G}]$. □

COMPARAISONS On peut aussi étendre la Proposition B.15

PROPOSITION B.43 *Si X est une variable aléatoire positive de $\mathfrak{L}_1(\Omega, \mathcal{F}, P)$, alors*

$$\mathbb{E}_P[X \mid \mathcal{G}] \geq 0 p.s.$$

On répète la démonstration de la Proposition B.15 (qui n'utilise pas le fait que $X \in \mathfrak{L}_2(P)$ mais une garantie sur l'existence de l'espérance conditionnelle).

THÉORÈMES DE CONVERGENCE CONDITIONNELS Les théorèmes de convergence de la théorie de la mesure (convergence monotone, Fatou, convergence dominée).

THÉORÈME B.44 CONVERGENCE MONOTONE. *Soit $(X_n)_n$ une suite de variables aléatoires, telles que $X_n \nearrow X$ p.s., X intégrable, alors pour toute suite de versions des l'espérances conditionnelles des v.a. X_n et X*

$$\mathbb{E}_P[X_n \mid \mathcal{G}] \nearrow \mathbb{E}_P[X \mid \mathcal{G}] \, p.s.$$

PREUVE. La suite $X - X_n$ est positive et décroît vers 0 presque sûrement. Il suffit de montrer que $\mathbb{E}_P[X - X_n \mid \mathcal{G}] \searrow 0$ p.s. Notons d'abord que la suite $\mathbb{E}_P[X - X_n \mid \mathcal{G}]$ converge p.s. vers une limite positive ou nulle. Il faut vérifier que cette limite est bien presque sûrement nulle.

Soit $A \in \mathcal{G}$:

$$\begin{aligned}
\mathbb{E}_P\left[\mathbf{I}_A \lim_n \mathbb{E}_P[X - X_n \mid \mathcal{G}]\right] &= \lim_n \mathbb{E}_P[\mathbf{I}_A \mathbb{E}_P[X - X_n \mid \mathcal{G}]] \\
&\quad \text{théorème de convergence monotone} \\
&= \lim_n \mathbb{E}_P[\mathbf{I}_A (X_n - X)] \\
&\quad \text{théorème convergence monotone} \\
&= 0.
\end{aligned}$$

□

THÉORÈME B.45 FATOU. *Soit $(X_n)_n$ une suite de variables aléatoires positives*

$$\mathbb{E}_P\left[\liminf_n X_n \mid \mathcal{G}\right] \leq \liminf_n \mathbb{E}_P[X_n \mid \mathcal{G}] \quad p.s.$$

PREUVE. Soit $B \in \mathcal{G}$. Notons $X = \liminf_n X_n$, X est une variable aléatoire positive inégrable. Notons $Y = \liminf_n \mathbb{E}_P[X_n \mid \mathcal{G}]$, Y est une variable aléatoire positive $\mathcal{G}$-mesurable. Le théorème compare $\mathbb{E}_P[X \mid \mathcal{G}]$ et Y. Comme dans la version classique, la démonstration consiste à se rammener à un argument de convergence monotone. Définissons $Z_k = \inf_{n \geq k} X_n$. On a $Z_k \nearrow_k \liminf_n X_n = X$. D'après le théorème B.44,

$$\mathbb{E}_P[Z_k \mid \mathcal{G}] \nearrow_k \mathbb{E}_P\left[\liminf_n X_n \mid \mathcal{G}\right] \quad \text{p.s.}$$

Par ailleurs, pour tout $n \geq k$, $X_n \geq Z_k$, donc d'après le théorème de comparaison

$$\forall n \geq k \quad \mathbb{E}_P[Z_k \mid \mathcal{G}] \leq \mathbb{E}_P[X_n \mid \mathcal{G}] \quad \text{p.s.}$$

car une réunion dénombrable d'ensembles P-négligeables est P-négligeable. Donc pour tout k,

$$\mathbb{E}_P[Z_k \mid \mathcal{G}] \leq \liminf_n \mathbb{E}_P[X_n \mid \mathcal{G}] \quad \text{p.s.}$$

Ce qui implique

$$\lim_k \mathbb{E}_P[Z_k \mid \mathcal{G}] \leq \liminf_n \mathbb{E}_P[X_n \mid \mathcal{G}] \quad \text{p.s.}$$

□

THÉORÈME B.46 CONVERGENCE DOMINÉE. *Soit V une variable aléatoire intégrable. Soit $(X_n)_n$ une suite de variables aléatoires telles que $|X_n| \leq V$ et $X_n \to X$ p.s., alors pour toute suite de versions des l'espérances conditionnelles des v.a. X_n et X*

$$\mathbb{E}_P[X_n \mid \mathcal{G}] \to \mathbb{E}_P[X \mid \mathcal{G}] \quad p.s.$$

PREUVE. Notons $Y_n = \inf_{m \geq n} X_m$ et $Z_n = \sup_{m \geq n} X_m$. Nous avons $Z_n \leq V$ et $-V \leq Y_n$. Comme Y_n est croissante de limite X et Z_n décroissante de limite X. Le théorème de convergence monotone conditionnelle nous permet de conclure que $\mathbb{E}_P[Y_n \mid \mathcal{G}] \nearrow \mathbb{E}_P[X \mid \mathcal{G}]$ et $\mathbb{E}_P[Z_n \mid \mathcal{G}] \searrow \mathbb{E}_P[X \mid \mathcal{G}]$p.s. On termine en observant que pour tout n

$$\mathbb{E}_P[Y_n \mid \mathcal{G}] \leq \mathbb{E}_P[X_n \mid \mathcal{G}] \leq \mathbb{E}_P[Z_n \mid \mathcal{G}] \quad \text{p.s.}$$

□

INÉGALITÉS L'inégalité de Jensen possède une version conditionnelle.

THÉORÈME B.47 INÉGALITÉ DE JENSEN. *Si g est une fonction convexe sur $\mathbb{R}$, telle que $\mathbb{E}[|g(X)|] < \infty$ alors*

$$g(\mathbb{E}[X \mid \mathcal{G}]) \leq \mathbb{E}[g(X) \mid \mathcal{G}] \; p.s..$$

PREUVE. Une fonction convexe semi-continue inférieurement peut être représentée comme un supremum dénombrable de fonctions affines : il existe une suite de paires de coefficients $(a_n, b_n)_n$ telle que pour tout x :

$$g(x) = \sup_n [a_n x + b_n].$$

$$\begin{aligned} g(\mathbb{E}[X \mid \mathcal{G}]) &= \sup_n [a_n \mathbb{E}[X \mid \mathcal{G}] + b_n] \\ &\quad \text{par linéarité de l'espérance conditionnelle} \\ &= \sup_n [\mathbb{E}[a_n X + b_n \mid \mathcal{G}]] \\ &\leq \mathbb{E}\left[\sup_n (a_n X + b_n) \mid \mathcal{G}\right] P\text{-p.s.} \\ &\quad \text{comparaison des espérances conditionnelles.} \end{aligned}$$

□

Exhiber une suite de coefficients convenables pour les fonctions $x \mapsto |x|$, $x \mapsto x^2$, $x \mapsto \exp(x)$.

Vérifier que toute fonction f qui admet une représentation comme un supremum de fonctions affines :

$$\text{pour tout } x : f(x) = \sup_n [a_n x + b_n]$$

est convexe. Est-elle aussi nécessairement continue ?

Vérifier que toute fonction convexe continue f admet une représentation comme supremum denombrable de fonctions affines.

CONDITIONNEMENT SUCCESSIFS Quand on conditionne par deux tribus imbriquées, c'est la plus petite tribu qui l'emporte.

PROPOSITION B.48 *Soient $\mathcal{G} \subseteq \mathcal{H}$ deux sous-tribus de $\mathcal{F}$.*

$$\mathbb{E}_P\left[\mathbb{E}_P\left[X \mid \mathcal{H}\right] \mid \mathcal{G}\right] = \mathbb{E}_P\left[\mathbb{E}_P\left[X \mid \mathcal{G}\right] \mid \mathcal{H}\right] = \mathbb{E}_P\left[X \mid \mathcal{G}\right].$$

PREUVE. La seconde égalité est immédiate car toute variable aléatoire $\mathcal{G}$-mesurable est aussi $\mathcal{H}$-mesurable.

Pour la première égalité : pour tout $B \in \mathcal{G}$,

$$\begin{aligned}\mathbb{E}_P\left[\mathbf{I}_B \mathbb{E}_P\left[\mathbb{E}_P\left[X \mid \mathcal{H}\right] \mid \mathcal{G}\right]\right] &= \mathbb{E}_P\left[\mathbf{I}_B \mathbb{E}_P\left[X \mid \mathcal{H}\right]\right] \\ &\quad \text{comme } B \in \mathcal{H} \\ &= \mathbb{E}_P\left[\mathbf{I}_B X\right].\end{aligned}$$

□

PROPOSITION B.49 *Si X est indépendant de $\mathcal{G}$, alors*

$$\mathbb{E}_P\left[X \mid \mathcal{G}\right] = \mathbb{E}_P\left[X\right].$$

INDÉPENDANCE PREUVE. Notons que $\mathbb{E}_P[X]$ est $\mathcal{G}$-mesurable. Soit $B \in \mathcal{G}$,

$$\begin{aligned}\mathbb{E}_P\left[\mathbf{I}_B X\right] &= \mathbb{E}_P\left[\mathbf{I}_B\right] \mathbb{E}_P\left[X\right] \\ &\quad \text{par indépendance} \\ &= \mathbb{E}_P\left[\mathbf{I}_B \times \mathbb{E}_P\left[X\right]\right].\end{aligned}$$

Donc $\mathbb{E}_P[X] = \mathbb{E}_P[X \mid \mathcal{G}]$. □

PROPOSITION B.50 *Si $\mathcal{H}$ est une sous-tribu indépendante de $\sigma(\mathcal{G}, \sigma(X))$ alors*

$$\mathbb{E}_P\left[X \mid \sigma(\mathcal{G}, \mathcal{H})\right] = \mathbb{E}_P\left[X \mid \mathcal{G}\right] \quad p.s.$$

PREUVE. Rappelons-nous que l'espérance conditionnelle par rapport à $\sigma(\mathcal{G}, \mathcal{H})$ peut être caractérisée en considérant un π-système contenant Ω et engendrant $\sigma(\mathcal{G}, \mathcal{H})$, par exemple $\mathcal{G} \times \mathcal{H}$. Soient $B \in \mathcal{G}$ et $C \in \mathcal{H}$,

$$\begin{aligned}\mathbb{E}_P\left[\mathbf{I}_B \mathbf{I}_C \mathbb{E}_P\left[X \mid \mathcal{G}\right]\right] &= \mathbb{E}_P\left[\mathbf{I}_B \mathbb{E}_P\left[X \mid \mathcal{G}\right]\right] \times \mathbb{E}_P\left[\mathbf{I}_C\right] \\ &\quad \text{par indépendance de } \mathcal{H} \text{ et de} \sigma(\mathcal{G}, \sigma(X)) \\ &= \mathbb{E}_P\left[\mathbf{I}_B X\right] \times \mathbb{E}_P\left[\mathbf{I}_C\right] \\ &= \mathbb{E}_P\left[\mathbf{I}_C \mathbf{I}_B X\right] \\ &\quad \text{par indépendance de } \mathcal{H} \text{ et de } \sigma(\mathcal{G}, \sigma(X)).\end{aligned}$$

□

Cas simple du conditionnement par une tribu discrète

On revient sur la situation élémentaire : $(\Omega, \mathcal{F}, P)$ désigne toujours un espace probabilisé. Et $\mathcal{G}$ désigne une sous-tribu discrète de $\mathcal{F}$ engendrée par une partition dénombrable $(A_n)_n$ de Ω.

À partir des espérances conditionnelles sachant A_n, ou des probabilités conditionnelles sachant les événements A_n, on peut définir une fonction N de $\Omega \times \mathcal{F}$ par

$$N(\omega, B) = \mathbb{E}_P[\mathbf{I}_B \mid \mathcal{G}] = P\{B \mid A_n\} \text{ avec } \omega \in A_n \,.$$

La fonction N possède deux propriétés remarquables :

i) Pour tout $\omega \in \Omega$, $N(\omega, \cdot)$ définit une probabilité sur $(\Omega, \mathcal{F})$.
ii) Pour tout événement $B \in \mathcal{F}$, la fonction $N(\cdot, B)$ est une fonction $\mathcal{G}$-mesurable.

Dans cette situation simple, on observe que s'il est naturel de définir l'espérance conditionnelle à partir des probabilités conditionnelles, on pourrait tout aussi bien construire les probabilités conditionnelles à partir des espérances conditionnelles.

Dans le cas général, on cherche à construire une notion de probabilité conditionnelle lorsque la tribu conditionnante n'est pas discrète.

Pour chaque $B \in \mathcal{F}$, nous avons l'assurance qu'il existe une variable aléatoire $\sigma(X)$-mesurable qui est P-p.s. une version de l'espérance conditionnelle de $\mathbf{I}_B$ sachant X. Pour n'importe quelle collection dénombrable d'événements $(B_n)_n$ de $\mathcal{F}$, nous sommes même assurés de l'existence d'une collection de variables aléatoires qui sur un événement de P-probabilité 1 constituent chacune une version de l'espérance de $\mathbf{I}_{B_n}$ sachant X. Si les B_n forment une suite croissante vers B, nous sommes même assurés (conséquence du théorème B.44) que sauf sur une ensemble de P-probabilité nulle, on aura

$$\mathbb{E}_P\left[\mathbf{I}_{B_n} \mid X\right] \nearrow \mathbb{E}_P\left[\mathbf{I}_B \mid X\right].$$

Il est donc très tentant de définir une probabilité conditionnelle relativement à X comme une fonction de $X(\Omega) \times \mathcal{F} \to [0,1]$, $(x, B) \mapsto \mathbb{E}_P\left[\mathbf{I}_B \mid \sigma(X)\right]$ à partir d'une version de l'espérance conditionnelle par rapport à $\sigma(X)$. Malheureusement, nous n'avons pas les moyens de garantir que cet objet possède les propriétés d'une probabilité sur $(\Omega, \mathcal{F})$ P-p.s. Le problème vient ici non pas du caractère diffus de la loi de X mais de la richesse de $\mathcal{F}$, notamment du fait qu'elle contient éventuellement une quantité non-dénombrable d'événements, et que l'on peut construire une quantité non-dénombrable de suite croissantes d'événements sur lesquelles il faudra vérifier la propriété de monotonie (une réunion non-dénombrable d'événements P-négligeables n'est pas forcément P-négligeable).

Dans la suite, allons d'abord passer en revue un autre cas facile, où on peut définir une probabilité conditionnelle qui possède même une densité par rapport à une mesure de référence, puis nous verrons que si Ω n'est pas trop gros, on peut garantir l'existence d'une probabilité conditionnelle.

Cas d'une densité jointe

Nous envisageons ici un cas simplifié qui capture néanmoins la difficulté du cas général.

On suppose maintenant que $\Omega = \mathbb{R}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$. On a $\omega = (x, y)$ et deux variables aléatoires $X(x, y) = x$ et $Y(x, y) = y$ (les projections coordonnées). La loi P du couple aléatoire (X, Y) admet une densité $p(x, y)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$. La variable Y est supposée de carré intégrable. On note p_X la densité marginale de la loi de X :

$$p_X(x) = \int_{\mathbb{R}} p(x, y) \mathrm{d}y.$$

Et on note $D = \{x : p_X(x) > 0\}$ le support de la densité p_X.

Exercice B.51 Vérifier que p_X est bien la densité de $P \circ X^{-1}$.

Le fait de disposer d'une densité jointe nous permet de calculer facilement l'espérance conditionnelle et de définir tout aussi facilement ce que nous appellerons une « probabilité conditionnelle de Y sachant X » .

PROPOSITION B.52 *Soit X, Y deux variables aléatoires dont la loi conjointe P admet une densité jointe $p(.,.)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$. La densité de la première marginale par rapport à la mesure de Lebesgue sur $\mathbb{R}$ est notée p_X. Soit f une fonction mesurable sur $\mathbb{R}$ On suppose de plus $f(Y)$ intégrable. Toute version de l'espérance conditionnelle de $f(Y)$ par rapport à X satisfait P-presque surement l'égalité*

$$\mathbb{E}_P\left[f(Y) \mid X\right] = \int_{\mathbb{R}} \frac{p(x,y)}{p_X(x)} f(y)\mathrm{d}y\,.$$

La démonstration s'appuie essentiellement sur le théorème de Tonelli-Fubini rappelée ici pour mémoire.

THÉORÈME B.53 *[Tonelli-Fubini] Soit $(\mathcal{X}, \mathcal{A})$ et $(\mathcal{Y}, \mathcal{B})$ deux espaces mesurables, μ et ν deux mesures σ-finie sur ces espaces, $\mu \otimes \nu$ la mesure produit, et f une fonction réelle $\mathcal{A} \otimes \mathcal{B}$-mesurable telle que $\int |f| \mathrm{d}\mu \otimes \nu < 0$. Les propriétés suivantes sont vérifiées :*

1. *$\forall x \in \mathcal{X}, \quad y \mapsto f(x,y)$ est $\mathcal{B}$-mesurable.*
2. *La fonction $x \mapsto \int_{\mathcal{Y}} f(x,y)\mathrm{d}\nu(y)$ est $\mathcal{A}$-mesurable, finie μ-presque partout et*

$$\int_{\mathcal{X}\times\mathcal{Y}} f \mathrm{d}\mu \otimes \nu = \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} f(x,y)\mathrm{d}\nu(y)\right] \mathrm{d}\mu(x).$$

PREUVE. On considère la fonction $(x,y) \mapsto f(y)p(x,y)$. Elle est par hypothèse Lebesgue-intégrable. Le théorème de Fubini nous permet d'affirmer que $x \mapsto \int f(y)p(x,y)\mathrm{d}y$ est définie Lebesgue presque-partout, mesurable, et que

$$\int_{\mathbb{R}^2} f(y)p(x,y)\mathrm{d}x\mathrm{d}y = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f(y)p(x,y)\mathrm{d}y\right] \mathrm{d}x.$$

La densité p_X définit la trace de la loi P sur $\sigma(X)$. Notons pour x tel que $p_X(x) > 0$,

$$h(x) = \int f(y)\frac{p(x,y)}{p_X(x)}\mathrm{d}y.$$

et décidons $h(x) = 0$ ailleurs. La fonction $h()$ est bien sûr $\sigma(X)$-mesurable (quotient de deux fonctions mesurables).

Soit $B \in \sigma(X) \subseteq \sigma(X,Y)$, il existe $A \subseteq \mathbb{R}$, tel que $B = X^{-1}(A)$.

$$\begin{aligned}
\mathbb{E}_P\left[f(Y)\mathbf{I}_B\right] &= \mathbb{E}_P\left[\mathbf{I}_{X\in A} f(Y)\right] \\
&= \int\int \mathbf{I}_{x\in A} f(y)p(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{p_X>0} \mathbf{I}_{x\in A}\left(\int f(y)p(x,y)\mathrm{d}y\right)\mathrm{d}x \quad \text{Fubini} \\
&= \int_{p_X>0} \mathbf{I}_{x\in A}\left(\int f(y)\frac{p(x,y)}{p_X(x)}\mathrm{d}y\right)p_X(x)\mathrm{d}x \\
&= \mathbb{E}_P\left[\mathbf{I}_{x\in A} h(X)\right] \\
&= \mathbb{E}_P\left[\mathbf{I}_B h(X)\right].
\end{aligned}$$

Ceci permet de conclure que $h(X)$ est bien une version de l'espérance conditionnelle de $f(Y)$ sachant X ou $\sigma(X)$ (la tribu engendrée par X). □

Le rapport $p(x,y)/P_X(x)$ permet de définir une probabilité conditionnelle de Y sachant X. Cette loi conditionnelle est définie par une densité conditionnelle.

THÉORÈME B.54 *Soit X, Y deux variables aléatoires dont la loi conjointe P admet une densité jointe $p(.,.)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$, et dont la densité de la première marginale est notée p_X par rapport à la mesure de Lebesgue sur $\mathbb{R}$.*

La fonction N de $\mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$ définie par

$$(x,y)\mapsto N(x,y)=\begin{cases}\frac{p(x,y)}{p_X(x)} & \text{si } p_X(x)>0\\ 0 & \text{sinon,}\end{cases}$$

vérifie les propriétés suivantes

1. *Pour chaque x tel que $p_X(x)>0$, la fonctionnelle de $\mathcal{B}(\mathbb{R})\to[0,1]$, définie par*

$$B\mapsto\int_B N(x,y)\mathrm{d}y$$

est une probabilité sur $(\mathbb{R},\mathcal{B}(\mathbb{R}))$.

2. *Pour chaque $B\in\mathcal{B}(\mathbb{R})$, la fonction $x\mapsto\int_B N(x,y)\mathrm{d}y$ est borélienne, et $\sigma(X)$-mesurable si on appelle X et Y les projections coordonnées. .*

3. *Pour chaque $A\in\mathcal{B}(\mathbb{R}\times\mathbb{R})$*

$$P\{(X,Y)\in A\}=\int\left(\int\mathbf{I}_A(x,y)N(x,y)\mathrm{d}y\right)p_X(x)\mathrm{d}x.$$

4. *Pour tout fonction P-intégrable f sur $\mathbb{R}^2$, la variable aléatoire définie en appliquant*

$$x\mapsto\int_{\mathbb{R}}f(x,y)N(x,y)\mathrm{d}y$$

à X est une version de l'espérance conditionnelle de $f(X,Y)$ par rapport à $\sigma(X)$.

PREUVE. Preuve de i). Notons

$$P_x\{B\}=\int_B N(x,y)\mathrm{d}y.$$

Le fait que la fonctionnelle P_x soit à valeur dans $[0,1]$ est immédiat. Idem pour le fait que $P_x\{\emptyset\}=0$ et $P_x\{\mathbb{R}\}=1$. Il en va de même pour l'additivité. Il reste à vérifier que si B_n est une suite croissante de boréliens qui tend vers une limite B alors

$$P_x\{B_n\}\nearrow P_x\{B\}.$$

C'est une conséquence immédiate du théorème de convergence monotone pour la mesure de Lebesque ($\mathbf{I}_{B_n}(y)N(x,y)\leq N(x,y)$).

Preuve de ii) Comme la fonction $(x,y)\mapsto p(x,y)\mathbf{I}_B(y)$ est $\mathcal{B}(\mathbb{R}^2)$mesurable et intégrable, c'est un corollaire du théorème de Fubini que

$$x\mapsto\int_B p(x,y)\mathbf{I}_B(y)\mathrm{d}y$$

est définie presque partout et Borel-mesurable.

Preuve de iii) C'est aussi une conséquence immédiate du théorème de Fubini.

Preuve de iv) Ce n'est qu'une généralisation de la Proposition précédente, et une conséquence du théorème de Fubini. □

Exercice B.55 On considère la loi uniforme sur la surface de $\mathbb{R}^2$ définie par $0 \leq x \leq y \leq 1$. Donner la densité jointe $p()$, la densité marginale p_X et le noyau $N(,)$.

Remarque B.56 Si on continue de désigner par X et Y les projections coordonnées, ce théorème met en évidence que la fonction de $\mathbb{R}^2 \times \mathcal{B}(\mathbb{R}^2)$ vers $[0,1]$ définie (P-presque partout) par

$$(x, B) \mapsto \int_{\mathbb{R}} \mathbf{I}_B(x,z) N(x,z) \mathrm{d}z$$

est une version de l'espérance conditionnelle de $\mathbf{I}_B$ sachant X. Et pour chaque (x,y) tel que $p_X(x) > 0$ (donc P-presque partout), la fonction

$$B \mapsto \int_{\mathbb{R}} \mathbf{I}_B(x,y) N(x,y) \mathrm{d}y$$

définit bien une probabilité. Nous somme dans une situation où la construction d'une « probabilité conditionnelle » à partir des espérances conditionnelles est possible.

Probabilités conditionnelles régulières, noyaux

Nous allons énoncer quelques résultats qui permettent de travailler dans un cadre plus général,

Définition B.57 Soient $(\Omega, \mathcal{F})$ un espace probabilisable, et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$.
On appelle *noyau de probabilité conditionnelle sachant* $\mathcal{G}$ une fonction N de $\Omega \times \mathcal{F} \to \mathbb{R}_+$ qui vérifie

i) Pour tout ω, $N(\omega, .)$ définit une probabilité sur $(\Omega, \mathcal{F})$.
ii) Pour tout $A \in \mathcal{F}$, $N(\cdot, A)$ est $\mathcal{G}$-mesurable

Définition B.58 Soient $(\Omega, \mathcal{F}, P)$ un espace probabilisé et $\mathcal{G}$ un sous-tribu de $\mathcal{F}$. Un noyau N sur Ω vers $\mathcal{F}$ est une probabilité conditionnelle régulière sachant $\mathcal{G}$ si et seulement si

i) Pour tout $B \in \mathcal{F}$, $\omega \mapsto N(\omega, B)$ est une version de l'espérance conditionnelle de $\mathbf{I}_B$ sachant $\mathcal{G}$ ($N(\cdot, B)$ est donc $\mathcal{G}$-mesurable) :

$$N(\cdot, B) = \mathbb{E}[\mathbb{I}_B \mid \mathcal{G}] \quad P - \text{p.s.}$$

ii) Pour P-presque tout $\omega \in \Omega$, $B \mapsto N(\omega, B)$ définit une probabilité sur $(\Omega, \mathcal{F})$.

Une probabilité conditionnelle régulière (lorsque elle existe) est donc définie à partir de versions des espérances conditionnelles. En retour une probabilité conditionnelle régulière peut fournir un moyen de « calculer » des espérances conditionnelles, voire des espérances.

Théorème B.59 *Soient $(\Omega, \mathcal{F}, P)$ un espace probabilisé, et $\mathcal{G}$ une sous-tribu de $\mathcal{F}$. Soit N un noyau de probabilité sur $(\Omega, \mathcal{F})$ conditionnel par rapport à $\mathcal{G}$. Les propriétés suivantes sont équivalentes*

i) Pour tout $B \in \mathcal{F}$, P-presque sûrement, $\mathbb{E}[\mathbf{I}_B \mid \mathcal{G}](\omega) = N(\omega, B)$. ($N(\cdot, \cdot)$) définit une probabilité conditionnelle régulière.
ii) Pour toute fonction f P-intégrable sur $(\Omega, \mathcal{F})$, P-presque sûrement :

$$\mathbb{E}_P[f \mid \mathcal{G}](\omega) = \mathbb{E}_{N(\omega,\cdot)}[f].$$

Existence de probabilités conditionnelles régulières lorsque $\Omega = \mathbb{R}$

Nous allons vérifier l'existence de probabilités conditionnelles régulières dans un cas non-trivial.

Théorème B.60 *Soit P une probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et $\mathcal{G} \subseteq \mathcal{B}(\mathbb{R})$, alors il existe une probabilité conditionnelle régulière relativement à $\mathcal{G}$.*

REMARQUE B.61 On va essentiellement utiliser le fait que les boréliens $\mathcal{B}(\mathbb{R})$ peuvent être engendrés par une famille dénombrable de parties.

PREUVE. Notons $\mathcal{C}$ l'ensemble formé par les demi-droites $(-\infty, q]$ avec q rationnel, l'ensemble vide $\emptyset$ et l'univers $\mathbb{R}$. Cet ensemble est un π-système.

Pour $q < q' \in \mathbb{Q}$, on peut choisir deux versions Y_q et $Y_{q'}$ des espérances conditionnelles de $\mathbf{I}_{(-\infty,q]}$ de $\mathbf{I}_{(-\infty,q']}$ telles que P-p.s.

$$Y_q < Y_{q'}$$

et $Y_{q'} - Y_q$ est aussi une version de l'espérance conditionnelle de $\mathbf{I}_{(q,q']}$ (Théorèmes B.43 et B.42).

En utilisant le fait qu'une réunion dénombrable d'événements P-négligeables est P-négligeable et que $\mathbb{Q}^2$ est dénombrable, on peut choisir des versions $(Y_q)_{q\in\mathbb{Q}}$ des espérances conditionnelles de $\mathbf{I}_{(-\infty,q]}$ telles que P-p.s.

$$q < q' \Rightarrow Y_q < Y_{q'},$$

Par la suite nous noterons Ω_0 l'événement P-presque sûr sur lequel les Y_q satisfont les bonnes propriétés.

Pour chaque $x \in \mathbb{R}$, on peut définir $\mathrm{Z}_x$pour chaque $\omega \in \mathbb{R}$ par

$$Z_x(\omega) = \inf\{Y_q(\omega) : q \in \mathbb{Q}, x < q\}.$$

Sur Ω_0, la fonction $x \mapsto Z_x(\omega)$ est croissante, elle possède une limite à gauche en chaque point et elle continue à droite en tout point (ce n'est pas forcémment vrai pour $q \mapsto Y_q(\omega)$...). La fonction $x \mapsto Z_x(\omega)$ tend vers 0 quand x tend vers $-\infty$, vers 1 quand x tend vers $+\infty$. Autrement dit sur Ω_0, $x \mapsto Z_x(\omega)$ est une fonction de répartition, elle définit donc (de manière unique) une loi de probabilité sur $\mathbb{R}$ que nous noterons $\nu(\omega, .)$.

Par ailleurs pour chaque x, Z_x est définie comme un infimum dénombrable de variables alétaoires $\mathcal{G}$-mesurables, Z_x est donc une variable aléatoire $\mathcal{G}$-mesurable.

Il reste à montrer que pour tout $B \in \mathcal{F}$, $\omega \mapsto \nu(\omega, B)$ (sur Ω_0, 0 ailleurs) définit bien une version de l'espérance conditionnelle de $\mathbf{I}_B$ sachant $\mathcal{G}$.

Cette propriété est vérifiée pour B appartenant au π-système $\mathcal{C}$. En effet, si on fixe $q \in \mathbb{Q}$, et une suite $(q_n)_n$ décroissante de $\mathbb{Q}$ de limite q, pour toute version des espérances conditionnelles de $\mathbf{I}_{(-\infty,q_n]}$ et de $\mathbf{I}_{(-\infty,q]}$ et donc en particulier pour (Y_{q_n}) et Y_q, on a d'après le théorème de convergence monotone conditionnelle P-p.s.

$$\lim_n Y_{q_n} = Y_q.$$

Donc P-p.s., $Y_q = Z_q = \nu(\cdot, \mathbf{I}_{(-\infty,q]})$.

Appelons $\mathcal{D}$ l'ensemble des événements pour lesquels $\omega \mapsto \nu(\omega, B)$ (sur Ω_0, 0 ailleurs) définit bien une version de l'espérance conditionnelle de $\mathbf{I}_B$ sachant $\mathcal{G}$. Nous allons montrer que $\mathcal{D}$ est un λ-système, c'est-à-dire que

i.) $\mathcal{D}$ contient $\emptyset$ et $\mathbb{R} = \Omega$.
ii.) Si B, B' appartiennent à $\mathcal{D}$, et $B \subseteq B'$ alors $B' \setminus B \in \mathcal{D}$.
iii.) Si $(B_n)_n$ est une suite croissante d'événements de $\mathcal{D}$, de limite B alors $B \in \mathcal{D}$.

La clause a) est assurée par construction.

Clause b) Si B et B' appartiennent à $\mathcal{D}$, alors d'après le théorème B.42, si $\mathbb{E}\left[\mathbf{I}_{B'\setminus B} \mid \mathcal{G}\right]$ est une version de l'espérance conditionnelle de $\mathbf{I}_{B'\setminus B}$ sachant $\mathcal{G}$, sur un événement $\Omega_1 \subseteq \Omega_0$ de probabilité 1 :

$$\mathbb{E}\left[\mathbf{I}_{B'\setminus B} \mid \mathcal{G}\right] = \mathbb{E}\left[\mathbf{I}_{B'} - \mathbf{I}_B \mid \mathcal{G}\right] = \mathbb{E}\left[\mathbf{I}_{B'} \mid \mathcal{G}\right] - \mathbb{E}\left[\mathbf{I}_B \mid \mathcal{G}\right] = \nu(\omega, B') - \nu(\omega, B) = \nu(\omega, B' \setminus B).$$

Clause c). Si $(B_n)_n$ est une suite croissante d'événements de $\mathcal{D}$, de limite B, si $\mathbb{E}\left[\mathbf{I}_B \mid \mathcal{G}\right]$ est une version de l'espérance conditionnelle de $\mathbf{I}_B$ sachant $\mathcal{G}$, sur un événement $\Omega_1 \subseteq \Omega_0$ de probabilité 1 :

$$\mathbb{E}\left[\mathbf{I}_B \mid \mathcal{G}\right] = \lim_n \mathbb{E}\left[\mathbf{I}_{B_n} \mid \mathcal{G}\right] = \lim_n \nu(\omega, B_n) = \nu(\omega, B).$$

Donc $B \in \mathcal{D}$.

Le théorème π-λ (Lemme des classes monotones) nous permet donc de conclure que $\mathcal{F} \subseteq \mathcal{D}$ et d'achever la preuve du théorème. □

REMARQUE B.62 Un raisonnement plus technique permet de montrer que l'existence de probabilités conditionnelles régulières est garantie dès que l'univers Ω peut être muni d'une structure d'espace métrique complet et séparable et que la tribu $\mathcal{F}$ est la tribu des Boréliens de la métrique en question.

Il est très fréquent de définir une loi à partir d'une loi marginale et d'un noyau.

PROPOSITION B.63 *Soient $(\Omega, \mathcal{F})$ un espace probabilisable et X une variable aléatoire de $(\Omega, \mathcal{F})$ vers $(\mathcal{X}, \mathcal{G})$. Soient P_X une probabilité sur $(\mathcal{X}, \mathcal{G})$ et N un noyau de transition de $\mathcal{X}$ vers $\mathcal{F}$. Alors il existe une et une seule probabilité P sur $(\Omega, \mathcal{F})$ telle que $P_X = P \circ X^{-1}$ et $N(x, \cdot)$ définit la loi conditionnelle sachant $X = x$.*

Le théorème suivant garantit l'existence d'une probabilité conditionnelle régulière dans tous les scenarii qui nous intéressent (notamment en inférence bayésienne).

THÉORÈME B.64 (EXISTENCE DE PROBABILITÉS CONDITIONNELLES) *Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé sur lequel vivent deux variables aléatoires X, Y. Soit $\mathcal{G}$ la sous-tribu engendrée par X, soit $(\mathcal{Y}, \mathcal{H})$ l'espace probabilisable image de $(\Omega, \mathcal{F})$ par Y.*

Si $\mathcal{Y}$ est un espace métrique complet séparable et $\mathcal{H}$ la tribu des boréliens associée, *alors il existe une probabilité conditionnelle régulière de Y sachant X (notée $P(\cdot \mid X)(\cdot)$).*

C.1 CONVERGENCES, DÉFINITION

Dans l'étude du comportement asymptotique des estimateurs, des régions de confiance et des tests, nous utiliserons abondamment des arguments de type « loi des grands nombres » ou « théorème central limite ».

DÉFINITION C.1 (CONVERGENCE EN PROBABILITÉ) Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathbb{R}^k$, vivant sur un même espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ *converge en probabilité* (en toute rigueur converge en $\mathbb{P}$-probabilité) vers une variable aléatoire X à valeurs dans $\mathbb{R}^k$, vivant sur cet espace probabilisé si et seulement si, pour tout $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\{\|X_n - X\| > \epsilon\} = 0\,.$$

THÉORÈME C.2 (LOI FORTE DES GRANDS NOMBRES) *Si $X_1, \ldots, X_n, \ldots$ sont des variables aléatoires à valeur dans $\mathbb{R}^k$, d'un même espace probabilisé, indépendamment identiquement distribuées, intégrables, d'espérance μ alors presque sûrement*

$$\lim_{n\to\infty} \overline{X}_n = \mu \qquad \textit{avec} \quad \overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La convergence a aussi lieu en probabilité.

La loi des grands nombres est l'ingrédient essentiel des preuves de consistance.

Pour étudier les vitesses d'estimation, il faut disposer de résultats de convergence en distribution.

DÉFINITION C.3 (CONVERGENCE ÉTROITE) Une suite de lois de probabilités $(P_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^k$ converge ÉTROITEMENT/FAIBLEMENT vers une loi de probabilité P (sur $\mathbb{R}^k$) si et seulement si pour toute fonction continue et bornée f de $\mathbb{R}^k$ dans $\mathbb{R}$, la suite $(\mathbb{E}_{P_n}[f])_{n\in\mathbb{N}}$ converge vers $\mathbb{E}_P[f]$.

Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathbb{R}^k$ définies sur une suite d'espaces probabilisés $(\Omega_n, \mathcal{F}_n, P_n)$ *converge en loi* (ou en distribution) si la suite des lois image $(P_n \circ X_n^{-1})_{n\in\mathbb{N}}$ converge étroitement. On note de façon abrégée cette convergence

$$X_n \rightsquigarrow X \qquad \text{ou} \qquad X_n \rightsquigarrow \mathcal{L}$$

($\mathcal{L}$ désigne une loi de probabilité), en sous-entendant les espaces probabilisés.

Pour établir la convergence en loi ou l'utiliser, on dispose d'une multitude de critères équivalents. En voici quelques uns. Voir cours Mesure, intégration et probabilités pour détails et justifications.

THÉORÈME C.4 (PORTEMANTEAU) *Une suite de lois de probabilités $(P_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^k$ converge* ÉTROITEMENT/FAIBLEMENT *vers une loi de probabilité P (sur $\mathbb{R}^k$) si et seulement si l'une des propriétés suivantes est vérifiée*

i) Pour toute fonction continue et bornée f de $\mathbb{R}^k$ dans $\mathbb{R}$, la suite $\mathbb{E}_{P_n}[f]$ converge vers $\mathbb{E}_P[f]$.
ii) Pour toute fonction uniformément continue et bornée f de $\mathbb{R}^k$ dans $\mathbb{R}$, la suite $\mathbb{E}_{P_n}[f]$ converge vers $\mathbb{E}_P[f]$.
iii) Pour toute fonction lipschitzienne et bornée f de $\mathbb{R}^k$ dans $\mathbb{R}$, la suite $\mathbb{E}_{P_n}[f]$ converge vers $\mathbb{E}_P[f]$.
iv) Pour tout sous-ensemble fermé F de $\mathbb{R}^k$, $\limsup P_n\{F\} \le P\{F\}$.
v) Pour tout sous-ensemble ouvert O de $\mathbb{R}^k$, $\liminf P_n\{O\} \ge P\{O\}$.

vi) Pour toute fonction continue P-presque sûrement continue et bornée f de $\mathbb{R}^k$ dans $\mathbb{R}$, la suite $(\mathbb{E}_{P_n}[f])$ converge vers $\mathbb{E}_P[f]$.
vii) Pour tout événement (borélien) A tel que $P(A^\circ) = P(\overline{A})$ (la frontière de A est de probabilité nulle), $\lim_n P_n(A) = P(A)$.

Remarque C.5 En français, comme en anglais, un portemanteau est un fourre-tout.

Le théorème suivant est un corollaire immédiat du théorème d'injectivité des fonctions caractéristiques (voir Théorème 2.19).

Théorème C.6 (Cramer-Wold) *La loi d'un vecteur aléatoire $(X_1, \ldots, X_n)^t$ à valeur dans $\mathbb{R}^n$ est complètement déterminée par l'ensemble des lois des variables $\sum_{i=1}^n t_i X_i$ lorsque $(t_1, \ldots, t_n)^t$ parcourt $\mathbb{R}^n$.*

C.2 extensions du théorème central limite

En statistique classique, on étudie souvent le comportement des expériences échantillonnées et on laisse la taille de l'échantillon tendre vers l'infini. Comme les quantités étudiées sont souvent bien approchées par des sommes de variables aléatoires indépendantes, le théorème central limite ou plutôt un théorème central limite permet de décrire la loi limite. La version de Lindeberg-Feller traite des sommes de vecteurs aléatoires indépendants mais pas forcément identiquement distribuées.

Théorème C.7 (théorème central limite Lindeberg-Feller) *Soit $(X_{i,n})_{i \leq n < \infty}$ un tableau triangulaire de variables aléatoires centrées de variances finies, telles que*

1. *Pour chaque n $(X_{i,n})_{i \leq n}$ forme une famille indépendante,*
2. *Pour tout $\epsilon > 0$, (en notant $\sigma_n^2 = \sum_{i \leq n} \mathbb{E} X_{i,n}^2$),*

$$\lim_{n \to \infty} \frac{1}{\sigma_n^2} \sum_{i \leq n} \mathbb{E}\big[X_{i,n}^2 \mathbb{I}_{|X_{i,n}| > \epsilon \sigma_n}\big] = 0$$

alors

$$\frac{1}{\sigma_n}\Big(\sum_{i \leq n} X_{i,n}\Big) \rightsquigarrow \mathcal{N}(0,1)\,.$$

Théorème C.8 (inégalité de Berry-Esseen) *Si $X_1, \ldots, X_n, \ldots$ sont i.i.d. centrées, de variance σ^2, et si $S_n := \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n X_i$ alors*

$$\sup_{x \in \mathbb{R}} |\mathbb{P}\{S_n \leq x\} - \Phi(x)| \leq \frac{\mathbb{E}[|X_1|^3]}{\sqrt{n}\sigma^3}$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$.

Dans la section D.3 nous donnons une démonstration combinée des deux théorèmes sous des conditions restrictives (existence d'un moment d'ordre quatre et meme kurtosis uniformément borné).

Si on dispose de deux suites $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ de variables aléatoires telles que $X_n \rightsquigarrow X$, $Y_n \rightsquigarrow Y$, on ne peut rien dire en général sur la suite $(X_nY_n)_n$, on ne peut pas affirmer à coup sûr que $X_nY_n \rightsquigarrow XY$. Mais, si Y est une variable aléatoire dégénérée presque sûrement égale à une constante y, alors on peut s'appuyer sur le lemme de Slutsky.

LEMME C.9 (LEMME DE SLUTSKY) *Si $(X_n)_n$ est une suite de variables aléatoires sur $(\Omega_n, \mathcal{F}_n, P_n)$ telle que $X_n \rightsquigarrow X$ et si (Y_n) est une autre suite de variables aléatoires sur $(\Omega_n, \mathcal{F}_n, P_n)$, et y une constante telle que $Y_n \rightsquigarrow y$, alors*

$$(Y_n, X_n) \rightsquigarrow (y, X)\,.$$

On invoque en général la forme suivante qui est prête à l'emploi.

THÉORÈME C.10 (Slutsky) *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^k$, X une autre variable aléatoire à valeur dans $\mathbb{R}^k$, soit $(Y_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^{k'}$. Si*

$$\begin{aligned} X_n &\rightsquigarrow X \\ Y_n &\rightsquigarrow y \in \mathbb{R}^{k'} \end{aligned}$$

alors si g est une fonction continue de $\mathbb{R}^k \times \mathbb{R}^{k'}$ dans $\mathbb{R}^{k''}$

$$g(X_n, Y_n) \rightsquigarrow g(X, y).$$

PREUVE. Dans la seconde formulation, il suffit (d'après le théorème portemanteau) de s'intéresser au cas des fonctions bornées et lipschitziennes. On suppose $\|g\|_\infty \leq b$ et g L-lipschitzienne.

$$\begin{aligned} &|\mathbb{E}\left[g(X_n, Y_n)\right] - \mathbb{E}\left[g(X, Y)\right]| \\ &\leq |\mathbb{E}\left[g(X_n, Y_n)\right] - \mathbb{E}\left[g(X_n, y)\right]| + |\mathbb{E}\left[g(X_n, y)\right] - \mathbb{E}\left[g(X, y)\right]| \end{aligned}$$

La convergence en loi de la suite (X_n) vers X garantit que

$$\lim_n |\mathbb{E}\left[g(X_n, y)\right] - \mathbb{E}\left[g(X, y)\right]| = 0\,.$$

Par ailleurs, les hypothèses sur g garantissent pour tout $\epsilon > 0$,

$$|g(X_n, Y_n) - g(X_n, y)| \leq 2\mathbb{I}_{d(Y_n, y)>\epsilon}\|g\|_\infty + L\epsilon\,.$$

La convergence en loi de $(Y_n)_n$ vers y implique la convergence en probabilité, donc

$$\lim_n \mathbb{E}\mathbb{I}_{d(Y_n, y)>\epsilon} = 0\,.$$

□

Le lemme de Slutsky est un élément de la preuve du théorème suivant appelé *méthode delta*.

THÉORÈME C.11 (MÉTHODE DELTA) *Soient*

i) $(a_n)_n$ une suite positive qui tend vers l'infini.
ii) $(X_n)_n$ une suite de variables aléatoires à valeur dans $\mathbb{R}^k$.

iii) X *une variable aléatoire à valeur dans* $\mathbb{R}^k$.
iv) $\vec{x} \in \mathbb{R}^k$.
v) f *une fonction de* $\mathbb{R}^k$ *dans* $\mathbb{R}^m$, *différentiable en* $\vec{x}$, *on note* Df *sa différentielle en* $\vec{x}$ *(c'est une fonction linéaire de* $\mathbb{R}^k$ *dans* $\mathbb{R}^m$*).*

Si

$$a_n (X_n - \vec{x}) \rightsquigarrow X$$

alors

i) $X_n \xrightarrow{P} \vec{x}$;

ii) $a_n (f(X_n) - f(\vec{x})) - a_n Df (X_n - \vec{x}) \xrightarrow{P} 0$;

iii) $a_n (f(X_n) - f(\vec{x})) \rightsquigarrow DfX$.

La notion de famille de lois de probabilités *tendue* est définie par le critère suivant.

Définition C.12 Une famille $\mathcal{C}$ de lois de probabilités sur $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ est dite tendue si et seulement si pour tout $\epsilon > 0$, il existe un compact $K(\epsilon) \subseteq \mathbb{R}^k$ tel que

$$\forall P \in \mathcal{C}, \qquad P\{K(\epsilon)\} \geq 1 - \epsilon .$$

Le critère de Prokhorov relie tension et convergence étroite (en distribution).

Théorème C.13 *Une famille* $\mathcal{C}$ *de lois de probabilités sur* $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ *est tendue si et seulement si sa fermeture est relativement compacte pour la topologie de la convergence étroite.*

C.4 remarques bibliographiques

L'espérance conditionnelle et les lois conditionnelles sont étudiées dans tous les livres de probabilités. Les notes classiques de Le Gall [1] fournissent plus que le nécessaire pour suivre un cours de statistique.

Dudley [6], part de l'existence de la dérivée de Radon-Nykodim (rigoureusement établie) pour construire et étudier l'espérance conditionnelle, puis les lois conditionnelles. Williams [13] construit l'espérance conditionnelle à partir du point de vue prédictif, d'abord dans $L_2(P)$ puis par aproximation dans $L_1(P)$. La construction de la dérivée de Radon-Nykodim s'appuie alors sur un argument de martingale. Une présentation intuitive de la construction de probabilités conditionnelles régulières sur $\mathbb{R}$ est tirée de [7]. Pollard [11] propose une étude à la fois érudite et progressive de ces questions en visant spécialement les usagers de la théorie.

La méthode delta est décrite, étendue et illustrée dans A. van der Vaart. **Asymptotic statistics**. Cambridge University Press, 1998. Elle est généralement attribuée à H. Cramér. **Mathematical Methods of Statistics**. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946, p. xvi+575. mr : 0016588.

D.1 MÉTRISATIONS DE LA CONVERGENCE EN LOI

La convergence en loi peut, lorsqu'on parle de lois sur des espaces métriques complets séparables, être métrisée, autrement dit définie par rapport à une métrique. En fait, plusieurs métriques sont possibles et on peut selon les besoins, choisir la métrique la plus intéressante. Les métriques qui nous intéressent sont de la forme

$$\mathrm{d}_{\mathcal{H}}(P,Q) = \sup_{h\in\mathcal{H}} \int h\mathrm{d}P - \int h\mathrm{d}Q$$

où $\mathcal{H}$ est un ensemble bien choisi de fonctions mesurables.

La métrique de Kolmogorov-Lévy est utilisée implicitement dans la formulation du théorème de Glivenko-Cantelli, elle est controlée dans l'inégalité de Dvoretzky-Kiefer-Wolfowitz.

DÉFINITION D.1 (« DISTANCE » DE KOLMOGOROV-LÉVY) Si W et Z sont deux variables aléatoires réelles, la distance de Kolmogorov-Lévy entre les lois de ces deux variables aléatoires est définie par

$$\mathrm{d}_K(\mathcal{L}(W),\mathcal{L}(Z)) = \sup_{x\in\mathbb{R}} |\mathbb{P}\{W\le x\} - \mathbb{P}\{Z\le x\}| \ .$$

Le fait que la distance de Kolmogorov-Lévy métrise la convergence étroite vers une loi dont la fonction de répartition est continue se vérifie avec les arguments utilisés dans les preuves classiques du théorème de Glivenko-Cantelli.

REMARQUE D.2 Si $\lim_n x_n = x$, la suite des masse δ_{x_n} converge étroitement vers la masse δ_x, néanmoins, si $x_n \neq x$, $\mathrm{d}_K(\delta_{x_n},\delta_x) = 1$.

Pour métriser la convergence étroites des lois sur $\mathbb{R}$ en toute généralité, on peut utiliser la distance de Lévy-Prokhorov :

$$\inf\Big\{\epsilon : \epsilon>0, \qquad P(-\infty,x-\epsilon]-\epsilon \le Q(-\infty,x] \le P(-\infty,x-\epsilon]+\epsilon\Big\}\,.$$

La distance de Kolmogorov-Lévy représente un relâchement pertinent de la distance en variation (qui ne métrise pas la convergence en loi). Le fait que les fonctions test qui définissent la distance de Kolmogorov-Lévy ne soient pas absolument continues ne facilite pas la vie. C'est pourquoi, on travaille souvent avec une distance définie par des fonctions plus faciles à manier. Cette distance appartient à la famille des distances dites de Monge-Wasserstein, ou distance de transport.

DÉFINITION D.3 (DISTANCE DE WASSERSTEIN) Soient W et Z de variables aléatoires de lois $\mathcal{L}(W)$ et $\mathcal{L}(Z)$ à valeur dans $\mathbb{R}^k$,

$$\mathrm{d}_M(\mathcal{L}(W),\mathcal{L}(Z)) = \sup_{h:1-\mathrm{Lipschitz}} |\mathbb{E}h(W) - \mathbb{E}h(Z)| \ .$$

Lorsqu'on veut mesurer la distance à une loi absolument continue comme la loi normale, le fait d'utiliser une métrique plutôt qu'une autre n'est pas très important.

PROPOSITION D.4 *Si P et Q sont deux lois sur $\mathbb{R}$, et si Q possède une densité majorée par C vis-à-vis de la mesure de Lebesgue, alors*

$$\mathrm{d}_K(P,Q) \le \sqrt{2C\mathrm{d}_M(P,Q)}\,.$$

PREUVE. Pour $x\in\mathbb{R}$, on note $h_x = \mathbb{I}_{(-\infty,x]}$ et pour $\epsilon>0$, on définit $h_{x,\epsilon}$ par

$$h_{x,\epsilon}(w) = \begin{cases} 1 & \text{si } x\ge w \\ 0 & \text{si } w > x+\epsilon \\ 1-\frac{w-x}{\epsilon} & \text{si } x\le w\le x+\epsilon \end{cases}$$

On note q la densité de Q par rapport à la mesure de Lebesgue. La fonction $h_{x,\epsilon}$ est $1/\epsilon$-Lipschitz et elle approche h_x. Si $W \sim P$ et $Z \sim Q$,

$$\mathbb{E}h_x(W) - \mathbb{E}h_x(Z) = \underbrace{\mathbb{E}h_x(W) - \mathbb{E}h_{x,\epsilon}(Z)}_{\text{(I)}} + \underbrace{\mathbb{E}h_{x,\epsilon}(Z) - \mathbb{E}h_x(Z)}_{\text{(II)}} .$$

On majore ensuite (simplement) les deux expressions (I) et (II).

$$\begin{aligned}\text{(I)} &\leq \mathbb{E}h_{x,\epsilon}(W) - \mathbb{E}h_{x,\epsilon}(Z) \\ &\leq \frac{1}{\epsilon}\mathrm{d}_M(P,Q)\end{aligned}$$

et

$$\begin{aligned}\text{(II)} &\leq \int_x^{x+\epsilon} \left(1 - \frac{w-x}{\epsilon}\right) q(z)\mathrm{d}z \\ &\leq C\frac{\epsilon}{2} .\end{aligned}$$

En choisissant $\epsilon = \sqrt{2\mathrm{d}_M(P,Q)/C}$, et en optimisant le choix de x, on obtient le résultat désiré. □

D.2 MÉTHODE DE STEIN

Dans l'approche de Stein, on s'intéresse à l'approximation d'une loi cible (loi normale, Poisson, χ^2_k, ...) par la loi d'une variable aléatoire dont on connait le mode de fabrication (somme de variables aléatoires indépendantes, statistiques d'ordre, logarithme de rapport de vraisemblance, ...). La loi cible est caractérisée par une identité. Dans le cas de la loi normale, il s'agit de l'identité de Stein.

PROPOSITION D.5 *(Identité de Stein) Soit Z une variable aléatoire réelle de loi Q. Si pour toute fonction absolument continue f telle que $f'(Z)$ soit Q-intégrable, on a*

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$$

alors

$$Q = \mathcal{N}(0,1) .$$

La réciproque est vraie.

PREUVE. Voir chapitre sur les vecteurs gaussiens, Lemmes 2.1 et 2.3. □

Cette identité nous conduit à définir un opérateur $\mathcal{A}$ qui agit sur les fonctions absolument continues : $\mathcal{A}f(x) = f'(x) - xf(x)$, qui permet de caractériser $\mathcal{N}(0,1)$:

$$Q = \mathcal{N}(0,1) \Longleftrightarrow \forall f \text{a.c} \quad 0 = \int \mathcal{A}f(z)Q(\mathrm{d}z) .$$

Maintenant pour la fonction $h \in \mathcal{H}$, si on peut définir f_h tel que

$$\mathcal{A}f_h(w) = h(w) - \int h(z)\phi(z)\mathrm{d}z ,$$

on a

$$\mathbb{E}_Q h(W) - \mathbb{E}_{\mathcal{N}(0,1)}h(Z) = \mathbb{E}_Q \mathcal{A}f_h(W) .$$

Pour majorer $d_{\mathcal{H}}(Q,\mathcal{N}(0,1))$, il suffit d'etre capable de majorer les espérances de la forme

$$\mathbb{E}_Q \mathcal{A}f_h(W) \qquad \text{pour} h \in \mathcal{H}, \text{ et } \mathcal{A}f_h(w) = h(w) - \int h(z)\phi(z)\mathrm{d}z .$$

Ce dernier point a conduit à développer des méthodes de calcul approximatif d'intégrales très créatives. Pour comprendre ce qui est en jeu, il faut d'abord étudier les propriétés de régularité et de bornitude des fonctions $\mathcal{A}f_h, h \in \mathcal{H}$, lorsque $\mathcal{H}$ est elle-meme définie par des propriétés de régularité.

Dans la suite on abrège $\int_{\mathbb{R}} h(z)\phi(z)\mathrm{d}z$ en $\Phi(h)$.

LEMME D.6 *Soit f_h la solution de l'équation différentielle de*

$$f_h'(w) - w f_h(w) = h(w) - \Phi(h)\,.$$

donnée par

$$f_h(w) = \begin{cases} \mathrm{e}^{w^2/2} \int_w^\infty \mathrm{e}^{-t^2/2}(\Phi(h) - h(t))\mathrm{d}t & \textit{si } w \geq 0 \\ \mathrm{e}^{w^2/2} \int_{-\infty}^w \mathrm{e}^{-t^2/2}\,(h(t) - \Phi(h))\,\mathrm{d}t & \textit{si } w \leq 0\,. \end{cases}$$

Si h est absolument continue alors f_h est la seule solution bornée de l'équation différentielle et

$$\|f_h\|_\infty \leq 2\|h'\|_\infty, \qquad \|f_h'\|_\infty \leq \sqrt{\frac{2}{\pi}}\|h'\|_\infty, \qquad \|f_h''\|_\infty \leq 2\|h'\|_\infty\,.$$

PREUVE.(esquisse) Vérification que f_h est l'unique solution bornée de l'équation différentielle.

Notons qu'ajouter une constante à h ne modifie pas $h - \Phi(h)$. On postulera que $h(0) = 0$. On aura $|h(w)| \leq \|h'\|_\infty |w|$.

Pour $w > 0$, observons

$$\mathrm{e}^{w^2/2} \int_w^\infty t\,\mathrm{e}^{-t^2/2}\mathrm{d}t = 1 \qquad \Phi(|h|) \leq \|h'\|_\infty \sqrt{\frac{2}{\pi}}\,.$$

La fonction $w \mapsto \mathrm{e}^{w^2/2}$ est solution de l'équation différentielle sans second membre. On cherche une solution bornée de la forme

$$f(w) = g(w)\mathrm{e}^{w^2/2}\,.$$

L'équation différentielle se traduit par

$$g'(w) = \mathrm{e}^{-w^2/2}\,(h(w) - \Phi(h))\,.$$

La fonction

$$g(w) = \begin{cases} \int_w^\infty \mathrm{e}^{-t^2/2}\,(\Phi(h) - h(t))\,\mathrm{d}t & \text{si } w \geq 0 \\ \int_{-\infty}^w \mathrm{e}^{-t^2/2}\,(h(t) - \Phi(h))\,\mathrm{d}t & \text{si } w \leq 0\,. \end{cases}$$

est solution (les hypothèses d'intégrabilité sur h garantissent que g est bien définie), donc f_h est solution de l'équation différentielle énoncée dans le lemme.

Par ailleurs, pour $w > 0$

$$f_h(w) = g(w)\mathrm{e}^{w^2/2} \leq \|h'\|_\infty \left(\sqrt{\frac{2}{\pi}} \frac{\overline{\Phi}(w)}{\phi(w)} + 1 \right)\,,$$

on vérifie que le rapport $\frac{\overline{\Phi}(w)}{\phi(w)}$ est décroissant sur $[0, \infty)$, il est toujours inférieur à $\sqrt{\frac{\pi}{2}}$. On en déduit la bornitude de f_h :

$$\|f_h\|_\infty \leq 2\|h'\|_\infty\,.$$

Les autres solutions de l'équation différentielle sont de la forme $f_h + c\mathrm{e}^{w^2/2}$. Elles ne sont pas bornées.

Nous pouvons maintenant borner les dérivées d'ordre 1 et 2 de f_h.

On vérifie d'abord

$$h(w) - \Phi(h) = \int_{-\infty}^w h'(t)\Phi(t)\mathrm{d}t - \int_w^\infty h'(t)\overline{\Phi}(t)\mathrm{d}t\,.$$

En substituant dans la définition de f_h, on obtient :

$$\begin{aligned} f_h(w) = &-\sqrt{2\pi}\mathrm{e}^{w^2/2}(1 - \Phi(w)) \int_{-\infty}^w h'(t)\Phi(t)\mathrm{d}t \\ &- \sqrt{2\pi}\mathrm{e}^{w^2/2}\Phi(w) \int_w^\infty h'(t)(1 - \Phi(t))\mathrm{d}t \end{aligned}$$

$$
\begin{aligned}
f_h'(w) &= wf_h(w) + h(w) - \Phi(h) \\
&= (1 - \sqrt{2\pi}\mathrm{e}^{w^2/2}(1-\Phi(w))) \int_{-\infty}^{w} h'(t)\Phi(t)\mathrm{d}t \\
&\quad - (1 + \sqrt{2\pi}\mathrm{e}^{w^2/2}\Phi(w)) \int_{w}^{\infty} h'(t)(1-\Phi(t))\mathrm{d}t
\end{aligned}
$$

d'où

$$
\begin{aligned}
|f_h'(w)| \leq \|h'\|_\infty \sup_{w\in\mathbb{R}} \Big(&|1 - \sqrt{2\pi}\mathrm{e}^{w^2/2}(1-\Phi(w))| \int_{\infty}^{w} \Phi(t)\mathrm{d}t \\
&|1 + \sqrt{2\pi}\mathrm{e}^{w^2/2}\Phi(w)| \int_{w}^{\infty} (1-\Phi(t)) \Big)
\end{aligned}
$$

Pour terminer la majoration de $\|f_h'\|_\infty$, il suffit de majorer le supremum en w.

Pour établir l'existence et la bornitude de la dérivée seconde, on différencie à droite $f_h'(w) = wf_h(w) + h(w) - \Phi(h)$ pour obtenir

$$
\begin{aligned}
f_h''(w) &= f_h(w) + wf_h'(w) + h'(w) \\
&= (1+w^2)f_h(w) + ww(h(w) - \Phi(h)) + h'(w)\,.
\end{aligned}
$$

On réutilise les calculs esquissés précédemment pour majorer $\|f_h''\|_\infty$. □

Le fait que si h est 1-Lipschitz, f_h soit deux fois dérivable et de dérivée seconde bornée est particulièrement pratique lorsqu'il faut majorer $\mathbb{E}\mathcal{A}f_h(W)$. La façon dont on peut tirer avantage de cette régularité dépend de ce qu'on sait sur la structure de W. Le théorème suivant qui peut être vu comme une combinaison des théorèmes de Berry-Esseen et de Lindeberg-Feller, montre ce qui peut etre obtenu lorsque W est une somme de variables aléatoires indépendantes pas nécessairement identique distribuées, centrées et suffisamment intégrables.

D.3 UNE BORNE DE TYPE BERRY-ESSEEN POUR UN TLC DE TYPE LINDEBERG-FELLER

Le théorème suivant peut etre lu comme un résultat intermédiaire entre le théorème de Berry-Esseen C.8 (il donne une distance à la loi normale) et le théorème central limite de Lindeberg-Feller C.7 (il traite de sommes de variables aléatoires indépendantes mais pas nécessairement identiquement distribuées). Les conditions utilisées ici sont plus fortes que celles décrites dans les deux théorèmes précédents. Dans le théorème de Lindeberg-Feller, la condition d'intégrabilité uniforme est minimale. Dans l'énoncé du théorème de Berry-Esseen, on se contente de postuler que les sommants ont un moment d'ordre 3. Ici on suppose un peu plus : les sommants ont un kurtosis bornée. Rappelons que le kurtosis d'une loi est défini par

$$
\frac{\mathbb{E}[(X - \mathbb{E}X)^4]}{(\mathbb{E}[(X-\mathbb{E}X)^2])^2}\,.
$$

Le kurtosis des lois gaussiennes est toujours égal à 3. Pour les lois Gamma, le kurtosis ne dépend que du paramètre de forme p (il vaut $3 + 6/p$).

THÉORÈME D.7 *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes centrées, de kurtosis majoré par κ. On définit $\sigma^2 = \sum_{i=1}^n \mathrm{var}\,(X_i)$ et $W = \sum_{i=1}^n X_i/\sigma$.*

$$
\mathrm{d}_M(\mathcal{L}(W), \mathcal{N}(0,1)) \leq \sqrt{\frac{2}{\pi}} \left(\frac{1}{\sigma^3} \sum_{i=1}^n \mathbb{E}|X_i|^3 + \sqrt{\kappa \sum_{i=1}^n \frac{\mathrm{var}(X_i)^2}{\left(\sum_{i=1}^n \mathrm{var}(X_i)\right)^2}} \right)\,.
$$

PREUVE.(Preuve du théorème D.7) Dans la preuve h est une fonction 1-Lipschitzienne, f la solution bornée $\mathcal{A}f = h - \Phi(h)$. Pour majorer $\mathbb{E}h(W) - \Phi(h)$, on va majorer $|\mathbb{E}[f_h'(W) - Wf(W)]|$.

Dans la suite, pour $i \in 1, \dots, n$, $W_i = \sum_{j \neq i} X_j/\sigma$.
Observons d'abord que comme W_i et X_i sont indépendantes et centrées

$$\mathbb{E}[X_i f(W_i)] = 0\,.$$

$$\begin{aligned}
\mathbb{E}[Wf(W)] &= \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i f(W)\right] \\
&= \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n (X_i f(W) - X_i f(W_i))\right] \\
&= \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i (f(W) - f(W_i))\right].
\end{aligned}$$

On utilisera par la suite

$$\begin{aligned}
\mathbb{E}[Wf(W)] = \mathbb{E}&\left[\frac{1}{\sigma}\sum_{i=1}^n X_i (f(W) - f(W_i) - (W - W_i)f'(W))\right] \\
&+ \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i ((W - W_i)f'(W))\right].
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}[f'(W) - Wf(W)]| \leq &\underbrace{\left|\mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i (f(W) - f(W_i) - (W - W_i)f'(W))\right]\right|}_{\text{(I)}} \\
&+ \underbrace{\left|\mathbb{E}\left[\left(1 - \frac{1}{\sigma}\sum_{i=1}^n X_i (W - W_i)\right) f'(W)\right]\right|}_{\text{(I)}}.
\end{aligned}$$

Pour majorer (I), l'inégalité des accroissements finis suffit

$$\begin{aligned}
\text{(I)} &\leq \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n |X_i|\,|f(W) - f(W_i) - (W - W_i)f'(W)|\right] \\
&\leq \frac{1}{\sigma}\sum_{i=1}^n \mathbb{E}\left[|X_i|\frac{|X_i|^2}{\sigma^2}\|f''\|_\infty\right] \\
&\leq \|f''\|_\infty \frac{\sum_{i=1}^n \mathbb{E}|X_i|^3}{\sigma^3}\,.
\end{aligned}$$

Pour majorer (II), on utilise Cauchy-Schwarz et l'hypothèse de kurtosis.

$$\begin{aligned}
\text{(II)} &\leq \mathbb{E}\left[\left|\left(1 - \frac{1}{\sigma^2}\sum_{i=1}^n X_i^2\right)\right| |f'(W)|\right] \\
&\leq \|f'\|_\infty \mathbb{E}\left[\left|\left(1 - \frac{1}{\sigma^2}\sum_{i=1}^n X_i^2\right)\right|\right] \\
&\leq \|f'\|_\infty \frac{1}{\sigma^2}\left(\sum_{i=1}^n \mathrm{var}(X_i^2)\right)^{1/2} \\
&\leq \|f'\|_\infty \frac{1}{\sigma^2}\left(\kappa\sum_{i=1}^n \mathrm{var}(X_i)^2\right)^{1/2} \\
&\leq \|f'\|_\infty \left(\kappa\sum_{i=1}^n \left(\frac{\mathrm{var}(X_i)}{\sum_{i=1}^n \mathrm{var}(X_i)}\right)^2\right)^{1/2}
\end{aligned}$$

□

Remarque D.8 En utilisant le fait que pour X centrée, de kurtosis κ

$$\mathbb{E}|X_i|^3 \leq (\mathbb{E}|X_i|^4)^{3/4} \leq \kappa^{3/4}(\mathbb{E}|X_i|^2)^{3/2}\,,$$

le membre droit du majorant peut se majorer lui même en

$$\kappa^{3/4}\|f''\|_\infty \left(\sum_{i=1}^n \left(\frac{\mathrm{var}(X_i)}{\sum_{j=1}^n \mathrm{var}(X_j)}\right)^{3/2}\right) + \kappa^{1/2}\|f'\|_\infty \left(\sum_{i=1}^n \left(\frac{\mathrm{var}(X_i)}{\sum_{i=1}^n \mathrm{var}(X_i)}\right)^{2}\right)^{1/2}$$

Si les X_i sont identiquement distribuées, alors le majorant du théorème s'écrit

$$\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}\left(\kappa^{3/4} + \kappa^{1/2}\right)\,.$$

Remarque D.9

D.4 remarques bibliographiques

La possibilité de métriser la convergence en distribution (et la convergence en probabilité) est traitée avec beaucoup de rigueur et de clarté dans R. M. Dudley. **Real analysis and probability**. T. 74. Cambridge Studies in Advanced Mathematics. Cambridge : Cambridge University Press, 2002, p. x+555. mr : MR1932358(2003h:60001).

La méthode de Stein pour établir des versions précises et générales du théorème central limite est décrite dans N. Ross. « Fundamentals of Stein's method ». In : *ArXiv e-prints* (sept. 2011).

Un traitement plus approfondi et très lisible est fourni par L. H. Y. Chen, L. Goldstein et Q.-M. Shao. **Normal approximation by Stein's method**. Probability and its Applications (New York). Springer, Heidelberg, 2011, p. xii+405. isbn : 978-3-642-15006-7. mr : 2732624.

La présentation adoptée ici est tirée des premières pages de [2].

La démonstration complète du Lemme D.6 se trouve dans [4].

E.1 THÉORÈMES D'INVERSION

THÉORÈME E.1 (THÉORÈME D'INVERSION LOCALE) *Soit E un espace euclidien, $U \subseteq E$ ouvert, $f : U \to E$ continument différentiable, et $y_0 \in U$. Si la différentielle de f en y_0 $(Df(y_0))$ est inversible, d'inverse $(Df(y_0))^{-1}$ alors il existe un voisinage ouvert $V \subseteq U$ de y_0 sur lequel f est bijective vers un voisinage ouvert W de $x_0 = f(y_0)$ $W = f(V)$ et tel que l'inverse g de f soit différentiable en x_0, de différentielle $(Df(y_0))^{-1}$.*

Même si les conditions du théorème d'inversion locale sont vérifiées en tout point de U, cela ne garantit pas que la fonction f est un bijective de U dans $f(U)$, ni que son inverse est C^1. Il faut renforcer les hypothèses pour obtenir le théorème d'inversion globale.

THÉORÈME E.2 (THÉORÈME D'INVERSION GLOBALE) *Soit E un espace euclidien, $U \subseteq E$ ouvert, $f : U \to E$ continument différentiable (de classe C^1) et injective. Si la différentielle de f est inversible en tout point de U, alors $f(U)$ est ouvert dans E, et f admet une inverse continument différentiable sur $f(U)$ (f est un difféomorphisme de U dans $f(U)$)*

Références

[1] J.-F. LE GALL. « Intégration, probabilités et processus aléatoires ». In : *Ecole Normale Supérieure de Paris* (2006).

[2] N. ROSS. « Fundamentals of Stein's method ». In : *ArXiv e-prints* (sept. 2011).

[3] R. BHATIA. **Matrix analysis**. Springer-Verlag, 1997.

[4] L. H. Y. CHEN, L. GOLDSTEIN et Q.-M. SHAO. **Normal approximation by Stein's method**. Probability and its Applications (New York). Springer, Heidelberg, 2011, p. xii+405. ISBN : 978-3-642-15006-7. MR : 2732624.

[5] H. CRAMÉR. **Mathematical Methods of Statistics**. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946, p. xvi+575. MR : 0016588.

[6] R. M. DUDLEY. **Real analysis and probability**. T. 74. Cambridge Studies in Advanced Mathematics. Cambridge : Cambridge University Press, 2002, p. x+555. MR : MR1932358(2003h:60001).

[7] R. DURRETT. **Probability : theory and examples**. Cambridge University Press, 2010.

[8] L. ELDÉN. **Matrix methods in data mining and pattern recognition**. T. 4. Fundamentals of Algorithms. Society for Industrial et Applied Mathematics (SIAM), Philadelphia, PA, 2007, p. x+224. ISBN : 978-0-898716-26-9. MR : 2314399.

[9] G. H. GOLUB et C. F. VAN LOAN. **Matrix computations**. Third. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 1996, p. xxx+698.

[10] R. HORN et C. JOHNSON. **Matrix analysis**. Cambridge University Press, 1990.

[11] D. POLLARD. **A user's guide to measure theoretic probability**. T. 8. Cambridge University Press, 2002.

[12] A. VAN DER VAART. **Asymptotic statistics**. Cambridge University Press, 1998.

[13] D. WILLIAMS. **Probability with martingales**. Cambridge university press, 1991.