

Statistiques 2017-19

FIMFA ENS ULM

S. Boucheron, C. Boyer et R. Ryder

8 décembre 2017

TABLE DES MATIÈRES

Table des matières	i
Table des figures	iv
1 Problèmes d'inférence statistique	1
1.1 Un problème jouet, trois questions	1
1.2 Quelques définitions : expérience statistique, échantillon, statistique, estimateur	1
1.3 Propriétés des estimateurs	3
1.4 Intervalles de confiance	4
1.5 Tests	7
1.6 Références	10
2 Vecteurs gaussiens	11
2.1 Motivations	11
2.2 Loi(s) normale(s)	11
2.3 Vecteurs gaussiens	14
2.4 Convergence de vecteurs gaussiens	18
2.5 Conditionnement gaussien	18
2.6 Compléments sur les lois Gamma	21
2.7 Normes de vecteurs gaussiens centrés	22
2.8 Normes de vecteurs gaussiens décentrés	23
2.9 Théorème de Cochran et conséquences	24
2.10 Concentration gaussienne	25
2.11 Remarques bibliographiques	27
3 Estimation dans les modèles gaussiens	29
3.1 Exemples de modèles gaussiens	29
3.2 Modèles dominés et vraisemblance, estimation de vecteurs gaussiens (décalage gaussien)	29
3.3 Régression avec design fixe et bruit gaussien homoschédistique	31
3.4 Tests d'hypothèses linéaires	33
3.5 Données whiteside	34
3.6 Remarques bibliographiques	37
4 Méthodes d'estimation : les moments illustrés sur les modèles exponentiels	39
4.1 Introduction	39
4.2 Méthode des moments	39
4.3 Modèles exponentiels	41
4.4 Performance de la méthode des moments dans les modèles exponentiels	44
4.5 Remarques bibliographiques	45

5	Maximisation de la vraisemblance	47
5.1	Modèles dominés et vraisemblance (bis)	47
5.2	Maximisation de la vraisemblance	47
5.3	Maximum de vraisemblance dans les modèles exponentiels canoniques	48
5.4	Phénomène de Wilks et régions de confiance	51
5.5	Calcul du maximum de vraisemblance	54
5.6	Remarques bibliographiques	55
6	Tests	57
6.1	Introduction	57
6.2	Lemme de Neyman-Pearson	57
6.3	Séparation, inégalité de Pinsker	58
6.4	Tests et distance de Hellinger	62
6.5	Remarques bibliographiques	66
7	Tests du χ^2	67
7.1	Introduction	67
7.2	Problèmes	68
7.3	Test du χ^2 d'adéquation	69
7.4	Puissance du test du χ^2	70
7.5	Hypothèse nulle composite	72
7.6	Le test d'indépendance	78
7.7	Remarques bibliographiques	78
8	Tests non-paramétriques	79
8.1	Un problème	79
8.2	Le principe du test de Kolmogorov et Smirnov	79
8.3	La transformation quantile	80
8.4	Une loi des grands nombres « fonctionnelle » : le théorème de Glivenko-Cantelli	83
8.5	Inégalités pour la statistique de Kolmogorov-Smirnov,	85
8.6	Adéquation à une famille de lois	88
8.7	Remarques bibliographiques	89
9	Décision, risque, efficacité	91
9.1	Le programme de Fisher	91
9.2	Le phénomène de Stein	92
9.3	Pertes, risques, risque minimax	95
9.4	Risque Bayésien	96
9.5	Liens entre estimateurs bayésiens et minimax	97
9.6	Rappels sur les lois conditionnelles	98
9.7	Construction d'estimateurs bayésiens	99
9.8	Limites de l'approche minimax	103
9.9	Inégalité de Van Trees	104
9.10	Remarques bibliographiques	107
10	Aspects pratiques des méthodes bayésiennes	109
10.1	Introduction	109
10.2	Choix de la loi a priori	110
10.3	Conjugaison dans les modèles exponentiels	112
10.4	Facteur de Bayes	113
10.5	Monte-Carlo	114
10.6	Algorithme de Metropolis-Hastings	115
10.7	Remarques bibliographiques	117
11	Estimation non-paramétrique : estimation de densité	119
11.1	Problème	119
11.2	Fonctions de perte	119
11.3	Estimation par histogrammes	120
11.4	Fenêtres glissantes	121

11.5	Noyaux et outils	123
11.6	Consistance universelle des méthodes de noyau	124
11.7	Vitesse de convergence	128
11.8	Remarques bibliographiques	133
A	Outils matriciels	135
A.1	Factorisation de Cholesky	135
A.2	Décomposition en valeurs singulières et décomposition spectrale	136
A.3	Meilleure approximation de rang donné par rapport aux normes de Hilbert-Schmidt et d'opérateur	137
A.4	Pseudo-inverse	139
A.5	Remarques bibliographiques	140
B	Outils probabilistes	141
B.1	Calcul de densité image	141
B.2	Convergences	141
B.3	Métrisations de la convergence en loi	142
B.4	Extensions du théorème central limite	143
B.5	Méthode delta	143
B.6	Remarques bibliographiques	145
C	Outils d'analyse	147
C.1	Théorèmes d'inversion	147
C.2	Convexité	147

TABLE DES FIGURES

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 $\theta = .4$	3
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.	9
2.1	Densités de $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 2)$ (tirets), $\mathcal{N}(1, 1/4)$ (pointillés)	11
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.	36
3.2	Visualition des droites de régression correspondant aux deux conditions d'isolation.	37
7.1	Les points représentent des données du docteur Geissler, les triangles représentent les effectifs attendus dans un échantillon de 6115 réalisations indépendantes de la loi binomiale de paramètres 12 et .52. 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 peut tester d'autre hypothèses. L'une des plus simples postule que nous avons affaire à un <i>mélange</i> 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.	73
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 la loi uniforme.	80
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.	86
10.1	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.	111
10.2	Estimation d'un Bayes factor par Monte-Carlo et par Échantillonnage d'importance(échantillonnage préférentiel). En rouge, la valeur analytique. Notez que l'échelle des ordonnées n'est pas la même sur les deux figures.	115
11.1	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	132
11.2	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	133

1.1 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, \dots, x_n . Ces résultats constituent « les données » ou l'« échantillon ». On construit à partir de ces données une « estimation » $\hat{\theta}_n$ de θ , cette estimation est une fonction des données, pas de l'estimande θ qui reste inconnue. On espère que $\hat{\theta}_n$ 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 $\underline{\theta}_n, \bar{\theta}_n$ telles qu'avec une forte probabilité, l'estimande θ appartient à l'intervalle aléatoire

$$[\underline{\theta}_n(x_1, \dots, x_n), \bar{\theta}_n(x_1, \dots, x_n)] =: [\underline{\theta}_n, \bar{\theta}_n].$$

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 $\bar{\theta}_n - \underline{\theta}_n$ et la probabilité de couverture, c'est à dire la probabilité que $\theta \in [\underline{\theta}_n, \bar{\theta}_n]$.

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 ».

1.2 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 (Ω, \mathcal{F}) (l'univers et une tribu de parties). Ici $\Omega = \{\text{pile, face}\}$ 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 $\{\text{pile, face}\}$ (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 coûteux 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 $X : \Omega \rightarrow \mathcal{X}$ 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 $X = \text{Id}$.

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 $(\Omega^n, \sigma(\times_{i=1}^n \mathcal{F}), \mathcal{P}_n := \{P^{\otimes n}, P \in \mathcal{P}\}, \Theta, \mathcal{X}^n, \sigma(\times_{i=1}^n \mathcal{G}), X)$.

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

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

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

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

la variance empirique

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

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'échantillonnage.

Par exemple, dans notre problème jouet, on peut chercher à estimer $P_\theta\{\text{Face}\} = \theta$, par \bar{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 θ, ψ, \dots , on utilise souvent le raccourci $\hat{\theta}$ ou $\hat{\theta}_n, \hat{\psi}_n, \dots$ pour désigner l'estimateur (plutôt que $\hat{\theta}(X_1, \dots, X_n)$ ou $\hat{\psi}(X_1, \dots, X_n)$).

Un estimateur est une variable aléatoire. On peut visualiser ses fluctuations à l'aide de maintes techniques graphiques comme les histogrammes, voir Figure 1.2.

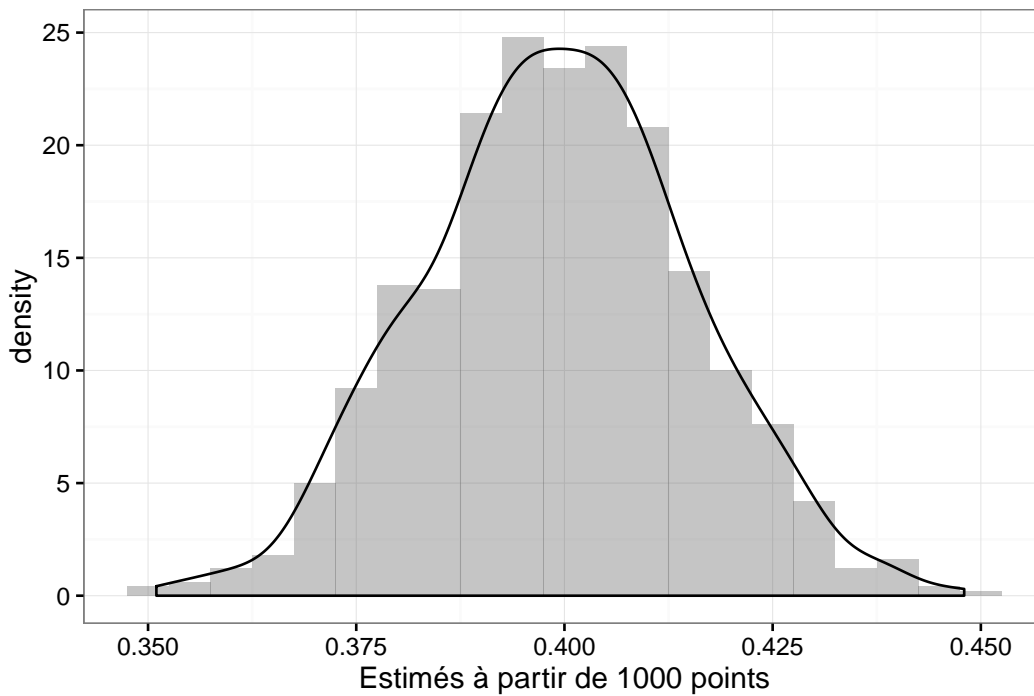


FIG. 1.1 : Histogramme d'un échantillon de 1000 estimations obtenues à partir de 1000 échantillons de taille 1000 de tirages de Bernoulli de probabilité de succès $\theta = .4$.

1.3 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

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

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 $\psi(P)$:

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

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

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

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

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

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

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

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 $(\hat{\theta}_n)$ est *consistante* (pour l'estimation de θ) si

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

La suite est *fortement consistante* si

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

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

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

et que donc

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

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

1.4 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 $\underline{\theta}_n, \bar{\theta}_n$ telles que

$$\forall \theta \in \Theta, \quad P_\theta^{\otimes n} \{ \theta \in [\underline{\theta}_n, \bar{\theta}_n] \} \geq 1 - \alpha.$$

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 $\underline{\theta}_n, \bar{\theta}_n$ telles que

$$\forall \theta \in \Theta, \quad \lim_n P_\theta^{\otimes n} \{ \theta \in [\underline{\theta}_n, \bar{\theta}_n] \} = 1 - \alpha.$$

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 $1 - \alpha$:

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

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 :

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

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 $\hat{\theta}_n$ sont distribués selon $P_{\theta}^{\otimes n}$,

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

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

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

Le lemme de Slutsky (voir Appendice B.5), et le fait que $\hat{\theta}_n/\theta$ converge en probabilité vers 1 lorsque $n \rightarrow \infty$, permet d'écrire pour tout $\alpha \in]0, 1[$,

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

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

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

Un raffinement du théorème central limite, le théorème de Berry-Esseen (voir Appendice B.13), 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 $\lambda \geq 0$,*

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

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

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

Et notons Q_{λ} la loi de densité $\exp(\lambda x - F(\lambda))$ par rapport à Q . On vérifie que

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

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 $t > 0$

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

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

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

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

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

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

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 \bar{X}_n autour de son espérance sont précis, plus petit est l'intervalle de confiance.

Construction calculatoire

Dans cette section, on note $\text{qb}(\alpha, n, \theta)$ le quantile d'ordre $1 - \alpha$ de la loi binomiale de paramètres n, θ (cela correspond à la fonction `qbinom()` de R). On définit la région empirique

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

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

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

et

$$\bar{\theta} = \sup\{\theta' : \text{qb}(\alpha/2, n, \theta') \leq n\hat{\theta}_n\}.$$

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

1.5 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 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

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

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 $P_\theta^{\otimes n}$:

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

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 $\bar{X}_n = \sum_{i=1}^n X_i/n = \hat{\theta}_n$ (ce n'est pas du tout une simple coïncidence).

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

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

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

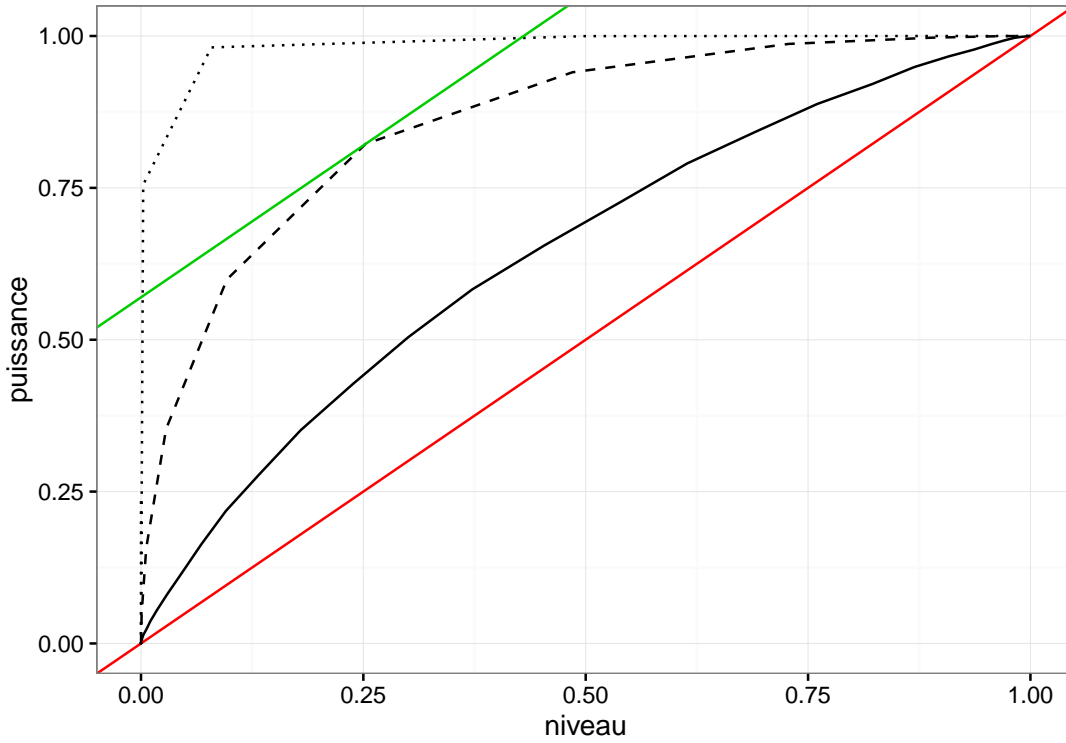


FIG. 1.2 : Pour une grille de seuils, on évalue niveau et puissance pour différentes tailles d'échantillon (100, 1000, 5000). Sur un même graphique on représente la courbe niveau/puissance pour ces trois tailles d'échantillon (lignes brisées noires). Pour chaque courbe, le meilleur compromis erreur de 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.

La preuve du lemme 1.16 se réduit alors à :

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

□

1.6 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, BOUCHERON, LUGOSI et MASSART [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

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

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

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

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

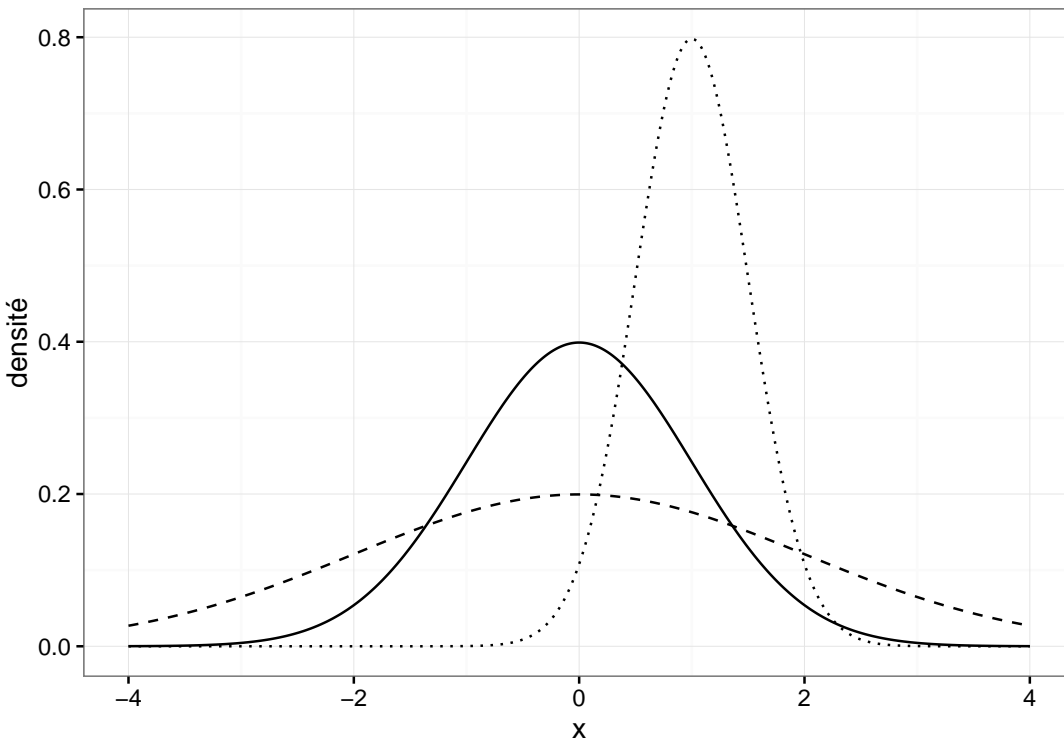


FIG. 2.1 : Densités de $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 2)$ (tirets), $\mathcal{N}(1, 1/4)$ (pointillés)

Toute transformation affine d'une gaussienne centrée réduite suit une loi gaussienne. Si $X \sim \mathcal{N}(0, 1)$ alors $\sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$ 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 dérivable telle que $\mathbb{E}[g'(X)] < \infty$ et $\mathbb{E}[|Xg(X)|] < \infty$, alors

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

PREUVE.

$$\begin{aligned} \mathbb{E}[g'(X)] &= \int_{\mathbb{R}} \phi(x)g'(x)dx \\ &= [g(x)\phi(x)]_{-\infty}^{\infty} - \int_{\mathbb{R}} -x\phi(x)g(x)dx \\ &= 0 + \mathbb{E}[Xg(X)]. \end{aligned}$$

□

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

$$\widehat{\Phi}(t) := \mathbb{E}[e^{itX}] = e^{it\mu - \frac{t^2\sigma^2}{2}}.$$

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

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

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

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

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

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

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

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

□

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 $\widehat{F}(t) = e^{-t^2/2}$ et $\widehat{G}(t) = 0$. On vient de vérifier que la fonction caractéristique de la loi de X est celle de $\mathcal{N}(0, 1)$. \square

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 :

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

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 $x \geq 0$,

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

PREUVE. Essentiellement une intégration par parties

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

Comme le second terme est négatif, on a :

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

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 $\int_x^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{u^2} e^{-\frac{u^2}{2}} du$.

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

\square

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

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

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

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

D'où, par récurrence sur k ,

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

□

On note au passage que $(2k)!/(2^k k!)$ est égal au nombre de partitions de $\{1, \dots, 2k\}$ 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 :

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

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 $(X_1, \dots, X_n)^t$ est un *vecteur gaussien* si et seulement si pour toute collection de réels $(\lambda_1, \lambda_2, \dots, \lambda_n)$, 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, \dots, X_n est une suite de variables gaussiennes indépendantes, alors $(X_1, \dots, X_n)^t$ est un vecteur gaussien.*

Dans la suite, on appellera *vecteur gaussien standard*, 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 $X = (X_1, \dots, X_n)^t$ la matrice K de dimensions $n \times n$ dont les coefficients sont donnés par :

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

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

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

(ce n'est pas un résultat gaussien).

En effet

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

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

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

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

PROPOSITION 2.9 Si $Y = (Y_1, \dots, Y_n)^t$ 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

$$\lambda^t AKA^t \lambda,$$

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 ,

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

On notera $DP(k)$ et $SDP(k)$ 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 $\lambda \in \mathbb{R}^n$,

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

soit $\lambda^t K \lambda = \text{var}(\langle \lambda, X \rangle)$. 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 : une matrice définie positive K admet une *décomposition de Cholesky*: 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 ($A \in \text{DP}(n)$), 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 . \square

DÉFINITION 2.14 ESPACE GAUSSIEN Si $(X_1, \dots, X_n)^t$ 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 $(X_1, \dots, X_n)^t$).

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

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 :

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

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

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

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, \dots, 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

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

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

□

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, \dots, 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 $\vec{s} \in \mathbb{R}^k$:

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

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 $\vec{s} \in \mathbb{R}^k$:

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

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

$$K = \begin{bmatrix} A & B^t \\ B & W \end{bmatrix}$$

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

On s'intéresse à l'espérance conditionnelle de $(X_1, \dots, X_k)^t$ sachant (X_{k+1}, \dots, X_n) et à la loi conditionnelle de $(X_1, \dots, X_k)^t$ sachant (X_{k+1}, \dots, 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 $(X_{k+1}, \dots, X_n)^t$ sachant $(X_1, \dots, X_k)^t$ est une transformation affine de $(X_1, \dots, X_k)^t$:*

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

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

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'un 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 $(X_1, \dots, X_n)^t$ 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 $n \times n$*

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

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 $A = L_1L_1^t$, $W - BA^{-1}B^t = L_2L_2^t$ où L_1, L_2 sont triangulaires inférieures, et la factorisation de Cholesky de K s'écrit :

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

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

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

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

Dans la suite, on suppose que $(X_1, \dots, X_n)^t$ et $(Y_1, \dots, Y_n)^t$ vivent sur le même espace probabilisé. Comme L_1 est inversible, les tribus engendrées par $(X_1, \dots, X_k)^t$ et $(Y_1, \dots, Y_k)^t$ 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 $(Y_{k+1}, \dots, Y_n)^t$ 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 $(X_{k+1}, \dots, X_n)^t$ sachant $(X_1, \dots, X_k)^t$ 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 $(Y_1, \dots, Y_k)^t$. Comme $(Y_1, \dots, Y_k)^t = L_1^{-1}(X_1, \dots, X_k)^t$, 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 $\text{var}(X) = \sigma_x^2$, $\text{var}(Y) = \sigma_y^2$ et $\text{cov}(X, Y) = \rho\sigma_x\sigma_y$, la loi conditionnelle de Y sachant X est

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

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 $\mathbb{E}[\|Y - Z\|^2]$).

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

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

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

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

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.

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

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.

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

On a établi au passage que

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

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_k^2) 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 $2e^{-x^2/2}/\sqrt{2\pi}$. 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

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

□

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^tMX 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 $OO^t = \text{Id}$).

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^tMX est donc distribuée comme Y^tL^tMLY . Il existe une transformation orthogonale O telle que $L^tML = O^t \text{diag}(\lambda_i)O$. Le vecteur OY est distribué selon $\mathcal{N}(0, I_n)$ car les lois gaussiennes standards sont invariantes par transformation orthogonale. \square

2.8 NORMES DE VECTEURS GAUSSIENS DÉCENTRÉS

La loi du carré de la norme d'un vecteur gaussien général de matrice de covariance $\sigma^2 \text{Id}$, 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

$$\mathbb{P}\{X \leq Y\} = 1.$$

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 $x \in \mathbb{R}$, $F(x) \geq G(x)$. Les fonctions quantiles $F^{\leftarrow}, G^{\leftarrow}$ vérifient $F^{\leftarrow}(p) \leq G^{\leftarrow}(p)$ pour $p \in (0, 1)$.

Réciproquement.

PROPOSITION 2.35 Si F et G sont deux fonctions de répartition qui vérifient pour tout $x \in \mathbb{R}$ $F(x) \geq G(x)$ 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 :

$$\mathbb{P}\{X \leq Y\} = 1.$$

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

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

\square

THÉORÈME 2.36 Si $X \sim \mathcal{N}(0, \sigma^2 \text{Id})$ et $Y \sim \mathcal{N}(\theta, \sigma^2 \text{Id})$ 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 $x \geq 0$,

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

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 X est distribué comme $(Y_1 + \|\theta\|_2)^2 + \sum_{i=2}^d Y_i^2$. Ceci prouve la première partie de l'énoncé.

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

$$\mathbb{P} \{ (Y_1 + \|\theta\|_2)^2 \leq x \} \leq \mathbb{P} \{ Y_1^2 \leq x \},$$

soit

$$\mathbb{P} \{ |Y_1 + \|\theta\|_2| \leq \sqrt{x} \} \leq \mathbb{P} \{ |Y_1| \leq \sqrt{x} \},$$

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. \square

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

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

où $X \sim \mathcal{N}(0, \text{Id}_d)$, $\theta \in \mathbb{R}^d$ 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 $\mathbb{R}^n = \bigoplus_{j=1}^k E_j$ 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 $(\pi_{E_j} X)_{j \leq k}$ est une famille indépendante et pour chaque j

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

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 de la proposition 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 $(\alpha_j)_{j \in \mathcal{I}}, (\beta_j)_{j \in \mathcal{J}}$, 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.

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

\square

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 $\bar{X}_n := \sum_{i=1}^n X_i/n$ et $V := \sum_{i=1}^n (X_i - \bar{X}_n)^2$, alors

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

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

Comme

$$\begin{pmatrix} \bar{X}_n \\ \vdots \\ \bar{X}_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \times (1 \quad \dots \quad 1) X$$

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

Le vecteur $(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^t$ 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 $(\bar{X}_n, \dots, \bar{X}_n)^t$, et $(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^t$ sont indépendants.

La loi de \bar{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 $X \sim \mathcal{N}(0, 1)$, $Y \sim \chi_p^2$ 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 ,

$$\text{var}(f(X)) \leq \mathbb{E}\|\nabla f\|^2 \quad (\text{inégalité de Poincaré})$$

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

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

et pour $\lambda > 0$

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

Pour tout $t \geq 0$,

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

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} .*

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

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

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

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

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

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

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

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

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

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

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

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

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*

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

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 $(\mathbb{E}\|X\|_2)^2 = \mathbb{E}\|X\|_2^2 - \text{var}(\|X\|_2) = d - \text{var}(\|X\|_2)$ et de la majoration de variance obtenue de l'inégalité de Poincaré. \square

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

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

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

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

et que

$$\mathbb{P}\{\|X - Y\| - \mathbb{E}\|X - Y\| \geq t\} \leq e^{-t^2}.$$

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

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

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 ,

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

On convient de noter $B_p^d(r) = \{\theta : \theta \in \mathbb{R}^d, \sum_{i=1}^d |\theta_i|^p \leq r^p\}$. 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, \dots, X_n et la suite $\langle X_1, \theta \rangle, \dots, \langle X_n, \theta \rangle$. 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 $(\Omega, \mathcal{F}, \mathcal{P}, \dots)$ est *dominé* si toutes les lois de \mathcal{P} sont absolument continues par rapport à une mesure σ -finie sur (Ω, \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 $\theta = (\mu, \Sigma), x = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$ 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 \bar{X}_n)

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

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

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

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 $(\bar{X}_n, \widehat{\Sigma})$ 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 :

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

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

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

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 $\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \text{DP}(d)\}$, 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 $\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \text{DP}(d)\}$, la covariance empirique est ancillaire par rapport à l'espérance. De fait, l'estimateur de choix (\bar{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 *pivotal* 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 \bar{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 ,

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

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

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

est une statistique suffisante minimale ($X_{(i)}$ est la $i^{\text{ème}}$ 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 $(\bar{X}_n, \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X}_n)^2)$ (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

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

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

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

où le paramètre inconnu est θ_0 dans \mathbb{R}^p . Selon les circonstances, on connaît ou on ne connaît 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 moindres carrés ordinaires*, MCO)

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

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 :

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

où \hat{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 $\hat{y} = \mathbf{X}\theta$. Comme

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

cette solution est

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

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

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

et

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

Comme

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

on constate que l'estimateur des moindres carrés $\hat{\theta}$ 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 moindres 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 moindres 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 homoschéastique, 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 homoschéastique,

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

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

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

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

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

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 $\hat{\theta} - \theta_0$ par $\mathbf{O}\mathbf{D}^{1/2}\mathbf{O}^t = (\mathbf{X}^t\mathbf{X})^{1/2}$. \square

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

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

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 $\hat{\epsilon} := \hat{y} - y$. La loi du vecteur des résidus est facilement déduite de l'observation :

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

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 homoschédistique,*

- i) $\hat{Y} - Y$ est indépendant de $\mathbf{X}(\hat{\theta} - \theta_0)$.
- ii) $\frac{\|\hat{Y} - Y\|^2}{\sigma^2} \sim \chi_{n-p}^2$.
- iii) $\frac{\|\mathbf{X}(\hat{\theta} - \theta_0)\|^2}{\sigma^2} \sim \chi_p^2$.

PREUVE. On observe $\hat{Y} - Y = \sigma(\mathbf{H} - \mathbf{Id})\epsilon$ et $\mathbf{X}(\hat{\theta} - \theta_0) = \sigma\mathbf{H}\epsilon$. Au facteur σ près, on obtient $\hat{Y} - Y$ et $\mathbf{X}(\hat{\theta} - \theta_0)$ 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. \square

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

DÉFINITION 3.14 (DISTRIBUTION DE FISHER) Si $U \sim \chi_j^2$ est indépendante de $V \sim \chi_k^2$ 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 $\frac{\|\mathbf{H}\epsilon\|^2}{\|(\mathbf{Id} - \mathbf{H})\epsilon\|^2}$ 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 homoschédistique, l'ellipsoïde*

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

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 ellipsoïde 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, \dots, 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 $\mathcal{L}(\mathbf{X}^0)$, $\hat{y}^0 = \mathbf{H}^0 y = \mathbf{H}^0 \times \mathbf{H} y$. On note θ^0 le vecteur de dimension k formé par les k premiers coefficients de θ .

Sous l'hypothèse nulle Hyp_0 ,

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

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

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

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 $\mathcal{N}(\mathbf{X}^0 \theta^0 + \mathbf{X}^1 \theta^1, \sigma^2 \mathbf{I}_n)$.

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

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

est maintenant distribuée selon une loi du χ^2 à $p - k$ degrés de liberté décentrée de $\|(\mathbf{I}_d - \mathbf{H}^0) \mathbf{X}^1 \theta^1\|_2$. La quantité aléatoire $\|\hat{Y} - Y\|^2 / \sigma_0^2$ reste distribuée selon χ_{n-p}^2 . 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 2.8).

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 ($Y = \text{Gas}$) 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 liberté $n = 56, p = 2, k = 1$. 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 homoschéastique. 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 interpréter 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|)
## (Intercept)   6.85383     0.13596   50.409 < 2e-16 ***
```

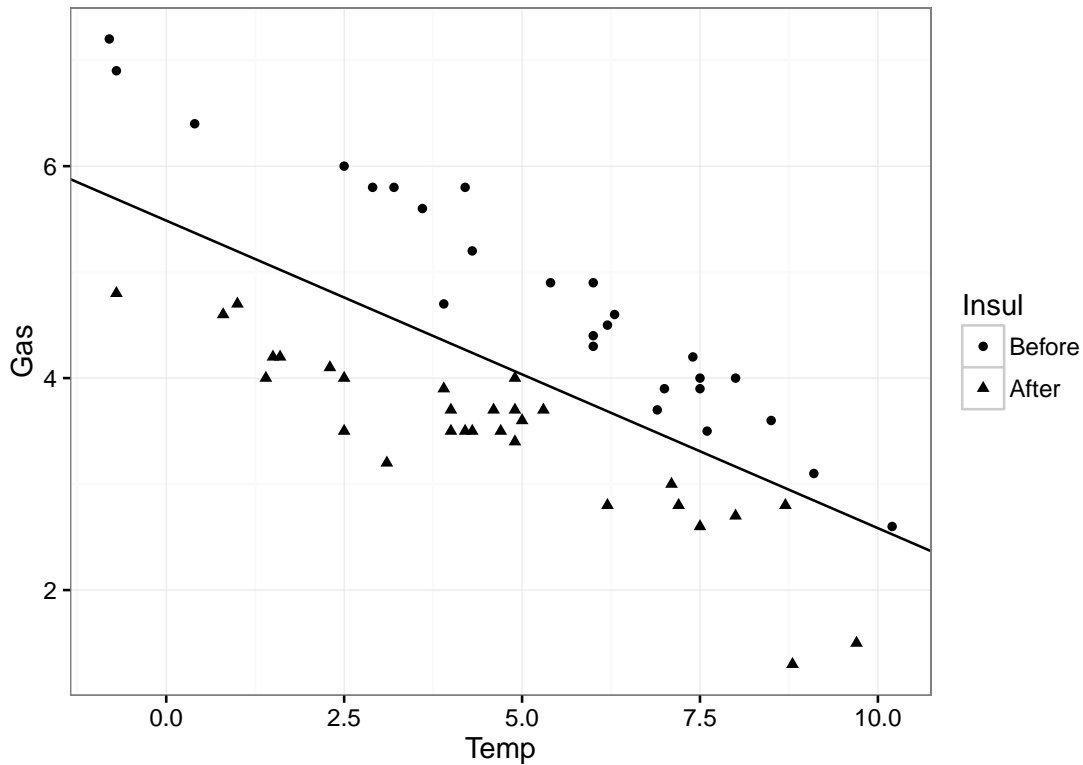


FIG. 3.1 : Les points correspondant à l'hiver précédent l'isolation se situent au dessus de la droite de régression et les autres au dessous.

```
## 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
```

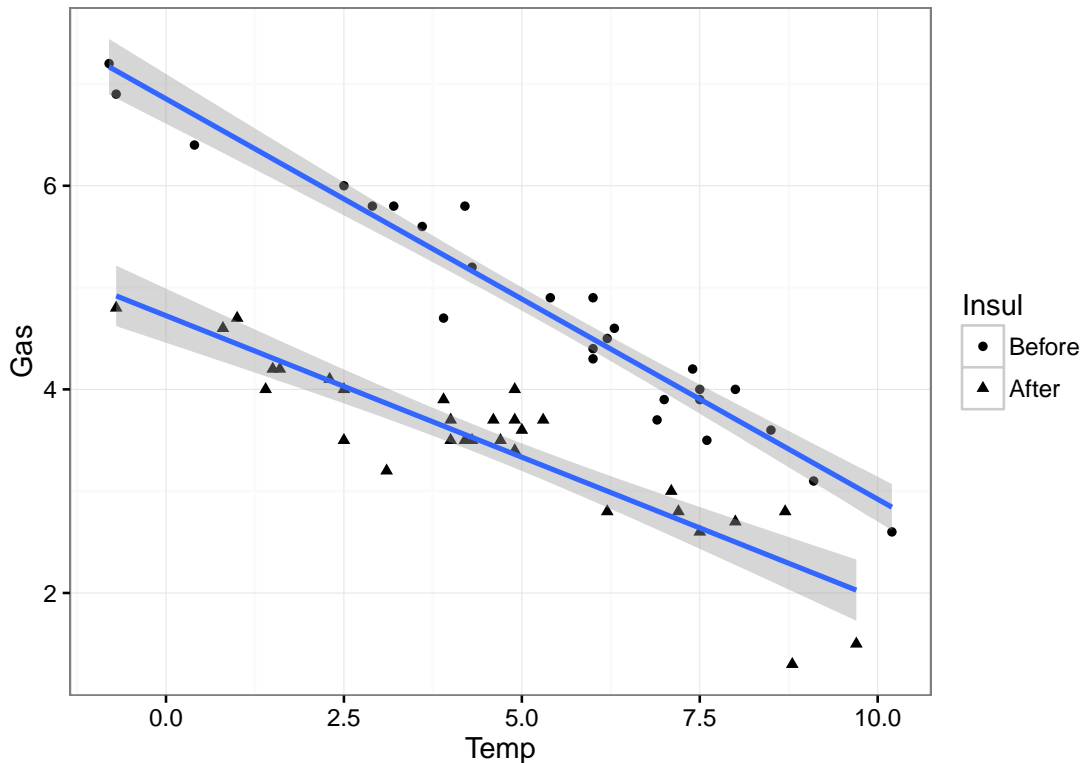



FIG. 3.2 : Visualisation des droites de régression correspondant aux deux conditions d'isolation.

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 moindres 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
- ii) Maximisation de la vraisemblance
- iii) Minimisation de contraste

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 :

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

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 $\text{Id}(X)$. 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 \bar{X}_n de l'échantillon. Pour estimer le paramètre originel, on inverse ψ : on estime θ par $\hat{\theta} = 1/\bar{X}_n$. L'estimateur est alors biaisé (inégalité de Jensen).

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

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

Outils d'analyse

Si on souhaite construire un estimateur pour la paramétrisation originelle (par exemple pour l'intensité dans le cadre des lois exponentielles), il est tentant d'utiliser l'estimateur $\psi^{-1}(\bar{T}_n)$. Si ψ^{-1} est continue, la convergence presque sûre de la suite $(\bar{T}_n)_{n \in \mathbb{N}}$ et le principe de l'image continue nous garantissent que la suite $(\psi^{-1}(\bar{T}_n))_{n \in \mathbb{N}}$ est une suite consistante d'estimateurs de θ . Pour caractériser l'éventuelle normalité asymptotique de $(\sqrt{n}(\psi^{-1}(\bar{T}_n) - \theta))_{n \in \mathbb{N}}$, il faut disposer d'informations supplémentaires sur la régularité de ψ^{-1} . Dans le cadre de ce chapitre, nous utiliserons le *théorème d'inversion globale*. La version suivante suffira à nos besoins.

THÉORÈME 4.2 (THÉORÈME D'INVERSION GLOBALE) *Soit E un espace euclidien, $U \subseteq E$ ouvert, $f : U \rightarrow E$ 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 $f(U)$)*

L'appendice C apporte quelques rappels d'analyse.

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

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

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

est un difféomorphisme 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

$$\psi(\theta) = \bar{T}_n$$

admet une solution $\hat{\theta}_n \in \Theta$.

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

iii) La suite $(\sqrt{n}(\psi^{-1}(\bar{T}_n) - \theta))_{n \in \mathbb{N}}$ est asymptotiquement normale centrée. La covariance de la loi limite est

$$(D\psi|_\theta)^{-1} \text{cov}_\theta(T) (D\psi|_\theta)^{-1}$$

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

Entropie relative

Dans la suite, on notera $P \not\leq Q$ 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 (Ω, \mathcal{F}) , l'entropie relative de P par rapport à Q , notée $D(P\|Q)$ est définie par

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

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 $P \trianglelefteq Q$, $\mathbb{E}_P \left[\log \frac{dP}{dQ} \right] = \mathbb{E}_Q \left[\frac{dP}{dQ} \log \frac{dP}{dQ} \right]$ est toujours bien défini, même si cette quantité n'est pas finie. En effet $x \log x \geq -e^{-1}$ sur $]0, \infty[$.
- ii) Si $P \trianglelefteq Q$ (P est absolument continue par rapport à Q), $\mathbb{E}_P \left[\log \frac{dP}{dQ} \right] = \mathbb{E}_Q \left[\frac{dP}{dQ} \log \frac{dP}{dQ} \right]$, et la convexité de $x \mapsto x \log x$, nous indique 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 $x \mapsto x \log x$).

4.3 MODÈLES EXPONENTIELS

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

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

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

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

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 reparamétrisation, on peut vérifier que les gaussiennes univariées s'inscrivent dans ce schéma :

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

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

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

THÉORÈME 4.6 Soient ν est une mesure dominante sur $(\mathcal{X}, \mathcal{F})$, et $T : \mathcal{X} \rightarrow \mathbb{R}^d$ \mathcal{F} mesurable. S'il n'est pas vide, l'ensemble

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

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

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

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

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

On a donc

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

□

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 supposons 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'irredundance des statistiques $T(X)$.

DÉFINITION 4.7 (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 4.8 *Un modèle exponentiel en forme canonique est minimal (ou de plein rang) si et seulement si il est identifiable.*

PREUVE.

⇒) 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.

⇐) 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 4.9 *La fonction $\theta \mapsto \log Z(\theta)$ est C^∞ sur $\Theta = \tilde{\Theta}^\circ$.*

De plus

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

et

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

PREUVE. Les preuves sont des exercices de convergence dominée.

□

Entropie relative dans un modèle exponentiel

L'entropie relative permet 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 $\log Z(\cdot)$:

$$D(P_\theta \| P_{\theta'}) = \log Z(\theta') - \log Z(\theta) + \langle \nabla \log Z(\theta), \theta - \theta' \rangle. \quad (4.1)$$

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 injectivité de $\theta \mapsto \nabla \log Z(\theta)$.

PROPOSITION 4.10 *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 (4.1),

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

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

THÉORÈME 4.11 *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*

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

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

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

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 .

PREUVE. D'après le théorème d'inversion globale, la différentiabilité à tous les ordres de $\theta \mapsto \log Z(\theta)$ dans Θ , et l'injectivité de la fonction Λ sur Θ , il ne reste qu'à vérifier qu'en tout $\theta \in \Theta$, la différentielle représentée par $\nabla^2 \log Z(\theta)$ est inversible, autrement dit que $\nabla^2 \log Z(\theta)$ est une matrice (symétrique) définie positive.

Pour tout vecteur non nul $c \in \mathbb{R}^k$,

$$c^t \nabla^2 \log Z(\theta) c = \text{var}_{P_\theta} (\langle c, T(X) \rangle) .$$

Si cette variance est nulle alors il existe $d \in \mathbb{R}$ tel que P_θ -presque sûrement, $\langle c, T(X) \rangle = d$, donc $\langle c, T(X) \rangle = d$ ν -presque partout, ce qui implique que le modèle n'est pas identifiable. \square

Le théorème 4.11, 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

$$\bar{T}_n = \nabla \log Z(\theta)$$

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

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 4.12 Si on considère 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 $x \in]0, \infty)$, on peut choisir comme dominante la mesure de densité $1/x$ sur $]0, \infty)$. Il s'agit d'un modèle exponentiel en forme canonique avec pour statistiques suffisantes $(\log X, -X)^t$. On a $\log Z(p, \lambda) = \log \Gamma(p) - p \log \lambda$ et

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

Le système d'équations

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

n'admet pas de solution explicite.

Dans la section suivante, nous allons voir la résolution de l'équation $\bar{T}_n = \nabla \log Z(\theta)$ 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.

4.4 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 $(\Lambda^{-1}(\bar{T}_n))$ 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 , $\nabla \ell_n(\theta) = n(\bar{T}_n - \nabla \log Z(\theta))$.

On note au passage que pour tout $\theta \in \Theta$,

$$\nabla \ell_n(\theta) = n(\bar{T}_n - \nabla \log Z(\theta)) .$$

S'il existe θ tel que $\nabla \ell_n(\theta) = 0$, alors par convexité de $-\ell_n(\cdot)$, ce θ est un maximum global de la vraisemblance. Si $\log Z(\cdot)$ est strictement convexe, ce maximum de vraisemblance est unique. Le maximum de vraisemblance est solution de l'équation $\bar{T}_n = \nabla \log Z(\theta)$. 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 $\theta \in \Theta$,

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

Sous $P_\theta^{\otimes n}$, la fonction score évaluée en θ est centrée. On peut en calculer la covariance, qu'on appelle l'*information de Fisher* pour le modèle échantillonné n fois en θ :

$$I_n(\theta) := \mathbb{E}_\theta [\nabla \ell_n(\theta) \nabla \ell_n(\theta)^t] = nI_1(\theta).$$

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

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

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

Si on note ψ la fonction $\theta \mapsto \nabla \log Z(\theta)$, l'estimateur au maximum de vraisemblance de θ est donné par $\psi^{-1}(\bar{T}_n)$. Le fait que ψ soit un difféomorphisme et la méthode delta indiquent que $\sqrt{n}(\hat{\theta}_n - \theta)$ 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 4.13 (NORMALITÉ ASYMPTOTIQUE) *Si un modèle exponentiel canonique est minimal, alors l'estimateur du maximum de vraisemblance est aussi un estimateur des moments et il est asymptotiquement normal*

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, I(\theta)^{-1}).$$

4.5 REMARQUES BIBLIOGRAPHIQUES

L'étude de la méthode des moments décrite ici est inspirée de l'ouvrage

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

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 suivant une exploration en profondeur des liens entre ces divergences d'information.

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

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 dans

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

L'introduction aux modèles exponentiels présentée ici est inspirée de

P. BICKEL et K. DOKSUM. **Mathematical statistics**. San Francisco, Calif. : Holden-Day Inc., 1976.

MR : 56\#1513

5.1 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 (Ω, \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 $\nu' \leq \nu$.

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

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

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 (Ω, \mathcal{F}) est *dominée* si il existe une mesure σ -finie ν sur (Ω, \mathcal{F}) telle que tout $P \in \mathcal{P}$ soit *absolument continue* par rapport à ν ($\forall P \in \mathcal{P}, P \leq \nu$).

Certaines collections de lois ne sont 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. Là encore, ce « choix » n'a pas d'effet sur les performances des méthodes. 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 $\Theta \times \Omega \times \dots \times \Omega$ qui à θ, x_1, \dots, 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 $(\theta = (\mu, \sigma), x_1, \dots, x_n)$ s'écrira

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

ou encore

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

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

5.2 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 $\hat{\theta}$ qui maximise $p_\theta(x_1, \dots, x_n)$ dans Θ :

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

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

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

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,

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

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.

5.3 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. Dans la suite on définit la *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 convexe, (en raison de l'absolue continuité des lois les unes par rapport aux autres). 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 $\theta \mapsto \nabla \log Z(\theta)$. 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

$$\theta \mapsto \ell_n(\theta) = n (\langle \theta, \bar{T}_n(X) \rangle - \log Z(\theta)) ,$$

le problème de la maximisation de la vraisemblance se décrit comme

Entrée : $t \in \mathbb{R}^k$

Résultat : $\hat{\theta} = \arg \max_{\theta \in \Theta} [\langle \theta, t \rangle - \log Z(\theta)]$ si existe.

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

EXEMPLE 5.4 Considérons le modèle des lois géométriques avec une paramétrisation canonique :

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

Si l'échantillon est une suite de 1 (ce qui se produit avec probabilité > 0 sous $P_{\theta_0}, \theta_0 \in \Theta$), la log-vraisemblance s'écrit

$$\ell_n(\theta) = n \log(e^\theta - 1) - n\theta = n \log(1 - e^{-\theta}).$$

Elle est concave et croissante sur Θ . Elle tend vers son supremum (fini) lorsque $\theta \rightarrow \infty$, 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, $\log Z(\theta) = -\log(e^\theta - 1)$ et $\nabla \log Z(\theta) = -e^\theta / (e^\theta - 1)$. La moyenne empirique de la statistique suffisante $-\bar{X}_n$ appartient à $(-\infty, -1]$. Il existe $\theta \in \Theta$ tel que $\nabla \log Z(\theta) = -\bar{X}_n$ si et seulement si $\bar{X}_n > 1$. Lorsqu'elle existe la solution de $\nabla \log Z(\theta) = -\bar{X}_n$ 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 5.5 *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 5.6

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

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 x 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 ,

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

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

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

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

Ceci signifie que t est un sous-gradient de $\theta \mapsto \log Z(\theta)$ en $\hat{\theta}$, qui est réduit à $\nabla \log Z(\hat{\theta})$ 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 $\hat{\theta} \in \Theta$ tel que $t = \nabla \log Z(\hat{\theta})$, 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 $\partial\Theta = \bar{\Theta} - \Theta$ 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 $\theta_n \rightarrow \partial\Theta$ de la façon suivante. On a soit $\theta_n \rightarrow \theta$ avec $\theta \notin \Theta$, soit $\|\theta_n\| \rightarrow \infty$.

Rappelons que si $\Theta \subseteq \mathbb{R}^d$ est ouvert et si $f : \Theta \rightarrow \mathbb{R}$ est continue et vérifie $\lim_{\theta \rightarrow \partial\Theta} f(\theta) = -\infty$, 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 5.7 *Si la fonction log-vraisemblance $\theta \mapsto \ell_n(\theta)$ est strictement concave et si elle tend vers $-\infty$ lorsque θ tend vers $\partial\Theta = \bar{\Theta} \setminus \Theta$, alors le maximum de vraisemblance est atteint en un point unique de Θ .*

EXERCICE 5.8 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. [1].

THÉORÈME 5.9 (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 5.10 *Dans un modèle exponentiel canonique minimal où $\Theta = \tilde{\Theta}$ (le domaine $\tilde{\Theta}$ de la fonction de partition est un ouvert), si $t_0 \in \mathbb{R}^k$ appartient à l'intérieur du support convexe de la mesure dominante, la fonction*

$$\theta \mapsto \langle \theta, t_0 \rangle - \log Z(\theta)$$

est strictement concave et elle tend vers $-\infty$ lorsque θ tend vers $\partial\Theta = \bar{\Theta} \setminus \Theta$.

REMARQUE 5.11 Dans un modèle exponentiel canonique minimal, $\tilde{\Theta}$ 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 $\tilde{\Theta}$ est alors $(-\infty, 1]$. Le support de la dominante est \mathbb{R} . L'ensemble des valeurs possibles de $\frac{d \log Z(\theta)}{d\theta}$ est borné.

PREUVE. La stricte concavité a déjà été démontrée.

Commençons par expliciter la condition « t appartient à l'intérieur C° du support convexe C de la mesure dominante ».

Si $t \notin C^\circ$, il existe un hyperplan défini par une direction $c \in \mathbb{R}^d$ tel que $\langle c, t \rangle \geq \langle c, u \rangle$ pour tout $u \in C$. On a donc pour tout $\theta \in \Theta$,

$$P_\theta \{ \langle c, t \rangle < \langle c, T(X) \rangle \} = 0.$$

A contrario, si pour toute direction $c \in \mathbb{R}^d$, pour tout $\theta \in \Theta$,

$$P_\theta \{ \langle c, t \rangle < \langle c, T(X) \rangle \} > 0$$

alors $t \in C^\circ$ (on a utilisé le fait que les lois P_θ sont mutuellement absolument continues).

Comme t_0 appartient à l'intérieur du support convexe C de la mesure dominante, pour toute direction c et tout $\theta \in \Theta$,

$$P_\theta \{ \langle c, t_0 \rangle < \langle c, T(X) \rangle \} > 0$$

Considérons une suite $(\theta_m)_{m \in \mathbb{N}}$ d'éléments de $\Theta \subset \mathbb{R}^k$, qui tend vers $\partial\Theta$.

a) Si $\theta_n \rightarrow \eta \in \partial\Theta \cap \mathbb{R}^k$ ($\eta \notin \Theta$, et donc comme $\tilde{\Theta}$ 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 ∞ . Pour alléger les notations on abrège $\log Z(\theta)$ en $\Lambda(\theta)$. On suppose sans perdre en généralité que la mesure dominante est une probabilité P_0 (voir page 42).

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 $\Lambda(\eta) < \infty$ (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

$$\mathbb{P}_0 \{ \langle \eta, T(X) \rangle \geq \langle \eta, t_0 \rangle + \delta \} > 0.$$

L'existence de δ est garantie par l'hypothèse que $t_0 \in C^\circ$.

$$\begin{aligned} \Lambda(\theta_m) &= \log \mathbb{E}_0 \left[e^{\langle \theta_m, T(X) \rangle} \right] \\ &\geq \log \mathbb{E}_0 \left[e^{\|\theta_m\| \langle \frac{\theta_m}{\|\theta_m\|}, T(X) \rangle} \mathbb{I}_{\langle \frac{\theta_m}{\|\theta_m\|}, T \rangle \geq \langle \frac{\theta_m}{\|\theta_m\|}, t_0 \rangle + \delta} \right] \\ &\geq \langle \theta_m, t_0 \rangle + \|\theta_m\| \delta + \log P_0 \left\{ \langle \frac{\theta_m}{\|\theta_m\|}, T \rangle \geq \langle \frac{\theta_m}{\|\theta_m\|}, t_0 \rangle + \delta \right\}. \end{aligned}$$

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, $\frac{\theta_m}{\|\theta_m\|} \rightarrow \eta \in \Theta$ 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 5.10 entraîne le corollaire suivant.

COROLLAIRE 5.12 *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.*

5.4 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 recours à la notion d'*information de Fisher*. La matrice d'information de Fisher $I_n(\theta^0)$ est définie par

$$I_n(\theta^0) := \mathbb{E}_{\theta^0} \left[\nabla \ell_n(\theta^0) \nabla \ell_n(\theta^0)^t \right]$$

où le gradient de la log-vraisemblance $\ell_n(\theta, x_1, \dots, x_n)$ est pris par rapport à θ en θ_0 . Ce gradient est une fonction de \bar{T}_n , il vaut

$$n (\bar{T}_n - \nabla \log Z(\theta^0)).$$

Comme $\mathbb{E}_{\theta^0}[\bar{T}_n] = \nabla \log Z(\theta^0)$,

$$I_n(\theta^0) := n^2 \text{cov}_{\theta^0}(\bar{T}_n) = n \text{cov}_{\theta^0}(T) = n \nabla^2 \log Z(\theta^0) = nI(\theta^0).$$

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

$$\sqrt{n}I(\theta)^{1/2}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, \text{Id}_k)$$

conduit naturellement à une construction de région de confiance,

$$\left\{ \theta' : n \left\| I(\theta')(\hat{\theta}_n - \theta') \right\|^2 \leq q_{k,1-\alpha} \right\}$$

où $q_{k,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi χ_k^2 . Cette région de confiance peut être rendue plus transparente en substituant à l'information $I(\theta')$ la quantité empirique $\hat{I}_n = I(\hat{q}_n)$. Le région de confiance devient alors un ellipsoïde centré en $\hat{\theta}_n$. 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 5.13 (PHÉNOMÈNE DE WILKS) *Dans un modèle exponentiel canonique minimal ($P_\theta, \theta \in \Theta \subseteq \mathbb{R}^k$), pour tout $\theta \in \Theta$, sous $P_\theta^{\otimes N}$,*

$$2nD(P_\theta \| P_{\hat{\theta}_n}) \rightsquigarrow \chi_k^2$$

et

$$nD(P_\theta \| P_{\hat{\theta}_n}) - (\ell_n(\hat{\theta}_n) - \ell_n(\theta)) \xrightarrow{P} 0.$$

L'ingrédient essentiel de la preuve est une variante de la méthode delta.

THÉORÈME 5.14 (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

$$a_n(X_n - x) \rightsquigarrow X$$

alors

- i) $X_n \xrightarrow{P} x$;
- ii) $a_n^2(f(X_n) - f(x)) - a_n^2 \frac{1}{2}(X_n - x)^t \nabla^2 f(X_n - x) \xrightarrow{P} 0$;
- iii) $a_n^2(f(X_n) - f(x)) \rightsquigarrow \frac{1}{2}X^t \nabla^2 f X$.

PREUVE. [Preuve du Théorème 5.14] 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 entraînent 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,

$$(f(X_n) - f(x)) - \frac{1}{2}(X_n - x)^t \nabla^2 f(X_n - x) = R(X_n - x)$$

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 B.18) de la suite $(a_n(X_n - x))_{n \in \mathbb{N}}$ entraîne que pour tout $\eta > 0$, il existe $M(\eta) > 0$, tel que pour tout n

$$\mathbb{P} \{a_n^2 \|X_n - x\|^2 \geq M(\eta)\} \leq \eta.$$

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

$$\mathbb{P} \{a_n^2 R(X_n - x) \geq \delta^2 M(\eta)\} \leq \eta.$$

Si on choisit δ de façon à ce que $\delta^2 M(\eta) < \eta$, on a donc pour n assez grand

$$\mathbb{P} \{a_n^2 R(X_n - x) \geq \eta\} \leq \eta.$$

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. \square

Nous pouvons maintenant passer à la preuve du Théorème 5.13.

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 entraînent que sous $P_{\theta}^{\otimes \mathbb{N}}$

$$nD(P_{\theta} \| P_{\hat{\theta}_n}) - \frac{n}{2} (\hat{\theta}_n - \theta)^t I(\theta) (\hat{\theta}_n - \theta) \xrightarrow{P} 0.$$

Ceci établit la première partie du théorème.

Notons tout d'abord que lorsque le maximum de vraisemblance est bien défini,

$$\ell_n(\hat{\theta}_n) - \ell_n(\theta) = nD(P_{\hat{\theta}_n} \| P_{\theta}).$$

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 entraînent que sous $P_{\theta}^{\otimes \mathbb{N}}$

$$nD(P_{\hat{\theta}_n} \| P_{\theta}) - \frac{n}{2} (\hat{\theta}_n - \theta)^t I(\theta) (\hat{\theta}_n - \theta) \xrightarrow{P} 0.$$

Ceci établit la seconde partie du théorème. \square

Considérons maintenant la région aléatoire définie par

$$\hat{A}_n(\alpha) := \left\{ \theta' : \theta' \in \Theta, \ell_n(\hat{\theta}_n) - \ell_n(\theta') \leq \frac{1}{2} q_{k, 1-\alpha} \right\}.$$

Un corollaire immédiat du Théorème 5.13 suit.

COROLLAIRE 5.15 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 $1 - \alpha$:

$$\lim_n P_\theta^n \left\{ \theta \in \widehat{A}_n(\alpha) \right\} = 1 - \alpha.$$

5.5 REMARQUES BIBLIOGRAPHIQUES

Au delà des modèles exponentiels les plus simples et les plus classiques, les modèles graphiques (graphical models) [4].

L'optimisation convexe joue un rôle croissant en statistique computationnelle et en théorie de l'apprentissage statistique [1]. Une panorama des résultats et méthodes d'optimisation utiles aux statisticiens peut être lu dans [2].

La méthode delta est exposée, illustrée et généralisée dans [3].

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

Nous revenons 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 1.5). 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 relierons 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 distance en variation est elle-même liée à l'entropie relative par l'inégalité de Pinsker. Cette relation importante permet de minorer les probabilités d'erreur dans les tests entre lois « produit ».

6.2 LEMME DE NEYMAN-PEARSON

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

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

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. \square

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 (Ω, \mathcal{F}) est définie par

$$d_{\text{TV}}(P, Q) := \sup_{A \in \mathcal{F}} P(A) - Q(A).$$

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 (Ω, \mathcal{F}) .*

La distance en variation est exigeante. Elle définit en général un topologie plus fine que la topologie de la convergence en loi. Là encore nous laissons cette proposition en exercice.

PROPOSITION 6.4 *Si une suite de lois (P_n) sur (Ω, \mathcal{F}) vérifie $\lim_n d_{\text{TV}}(P_n, Q) = 0$ où Q est une loi sur (Ω, \mathcal{F}) alors (P_n) converge étroitement vers Q .*

La réciproque est fautive : 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 $d_{\text{TV}}(P_n, \mathcal{N}(0, 1)) = 1$.

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 (Ω, \mathcal{F}) qui domine les lois P et Q , alors

$$d_{\text{TV}}(P, Q) = \frac{1}{2} \int_{\Omega} \left| \frac{dP}{d\nu}(\omega) - \frac{dQ}{d\nu}(\omega) \right| d\nu(\omega)$$

$$d_{\text{TV}}(P, Q) = 1 - \int_{\Omega} \frac{dP}{d\nu}(\omega) \wedge \frac{dQ}{d\nu}(\omega) d\nu(\omega).$$

PREUVE. On peut prendre comme mesure dominante $(P + Q)/2$. Et vérifier que l'ensemble

$$A := \left\{ \omega : \frac{dP}{d\nu}(\omega) > \frac{dQ}{d\nu}(\omega) \right\}$$

vérifie $d_{\text{TV}}(P, Q) = P(A) - Q(A)$. □

La connexion entre erreurs de test et distance en variation est résumée ici.

PROPOSITION 6.6 (AFFINITÉ DE TEST)

$$\inf_T (P_0\{T = 1\} + P_1\{T = 0\}) = 1 - d_{\text{TV}}(P_0, P_1) = \int_{\Omega} \frac{dP_1}{d\nu}(\omega) \wedge \frac{dP_0}{d\nu}(\omega) d\nu(\omega).$$

L'expression $1 - d_{\text{TV}}(P_0, P_1)$ est appelée *affinité de test*.

PREUVE. Soit A la région critique d'un test $T : A := \{\omega : T(\omega) = 1\}$.

$$P_0\{T = 1\} + P_1\{T = 0\} = 1 - (P_1\{A\} - P_0\{A\}).$$

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 :

$$d_{\text{TV}}(P^{\otimes n}, Q^{\otimes n}) \leq n d_{\text{TV}}(P, Q)$$

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 $P \leq Q$. 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

$$I_f(P\|Q) := \int_{\Omega} f\left(\frac{p(\omega)}{q(\omega)}\right) q(\omega) d\nu(\omega).$$

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)

$$d_{\text{TV}}(P, Q) \leq \sqrt{\frac{1}{2}D(P\|Q)}.$$

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 (Ω, \mathcal{F}) , l'entropie relative de P par rapport à Q vérifie*

$$D(P\|Q) = \sup_{f: \mathbb{E}_Q e^f < \infty} \{ \mathbb{E}_P f - \log \mathbb{E}_Q e^f \}.$$

PREUVE. Si $P \not\leq Q$, il existe A telle que $P(A) > 0$ et $Q(A) = 0$, en considérant $f_n = n\mathbb{1}_A$, on observe $\mathbb{E}_P f_n - \log \mathbb{E}_Q e^{f_n} = nP(A)$ qui tend vers ∞ avec n . Si $P \leq Q$, la représentation variationnelle est vérifiée.

Si $P \leq Q$, on vérifie d'abord que

$$D(P\|Q) = \sup_{U: \mathbb{E}_Q e^U = 1} \mathbb{E}_P U. \quad (6.1)$$

En effet, pour toute variable aléatoire U vérifiant $\mathbb{E}_Q e^U = 1$, on peut définir une autre loi Q' de densité e^U par rapport à Q , et $\frac{dP}{dQ} e^{-U}$ est une dérivée de Radon-Nykodym de P par rapport à Q' . On a alors

$$D(P\|Q) - \mathbb{E}_P U = \mathbb{E}_{Q'} \left[\frac{dP}{dQ} e^{-U} \log \left(\frac{dP}{dQ} e^{-U} \right) \right] \geq 0$$

où la dernière inégalité suit de l'inégalité de Jensen. Cela prouve que lorsque $P \leq Q$

$$D(P\|Q) \geq \sup_{U: \mathbb{E}_Q e^U = 1} \mathbb{E}_P U.$$

Comme il y a égalité lorsque $U = \log \frac{dP}{dQ}$, la première formule (6.1) est établie.

Si $\mathbb{E}_Q e^f < \infty$,

$$\mathbb{E}_P f - \log \mathbb{E}_Q e^f = \mathbb{E}_P \left[\log \frac{e^f}{\mathbb{E}_Q e^f} \right].$$

Le supremum dans l'énoncé de la formulation variationnelle n'est pas plus grand que celui de (6.1). \square

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

$$D(P\|Q) \geq \mathbb{E}_P [\lambda(\mathbb{I}_A - \mathbb{E}_Q \mathbb{I}_A)] - \log \mathbb{E}_Q \left[e^{\lambda(\mathbb{I}_A - \mathbb{E}_Q \mathbb{I}_A)} \right].$$

D'après le lemme de Hoeffding,

$$\log \mathbb{E}_Q \left[e^{\lambda(\mathbb{I}_A - \mathbb{E}_Q \mathbb{I}_A)} \right] \leq \frac{\lambda^2}{8}.$$

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

$$D(P\|Q) \geq \sup_{A \in \mathcal{F}} 2(P(A) - Q(A))^2 = 2d_{\text{TV}}(P, Q)^2.$$

□

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

$$D(P_1\|P_0) \leq \frac{1}{2n} \frac{h^2}{\theta(1-\theta)}.$$

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 décroît 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 $\tilde{\theta}_n$ 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 $\|\tilde{\theta}_n - \theta\| > \|\tilde{\theta}_n - \theta'\|$.

$$\begin{aligned} & \mathbb{E}_\theta \left[\|\tilde{\theta}_n - \theta\|^2 \right] + \mathbb{E}_{\theta'} \left[\|\tilde{\theta}_n - \theta'\|^2 \right] \\ & \geq \mathbb{E}_\theta \left[\|\tilde{\theta}_n - \theta\|^2 \mathbb{1}_{\|\tilde{\theta}_n - \theta\| > \|\tilde{\theta}_n - \theta'\|} \right] + \mathbb{E}_{\theta'} \left[\|\tilde{\theta}_n - \theta'\|^2 \mathbb{1}_{\|\tilde{\theta}_n - \theta\| < \|\tilde{\theta}_n - \theta'\|} \right] \\ & \geq \frac{\|\theta' - \theta\|^2}{4} (P_\theta^{\otimes n}\{T = 1\} + P_{\theta'}^{\otimes n}\{T = 0\}) \\ & \geq \frac{\|\theta' - \theta\|^2}{4} \left(1 - \sqrt{\frac{n}{2}D(P_\theta\|P_{\theta'})} \right) \end{aligned}$$

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

$$\begin{aligned} & \mathbb{E}_\theta \left[\|\tilde{\theta}_n - \theta\|^2 \right] + \mathbb{E}_{\theta+h/\sqrt{n}} \left[\|\tilde{\theta}_n - \theta - \frac{h}{\sqrt{n}}\|^2 \right] \\ & \geq \frac{\|h\|^2}{4n} \left(1 - \sqrt{\frac{h^t I(\theta) h}{8}} \right). \end{aligned}$$

Ce résultat peut être affiné. Les développements les plus intéressants s'appuient sur une autre f -divergence, la distance de Hellinger.

6.4 TESTS ET 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 (comme 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 développements actuels en statistique non-paramétrique.

DÉFINITION 6.11 (DISTANCE ET AFFINITÉ DE HELLINGER) Soient P et Q deux lois de probabilités sur (Ω, \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

$$H(P, Q)^2 := \frac{1}{2} \int_{\Omega} (\sqrt{p} - \sqrt{q})^2 d\nu.$$

L'affinité de Hellinger $\rho(P, Q)$ est définie par

$$\rho(P, Q) := 1 - H(P, Q)^2 = \int_{\Omega} \sqrt{pq} d\nu.$$

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 vérifie immédiatement que

$$H(P, Q)^2 \leq d_{\text{TV}}(P, Q).$$

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 paramétrer 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 (Ω, \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

$$L_n \geq z \iff \prod_{i=1}^n \sqrt{\frac{q(X_i)}{p(X_i)}} \geq e^{z/2}$$

l'inégalité de Markov entraine

$$\begin{aligned} P^n \{L_n \geq z\} &\leq e^{-z/2} \mathbb{E}_{P^n} \left[\prod_{i=1}^n \sqrt{\frac{q(X_i)}{p(X_i)}} \right] \\ &= e^{-z/2} \rho(P, Q)^n \\ &\leq \exp\left(-\frac{z}{2} - nH(P, Q)^2\right). \end{aligned}$$

La seconde inégalité se démontre de manière semblable. \square

En choisissant $z = 0$, et en définissant le test T_n par

$$T_n = \mathbb{I}_{L_n \geq 1},$$

on a simultanément

$$\max(\mathbb{E}_{P^n} T_n, \mathbb{E}_{Q^n} (1 - T_n)) \leq \exp(-nH(P, Q)^2).$$

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 (Ω, \mathcal{F}) , pour toute région critique d'un test de P^n contre Q^n (tout événement A dans $\mathcal{F}^{\otimes n}$),*

$$\max(P^n(A), Q^n(A^c)) \geq \frac{1}{4} - \frac{nH(P, Q)^2}{2}.$$

PREUVE.

$$\begin{aligned} \rho(P^n, Q^n) &= \int_{\Omega^n} \sqrt{p^n q^n} \mathbb{I}_A d\nu + \int_{\Omega^n} \sqrt{p^n q^n} \mathbb{I}_{A^c} d\nu \\ &\leq \left(\int_{\Omega^n} p^n \mathbb{I}_A d\nu \right)^{1/2} \left(\int_{\Omega^n} q^n d\nu \right)^{1/2} + \left(\int_{\Omega^n} p^n d\nu \right)^{1/2} \left(\int_{\Omega^n} \mathbb{I}_{A^c} q^n d\nu \right)^{1/2} \\ &= P^n(A)^{1/2} + Q^n(A^c)^{1/2}. \end{aligned}$$

Par ailleurs, $\rho(P^n, Q^n) = (1 - H(P, Q)^2)^n$. Finalement

$$\max(P^{\otimes n}(A), Q^{\otimes n}(A^c)) \geq \frac{(1 - H(P, Q)^2)^{2n}}{4} \geq \frac{1}{4} - \frac{nH(P, Q)^2}{2}.$$

\square

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

$$\mathcal{B}(P_0, \epsilon) = \{P : H(P, P_0) \leq \epsilon\} \quad \text{et} \quad \mathcal{B}(Q_0, \epsilon) = \{Q : H(Q, Q_0) \leq \epsilon\}$$

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 :

$$\cos(\omega) = \rho(P_0, Q_0) \quad \text{soit} \quad H(P_0, Q_0)^2 = 2 \sin(\omega/2)^2.$$

L'arc de cercle qui relie les points correspondants à P_0 et Q_0 sur la sphère unité 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

$$v_\beta = \frac{1}{\sin(\omega)} (\sin(\beta\omega)\sqrt{q_0} + \sin((1-\beta)\omega)\sqrt{p_0}).$$

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 $\mathcal{B}(P_0, H(P_0, P_1))$ et $\mathcal{B}(Q_0, H(Q_0, Q_1))$ 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

$$\prod_{i=1}^n \sqrt{\frac{q_1(X_i)}{p_1(X_i)}}.$$

Les performances du test.

THÉORÈME 6.14 *Soit P_0, Q_0 deux lois de probabilités sur \mathcal{X} , il existe une statistique $\psi : \mathcal{X} \rightarrow [0, \infty)$ telle que pour toute loi R sur \mathcal{X} ,*

$$\begin{aligned} \mathbb{E}_R \psi(X) &\leq 1 - \frac{5}{24} H(P_0, Q_0)^2 & \text{si } H(P_0, R) &\leq \frac{1}{4} H(P_0, Q_0) \\ \mathbb{E}_R \frac{1}{\psi(X)} &\leq 1 - \frac{5}{24} H(P_0, Q_0)^2 & \text{si } H(Q_0, R) &\leq \frac{1}{4} H(P_0, Q_0). \end{aligned}$$

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 .

$$\mathbb{E}_R \psi(X) = \int_{\mathcal{X}} r \sqrt{\frac{s}{t}} d\nu = \int_{\mathcal{X}} \sqrt{\frac{s}{t}} (\sqrt{r} - \sqrt{t})^2 d\nu + 2 \int_{\mathcal{X}} \sqrt{sr} d\nu - \int_{\mathcal{X}} \sqrt{st} d\nu.$$

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 entraîne

$$\mathbb{E}_R \psi(X) \leq 4 \cos(\omega/3) H(R, P_1)^2 + 2\rho(R, Q_1) - \rho(P_1, Q_1).$$

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 $\int u^2 d\nu + \theta^2 = 1$. Cette décomposition permet de réécrire

$$\rho(R, Q_1) = \theta \cos((\gamma - 2/3)\omega) \quad \rho(R, P_1) = \theta \cos((\gamma - 1/3)\omega).$$

Dans la suite on utilisera

$$H(P_1, P_0)^2 = 1 - \cos(\omega/3) = 2 \sin(\omega/6)^2 \quad H(P_0, Q_0)^2 = 2 \sin(\omega/2)^2$$

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 :

$$\begin{aligned} \mathbb{E}_R \psi(X) &\leq 4 \cos(\omega/3) (1 - \theta \cos((\gamma - 1/3)\omega)) + 2\theta \cos((\gamma - 2/3)\omega) - \cos(\omega/3) \\ &\leq 3 \cos(\omega/3) - 2\theta \cos(\gamma\omega) \\ &= 3\rho(P_1, P_0) - 2\rho(R, P_0) \\ &= 1 - 3H(P_1, P_0)^2 + 2H(R, P_0)^2 \\ &\leq 1 - \frac{1}{3} H(Q_0, P_0)^2 + 2H(R, P_0)^2. \end{aligned}$$

Si $H(R, P_0) \leq \frac{1}{4} H(Q_0, P_0)$,

$$\mathbb{E}_R \psi(X) \leq 1 - \frac{5}{24} H(Q_0, P_0)^2$$

Sous R^n ,

$$\begin{aligned} \mathbb{P} \left\{ \prod_{i=1}^n \psi(X_i) \geq 1 \right\} \\ &\leq (\mathbb{E}_R \psi(X))^n \\ &\leq \exp \left(-n \frac{5}{24} H(Q_0, P_0)^2 \right). \end{aligned}$$

□

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

$$T_n(X_1, \dots, X_n) = \begin{cases} 1 & \text{si } \prod_{i=1}^n \psi(X_i) \geq 1 \\ 0 & \text{sinon.} \end{cases}$$

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*

$$\max \left(\sup_{R \in \mathcal{B}(P_0, \frac{\epsilon}{4})} \mathbb{E}_R T_n, \sup_{R \in \mathcal{B}(Q_0, \frac{\epsilon}{4})} \mathbb{E}_R (1 - T_n) \right) \leq \exp \left(-n \frac{5\epsilon^2}{24} \right).$$

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. Après 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 INTRODUCTION

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 $\theta[0] = 1 - \sum_{\alpha=0}^p \theta[\alpha]$.

Un modèle multinomial est un modèle exponentiel minimal. Nous n'utilisons pas ici la paramétrisation canonique qui serait

$$\left(\log \frac{\theta[\alpha]}{1 - \sum_{\alpha>0} \theta[\alpha]} \right)_{\alpha \in \{1, \dots, p\}}.$$

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}_θ ,

$$\frac{1}{n} \text{cov}(N^n(\alpha), N^n(\beta)) = \begin{cases} \mathbf{p}_\theta(\alpha) - \mathbf{p}_\theta(\alpha)^2 & \text{si } \alpha = \beta \\ -\mathbf{p}_\theta(\alpha)\mathbf{p}_\theta(\beta) & \text{sinon.} \end{cases}$$

Nous noterons $\text{diag} \left(\frac{1}{\sqrt{n\mathbf{p}_\theta}} \right)$ 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

$$\frac{1}{n} \text{cov}(N^n) = \text{diag}(\mathbf{p}_\theta) - \begin{pmatrix} \mathbf{p}_\theta(0) \\ \vdots \\ \mathbf{p}_\theta(p) \end{pmatrix} \times (\mathbf{p}_\theta(0), \dots, \mathbf{p}_\theta(p)).$$

En multipliant le vecteur aléatoire $(N^n - n\mathbf{p}_\theta)/\sqrt{n}$ par $\text{diag}(1/\sqrt{\mathbf{p}_\theta})$ on obtient un vecteur aléatoire de matrice de covariance

$$\text{cov} \left(\left(\frac{N^n(\alpha) - n\mathbf{p}_\theta(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}} \right)_{\alpha \leq p} \right) = \text{Id}_{p+1} - \begin{pmatrix} \sqrt{\mathbf{p}_\theta(0)} \\ \vdots \\ \sqrt{\mathbf{p}_\theta(p)} \end{pmatrix} \times (\sqrt{\mathbf{p}_\theta(0)}, \dots, \sqrt{\mathbf{p}_\theta(p)}).$$

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}_θ de la fonction score évaluée en θ , elle est définie par :

$$I(\theta) \equiv \mathbb{E}_\theta [\nabla \log \mathbf{p}_\theta \times \nabla^t \log \mathbf{p}_\theta] = \left[\mathbb{E}_\theta \left[\frac{\partial_i p_\theta}{p_\theta} \frac{\partial_{i'} p_\theta}{p_\theta} \right] \right]_{i, i' \leq r} = \sum_{\alpha=0}^p \frac{\nabla \mathbf{p}_\theta(\alpha)}{\sqrt{\mathbf{p}_\theta(\alpha)}} \frac{\nabla \mathbf{p}_\theta(\alpha)^t}{\sqrt{\mathbf{p}_\theta(\alpha)}}.$$

La matrice d'information de Fisher en θ s'écrit

$$I(\theta) = \text{diag} \begin{pmatrix} \frac{1}{\theta[1]} \\ \vdots \\ \frac{1}{\theta[p]} \end{pmatrix} + \frac{1}{\theta[0]} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \times (1 \quad \dots \quad 1).$$

La relation entre la matrice d'information et la matrice de covariance est facilement vérifiée à l'aide de la formule de Sherman-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

$$I(\theta)^{-1} = \frac{1}{n} \text{cov} \left((N^n(\alpha))_{1 \leq \alpha \leq p} \right).$$

7.2 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 7.1 Test d'ADÉQUATION (AJUSTEMENT) à une hypothèse simple. Soit P une loi de probabilité sur un ensemble fini Ω et (x_1, \dots, x_n) un échantillon de n points de Ω .

L'échantillon (x_1, \dots, x_n) a-t-il été engendré par n tirages indépendants selon la loi P ?

EXEMPLE 7.2 En 1889, le docteur Geissler étudia les registres des hôpitaux de Saxe et nota le nombre de garçons dans 6115 familles de 12 enfants.

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

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 7.3 Test d'HOMOGENÉITÉ. Soit (x_1, \dots, x_n) et (y_1, \dots, 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) ?

EXEMPLE 7.4 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 7.5 (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 est plus confidentiel.

DÉFINITION 7.6 (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 $A, B \subseteq \Omega$?

7.3 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 5.4).

Nous allons reprendre l'approche suivie par Pearson au début du vingtième siècle.

DÉFINITION 7.7 (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

$$\sum_{\alpha=0}^p \frac{(N^n(\alpha) - n\mathbf{p}_\theta(\alpha))^2}{n\mathbf{p}_\theta(\alpha)} = \left\| \left(\frac{N^n(\alpha) - n\mathbf{p}_\theta(\alpha)}{\sqrt{n\mathbf{p}_\theta(\alpha)}} \right)_{\alpha \leq p} \right\|^2.$$

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

$$\chi^2(P | Q) := \int \frac{dQ}{d\nu} \left(\frac{dP}{dQ} - 1 \right)^2 d\nu = \int \frac{dP}{d\nu} \left(\frac{dP}{dQ} - 1 \right) d\nu.$$

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

$$n\chi^2(P_n | P).$$

Rappelons

$$\frac{1}{n} \text{cov}(N^n, N^n) = \text{diag}(\mathbf{p}_\theta) - \begin{pmatrix} \mathbf{p}_\theta(0) \\ \vdots \\ \mathbf{p}_\theta(p) \end{pmatrix} \times (\mathbf{p}_\theta(0), \dots, \mathbf{p}_\theta(p)).$$

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_p^2 .

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 au quantile d'ordre $1 - \alpha$ de la loi χ_p^2 .

EXEMPLE 7.8 (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 aux quantiles de la loi χ_{12}^2 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} .

7.4 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 $\alpha \in \{0, \dots, p\}$ 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 $\{0, \dots, p\}$. La divergence du χ^2 entre P_{θ_n} et P_θ se calcule simplement :

$$\chi^2(P_{\theta_n} | P_\theta) = \frac{1}{n} \sum_{\alpha=0}^p \frac{h_\alpha^2}{\theta_\alpha} = \frac{1}{n} h^t I(\theta) h.$$

Elle permet de quantifier la vitesse à laquelle P_{θ_n} s'approche de P_θ . On peut aussi calculer un équivalent de la distance de Hellinger entre P_{θ_n} et P_θ :

$$\lim_{n \rightarrow \infty} nH(P_{\theta_n}, P_\theta)^2 = \frac{1}{4} \lim_n n\chi^2(P_{\theta_n} | P_\theta).$$

La limite des lois images de $P_{\theta_n}^{\otimes n}$ par les statistiques de Pearson $n\chi^2(P_n | P_\theta)$ se détermine.

PROPOSITION 7.9 La limite des lois de la statistique de Pearson sous $(P_{\theta_n}^{\otimes n})$ est la loi du χ^2 à p degrés de liberté décentrée de paramètre de décentrage $h^t I(\theta) h = \sum_{\alpha=0}^p h_\alpha^2 / \theta_\alpha = n\chi^2(P_{\theta_n} | P_\theta)$ (loi 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 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 B.12) est justement valable pour les tableaux triangulaires. On vérifie facilement que les conditions sont remplies dans 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 $\tau > 0$

$$\max(P_{\theta}^{\otimes n}\{C(n, \mathbf{p}_{\theta}) \geq \tau\}, P_{\theta_n}^{\otimes n}\{C(n, \mathbf{p}_{\theta}) < \tau\}) \geq \frac{1}{4} - \frac{1}{2}nH(P_{\theta}, P_{\theta_n})^2.$$

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

$$\delta = \sum_{\alpha} h(\alpha)^2 / \mathbf{p}_{\theta}(\alpha) = h^t I(\theta) h$$

(on choisit bien sûr h de façon à avoir $\delta < 2$), la limite du membre gauche est

$$\max(\alpha, F_{p,\delta}(F_p^{\leftarrow}(1 - \alpha)))$$

alors que celle du membre droit est $\frac{1}{4} - \frac{\delta}{8}$. Pour $\delta = 1$, $p = 12$, le minimum du membre gauche est atteint autour de $\alpha \approx .45$ 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 α et pour toute suite $(\theta_n)_n$ d'alternatives contigue vérifiant $n\chi^2(P_{\theta_n} | P_{\theta}) \rightarrow \delta$, de niveau asymptotique supérieur à $F_{p,\delta}(F_p^{\leftarrow}(1 - \alpha))$?

7.5 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 (continument) différentiable de Θ_0 dans l'ensemble des fonctions de masse de probabilité sur $\{0, \dots, p\}$. On notera encore \mathbf{p}_{θ} 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, \dots, x_n a été tiré selon une loi P inconnue sur $\{0, \dots, p\}$, tester

$H_0 : \exists \theta \in \Theta_0$ tel que la loi P est définie par \mathbf{p}_{θ}

$H_1 : \nexists \theta \in \Theta_0, P$ est définie par \mathbf{p}_{θ} .

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) $\hat{\theta}$ pour ce modèle, on peut aussi procéder de la façon suivante.

DÉFINITION 7.10 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 :

$$C(n, \mathbf{p}_{\hat{\theta}}) \equiv \sum_{\alpha=0}^p \left(\frac{N^n(\alpha) - n\mathbf{p}_{\hat{\theta}}(\alpha)}{\sqrt{n\mathbf{p}_{\hat{\theta}}\{\alpha\}}} \right)^2$$

où $\hat{\theta}$ 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 7.11 (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
```

FIG. 7.1 : Les points représentent des données du docteur Geissler, les triangles représentent les effectifs attendus dans un échantillon de 6115 réalisations indépendantes de la loi binomiale de paramètres 12 et .52. 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 peut tester d'autres 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.

```
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 \mathbf{p} est défini à partir de $\hat{\theta}_n \approx .519$ calculé sur les données. Cette statistique n'est pas comparée au quantile de la loi χ_{11}^2 mais au quantile de la loi χ_{12}^2 . 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 $(N_\alpha - n\mathbf{p}_{\hat{\theta}_n}(\alpha))/\sqrt{n\mathbf{p}_{\hat{\theta}_n}}$ n'est pas approximativement normale et le raisonnement asymptotique qui justifie la comparaison de la statistique de Pearson aux quantiles de la loi χ_{11}^2 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_0^2 . La probabilité qu'une variable χ_0^2 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 $C(n, \mathbf{p}_{\hat{\theta}})$ sous l'hypothèse nulle lorsque Θ_0 est un modèle « régulier » et $\hat{\theta}$ 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 appelés ici.

HYPOTHÈSES TECHNIQUES : MODÈLES MULTINOMIAUX RÉGULIERS

On suppose que

- i) l'espace des paramètres $\Theta \subset \mathbb{R}^r$ est ouvert, que la fonction de $\Theta_0 \rightarrow]0, 1[^{p+1}$, $\theta \mapsto \mathbf{p}_\theta$ est deux fois continument différentiable en θ .
- ii) La matrice d'information de Fisher est encore définie par :

$$I(\theta) \equiv \mathbb{E}_\theta [\nabla \log \mathbf{p}_\theta \times \nabla^t \log \mathbf{p}_\theta] = \left[\mathbb{E}_\theta \left[\frac{\partial_i \mathbf{p}_\theta}{\mathbf{p}_\theta} \frac{\partial_{i'} \mathbf{p}_\theta}{\mathbf{p}_\theta} \right] \right]_{i, i' \leq r} = \sum_{\alpha=0}^p \frac{\nabla \mathbf{p}_\theta(\alpha)}{\sqrt{\mathbf{p}_\theta(\alpha)}} \frac{\nabla \mathbf{p}_\theta(\alpha)^t}{\sqrt{\mathbf{p}_\theta(\alpha)}}.$$

Elle est supposée inversible, donc de rang r , en tout $\theta \in \Theta_0$.

On peut définir une matrice $B(\theta)$ de dimensions $r \times (p+1)$ et factoriser la matrice d'information $I(\theta)$:

$$B(\theta) := \left(\frac{\partial_j \mathbf{p}_\theta(\alpha)}{\sqrt{\mathbf{p}_\theta(\alpha)}} \right)_{\substack{1 \leq j \leq r \\ 0 \leq \alpha \leq p}} \quad \text{et} \quad I(\theta) := B(\theta) \times B^t(\theta). \quad (7.1)$$

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

$$B(\theta) \times \begin{pmatrix} \sqrt{\mathbf{p}_\theta(0)} \\ \vdots \\ \sqrt{\mathbf{p}_\theta(p)} \end{pmatrix} = \left(\sum_{\alpha=0}^p \partial_j \mathbf{p}_\theta(\alpha) \right)_{1 \leq j \leq r} = 0.$$

REMARQUE 7.12 On a aussi :

$$I(\theta) = \left(-\mathbb{E}_\theta [\partial_{j,j'} \log \mathbf{p}_\theta] \right)_{j,j' \leq r} = -\mathbb{E} [\nabla^2 \log \mathbf{p}_\theta].$$

Les coefficients de la matrice d'information peuvent en effet s'écrire

$$-\sum_{\alpha=0}^p \frac{\partial_{j,j'} \mathbf{p}_\theta\{\alpha\} \times \mathbf{p}_\theta\{\alpha\} - \partial_j \mathbf{p}_\theta\{\alpha\} \times \partial_{j'} \mathbf{p}_\theta\{\alpha\}}{\mathbf{p}_\theta\{\alpha\}} \quad \text{pour } 1 \leq j, j' \leq r.$$

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

$$\Gamma(\theta) := \text{Id}_{p+1} - \sqrt{\mathbf{p}_\theta} \sqrt{\mathbf{p}_\theta}^t. \quad (7.2)$$

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 $\text{diag} \left(\frac{1}{\sqrt{n\mathbf{p}_\theta}} \right)$ 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

$$Z^n := \left(\frac{N^n(\alpha) - n\mathbf{p}_\theta\{\alpha\}}{\sqrt{n\mathbf{p}_\theta\{\alpha\}}} \right)_{\alpha \leq p} = \text{diag} \left(\frac{1}{\sqrt{n\mathbf{p}_\theta}} \right) \times (N^n - n\mathbf{p}_\theta).$$

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 $j \leq r$,

$$\partial_j \ell_n(\theta') = \sum_{l=1}^n \frac{\partial_j \mathbf{p}_{\theta'}\{x_l\}}{\mathbf{p}_{\theta'}\{x_l\}}.$$

La fonction score est définie comme le gradient de la log-vraisemblance : $\nabla \ell_n(\theta') = (\partial_j \ell_n(\theta'))_{j \leq r}$, c'est un vecteur colonne.

On définit la matrice symétrique $A(\theta)$ par

$$A(\theta) := \begin{pmatrix} \Gamma(\theta) & \vdots & B^\dagger(\theta)I^{-1}(\theta) \\ \dots & \dots & \dots \\ I^{-1}(\theta)B(\theta) & \vdots & I^{-1}(\theta) \end{pmatrix}$$

où $\Gamma(\theta)$ est défini par (7.2) et $B(\theta)$ et $I(\theta)$ sont définis par (7.1). On note la factorisation

$$A(\theta) = \begin{pmatrix} \text{Id}_{p+1} \\ \dots \\ I^{-1}(\theta)B(\theta) \end{pmatrix} \times \Gamma(\theta) \times \begin{pmatrix} \text{Id}_{p+1} & \vdots & B(\theta)^t I^{-1}(\theta) \end{pmatrix}$$

qui se vérifie en n'oubliant pas que $B(\theta)^t I^{-1}(\theta) = B(\theta)^t$.

Le résultat essentiel sur le comportement asymptotique de la statistique de Pearson (étendue) est donné dans le théorème suivant.

THÉORÈME 7.13 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}_θ

$$C(n, \mathbf{p}_{\hat{\theta}}) \rightsquigarrow X \quad \text{où} \quad X \sim \chi_{p-r}^2.$$

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 7.14 Sous les conditions du théorème 7.13, pour tout $\theta \in \Theta_0$, sous \mathbf{p}_θ ,

$$I(\theta)\sqrt{n}(\hat{\theta}_n - \theta) - \frac{1}{\sqrt{n}}\nabla\ell_n(\theta) \xrightarrow{\mathbf{P}_\theta} 0. \quad (7.3)$$

REMARQUE 7.15 Ce couplage asymptotique entre $I(\theta)^{-1}\frac{1}{\sqrt{n}}\nabla\ell_n(\theta)$ et $\sqrt{n}(\hat{\theta}_n - \theta)$ n'est pas une propriété spécifique des modèles multinomiaux réguliers mais des modèles dits réguliers en général. Il peut être établi dans des conditions beaucoup moins exigeantes que celles énoncées ici.

LEMME 7.16 Sous les conditions du théorème 7.13, pour tout $\theta \in \Theta_0$, sous \mathbf{p}_θ , la suite de vecteurs aléatoires de dimension $p + 1 + r$

$$X^n := \begin{pmatrix} Z_n \\ \vdots \\ \sqrt{n}(\hat{\theta} - \theta) \end{pmatrix}$$

converge en loi vers un vecteur gaussien centré de matrice de covariance $A(\theta)$.

La démonstration du Lemme 7.14 commence par vérifier que le score (le gradient de la log-vraisemblance $\ell_n(\cdot)$) en θ est une transformation linéaire de Z^n .

PREUVE. (LEMME 7.14) 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 .

$$\begin{aligned} \frac{1}{\sqrt{n}}\nabla\ell_n(\theta) &= \frac{1}{\sqrt{n}} \sum_{\alpha=0}^p N^n(\alpha) \frac{\nabla\mathbf{p}_\theta\{\alpha\}}{\mathbf{p}_\theta\{\alpha\}} \\ &\quad \text{comme } 0 = \sum_{\alpha=0}^p \nabla\mathbf{p}_\theta\{\alpha\}, \\ &= \sum_{\alpha=0}^p \frac{1}{\sqrt{n}} \frac{(N^n(\alpha) - n\mathbf{p}_\theta\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \times \frac{\nabla\mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \\ &= B(\theta) \times Z^n. \end{aligned}$$

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 $i \leq r$,

$$0 = \partial_i \ell_n(\widehat{\theta}_n) = \partial_i \ell_n(\theta) + \sum_{j=1}^r (\widehat{\theta}_n[j] - \theta[j]) \int_0^1 \partial_{j,i} \ell_n(\theta + u(\widehat{\theta}_n - \theta)) du$$

soit en divisant les 2 membres par $1/\sqrt{n}$:

$$0 = \sum_{\alpha=0}^p \frac{\partial_i \mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \frac{(N^n(\alpha) - n\mathbf{p}_\theta\{\alpha\})}{\sqrt{n\mathbf{p}_\theta\{\alpha\}}} + \sum_{j=1}^r \sqrt{n}(\widehat{\theta}_n[j] - \theta[j]) \frac{1}{n} \int_0^1 \partial_{j,i} \ell_n(\theta + u(\widehat{\theta}_n - \theta)) du.$$

Vérifions maintenant que la matrice aléatoire de terme général

$$\left(\frac{1}{n} \int_0^1 \partial_{i,j} \ell_n(\theta + u(\widehat{\theta}_n - \theta)) du \right)$$

converge en probabilité vers $-I(\theta)$ dont le terme général est aussi la limite en probabilité de

$$\frac{1}{n} \partial_{i,j} \ell_n(\theta) = \frac{1}{n} \int_0^1 \partial_{i,j} \ell_n(\theta) du.$$

Nous voulons établir

$$\frac{1}{n} \int_0^1 \left[\partial_{i,j} \ell_n(\theta + u(\widehat{\theta}_n - \theta)) - \partial_{i,j} \ell_n(\theta) \right] du \xrightarrow{P} 0.$$

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

$$\left| \frac{1}{n} \int_0^1 \left[\partial_{i,j} \ell_n(\theta + u(\widehat{\theta}_n - \theta)) - \partial_{i,j} \ell_n(\theta) \right] du \right| \leq w(\|\theta - \widehat{\theta}_n\|).$$

L'hypothèse de consistance de l'estimateur $\widehat{\theta}_n$ permet alors d'établir la convergence en probabilité recherchée.

Par ailleurs la loi des grands nombres permet de conclure que

$$\frac{1}{n} \partial_{i,j} \ell_n(\theta) \xrightarrow{P} \mathbb{E}_\theta [\partial_{i,j} \log \mathbf{p}_\theta].$$

□

PREUVE. (LEMME 7.16) D'après le lemme 7.14, 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

$$\begin{pmatrix} \text{Id}_{p+1} \\ \vdots \\ I^{-1}(\theta)B(\theta) \end{pmatrix}.$$

La limite est un vecteur gaussien centré. La covariance de cette limite est donnée par

$$\begin{pmatrix} \text{Id}_{p+1} \\ \dots \\ I^{-1}(\theta)B(\theta) \end{pmatrix} \times \Gamma(\theta) \times \begin{pmatrix} \text{Id}_{p+1} & \vdots & B^t(\theta)I^{-1}(\theta) \end{pmatrix} = A(\theta).$$

□

Nous pouvons passer à la démonstration du théorème.

PREUVE. (THÉORÈME 7.13) 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 7.16, on en conclut que

$$\sqrt{n} \left(\frac{N^n(\alpha) - \mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}}, \frac{(\mathbf{p}_\theta\{\alpha\} - \mathbf{p}_{\hat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \right)_{\alpha \leq p}$$

converge en loi vers un vecteur Gaussien de matrice de covariance

$$\begin{pmatrix} \Gamma(\theta) & \vdots & B^\dagger(\theta)I^{-1}(\theta)B(\theta) \\ \cdots & & \cdots \\ B^\dagger(\theta)I^{-1}(\theta)B(\theta) & \vdots & B^\dagger(\theta)I^{-1}(\theta)B(\theta) \end{pmatrix} \\ = \begin{pmatrix} \text{Id}_k & \vdots & B^\dagger(\theta) \end{pmatrix} \begin{pmatrix} \Gamma(\theta) & \vdots & B^\dagger(\theta)I^{-1}(\theta) \\ \cdots & & \cdots \\ I^{-1}(\theta)B(\theta) & \vdots & I^{-1}(\theta) \end{pmatrix} \begin{pmatrix} \text{Id}_k \\ \cdots \\ B(\theta) \end{pmatrix}.$$

On déduit que la suite de vecteurs aléatoires $(Y^n)_{n \in \mathbb{N}}$

$$Y^n := \sqrt{n} \left(\frac{N^n(\alpha) - \mathbf{p}_\theta\{\alpha\}}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} - \frac{(\mathbf{p}_\theta\{\alpha\} - \mathbf{p}_{\hat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \right)_{\alpha \leq p}$$

converge en loi vers un vecteur Gaussien de matrice de covariance

$$\begin{aligned} & \Gamma(\theta) - B^\dagger(\theta)I^{-1}(\theta)B(\theta) \\ &= \text{Id} - \begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_k} \end{pmatrix} \begin{pmatrix} \sqrt{p_1} & \sqrt{p_2} & \cdots & \sqrt{p_k} \end{pmatrix} - B^\dagger(\theta) (B(\theta) \times B^\dagger(\theta))^{-1} B(\theta). \end{aligned}$$

La matrice symétrique

$$D(\theta) \equiv B^\dagger(\theta) (B(\theta) \times B^\dagger(\theta))^{-1} B(\theta) = B^\dagger(\theta)I^{-1}(\theta)B(\theta)$$

est idempotente. C'est une matrice de projection orthogonale de rang r .

De plus, le vecteur $(\sqrt{p_\theta\{1\}} \ \sqrt{p_\theta\{2\}} \ \cdots \ \sqrt{p_\theta\{k\}})^\dagger$ appartient à son noyau.

Sous \mathbf{p}_θ , les limites en loi de

$$C(n, \mathbf{p}_{\hat{\theta}}) \equiv \sum_{\alpha=0}^p \left(\frac{\sqrt{n}(N^n(\alpha)/n - \mathbf{p}_{\hat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_{\hat{\theta}_n}\{\alpha\}}} \right)^2$$

et de

$$\sum_{\alpha=0}^p \left(\frac{\sqrt{n}(N^n(\alpha)/n - \mathbf{p}_{\hat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \right)^2$$

sont identiques. La dernière quantité s'écrit

$$\sum_{\alpha=0}^p \left(\frac{\sqrt{n}(N^n(\alpha)/n - \mathbf{p}_\theta\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} - \frac{\sqrt{n}(\mathbf{p}_\theta\{\alpha\} - \mathbf{p}_{\hat{\theta}_n}\{\alpha\})}{\sqrt{\mathbf{p}_\theta\{\alpha\}}} \right)^2,$$

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 pondérée 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 $r + 1$).

□

7.6 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 7.13 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

$$\mathbf{p}_\theta\{\langle \alpha', \alpha'' \rangle\} = \begin{cases} \theta'[\alpha'] \times \theta''[\alpha''] & \text{si } 0 < \alpha' \leq k' \text{ et } 0 < \alpha'' \leq k'' \\ \theta'[\alpha'] \times \left(1 - \sum_{\beta''=1}^{k''} \theta''[\beta'']\right) & \text{si } 0 < \alpha' \leq k' \text{ et } \alpha'' = 0 \\ \left(1 - \sum_{\beta'=1}^{k'} \theta'[\beta']\right) \times \theta''[\alpha''] & \text{si } \alpha' = 0 \text{ et } 0 < \alpha'' \leq k'' \\ \left(1 - \sum_{\beta'=1}^{k'} \theta'[\beta']\right) \times \left(1 - \sum_{\beta''=1}^{k''} \theta''[\beta'']\right) & \text{si } \alpha' = 0 \text{ et } \alpha'' = 0. \end{cases}$$

L'estimateur au maximum de vraisemblance est simple : si 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 7.13 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.

7.7 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 7.14.

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 $a < b : \int_{[a,b]} f(x)d(x) = F(b) - F(a) = P[a, b]$).

Dans ce type de situation, on dispose d'un « résumé » de l'échantillon $S_n = (X_1, X_2, \dots, X_n)$, l'échantillon trié en ordre croissant : $(X_{(1)}, \dots, X_{(n)})$.

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, \dots, x_n) un échantillon de n points de \mathbb{R} .

- H_0 : L'échantillon (x_1, \dots, x_n) a été engendré par n tirages indépendants selon la loi P de densité f ?
- H_1 : L'échantillon (x_1, \dots, 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, \dots, 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, \dots, 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).

DÉFINITION 8.3 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.

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 $M > 0$

$$P' \{D_n > M\} \rightarrow_n 1.$$

Le test est donc consistant.

- iv) Pour terminer nous donnerons quelques résultats qui suggèrerons (mais ne montrerons 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).

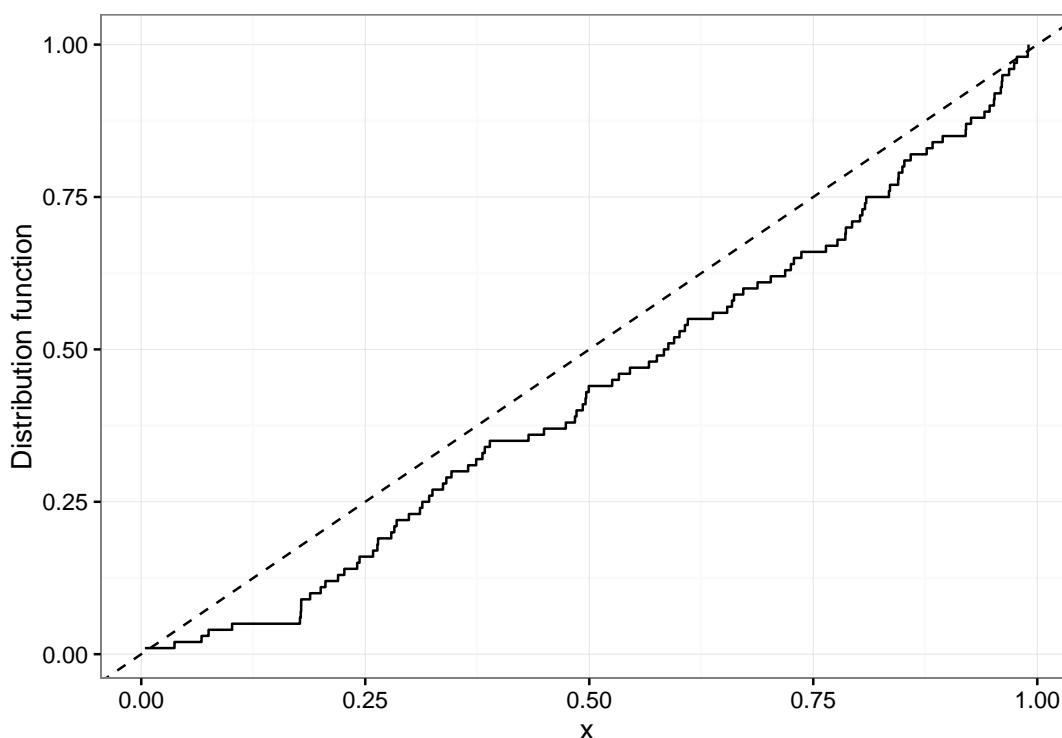


FIG. 8.1 : Fonction de répartition empirique F_n d'un échantillon de 100 points collectés indépendamment selon des tirages uniformes sur $[0, 1]$. La courbe en tirets est la fonction de répartition de la loi uniforme.

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 uniformément 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 8.4 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 8.5 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 , $F \circ F^{-1}(p) = p$.

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 1). 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 $F \circ F^{-1}(p) = p$. \square

A partir des propriétés de la fonction quantile, il est aisé de déduire le résultat suivant.

COROLLAIRE 8.6 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 $F(X_1), \dots, F(X_n)$ est distribué de la même façon que U_1, \dots, U_n où les $U_i \sim \text{Uniforme}[0, 1]$.

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 $x_{(1)}, \dots, x_{(n)}$ désigne la version triée en ordre croissant de (x_1, \dots, x_n) , les variables aléatoires $X_{(i)}$ sont appelées les statistiques d'ordre de l'échantillon).

LEMME 8.7 *Le supremum dans la définition de D_n peut être calculé efficacement à partir de l'échantillon trié en ordre croissant $(x_{(1)}, \dots, x_{(n)})$:*

$$D_n = \sqrt{n} \max_{0 \leq i < n} \max \left(\left| \frac{i}{n} - F(x_{(i)}) \right|, \left| \frac{i}{n} - F(x_{(i+1)}) \right| \right),$$

avec la convention $x_{(0)} = -\infty$.

PREUVE. Considérons un réel t compris entre $x_{(i)}$ et $x_{(i+1)}$. On a

$$F_n(t) = \frac{i}{n} = F_n(x_{(i)}) < F_n(x_{(i+1)}) = \frac{i+1}{n},$$

et

$$F(x_{(i)}) \leq F(t) \leq F(x_{(i+1)}).$$

On a donc

$$F_n(x_{(i)}) - F(x_{(i+1)}) \leq F_n(t) - F(t) \leq F_n(x_{(i)}) - F(x_{(i)})$$

et

$$F(x_{(i)}) - F_n(x_{(i)}) \leq F(t) - F_n(t) \leq F(x_{(i+1)}) - F_n(x_{(i)})$$

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 8.8 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 8.9 *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 $U_{(1)} \leq U_{(2)} \leq \dots \leq U_{(n)}$ 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 8.10 [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:*

$$\forall \varepsilon > 0, \quad P \{ \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| > \varepsilon \} \rightarrow_n 0.$$

REMARQUE 8.11 Pour chaque valeur particulière de x , on peut écrire

$$F_n(t) - F(t) = \frac{1}{n} \sum_{i \leq n} (\mathbf{1}_{x_i \leq t} - \mathbb{E}[\mathbf{1}_{X \leq t}]).$$

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

$$\forall \varepsilon > 0, \forall t \in \mathbb{R}, \quad P \{ |F_n(t) - F(t)| > \varepsilon \} \rightarrow_n 0.$$

Le théorème de Glivenko-Cantelli constitue une loi des grands nombres *uniforme*.

PREUVE. Soit ε un réel strictement positif. Pour $k \in \mathbb{N}$, $1 \leq k \leq \lfloor \frac{1}{\varepsilon} \rfloor$, on définit

$$y_k^\varepsilon \equiv F^{-1}(k\varepsilon).$$

Etant donné un échantillon x_1, \dots, x_n , et la fonction de répartition empirique F_n associée, si $y_k^\varepsilon \leq x \leq y_{k+1}^\varepsilon$, alors

$$\begin{aligned} F(y_k^\varepsilon) &\leq F(x) \leq F(y_{k+1}^\varepsilon) \\ F_n(y_k^\varepsilon) &\leq F_n(x) \leq F_n(y_{k+1}^\varepsilon). \end{aligned}$$

Donc

$$F(y_k^\varepsilon) - F_n(y_{k+1}^\varepsilon) \leq F(x) - F_n(x) \leq F(y_{k+1}^\varepsilon) - F_n(y_k^\varepsilon)$$

ce qui entraîne

$$|F(x) - F_n(x)| \leq \max(|F(y_k^\varepsilon) - F_n(y_k^\varepsilon)|, |F(y_{k+1}^\varepsilon) - F_n(y_{k+1}^\varepsilon)|) + \varepsilon$$

soit

$$\sup_{y_1^\varepsilon \leq x \leq y_{\lfloor 1/\varepsilon \rfloor}^\varepsilon} |F(x) - F_n(x)| \leq \max_{1 \leq k \leq \lfloor 1/\varepsilon \rfloor} \max(|F(y_k^\varepsilon) - F_n(y_k^\varepsilon)|) + \varepsilon.$$

Par ailleurs, pour $x \leq y_1^\varepsilon$,

$$|F(x) - F_n(x)| \leq |F_n(y_1^\varepsilon) - F(y_1^\varepsilon)| + \varepsilon$$

et de même pour $x \geq y_{\lfloor 1/\epsilon \rfloor}^\epsilon$

$$|F(x) - F_n(x)| \leq \left| F_n(y_{\lfloor 1/\epsilon \rfloor}^\epsilon) - F(y_{\lfloor 1/\epsilon \rfloor}^\epsilon) \right| + \epsilon.$$

En récapitulant

$$\frac{D_n}{\sqrt{n}} \leq \max_{1 \leq k \leq \lfloor 1/\epsilon \rfloor} \max (|F(y_k^\epsilon) - F_n(y_k^\epsilon)|) + \epsilon.$$

On peut maintenant observer que $F(y_k^\epsilon) - F_n(y_k^\epsilon)$ est une 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

$$P^n \{ |F(y_k^\epsilon) - F_n(y_k^\epsilon)| \geq u \} \leq \frac{k\epsilon(1 - k\epsilon)}{nu^2} \leq \frac{1}{4nu^2}.$$

Et

$$P^n \left\{ \max_{k \leq \lfloor 1/\epsilon \rfloor} |F(y_k^\epsilon) - F_n(y_k^\epsilon)| \geq u \right\} \leq \frac{\lfloor 1/\epsilon \rfloor}{4nu^2}.$$

En choisissant $u = \epsilon$, on peut en déduire que, lorsque n tend vers l'infini :

$$P^n \left\{ \left| \frac{D_n}{\sqrt{n}} \right| \geq 2\epsilon \right\} \rightarrow 0.$$

Comme on peut choisir ϵ arbitrairement petit, la convergence en probabilité est démontrée. □

REMARQUE 8.12 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 8.13 Si (X_1, \dots, 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

$$\sqrt{n} \sup_t |F_n(t) - F(t)| \geq \sqrt{n} (F_n(x) - F(x)) = \sqrt{n} (F_n(x) - F'(x)) + \sqrt{n} (F'(x) - F(x)),$$

on a donc

$$D_n \geq \sqrt{n} (F_n(x) - F'(x)) + \sqrt{n}\epsilon.$$

La loi des grands nombres, indique que sous P' , $|F'(x) - F_n(x)| \xrightarrow{P'} 0$. Donc

$$P' \{ F_n(x) - F'(x) < -\epsilon/2 \} \rightarrow_n 0,$$

soit pour tout $M > 0$

$$P' \left\{ \sqrt{n} \sup_t |F_n(t) - F(t)| > M \right\} \rightarrow 1.$$

Ceci implique que quelque soit le choix d'un seuil $\tau > 0$, le test qui rejette l'hypothèse nulle lorsque $D_n \geq \tau$, la probabilité de rejeter lorsque l'hypothèse nulle n'est pas vérifiée tend vers 1 lorsque la taille de l'échantillon tend vers l'infini. □

8.5 INÉGALITÉS POUR LA STATISTIQUE DE KOLMOGOROV-SMIRNOV,

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 8.14 *Sous l'hypothèse nulle, la statistique de Kolmogorov-Smirnov converge en loi. Pour $x > 0$,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \{D_n \geq x\} = 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 x^2} .$$

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 8.15 *Pour tout entier $n > 1$, tout $\epsilon > 0$,*

$$\mathbb{P} \{D_n \geq \epsilon\} \leq 2e^{-2\epsilon^2} . \tag{8.1}$$

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

$$\mathbb{P} \{ \sqrt{n} |F_n(x) - F(x)| \geq \epsilon \} \leq 2e^{-2\epsilon^2} .$$

La proposition 8.15 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 8.16 *Pour tout entier $n > 0$, tout $\epsilon > 0$*

$$\mathbb{P} \{D_n \geq \epsilon\} \leq 4e^{-\frac{\epsilon^2}{8}} .$$

La preuve de la Proposition 8.16 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 $1/2$.

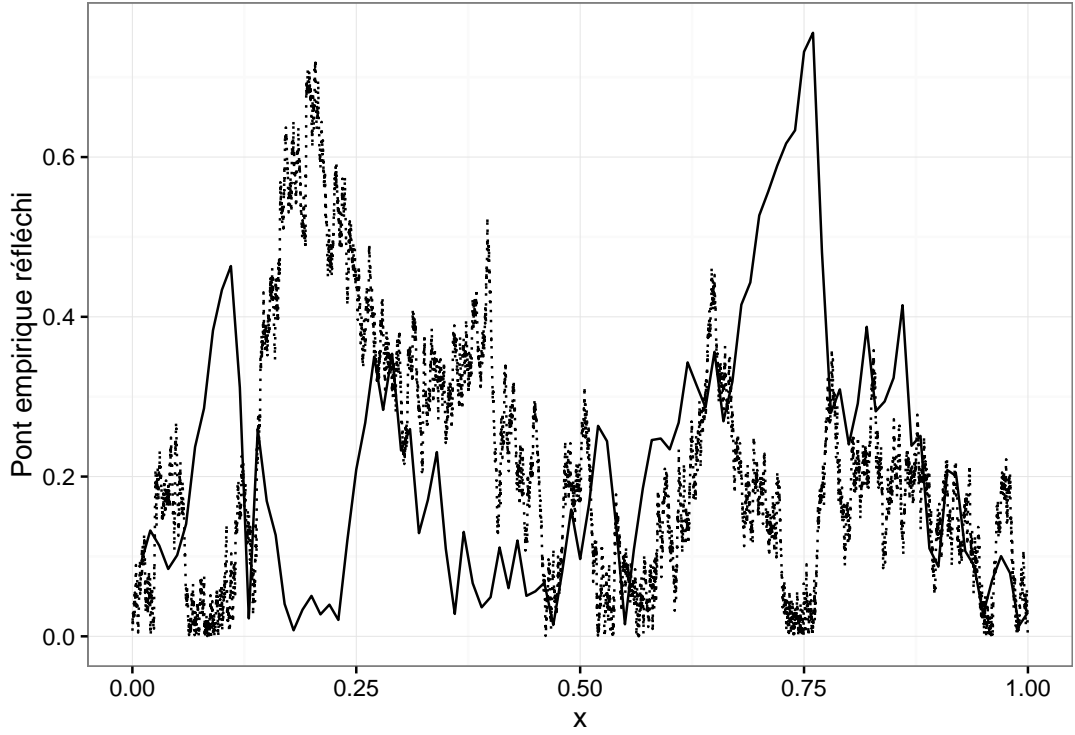


FIG. 8.2 : Graphe de la fonction $s \mapsto \sqrt{n}|F_n(s) - F(s)|$ pour un échantillon de taille 100 de la loi uniforme. Le processus illustré est souvent appelé le pont empirique. La limite en loi du pont empirique est la loi du pont brownien (mouvement brownien conditionné à valoir 0 au temps 1.) La ligne pointillée est calculée à partir d'un échantillon de taille 10000.

LEMME 8.17 Soit X_1, \dots, X_n une suite de vecteurs aléatoires indépendants indexés par \mathcal{T} . Soient $\epsilon_1, \dots, \epsilon_n$ une suite de variables de Rademacher indépendantes entre elles et indépendantes des X_1, \dots, X_n . Si $\Psi : \mathbb{R} \rightarrow \mathbb{R}$ est convexe et croissante, alors

$$\mathbb{E} \left[\Psi \left(\sup_{t \in \mathcal{T}} \left| \sum_{i=1}^n X_{i,t} - \mathbb{E}X_{i,t} \right| \right) \right] \leq \mathbb{E} \left[\Psi \left(2 \sup_{t \in \mathcal{T}} \left| \sum_{i=1}^n \epsilon_i X_{i,t} \right| \right) \right].$$

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

$$\mathbb{E} \left[\Psi \left(2 \sup_{t \in \mathcal{T}} \left| \sum_{i=1}^n \epsilon_i X_{i,t} \right| \right) \mid X_1, \dots, X_n \right]$$

où l'espérance conditionnelle n'est que l'espérance calculée par rapport aux variables de Rademacher (c'est une simple somme).

PREUVE. 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 $X'_{i,s} = \mathbb{1}_{Y'_i \leq s}$. 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 $X_1 - X'_1$ est *symétrique* : elle a même loi que son opposé. Dans la suite les espérances sont prises par rapport à $X_1, \dots, X_n, X'_1, \dots, X'_n, \epsilon_1, \dots, \epsilon_n$.

$$\begin{aligned} \mathbb{E} [e^{\lambda Z}] &= \mathbb{E} \left[\sup_{s \in [0,1]} e^{\lambda |\sum_{i=1}^n X_{i,s} - \mathbb{E} X'_{i,s}|} \right] \\ &\leq \mathbb{E} \left[\sup_{s \in [0,1]} e^{\lambda |\sum_{i=1}^n X_{i,s} - X'_{i,s}|} \right] && \text{(Inégalité de Jensen)} \\ &= \mathbb{E} \left[\sup_{s \in [0,1]} e^{\lambda |\sum_{i=1}^n \epsilon_i (X_{i,s} - X'_{i,s})|} \right] && (X_i - X'_i \text{ est symétrique)} \\ &\leq \mathbb{E} \left[\sup_{s \in [0,1]} \frac{1}{2} \left(e^{2\lambda |\sum_{i=1}^n \epsilon_i X_{i,s}|} + e^{2\lambda |-\sum_{i=1}^n \epsilon_i X'_{i,s}|} \right) \right] && \text{(Inégalité de Jensen)} \\ &\leq \mathbb{E} \left[\sup_{s \in [0,1]} e^{2\lambda |\sum_{i=1}^n \epsilon_i X_{i,s}|} \right]. \end{aligned}$$

Nous venons d'établir une *inégalité de symétrisation* :

$$\mathbb{E} \left[\sup_{s \in [0,1]} e^{\lambda |\sum_{i=1}^n X_{i,s} - \mathbb{E} X'_{i,s}|} \right] \leq \mathbb{E} \left[\sup_{s \in [0,1]} e^{2\lambda |\sum_{i=1}^n \epsilon_i X_{i,s}|} \right].$$

□

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. Nous allons maintenant exploiter le fait que nous avons affaire à des indicatrices de demi-droites.

PREUVE. Presque sûrement, les Y_i sont deux à deux distincts. Pour une réalisation des ϵ_i et des Y_i , on a

$$\begin{aligned} \sup_{s \in [0,1]} e^{2\lambda |\sum_{i=1}^n \epsilon_i X_{i,s}|} &= \sup_{s \in [0,1]} e^{2\lambda |\sum_{i=1}^n \epsilon_i \mathbb{1}_{Y_i \leq s}|} \\ &= \max_{k \leq n} e^{2\lambda |\sum_{i=1}^n \epsilon_i \mathbb{1}_{Y_i \leq Y_{(k)}}|} \\ &= \max_{k \leq n} e^{2\lambda |\sum_{i=1}^k \epsilon_i|}. \end{aligned}$$

Donc

$$\mathbb{E} \left[\sup_{s \in [0,1]} e^{2\lambda |\sum_{i=1}^n \epsilon_i X_{i,s}|} \mid Y_1, \dots, Y_n \right] = \mathbb{E} \left[\max_{k \leq n} e^{2\lambda |\sum_{i \leq k} \epsilon_i|} \right]. \quad (8.2)$$

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.

$$\mathbb{E} \left[\max_{k \leq n} e^{2\lambda |\sum_{i \leq k} \epsilon_i|} \right] = \int_0^\infty \mathbb{P} \left\{ \max_{k \leq n} e^{2\lambda |\sum_{i \leq k} \epsilon_i|} \geq a \right\} da.$$

On note

$$A_k = \left\{ e^{2\lambda |\sum_{i \leq k} \epsilon_i|} \geq a, \text{ et } e^{2\lambda |\sum_{i \leq j} \epsilon_i|} < a \text{ pour } j < k \right\}.$$

Les A_k forment une famille d'événements deux à deux disjoints. D'une part,

$$\left\{ \max_{k \leq n} e^{2\lambda |\sum_{i \leq k} \epsilon_i|} \geq a \right\} = \bigcup_{k \leq n} A_k$$

et d'autre part, pour chaque $k \in \{1, \dots, n\}$,

$$\mathbb{P} \left\{ e^{2\lambda |\sum_{i \leq n} \epsilon_i|} \geq a \mid A_k \right\} \geq \frac{1}{2},$$

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 :

$$\begin{aligned} \mathbb{P} \left\{ \max_{k \leq n} e^{2\lambda |\sum_{i \leq k} \epsilon_i|} \geq a \right\} &\leq 2 \sum_{k \leq n} \mathbb{P}\{A_k\} \mathbb{P} \left\{ e^{2\lambda |\sum_{i \leq n} \epsilon_i|} \geq a \mid A_k \right\} \\ &= 2\mathbb{P} \left\{ e^{2\lambda |\sum_{i \leq n} \epsilon_i|} \geq a \right\} \end{aligned}$$

D'où

$$\begin{aligned} \mathbb{E} \left[\max_{k \leq n} e^{2\lambda |\sum_{i \leq k} \epsilon_i|} \right] &\leq 2\mathbb{E} \left[e^{2\lambda |\sum_{i \leq n} \epsilon_i|} \right] \\ &\leq 2\mathbb{E} \left[e^{2\lambda \sum_{i \leq n} \epsilon_i} + e^{-2\lambda \sum_{i \leq n} \epsilon_i} \right] \\ &\leq 4\mathbb{E} \left[e^{2\lambda \sum_{i \leq n} \epsilon_i} \right] && \text{(symétrie)} \\ &\leq 4e^{n \frac{(2\lambda)^2}{2}} && \text{(Lemme de Hoeffding).} \end{aligned}$$

On aboutit à

$$\mathbb{E} [e^{\lambda Z}] \leq 4e^{2n\lambda^2}$$

et la proposition 8.16 s'obtient en choisissant $\lambda = \epsilon/\sqrt{n}$ 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.

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| F_n(x) - \left(1 - e^{-x/\hat{\sigma}} \right) \right|$$

où $\hat{\sigma} = \bar{X}_n$. On peut vérifier que sous l'hypothèse nulle, la loi de cette statistique ne dépend pas du paramètre σ (la loi de (X_i/\bar{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

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\frac{X_i}{\bar{X}_n} \leq x} - (1 - e^{-x}) \right|.$$

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 8.18 Soient Y_1, \dots, 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, \dots, 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 8.19 Soient Y_1, \dots, 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 $Z_j = \sum_{k=1}^j Y_k / S_n$ 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, \dots, Z_{n-1} et à comparer cette statistique aux quantiles de la loi de la statistique de Kolmogorov-Smirnov pour les échantillons de taille $n - 1$.

Cette démarche peut sembler inattaquable. Elle présente pourtant quelques défauts.

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 aléatoires $(X'_i)_i$ 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 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 , $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, I(\theta)^{-1})$.

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 9.1 (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 $\dot{p} : \Theta \times \mathcal{X} \rightarrow \mathbb{R}^d$ (avec $\|\dot{p}(\theta, \cdot)\| \in L_2(P_\theta)$ pour tout $\theta \in \Theta$) telle que

$$\int_{\mathcal{X}} \left(\sqrt{p(\theta', x)} - \sqrt{p(\theta, x)} - \frac{\sqrt{p(\theta, x)}}{2}(\theta' - \theta)^t \dot{p}(\theta, x) \right)^2 d\nu(x) = o(\|\theta - \theta'\|^2) .$$

au voisinage de θ La fonction $\dot{p}(\theta, \cdot)$ est appelée fonction score. L'information de Fisher du modèle en θ est

$$\mathbb{E}_\theta [\dot{p}(\theta, X)\dot{p}(\theta, X)^t] .$$

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 9.2 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, $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, 1)$.

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 9.3 On se place dans un 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 $P_\theta^{\mathbb{N}}$

$$\sqrt{n}(T_n - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2(\theta))$$

et aussi qu'il existe $\theta_0, \delta > 0$ tels que

$$\sup_n \mathbb{E}_{\theta_0} \left[(\sqrt{n}(T_n - \theta_0))^{2+\delta} \right] \leq \infty.$$

Hodge étudie alors la suite d'estimateurs $(\tilde{T}_n)_n$ définie par

$$\tilde{T}_n := \begin{cases} T_n, & \text{si } |T_n - \theta_0| > n^{-1/4} \\ 0 & \text{sinon.} \end{cases}$$

On vérifie alors que pour $\theta \neq \theta_0$, sous $P_\theta^{\mathbb{N}}$

$$\sqrt{n}(\tilde{T}_n - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2(\theta))$$

alors que sous $P_{\theta_0}^{\mathbb{N}}$,

$$\sqrt{n}(\tilde{T}_n - \theta_0) \rightsquigarrow \delta_0.$$

Autrement $(\tilde{T}_n)_n$ 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 \tilde{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.

9.2 LE PHÉNOMÈNE DE STEIN

On se place dans le modèle de la localisation Gaussienne : $Y \sim \mathcal{N}(\theta, \sigma^2 \text{Id})$ avec $\theta \in \mathbb{R}^d$ le paramètre à estimer. L'information de Fisher est constante $I(\theta) = \frac{1}{\sigma^2} \text{Id}$. On s'intéresse au cas où $d \geq 3$.

On définit l'estimateur par contraction (appelé estimateur de James-Stein) :

$$\hat{\theta} := \left(1 - \sigma^2 \frac{d-2}{\|Y\|^2} \right) Y.$$

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 9.4 Pour tout $\theta \in \mathbb{R}^d$,

$$\mathbb{E}_\theta \left[\|\hat{\theta} - \theta\|^2 \right] \leq \mathbb{E}_\theta \left[\|Y - \theta\|^2 \right] = d\sigma^2.$$

Pour $\theta = 0$,

$$\mathbb{E}_\theta \left[\|\hat{\theta} - \theta\|^2 \right] = 2\sigma^2.$$

Pour $\theta \neq 0$,

$$\lim_{\sigma \rightarrow 0} \frac{\mathbb{E}_\theta \left[\|\hat{\theta} - \theta\|^2 \right]}{\mathbb{E}_\theta \left[\|Y - \theta\|^2 \right]} = 1.$$

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 $g : \mathbb{R} \rightarrow \mathbb{R}$ est absolument continue sur $[a, b]$ si et seulement si il existe une fonction g' intégrable de $[a, b]$ dans \mathbb{R} telle que $g(x) - g(a) = \int_{[a,x]} g'(y)dy$ (la fonction de répartition d'une loi absolument continue par rapport à la mesure de Lebesgue est une fonction absolument continue).

Une fonction $f : \mathbb{R}^d \rightarrow \mathbb{R}$, est dite presque partout absolument continue en chaque coordonnée si et seulement si pour tout i , pour presque tout $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, la fonction

$$x \mapsto f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n)$$

est absolument continue. On note en abrégé $\partial_i f$ une fonction qui vérifie

$$\begin{aligned} f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, a, x_{i+1}, \dots, x_n) \\ = \int_a^x \partial_i f(x_1, \dots, x_{i-1}, u, x_{i+1}, \dots, x_n) du \end{aligned}$$

(elle est implicitement paramétrée par $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$).

PROPOSITION 9.5 (IDENTITÉ DE STEIN) Soit $Y \sim \mathcal{N}(\theta, \sigma^2 \text{Id}_d)$ avec $\theta \in \mathbb{R}^d$. Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$, presque partout absolument continue en chaque coordonnée, vérifiant pour chaque $i \leq d$

$$\mathbb{E} [|\partial_i f|] := \mathbb{E} [|\partial_i f(Y)|] < \infty.$$

Alors pour chaque $i \leq d$,

$$\mathbb{E} [(\theta_i - Y_i) f(Y)] = -\sigma^2 \mathbb{E} [\partial_i f].$$

L'identité de Stein caractérise les gaussiennes, voir lemmes 2.1 et 2.3.

PREUVE. L'identité de Stein peut se concevoir comme un corollaire de l'identité de covariance gaussienne (voir Section 2.10). Si $X \sim \mathcal{N}(0, \text{Id}_d)$, $\sigma X + \theta \sim Y$. Si on note $g(X) = -\sigma X_i$ et $h(X) = f(\theta + \sigma X)$ alors

$$\mathbb{E} [(\theta_i - Y_i) f(Y)] = \text{cov}((\theta_i - Y_i), f(Y)) = \text{cov}(g(X), h(X)).$$

Avec $Z \sim \mathcal{N}(0, \text{Id}_d)$ indépendant de X , l'identité de covariance gaussienne s'écrit alors

$$\begin{aligned} \text{cov}(g(X), h(X)) &= \int_0^1 \mathbb{E} [\langle \nabla g(X), \nabla h(\sqrt{1-u^2}X + uZ) \rangle] du \\ &= -\sigma \int_0^1 \mathbb{E} [\sigma \partial_i f(\sigma(\sqrt{1-u^2}X + uZ) + \theta)] du \\ &= -\sigma^2 \int_0^1 \mathbb{E} [\partial_i f(\sigma X + \theta)] du \\ &= -\sigma^2 \mathbb{E} [\partial_i f(Y)]. \end{aligned}$$

□

Le second ingrédient repose sur les propriétés des normes de vecteurs gaussiens.

PROPOSITION 9.6 (NORMES DE VECTEURS GAUSSIENS) Soit $Y \sim \mathcal{N}(\theta, \sigma^2 \text{Id}_d)$ 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 $Y \sim \mathcal{N}(\theta, \sigma^2 \text{Id}_d)$ et $Z \sim \mathcal{N}(0, \sigma^2 \text{Id}_d)$ 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 $d \geq 3$,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{\|Y\|^2} \right] &= \int_0^\infty \frac{1}{2x} \frac{(x/2)^{d/2-1}}{\Gamma(d/2)} e^{-x/2} dx \\ &= \frac{\Gamma((d-2)/2)}{2\Gamma(d/2)} \int_0^\infty \frac{(x/2)^{(d-2)/2-1}}{2\Gamma((d-2)/2)} e^{-x/2} dx \\ &= \frac{\Gamma((d-2)/2)}{2\Gamma(d/2)} \\ &= \frac{1}{d-2}. \end{aligned}$$

□

PREUVE.(théorème 9.4) On s'intéresse à des estimateurs de la forme $\hat{\theta} := g(Y)Y$ où $g : \mathbb{R}^d \rightarrow \mathbb{R}$. On attend de la fonction g qu'elle contracte l'estimateur au maximum de vraisemblance Y . On part du développement

$$\mathbb{E} \left[\|\hat{\theta} - \theta\|^2 \right] = \underbrace{\mathbb{E} [\|Y - \theta\|^2]}_{(a)} + 2 \underbrace{\mathbb{E} [\langle \theta - Y, (1 - g(Y))Y \rangle]}_{(b)} + \underbrace{\mathbb{E} [(1 - g(Y))^2 \|Y\|^2]}_{(c)}.$$

On a immédiatement

$$(a) = \sigma^2 d.$$

Pour traiter (b), on utilise l'identité de Stein, pour chaque $i \leq d$,

$$\mathbb{E} [(\theta_i - Y_i)(1 - g(Y))Y_i] = -\sigma^2 \mathbb{E} [(1 - g(Y)) - Y_i \partial_i g],$$

en sommant sur i ,

$$(b) = -2\sigma^2 \mathbb{E} [d(1 - g(Y)) - \langle Y, \nabla g \rangle].$$

Soit

$$(b) + (c) = \mathbb{E} [-2d\sigma^2(1 - g(Y)) + 2\sigma^2 \langle Y, \nabla g \rangle + \|Y\|^2(1 - g(Y))^2].$$

Nous allons maintenant chercher à minimiser (b) + (c) en nous concentrant sur des fonctions g de la forme $g(y) = 1 - c/\|y\|^2$, soit

$$\partial_i g(y) = \frac{2cy_i}{\|y\|^4}.$$

$$(b) + (c) = \mathbb{E} \left[-\frac{2dc\sigma^2}{\|Y\|^2} + 2\frac{2c\sigma^2}{\|Y\|^2} + \frac{c^2}{\|Y\|^2} \right] = (c^2 - 2c(d-2)\sigma^2) \mathbb{E} \left[\frac{1}{\|Y\|^2} \right].$$

Cette expression est minimisée en choisissant $c = (d-2)\sigma^2$, 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 9.6 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, \dots, Z_d sont i.i.d. $\mathcal{N}(0, 1)$. Lorsque $\sigma \rightarrow 0$, $\frac{\sigma^2(d-2)^2}{\|Y\|^2}$ 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, \dots, 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

$$\lim_{\sigma \rightarrow 0} \mathbb{E} \left[\frac{\sigma^2(d-2)^2}{\|Y\|^2} \right] < \infty$$

ce qui suffit à établir la dernière affirmation du théorème. \square

9.3 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 $g : \Theta \rightarrow E$. Dans la suite T désigne un estimateur de $g(\theta)$, c'est à dire une fonction des observations X_1, \dots, 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 9.7 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, 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 $\mathcal{Z} \rightarrow \mathcal{Y}$, l'erreur de classification de f en (z, y) est $\frac{1}{2}|f(z) - y|$ que l'on peut aussi écrire $(1 - f(z)y)/2$. 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) $((Z_1, Y_1), \dots, (Z_n, Y_n))$, la qualité du classifieur choisi \hat{f}_n est évaluée par son erreur. Le risque d'une méthode d'apprentissage sous une loi jointe sur $\mathcal{Z} \times \mathcal{Y}$:

$$\mathbb{E} \left[\mathbb{E} \left[\frac{1}{2} |\hat{f}_n(Z) - Y| \right] \right]$$

où l'espérance interne porte sur (Z, Y) et l'espérance externe sur $((Z_1, Y_1), \dots, (Z_n, Y_n))$. 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 $\bar{R}(\Theta)$ est défini par

$$\bar{R}(\Theta) := \inf_T R(T, \Theta) = \inf_T \sup_{\theta \in \Theta} R(T, \theta).$$

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.

9.4 RISQUE BAYÉSIEN

On suppose Θ muni d'une tribu \mathcal{A} qui en fait un espace probablisable. Cet espace probablisable 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 (Θ, \mathcal{A}) vers $(\mathcal{X}, \mathcal{F})$ (et plus généralement de (Θ, \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, \dots, X_n sera appelée *loi de mélange* définie par l'a priori Π et notée P :

$$P(A) := \int_{\Theta} \int_{\mathcal{X}^n} \mathbb{I}_{(x_1, \dots, x_n) \in A} dP_\theta(x_1) \dots dP_\theta(x_n) d\Pi(\theta).$$

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 $X_1 = x_1, \dots, X_n = x_n$ est appelée *loi a posteriori* $\Pi(\cdot | x_1, \dots, x_n)$.

Pour toute statistique T , pour tout $\omega \in \Theta$, la fonction $\Theta \rightarrow \mathbb{R}$ qui associe à une statistique $T(\omega)$, sa perte en ω contre $\theta : \theta \mapsto L(T(\omega), \theta)$, est \mathcal{A} -mesurable. On peut alors définir le risque intégré contre l'a priori Π

$$\int_{\Theta} R(T, \theta) \Pi(d\theta)$$

et lorsque Π est une loi de probabilité

$$\int_{\Theta} R(T, \theta) \Pi(d\theta) = \mathbb{E}_{\Pi} [\mathbb{E}_{P_\theta} [L(T, \theta)]] .$$

Le *risque maximin* est défini par

$$\underline{R}(\Theta) := \sup_{\Pi} \inf_T \int_{\Theta} R(T, \theta) \Pi(d\theta)$$

où le supremum à gauche est pris sur les lois de probabilité définies sur Π .

PROPOSITION 9.8 *Le risque maximin $\underline{R}(\Theta)$ est toujours inférieur ou égal au risque minimax :*

$$\underline{R}(\Theta) = \sup_{\Pi} \inf_T \int_{\Theta} R(T, \theta) \Pi(d\theta) \leq \inf_T \sup_{\theta \in \Theta} R(T, \theta) = \bar{R}(\Theta).$$

PREUVE. Pour une loi a priori Π et un estimateur T quelconques

$$\int_{\Theta} R(T, \theta) \Pi(d\theta) \leq \sup_{\theta \in \Theta} R(T, \theta),$$

donc, pour toute loi a priori Π , le risque bayésien est inférieur ou égal au risque minimax :

$$\inf_T \int_{\Theta} R(T, \theta) \Pi(d\theta) \leq \inf_T \sup_{\theta \in \Theta} R(T, \theta).$$

Si maintenant, on maximise le risque bayésien, celui ci reste inférieur au risque minimax. \square

Pour chaque problème de statistique, on peut se demander s'il y a égalité entre risque maximin et risque minimax.

9.5 LIENS ENTRE ESTIMATEURS BAYÉSIENS ET MINIMAX

THÉORÈME 9.9 *Soit T un estimateur de $g(\theta)$ tel que $\sup_{\theta \in \Theta} R(T, \theta) = r < \infty$. Soit $\Theta' \subseteq \Theta$ vérifiant*

$$\forall \theta \in \Theta', \quad R(T, \theta) = r.$$

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

$$\int_{\Theta} R(T', \theta) d\Pi(\theta) \leq r' < r = \int_{\Theta} R(T, \theta) d\Pi(\theta).$$

Ceci contredit l'hypothèse du caractère bayésien de l'estimateur T relativement à l'a priori Π . On a donc

$$\forall T', \quad r \leq \sup_{\theta \in \Theta'} R(T', \theta) \leq \sup_{\theta \in \Theta} R(T', \theta).$$

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. \square

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 9.10 *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*

$$\lim_{k \rightarrow \infty} \inf_{T'} \int_{\Theta} R(T', \theta) d\Pi_k(\theta) = r$$

alors T est un estimateur minimax.

PREUVE. Pour tout k , on note

$$r_k := \inf_{T'} \int_{\Theta} R(T', \theta) d\Pi_k(\theta),$$

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 . \square

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.

DÉFINITION 9.11 (MESURES σ -FINIES) Une mesure μ sur $(\mathcal{X}, \mathcal{F})$ est dite σ -finie si et seulement si \mathcal{X} est réunion dénombrable de parties \mathcal{F} -mesurables de μ -mesure finie.

La mesure de comptage sur \mathbb{R} est un exemple de mesure qui n'est pas σ -finie. La mesure de Lebesgue est un exemple de mesure σ -finie.

Envisageons la situation suivante, on dispose d'une loi de probabilité P sur (Ω, \mathcal{F}) et d'une sous-tribu \mathcal{G} . Si X est une variable aléatoire sur (Ω, \mathcal{F}) qui est P -intégrable, on sait définir l'espérance conditionnelle de X sachant \mathcal{G} notée $\mathbb{E}[X | \mathcal{G}]$: c'est une variable aléatoire Y , \mathcal{G} -mesurable telle que pour toute variable aléatoire Z , \mathcal{G} -mesurable et P -intégrable :

$$\mathbb{E}[XZ] = \mathbb{E}[ZY].$$

Deux versions de l'espérance conditionnelle sont P -presque sûrement égales.

Si maintenant on cherche à définir une notion de probabilité conditionnelle sachant \mathcal{G} , on est tenté de définir

$$P(A | \mathcal{G}) = \mathbb{E}[\mathbb{I}_A | \mathcal{G}].$$

On aura bien, à partir des propriétés classiques de l'espérance conditionnelle, P -p.s.

$$\begin{aligned} P(\Omega | \mathcal{G}) &= 1 \\ P(\emptyset | \mathcal{G}) &= 0 \\ P(\cup_n A_n | \mathcal{G}) &= \sum_n P(A_n | \mathcal{G}) \quad \text{pour } (A_n) \text{ disjoints deux à deux.} \end{aligned}$$

Il faut faire attention à la formulation de la dernière propriété. Elle est valable si on choisit la collection de parties disjointes avant de « choisir » une version de l'espérance conditionnelle. Mais une fois la version de l'espérance conditionnelle choisie, on ne peut pas garantir que cette propriété soit vérifiée P -presque sûrement pour l'infini (non-dénombrable) de collections d'événements $(A_n)_n$. A partir des seules (et très générales) propriétés de l'espérance conditionnelle, on ne peut pas garantir en toute généralité l'existence d'une version de l'espérance conditionnelle telle que P -p.s.

$$A \mapsto \mathbb{E}[\mathbb{I}_A | \mathcal{G}] \quad \text{pour } A \in \mathcal{F}$$

définisse une loi de probabilité sur (Ω, \mathcal{F}) .

Rappelons la définition de l'objet qui nous intéresse.

DÉFINITION 9.12 (PROBABILITÉ CONDITIONNELLE RÉGULIÈRE) Une fonction $P(\cdot | \mathcal{G})(\cdot)$ de $\mathcal{F} \times \Omega$ dans $[0, 1]$ est appelée probabilité conditionnelle régulière sachant \mathcal{G} si et seulement si

- i) pour tout $B \in \mathcal{F}$, $P(B | \mathcal{G})(\omega) = \mathbb{E}[\mathbb{I}_B | \mathcal{G}](\omega)$ P -p.s.
- ii) $P(\cdot | \mathcal{G})(\omega)$ est P -p.s. une loi de probabilité sur (Ω, \mathcal{F}) .

En statistique bayésienne, 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 9.12.

DÉFINITION 9.13 (PROBABILITÉ CONDITIONNELLE DE Y SACHANT X) Soit (Ω, \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 (Ω, \mathcal{F}) par Y . Une fonction $P(\cdot | X)(\cdot)$ de $\mathcal{H} \times X(\Omega)$ définie par

$$P(H | X)(X(\omega)) = P(Y^{-1}(H) | \mathcal{G})(\omega) \quad \text{pour } H \in \mathcal{H}, \omega \in \Omega,$$

est une probabilité conditionnelle de Y sachant X . On notera $P(\cdot | x)$ la fonction $H \mapsto P(H | X)(x)$ 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 9.14 Si (X, Y) sont des variables aléatoires sur (Ω, \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

$$\mathbb{E}[f(X, Y) | X](\omega) = \int_{\mathcal{Y}} f(x, y)P(dy | x) \quad \text{pour } x = X(\omega)$$

et

$$\mathbb{E}[f(X, Y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x, y)P(dy | x)P_X(dx)$$

où $P_X = P \circ X^{-1}$.

Le théorème suivant garantit l'existence d'une probabilité conditionnelle régulière dans tous les scenarii qui nous intéressent.

THÉORÈME 9.15 (EXISTENCE DE PROBABILITÉS CONDITIONNELLES) Soit (Ω, \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 (Ω, \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 $P(\cdot | X)(\cdot)$).

9.7 CONSTRUCTION D'ESTIMATEURS BAYÉSIENS

La construction des estimateurs bayésiens repose sur la notion de loi a posteriori.

La loi a priori Π sur (Θ, \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 :

$$\mathbb{P}(A \times E) = \int_{\Theta} \mathbb{I}_A(\theta) \times P_\theta(E) \Pi(d\theta)$$

pour $A \in \mathcal{A}, E \in \mathcal{F}$.

La loi jointe définit une loi marginale \mathbb{Q} sur $(\mathcal{X}, \mathcal{F})$:

$$\mathbb{Q}(E) = \int_{\Theta} P_\theta(E) \Pi(d\theta).$$

Toute famille de lois $(P(\cdot | x))_{x \in \mathcal{X}}$ sur (Θ, \mathcal{A}) telle que pour tout $A \in \mathcal{A}$, $x \mapsto P(A | x)$ est \mathcal{F} mesurable et

$$\mathbb{P}(A \times E) = \int_{\mathcal{X}} \mathbb{I}_E(x) \times P(A | x) \mathbb{Q}(dx)$$

pour tous $A \in \mathcal{A}, E \in \mathcal{F}$, est appelée famille de lois a posteriori pour l'a priori Π . La loi $P(\cdot | x)$ est 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 9.15).

PROPOSITION 9.16 (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 $\Pi(\cdot | X)$.

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 9.17 (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 $p(x | \theta)$ alors la mesure a posteriori étant donné X_1, \dots, 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 θ, X_1, \dots, X_n*

$$\pi(\theta | X_1, \dots, X_n) \propto \pi(\theta) \times \prod_{i=1}^n p(X_i | \theta).$$

PREUVE. On traite le cas d'une observation ($n = 1$).

On définit

$$p_X(x) = \int_{\Theta} p(x | \theta) \pi(\theta) d\theta,$$

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 Tonelli-Fubini), mais son calcul peut être délicat.

L'existence d'une loi a posteriori est garantie par le Théorème 9.15. Il suffit de vérifier que P_X presque sûrement, pour tout $A \in \mathcal{F}$

$$\mathbb{E}[\mathbb{I}_A | X](\omega) = \int_{\Theta} \mathbb{I}_A(\theta) \Pi(d\theta | x) \quad \text{avec } x = X(\omega)$$

est égal à

$$\int_{\Theta} \mathbb{I}_A(\omega) \frac{p(\theta | x) \pi(\theta)}{p_X(x)} d\theta.$$

Sans perdre en généralité, on peut se restreindre à l'ensemble $\{x : p_X(x) > 0\}$.

Pour tout $A \subseteq \Theta$ mesurable,

$$\begin{aligned} \mathbb{E} \mathbb{I}_{\theta \in A} &= \int_{\Theta} \int_{\mathcal{X}} \mathbb{I}_{p_X(x) > 0} \mathbb{I}_{\theta \in A} p(x | \theta) \pi(\theta) \nu(dx) d\theta \\ &= \int_{\mathcal{X}} p_X(x) \mathbb{I}_{p_X(x) > 0} \left(\int_{\Theta} \mathbb{I}_{\theta \in A} \frac{p(x | \theta) \pi(\theta)}{p_X(x)} d\theta \right) \mathbb{I}_{p_X(x) > 0} \nu(dx) \end{aligned}$$

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 computationnelles 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 9.18 (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

$$e^{-\theta} \frac{\theta^k}{k!}$$

(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 :

$$\mathbb{I}_{\theta > 0} \lambda \frac{(\lambda\theta)^{p-1}}{\Gamma(p)} e^{-\lambda\theta}.$$

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

$$\propto e^{-\theta} \frac{\theta^k}{k!} \times \mathbb{I}_{\theta>0} \lambda \frac{(\lambda\theta)^{p-1}}{\Gamma(p)} e^{-\lambda\theta} \propto \mathbb{I}_{\theta>0} (\lambda+1) \frac{((\lambda+1)\theta)^{p+k-1}}{\Gamma(p+k)} e^{-(\lambda+1)\theta}.$$

On reconnaît 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, \dots, X_n , en notant $S_n := \sum_{i=1}^n X_i$, on reconnaîtrait que la loi a posteriori $\Pi(\cdot | X_1, \dots, X_n)$ 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 suite des variances des lois 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 9.19 *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 $\prod_{i=1}^n p(x_i | \theta)$ se factorise en

$$\prod_{i=1}^n p(x_i | \theta) = q(s | \theta) \times h(x_1, \dots, x_n, s) \quad \text{pour } s = S_n(x_1, \dots, x_n)$$

où $q(\cdot | \cdot)$ et h sont des fonctions mesurables. La densité de la loi a posteriori est donc proportionnelle à

$$\pi(\theta | x_1, \dots, x_n) \propto \pi(\theta) q(s | \theta) h(x_1, \dots, x_n, s) \quad \text{pour } s = S_n(x_1, \dots, x_n).$$

En tant que fonction de θ elle est proportionnelle à $\pi(\theta) q(s | \theta)$. □

La notion de loi a posteriori est centrale dans la construction d'estimateurs bayésiens.

PROPOSITION 9.20 *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 $\Pi(\cdot | x_1, \dots, x_n)$:*

$$T(x_1, \dots, x_n) = \operatorname{argmin}_t \int_{\Theta} L(t, g(\theta')) d\Pi(\theta' | x_1, \dots, x_n).$$

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.

$$\begin{aligned} \int_{\Theta} \mathbb{E}_\theta [L(T(X_1, \dots, X_n), g(\theta))] d\Pi(\theta) \\ = \int_{\mathcal{X}^n} \int_{\Theta} L(T(x_1, \dots, x_n), g(\theta)) d\Pi(\theta | x_1, \dots, x_n) dP(x_1, \dots, x_n). \end{aligned}$$

Si un estimateur minimise l'intégrande $\int_{\Theta} L(T(x_1, \dots, x_n), g(\theta)) d\Pi(\theta | x_1, \dots, x_n)$, il minimise le risque bayésien. \square

Pour le risque quadratique, l'estimateur bayésien prend une forme particulièrement simple.

COROLLAIRE 9.21 *Si la fonction de perte est quadratique (carré d'une distance L_2), l'estimateur bayésien est la moyenne a posteriori.*

PREUVE. Si $T(X_1, \dots, X_n) = \int_{\Theta} g(\theta) d\Pi(\theta | X_1, \dots, X_n)$ P -p.s., alors pour tout autre estimateur T' ,

$$\int_{\Theta} (T'(x_1, \dots, x_n) - g(\theta))^2 d\Pi(\theta | x_1, \dots, x_n) \geq \int_{\Theta} (T(x_1, \dots, x_n) - g(\theta))^2 d\Pi(\theta | x_1, \dots, x_n)$$

car l'espérance minimise l'écart quadratique. Donc en intégrant par rapport à la loi de mélange,

$$\begin{aligned} \int_{\mathcal{X}^n} \int_{\Theta} (T'(x_1, \dots, x_n) - g(\theta))^2 d\Pi(\theta | x_1, \dots, x_n) dP(x_1, \dots, x_n) \geq \\ \int_{\mathcal{X}^n} \int_{\Theta} (T(x_1, \dots, x_n) - g(\theta))^2 d\Pi(\theta | x_1, \dots, x_n) dP(x_1, \dots, x_n). \end{aligned}$$

\square

PROPOSITION 9.22 *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 9.21, T est l'espérance de $g(\theta')$ lorsque θ' est distribuée sous la loi a posteriori :

$$T(X_1, \dots, X_n) = \int_{\Theta} g(\theta') d\Pi(\theta' | X_1, \dots, X_n).$$

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,

$$\begin{aligned}
& \int_{\Theta} R(T, \theta) d\Pi(\theta) \\
&= \int_{\Theta} \left[\int_{\mathcal{X}^n} (T(x_1, \dots, x_n) - g(\theta))^2 \prod_{i=1}^n p(x_i | \theta) d\nu(x_1) \dots d\nu(x_n) \right] d\Pi(\theta) \\
&= \int_{\Theta} \left[\int_{\mathcal{X}^n} (T^2 - g(\theta)^2) \prod_{i=1}^n p(x_i | \theta) d\nu(x_1) \dots d\nu(x_n) \right] d\Pi(\theta) \\
&\quad \text{car } T \text{ est sans biais} \\
&= \int_{\mathcal{X}^n} \int_{\Theta} (T^2 - g(\theta)^2) d\Pi(\theta | x_1, \dots, x_n) dP(x_1, \dots, x_n) \\
&\quad \text{par désintégration} \\
&= \int_{\mathcal{X}^n} \int_{\Theta} \left(\left(\int_{\Theta} g(\theta') d\Pi(\theta' | x_1, \dots, x_n) \right)^2 - g(\theta)^2 \right) d\Pi(\theta | x_1, \dots, x_n) dP(x_1, \dots, x_n) \\
&\quad T \text{ est Bayésien, donc moyenne sous loi a posteriori} \\
&\leq \int_{\mathcal{X}^n} \int_{\Theta} \left(\left(\int_{\Theta} g(\theta')^2 d\Pi(\theta' | x_1, \dots, x_n) \right) - g(\theta)^2 \right) d\Pi(\theta | x_1, \dots, x_n) dP(x_1, \dots, x_n) \\
&\quad \text{inégalité de Jensen} \\
&= \int_{\mathcal{X}^n} \int_{\Theta} g(\theta')^2 d\Pi(\theta' | x_1, \dots, x_n) dP(x_1, \dots, x_n) \\
&\quad - \int_{\mathcal{X}^n} \int_{\Theta} g(\theta)^2 d\Pi(\theta | x_1, \dots, x_n) dP(x_1, \dots, x_n) \\
&= 0,
\end{aligned}$$

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 trivialisait l'estimateur. \square

REMARQUE 9.23 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.

9.8 LIMITES DE L'APPROCHE MINIMAX

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 $n_1 := \sum_{i=1}^n X_i = n\bar{X}_n$ 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

$$\hat{\theta}_n := \frac{p + n_1}{p + q + n} = \frac{\bar{X}_n + \frac{p}{n}}{1 + \frac{p+q}{n}}$$

On peut rechercher parmi les lois Beta celle qui forme l'a priori le plus défavorable, celle qui maximise le risque Bayésien (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

$$\hat{\theta}_n := \frac{\bar{X}_n + \frac{1}{2\sqrt{n}}}{1 + \frac{1}{\sqrt{n}}},$$

son biais et sa variance en θ étant respectivement

$$\frac{1}{\sqrt{n}} \frac{\frac{1}{2} - \theta}{1 + \frac{1}{\sqrt{n}}} \quad \text{et} \quad \frac{\theta(1 - \theta)}{n \left(1 + \frac{1}{\sqrt{n}}\right)^2}.$$

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 9.9, 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 9.9, 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 9.22.

9.9 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 $p(x | \theta)$. La fonction $\theta \mapsto p(x | \theta)$ est supposée mesurable.

Pour tout θ , on suppose que pour tout θ , la fonction score $\frac{d}{d\theta} \log p(x | \theta)$ est dans $L_2(P_\theta)$, et qu'elle est centrée

$$\int_{\mathcal{X}} \frac{d}{d\theta} \log p(x | \theta) p(x | \theta) d\nu(x) = 0.$$

On note $I(\theta)$ la variance de la fonction score

$$I(\theta) := \int_{\mathcal{X}} \left(\frac{d}{d\theta} \log p(x | \theta) \right)^2 p(x | \theta) d\nu(x).$$

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 :

$$I(\Pi) := \int_{\Theta} \frac{\pi'(\theta)^2}{\pi(\theta)} d\theta < \infty.$$

ii) La fonction $g : \Theta \rightarrow \mathbb{R}$ est absolument continue, dans $L_2(\Pi)$ et

$$\int_{\Theta} g'(\theta)^2 d\Pi(\theta) < \infty.$$

iii) Les fonctions $\theta \mapsto \pi(\theta)$ et $\theta \mapsto \pi(\theta)g(\theta)$ 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)

$$\int_{]a, b[} \frac{d(p(x | \theta)\pi(\theta))}{d\theta} d\theta = [p(x | \theta)\pi(\theta)]_a^b = 0$$

b)

$$\int_{]a, b[} \theta \frac{d(p(x | \theta)\pi(\theta))}{d\theta} d\theta = [\theta p(x | \theta)\pi(\theta)]_a^b - \int_{]a, b[} p(x | \theta)\pi(\theta) d\theta = - \int_{]a, b[} p(x | \theta) d\Pi(\theta)$$

THÉORÈME 9.24 (INÉGALITÉ DE VAN TREES) *Sous les hypothèses énoncées ci-dessus, si $T = T(X)$ est une statistique,*

$$\int_{\Theta} \mathbb{E}_{\theta} [(T - \theta)^2] d\Pi(\theta) \geq \frac{1}{\int_{\Theta} I(\theta) d\Pi(\theta) + I(\Pi)}.$$

PREUVE. L'étape cruciale de la preuve consiste à vérifier

$$\int_{\Theta} \int_{\mathcal{X}} \left[(T(x) - \theta) \frac{d}{d\theta} (p(x | \theta) \times \pi(\theta)) \right] d\nu(x) d\theta = 1. \quad (9.1)$$

La vérification revient à invoquer les hypothèses du théorème à l'aide du théorème de Fubini. D'une part, on a

$$\begin{aligned} & \int_{\Theta} \int_{\mathcal{X}} T(x) \frac{d}{d\theta} (p(x | \theta) \times \pi(\theta)) d\nu(x) d\theta \\ &= \int_{\mathcal{X}} T(x) \int_{\Theta} \frac{d}{d\theta} (p(x | \theta) \times \pi(\theta)) d\theta d\nu(x) \\ &= \int_{\mathcal{X}} T(x) [p(x | \theta) \times \pi(\theta)]_a^b d\nu(x) \\ &= \int_{\mathcal{X}} 0 d\nu(x) \\ &= 0. \end{aligned}$$

Et, on a aussi

$$\begin{aligned} & \int_{\Theta} \int_{\mathcal{X}} \left[\theta \frac{d}{d\theta} (p(x | \theta) \times \pi(\theta)) \right] d\nu(x) d\theta \\ &= \int_{\mathcal{X}} \int_{\Theta} \theta \frac{d}{d\theta} [(p(x | \theta) \times \pi(\theta))] d\theta d\nu(x) \\ &= \int_{\mathcal{X}} \left([\theta p(x | \theta) \times \pi(\theta)]_a^b - \int_{\Theta} p(x | \theta) \times \pi(\theta) d\theta \right) d\nu(x) \\ &= - \int_{\Theta} \int_{\mathcal{X}} p(x | \theta) d\nu(x) \pi(\theta) d\theta \\ &= -1. \end{aligned}$$

A partir de (9.1), on peut réécrire

$$\begin{aligned}
1 &= \int_{\Theta} \int_{\mathcal{X}} \left[(T(x) - \theta) \frac{d}{d\theta} \log(p(X | \theta) \times \pi(\theta)) \right] p(X | \theta) \times \pi(\theta) d\nu(x) d\theta \\
&\leq \left(\int_{\Theta} \int_{\mathcal{X}} [(T(x) - \theta)^2] p(x | \theta) \times \pi(\theta) d\nu(x) d\theta \right)^{1/2} \\
&\quad \times \left(\int_{\Theta} \int_{\mathcal{X}} \left[\left(\frac{d}{d\theta} \log(p(x | \theta) \times \pi(\theta)) \right)^2 \right] p(x | \theta) \times \pi(\theta) d\nu(x) d\theta \right)^{1/2}
\end{aligned}$$

où l'inégalité est une application de l'inégalité de Cauchy-Schwarz.

$$\begin{aligned}
&\int_{\Theta} \int_{\mathcal{X}} \left[\left(\frac{d}{d\theta} \log(p(x | \theta) \times \pi(\theta)) \right)^2 \right] p(x | \theta) \times \pi(\theta) d\nu(x) d\theta \\
&= \int_{\Theta} \int_{\mathcal{X}} \left[\left(\frac{d}{d\theta} \log p(x | \theta) \right)^2 + \left(\frac{d}{d\theta} \log \pi(\theta) \right)^2 \right. \\
&\quad \left. + 2 \frac{d}{d\theta} \log p(x | \theta) \times \frac{d}{d\theta} \log \pi(\theta) \right] p(x | \theta) \times \pi(\theta) d\nu(x) d\theta \\
&= \int_{\Theta} I(\theta) \pi(\theta) d\theta + I(\pi) \\
&\quad + 2 \int_{\Theta} \left(\int_{\mathcal{X}} \frac{d}{d\theta} \log p(x | \theta) p(x | \theta) d\nu(x) \right) \pi(\theta) d\theta \\
&= \int_{\Theta} I(\theta) \pi(\theta) d\theta + I(\pi),
\end{aligned}$$

où la dernière égalité est une conséquence du centrage des fonctions scores dans les modèles réguliers.

□

On prouve de la même façon.

COROLLAIRE 9.25

$$\int_{\Theta} \mathbb{E}_{\theta} [(T(X_1, \dots, X_n) - g(\theta))^2] d\Pi(\theta) \geq \frac{(\int_{\Theta} g'(\theta) d\Pi(\theta))^2}{n \int_{\Theta} I(\theta) d\Pi(\theta) + I(\Pi)}.$$

PREUVE. Il suffit de vérifier que pour ν -presque tout x ,

$$\begin{aligned}
&\int_{\Theta} g(\theta) \frac{d}{d\theta} (p(x | \theta) \times \pi(\theta)) d\theta \\
&= [g(\theta) (p(x | \theta) \times \pi(\theta))]_a^b - \int_{\Theta} g'(\theta) p(x | \theta) \pi(\theta) d\theta \\
&= - \int_{\Theta} g'(\theta) p(x | \theta) \pi(\theta) d\theta.
\end{aligned}$$

pour remplacer (9.1) par

$$\int_{\Theta} \int_{\mathcal{X}} \left[(T(x) - g(\theta)) \frac{d}{d\theta} (p(x | \theta) \times \pi(\theta)) \right] d\nu(x) d\theta = \int_{\Theta} g'(\theta) d\Pi(\theta).$$

□

EXEMPLE 9.26 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 $(\bar{X}_n)_+^\alpha$. Il est consistant en tout $\theta \in \Theta$. Si $\theta > 0$, la méthode delta nous indique que

$$\sqrt{n} ((\bar{X}_n)_+^\alpha - \theta^\alpha) \rightsquigarrow \mathcal{N}(0, \alpha^2 \theta^{2\alpha-2}).$$

On doit se demander comment se comporte le risque quadratique en $\theta = 0$ et comment se comporte le risque minimax à n fixé. En $\theta = 0$, $((\bar{X}_n)_+^\alpha - \theta^\alpha) = (\bar{X}_n)_+^\alpha$ 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 décroît 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 ??, Pour tout estimateur T

$$\sup_{\theta \geq 0} \mathbb{E}_\theta [(T(X_1, \dots, X_n) - \theta^\alpha)^2] \geq \int_{\Theta} \mathbb{E}_\theta [(T(X_1, \dots, X_n) - \theta^\alpha)^2] \pi(\theta) d\theta \geq \frac{(\int_{\Theta} \alpha \theta^{\alpha-1} \pi(\theta))^2}{nI(\theta) + I(\pi)}.$$

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

$$\frac{a^{2(\alpha-1)} (\int_{\Theta} \alpha \theta^{\alpha-1} \pi(\theta))^2}{n + I(\pi)/a^2}.$$

On peut optimiser le facteur d'échelle en choisissant

$$a = \sqrt{\frac{\alpha I(\pi)}{n(1-\alpha)}}.$$

On obtient finalement

$$\sup_{\theta \geq 0} \mathbb{E}_\theta [(T(X_1, \dots, X_n) - \theta^\alpha)^2] \geq \frac{(I(\pi))^{\alpha-1} \alpha^\alpha (1-\alpha)^{1-\alpha}}{n^\alpha} \left(\int_{\Theta} \alpha \theta^{\alpha-1} \pi(\theta) \right)^2.$$

9.10 REMARQUES BIBLIOGRAPHIQUES

Le livre de Lehmann et Casella [5] est la source principale d'inspiration du cours. L'estimateur de James-Stein est aussi exposé dans [7].

L'étude systématique des modèles différentiables en moyenne quadratique (théorie de Hajek-Le Cam) est au coeur de [2, 8]. Elle est aussi illustrée sur les tests dans [6] et explorée en profondeur dans [4].

[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.

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 [MR3587782] 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] I. IBRAGIMOV et R. KHASHINSKII. **Statistical Estimation : Asymptotic Theory**. New York : Springer-Verlag, 1981.
- [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.
- [7] A. TSYBAKOV. **Introduction à l'estimation non-paramétrique**. T. 41. Mathématiques & Applications. Berlin : Springer-Verlag, 2004, p. x+175. MR : MR2013911 (2005a:62007).
- [8] A. VAN DER VAART. **Asymptotic statistics**. Cambridge University Press, 1998.

10.1 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 une paramètre $\theta \in \Theta \subseteq \mathbb{R}^k$ inconnu, des observations y et une fonction de vraisemblance $L(\theta; y) = f(y|\theta)$. On note $\pi(\theta)$ la (densité de la) loi a priori et $\pi(\theta|y) \propto \pi(\theta)L(y; \theta)$ 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 10.1 (ENSEMBLE DE CRÉDIBILITÉ) Soit $\Pi(\cdot | y)$ 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

$$\Pi[\theta \in C_y | y] \geq 1 - \alpha.$$

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 10.2 (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.

10.2 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 10.3 Considérons le modèle simple suivant : Y_1, \dots, 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 10.3 montre l'influence de la loi a priori pour différentes valeurs de n .

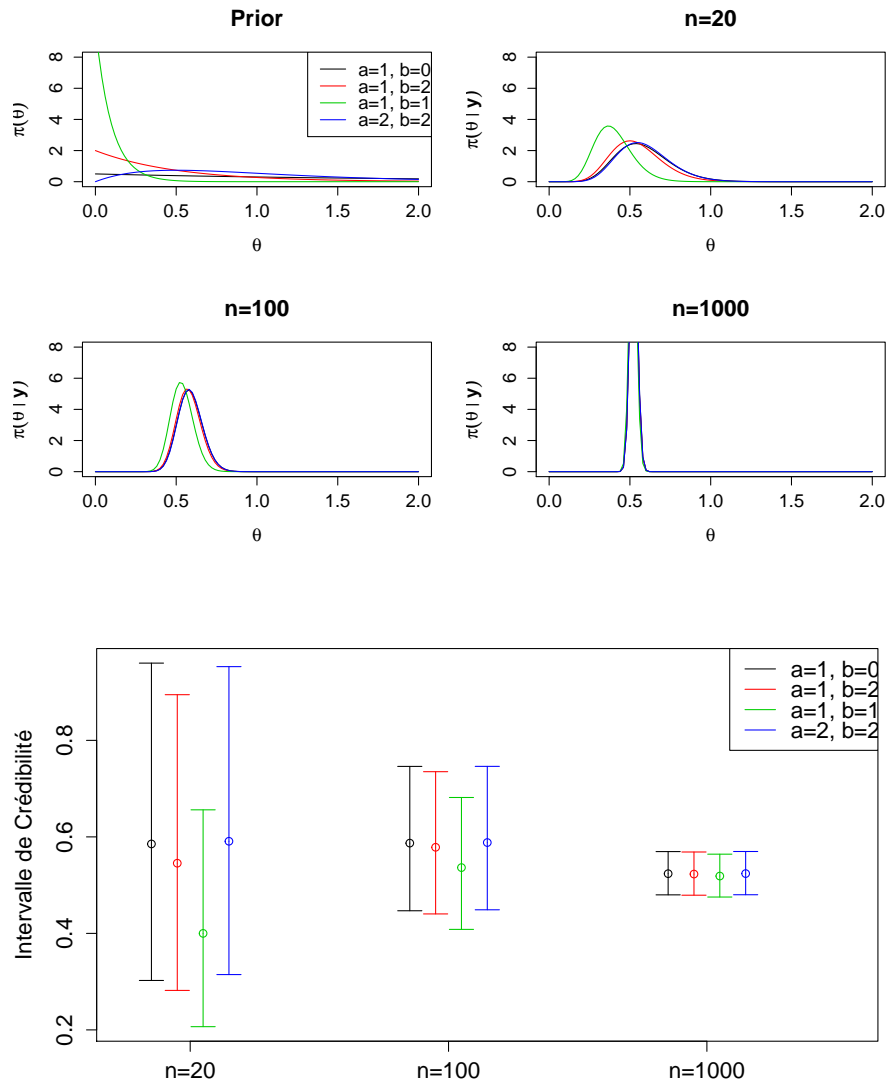


FIG. 10.1 : Pour un modèle de Poisson $\mathcal{P}(\lambda)$ et différentes lois a priori $\lambda \sim \Gamma(a, b)$, on observe que les lois a posteriori correspondantes (haut) se rapprochent au fur et à mesure que la taille d'échantillon n augmente, et notamment les intervalles de crédibilité (bas). Pour $n = 1000$, l'influence de la loi a priori est négligeable.

10.3 CONJUGAISON DANS LES MODÈLES EXPONENTIELS

Avec (Ω, \mathcal{F}) un espace probabilisable, si $T : \Omega \rightarrow \mathbb{R}^p$ est \mathcal{F} -mesurable, et ν une mesure σ -finie sur (Ω, \mathcal{F}) , le modèle exponentiel défini par ν et T est paramétré par l'intérieur de

$$\Theta := \left\{ \theta \in \mathbb{R}^p : Z(\theta) := \int_{\Omega} \exp(\langle \theta, T(\omega) \rangle) \nu(d\omega) < \infty \right\}$$

supposé non vide. Chaque loi P_{θ} est définie par sa densité par rapport à ν :

$$p(\omega | \theta) = \exp(\langle \theta, T(\omega) \rangle) / Z(\theta).$$

On peut définir une loi a priori sur Θ par sa densité par rapport à la mesure de Lebesgue

$$\pi(\theta) \propto \exp(\langle \theta, t \rangle - t_0 \log Z(\theta))$$

avec $t \in \mathbb{R}^p, t_0 \in \mathbb{R}$ tels que

$$\int_{\Theta^\circ} \exp(\langle \theta, t \rangle - t_0 \log Z(\theta)) d\theta < \infty.$$

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 $D_n = \omega_1, \dots, \omega_n$ est proportionnelle à

$$\exp\left(\langle \theta, t + \sum_{i=1}^n T(\omega_i) \rangle - (n + t_0) \log Z(\theta)\right),$$

la densité $\pi(\cdot | \omega_1, \dots, \omega_n)$ 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 $n \rightarrow \infty$,

$$\tilde{\theta}_n := \nabla \log Z^{-1} \left(\frac{t + \sum_{i=1}^n T(\omega_i)}{t_0 + n} \right).$$

Lorsque n tend vers l'infini, $\tilde{\theta}_n$ tend à s'approcher de l'EMV: $\hat{\theta}_n := \nabla \log Z^{-1} \left(\frac{\sum_{i=1}^n T(\omega_i)}{n} \right)$.

On peut se demander à quoi ressemble la loi a posteriori recentrée autour de $\tilde{\theta}_n$ et renormalisée par \sqrt{n} , soit la loi de densité

$$h \mapsto \frac{1}{\sqrt{n}} \pi \left(\hat{\theta}_n + \frac{h}{\sqrt{n}} | D_n \right)$$

La densité de la loi a posteriori recentrée autour de $\tilde{\theta}_n$ et renormalisée par \sqrt{n} est proportionnelle à

$$\exp \left(-\frac{t_0 + n}{2n} h^t \nabla^2 \log Z(\tilde{\theta}_n) h + \frac{(t_0 + n) \|h\|^2}{n} R_{\tilde{\theta}_n}(h/\sqrt{n}) \right)$$

où $R_{\tilde{\theta}_n}(x) = o(\|x\|)$.

La régularité de la fonction $\theta \mapsto \log Z(\theta)$ et la convergence presque sûre de $\tilde{\theta}_n$ vers θ_0 nous indiquent que $\nabla^2 \log Z(\tilde{\theta}_n)$ converge presque sûrement vers $\nabla^2 \log Z(\theta_0) = I(\theta_0)$. Un peu de travail supplémentaire nous permettrait de vérifier que presque sûrement

$$R_{\tilde{\theta}_n}(h/\sqrt{n}) \longrightarrow 0.$$

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 densités implique la convergence en distribution, et même la convergence au sens de la distance en variation (Lemme 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 $\sqrt{n}I(\hat{\theta}_0)^{1/2}(\hat{\theta}_n - \theta_0)$.

10.4 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 10.4 (FACTEUR DE BAYES)

On appelle facteur de Bayes la quantité

$$B_{01}^\pi(y) = \frac{\Pi[M_0|y]}{\Pi[M_1|y]} \bigg/ \frac{\Pi[M_0]}{\Pi[M_1]}.$$

On utilise le facteur de Bayes pour mesurer comment les données ont fait évoluer le rapport des chances. Intuitivement, si $B_{01}^\pi(y)$ est grand, alors les données y favorisent M_0 ; si $B_{01}^\pi(y)$ 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 :

$$B_{01}^\pi = \frac{m_0(y)}{m_1(y)} = \frac{\int \pi_0(\theta_0)L_0(\theta_0; y) d\theta_0}{\int \pi_1(\theta_1)L_1(\theta_1; y) d\theta_1}$$

où

$$m_i(y) = \mathbb{E}_{\pi_i}[L_i(\theta_i; y)] = \int \pi_i(\theta_i)L_i(\theta_i; y) d\theta_i$$

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 10.5

- 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 $\log_{10} B_{01}^x(y) > 0$, 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 10.6 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 :

$$\begin{aligned} \mathcal{M}_0 : \quad & Y_i \sim \mathcal{P}(\lambda) \quad \lambda \sim \Gamma(a, b) \\ \mathcal{M}_1 : \quad & Y_i | Z_i = k \sim \mathcal{P}(\lambda_k) \quad \lambda_1, \lambda_2 \sim \Gamma(a, b) \end{aligned}$$

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) :

$$\begin{aligned} m_0(y) &= E_{\pi_0} [L_0(y|\theta_0)] = \int_0^\infty \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} e^{-n\lambda} \lambda^{\sum y_i} \prod \frac{1}{y_i!} d\lambda = \frac{b^a}{\Gamma(a)} \prod \frac{1}{y_i!} \frac{\Gamma(a + \sum y_i)}{(b+n)^{a+\sum y_i}}; \\ m_1(y) &= \frac{b^{2a}}{\Gamma(a)^2} \frac{\Gamma(a + S_1)}{(b+n_1)^{a+S_1}} \frac{\Gamma(a + S_2)}{(b+n_2)^{a+S_2}} \prod \frac{1}{y_i!} \end{aligned}$$

d'où le facteur de Bayes

$$B_{01}(y) = \frac{\Gamma(a)}{b^a} \frac{\Gamma(a + S_1 + S_2)}{\Gamma(a + S_1)\Gamma(a + S_2)} \frac{(b+n_1)^{a+S_1} (b+n_2)^{a+S_2}}{(b+n)^{a+S_1+S_2}}$$

10.5 MONTE-CARLO

Beaucoup de quantités d'intérêt peuvent s'écrire sous la forme

$$I = \mathbb{E}_p[h(\theta)] = \int h(\theta)p(\theta) d\theta.$$

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

$$\hat{I}_T^{MC} = \frac{1}{T} \sum_{t=1}^T h(\theta_t).$$

Notons que \hat{I}_T^{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

$$\text{var}(\hat{I}_T^{MC}) = \frac{1}{T} \text{var}_p(h(\theta)).$$

Cette approche simple est rarement optimale : d'une part $\text{var}_p(h(\theta))$ 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,

$$I = \int \frac{h(\theta)p(\theta)}{\gamma(\theta)} \gamma(\theta) d\theta = \mathbb{E}_\gamma \left[\frac{h(\theta)p(\theta)}{\gamma(\theta)} \right]$$

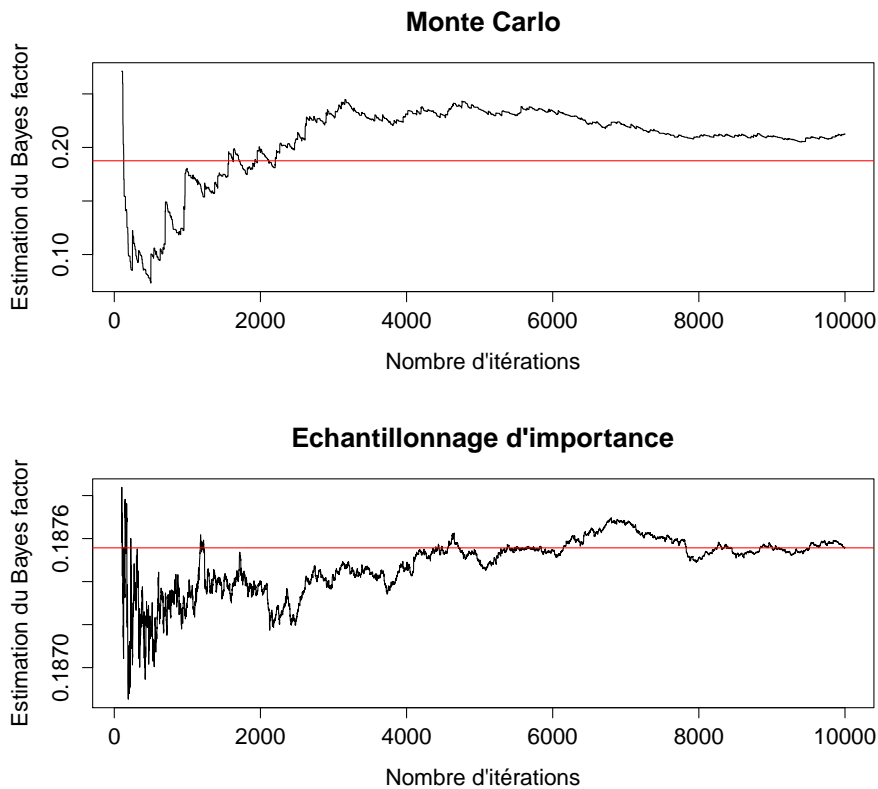


FIG. 10.2 : Estimation d'un Bayes factor par Monte-Carlo et par Échantillonnage d'importance (échantillonnage préférentiel). En rouge, la valeur analytique. Notez que l'échelle des ordonnées n'est pas la même sur les deux figures.

et on reprend l'idée précédente : on simule $\theta'_1, \dots, \theta'_T \sim \gamma$ et on pose

$$\hat{I}_T^{IS} = \frac{1}{T} \sum_{t=1}^T \frac{h(\theta_t)p(\theta_t)}{\gamma(\theta_t)}.$$

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

$$\text{var}(\hat{I}_T^{IS}) = \frac{1}{T} \text{var}_\gamma \left(\frac{h(\theta)p(\theta)}{\gamma(\theta)} \right)$$

on va chercher à minimiser cette variance, ce qui revient à prendre $\gamma \approx hp$ à une constante multiplicative près.

EXEMPLE 10.7 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 ci-dessous montre l'évolution des estimateurs \hat{I}_T^{MC} et \hat{I}_T^{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.

10.6 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 $Z_t \xrightarrow[t \rightarrow \infty]{\mathcal{L}} p$, 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 10.8 (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 $K(x \rightarrow x')$. On dit que (Z_t) vérifie le *detailed balance* pour p si

$$\forall x, x' \in \mathcal{X}, p(x)K(x \rightarrow x') = p(x')K(x' \rightarrow x).$$

On peut aussi dire que (Z_t) est *réversible*.

THÉORÈME 10.9 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

$$P[Z_{t+1} = i] = \sum_j P[Z_t = j]K(j \rightarrow i) = \sum_j p(j)K(j \rightarrow i) = \sum_j p(i)K(i \rightarrow j) = p(i)$$

et on a donc $Z_{t+1} \sim p$. □

DÉFINITION 10.10 (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

$$\alpha_t = \min \left(1, \frac{p(y_{t+1}) q(z_t|y_{t+1})}{p(z_t) q(y_{t+1}|z_t)} \right)$$

- 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 10.11 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 $x \rightarrow x'$, 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 $x' \rightarrow x$, la probabilité d'acceptation aurait été 1). On a

$$\frac{K(x \rightarrow x')}{K(x' \rightarrow x)} = \frac{q(x'|x)\alpha_t}{q(x|x')} = \frac{p(x')}{p(x)}$$

et le bilan détaillé est bien vérifié. \square

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^{MCMC} converge en probabilité. Étudions sa variance :

$$\text{var} \left(\hat{I}_T^{\text{MCMC}} \right) = \frac{1}{T} \text{var}(h(Z)) + \frac{2}{T^2} \sum_{t=1}^T \sum_{t'=1}^{t-1} \text{cov} (h(Z_t), h(Z_{t'})).$$

Comme Z_t converge en loi, on a à l'asymptote $\text{cov} (h(Z_t), h(Z_{t'})) = \text{cov} (h(Z_{t+k}), h(Z_{t'+k}))$ pour tout k et chaque terme apparaît donc environ T fois dans la double somme, d'où

$$\begin{aligned} \text{var} \left(\hat{I}_T^{\text{MCMC}} \right) &\approx \frac{1}{T} \text{var}(h(Z)) + \frac{2}{T} \sum_{k=1}^T \text{cov} (h(Z_t), h(Z_{t+k})) \text{ pour un } t \text{ arbitraire suffisamment grand} \\ &\approx \text{var}(h(Z)) \left(\frac{1}{T} + \frac{2}{T} \sum \rho_k \right) \text{ où } \rho_k = \text{cor}(h(Z_t), h(Z_{t+k})) \\ &\approx \frac{1}{M} \text{var}(h(Z)) \end{aligned}$$

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^{MCMC} est de même qualité qu'un estimateur Monte-Carlo \hat{I}_M^{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 $\sigma \rightarrow 0$, 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 $\sigma \rightarrow \infty$, 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, !REF! ont montré 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.

10.7 REMARQUES BIBLIOGRAPHIQUES

L'ouvrage [1] 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 [2].

En statistique non-paramétrique, l'étude fréquentiste des méthodes bayésiennes est un sujet très actif [MR3587782].

Les méthodes de Monte-Carlo jouent un rôle central en probabilités numériques [MR1842342]. Les propriétés de mélange des chaînes de Markov sont étudiées dans [MR2466937].

Références

- [1] 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).
- [2] A. VAN DER VAART. **Asymptotic statistics**. Cambridge University Press, 1998.

11.1 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 $\lim_{t \rightarrow \infty} \bar{F}(tx)/\bar{F}(t)$ 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 \hat{f}_n .

11.2 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 \hat{f}_n :

$$L(\hat{f}_n, f) := \|f - \hat{f}_n\|_1 = \int_{\mathbb{R}} |f(x) - \hat{f}_n(x)| dx.$$

Ce n'est pas la seule façon de faire, il existe une littérature abondante où on s'intéresse à la perte quadratique :

$$\int_{\mathbb{R}} |f(x) - \hat{f}_n(x)|^2 dx.$$

On peut aussi s'intéresser à la distance de Hellinger, ou plutôt à son carré :

$$\int_{\mathbb{R}} |\sqrt{f(x)} - \sqrt{\hat{f}_n(x)}|^2 dx.$$

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 11.1 Soient P, Q deux lois de probabilité sur un univers (Ω, \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 ,

$$\sup_{A \in \mathcal{F}} P(A) - Q(A) =: d_{\text{TV}}(P, Q) = \frac{1}{2} \int_{\Omega} |f(x) - g(x)| d\nu(x) = \int_{\Omega} (f(x) - g(x))_+ d\nu(x)$$

PREUVE. Comme $\int_{\Omega} f(x) d\nu(x) = \int_{\Omega} g(x) d\nu(x) = 1$, la seconde égalité est immédiate.

Si on définit A par $A := \{x : f(x) > g(x)\}$,

$$P(A) - Q(A) = \int_A f(x) - g(x) d\nu(x) = \int_{\Omega} (f(x) - g(x))_+ d\nu(x).$$

On peut donc déjà conclure que

$$d_{\text{TV}}(P, Q) \geq \int_{\Omega} (f(x) - g(x))_+ d\nu(x).$$

Soit B une partie mesurable

$$\begin{aligned} P(B) - Q(B) &= \int_{A \cap B} f(x) - g(x) d\nu(x) + \int_{A^c \cap B} f(x) - g(x) d\nu(x) \\ &= \int_{A \cap B} (f(x) - g(x))_+ d\nu(x) - \int_{A^c \cap B} (f(x) - g(x))_- d\nu(x) \\ &\leq \int_A (f(x) - g(x))_+ d\nu(x) \\ &= \int_{\Omega} (f(x) - g(x))_+ d\nu(x). \end{aligned}$$

Soit

$$d_{\text{TV}}(P, Q) \leq \int_{\Omega} (f(x) - g(x))_+ d\nu(x).$$

□

LEMME 11.2 (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

$$\lim_{n \rightarrow \infty} \int_{\Omega} (f(x) - f_n(x))_+ d\nu(x) = 0.$$

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. □

11.3 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 $b_1 \leq \dots \leq b_k$. Pour que la construction soit sans défauts, on se place dans le cas où aucun point de l'échantillon n'appartient à $(-\infty, b_1] \cup [b_k, \infty) \cup \{b_2, \dots, b_{k-1}\}$. 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.

11.4 FENÊTRES GLISSANTES

Un résultat de consistance

Dans la suite, un noyau sera une fonction intégrable K ($\int_{\mathbb{R}} |K(x)| dx < \infty$), on dira qu'il est positif si $K(x) \geq 0$ en tout x , et un noyau positif est une densité si $\int_{\mathbb{R}} K(x) dx = 1$. Pour la largeur de bande $h > 0$, $K_h(x) := \frac{1}{h} K(\frac{x}{h})$. On a bien sûr $\int_{\mathbb{R}} K_h(x) dx = \int_{\mathbb{R}} K(x) dx$.

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

$$\hat{f}_n(x) := \frac{1}{2h} (F_n(x+h) - F_n(x-h)).$$

Dans la suite, la largeur de bande notée h_n dépendra généralement de la taille de l'échantillon n et on conviendra de

$$\hat{f}_n(x) := \frac{1}{2h_n} (F_n(x+h_n) - F_n(x-h_n)).$$

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

$$F(x+h) - F(x-h) = \int_{x-h}^{x+h} f(y) dy.$$

Cela nous suffira par la suite. Si on veut étendre les résultats qui suivent à l'estimation par noyaux en dimension $d \geq 2$, il est utile d'avoir en tête le théorème suivant.

THÉORÈME 11.3 (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 $Q \in \mathcal{Q}$*

$$\frac{\text{Leb}(Q^*)}{\text{Leb}(Q)} \leq c$$

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 $x \in \mathbb{R}^d$,

$$\lim_{r \searrow 0} \sup_{Q \in \mathcal{Q}_r} \left| \frac{1}{\text{Leb}(Q)} \int_{x+Q} f(y) dy - f(x) \right| = 0.$$

L'ensemble des points où la convergence a lieu est appelé l'ensemble des points de Lebesgue.

En dimension $d \geq 2$, on peut choisir comme ensemble \mathcal{Q} , les boules pour la norme $L_p, p \in \{1, 2, \infty\}$.

PROPOSITION 11.4 *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*

$$\mathbb{E}_{P^{\otimes n}} \left[(\hat{f}_n(x) - f(x))^2 \right] = 0.$$

PREUVE. Le risque quadratique de $\hat{f}_n(x)$ admet une décomposition biais/variance.

$$\mathbb{E}_{P^{\otimes n}} \left[(\hat{f}_n(x) - f(x))^2 \right] = \text{var} \left(\hat{f}_n(x) \right) + \left(\mathbb{E} \left[\hat{f}_n(x) \right] - f(x) \right)^2.$$

La variance est traitée en notant que $\hat{f}_n(x)$ est une moyenne de variables aléatoires indépendantes identiquement distribuées comme $K_{h_n}(x - X)$,

$$\text{var} \left(\hat{f}_n(x) \right) = \frac{1}{n} \text{var} \left(K_{h_n}(x - X) \right).$$

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

$$\text{var} \left(\hat{f}_n(x) \right) = \frac{1}{2nh_n} \frac{F(x + h_n) - F(x - h_n)}{2h_n} (1 - F(x + h_n) - F(x - h_n)).$$

Comme x est supposé être un point de Lebesgue, si $\lim_n h_n = 0$,

$$\text{var} \left(\hat{f}_n(x) \right) \sim \frac{f(x)}{2nh_n} \quad \text{quand } n \rightarrow \infty.$$

Si de plus $\lim_n nh_n = \infty$, la variance tend vers O .

Comme $\mathbb{E} \hat{f}_n(x) = K_{h_n} * f(x)$, le biais s'écrit

$$\mathbb{E} \left[\hat{f}_n(x) \right] - f(x) = K_{h_n} * f(x) - f(x) = \frac{F(x + h_n) - F(x - h_n)}{2h_n} - f(x)$$

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 11.5 *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*

$$\lim_{n \rightarrow \infty} \mathbb{E}_{P^{\otimes n}} \left[\int_{\mathbb{R}} \left| \hat{f}_n(x) - f(x) \right| dx \right] = 0.$$

PREUVE.

$$\begin{aligned} \mathbb{E}_{P^{\otimes n}} \left[\int_{\mathbb{R}} \left| \hat{f}_n(x) - f(x) \right| dx \right] &= 2 \mathbb{E}_{P^{\otimes n}} \left[\int_{\mathbb{R}} \left(f(x) - \hat{f}_n(x) \right)_+ dx \right] \\ &= 2 \int_{\mathbb{R}} \mathbb{E}_{P^{\otimes n}} \left[\left(f(x) - \hat{f}_n(x) \right)_+ \right] dx. \end{aligned}$$

En tout x ,

$$\mathbb{E}_{P^{\otimes n}} \left[\left(f(x) - \hat{f}_n(x) \right)_+ \right] \leq f(x).$$

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 11.6 Ce résultat de consistance universelle doit être tempéré. La décroissance du risque peut être arbitrairement lente. La preuve souligne le fait que nous ne maîtrisons 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.

Fenêtres glissantes et histogrammes

11.5 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 11.7 (ESTIMATEUR À NOYAU) Soit K une fonction intégrable, $h > 0$ une largeur de bande, l'estimateur à noyau défini par K et h est

$$\widehat{f}_n(x) := \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K \left(\frac{x - X_i}{h} \right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i).$$

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, $K(x) = \frac{1}{2}e^{-|u|/\sqrt{2}} \sin \left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4} \right)$.

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 11.8 (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

$$f * g(x) := \int_{\mathbb{R}} f(x - y)g(y)dy = \int_{\mathbb{R}} f(y)g(x - y)dy.$$

La fonction $f * g$ est elle-même intégrable.

THÉORÈME 11.9 (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.

$$\begin{aligned}
 \|f * g\|_1 &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} f(x-y) \times g(y) dy \right| dx \\
 &\leq \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x-y) \times g(y)| dy dx \\
 &= \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x-y)| dx \times g(y) dy \\
 &= \int_{\mathbb{R}} \|f\|_1 \times g(y) dy \\
 &= \|f\|_1 \times \|g\|_1 .
 \end{aligned}$$

□

REMARQUE 11.10 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 11.11 *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 11.12 *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 11.13 (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

$$K * \mu(x) := \int_{\mathbb{R}} K(x-y) dF(y) .$$

Si K est une densité, $K * \mu$ est la densité de la loi de $X + Y$ où X et Y sont indépendantes, de lois de densité K et μ respectivement.

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

$$\widehat{f}_n(x) = P_n * K_h(x) .$$

11.6 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 $x \in \mathbb{R}$:

$$\mathbb{E} \left[\widehat{f}_n(x) \right] = K_{h_n} * f(x) = \int_{\mathbb{R}} \frac{1}{h_n} K \left(\frac{x-y}{h_n} \right) f(y) dy .$$

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 11.14 (LIMITES DE CONVOLUÉS) *Si le noyau $K \in L_1$ avec $\int_{\mathbb{R}} |K(x)| dx = 1$ et $K_h(x) := \frac{1}{h} K(\frac{x}{h})$, alors pour tout $f \in L_1$,*

$$\lim_{h \rightarrow 0} \int_{\mathbb{R}} |f * K_h(x) - f(x)| dx = 0 .$$

Pour établir ce théorème, nous partirons des deux propositions suivantes :

PROPOSITION 11.15 *Si le noyau $K \in L_1$ avec $\int_{\mathbb{R}} |K(x)| dx = 1$ et $K_h(x) := \frac{1}{h} K(\frac{x}{h})$, alors pour toute fonction g lipschtzienne et à support borné*

$$\lim_{h \rightarrow 0} \int_{\mathbb{R}} |g * K_h(x) - g(x)| dx = 0 .$$

PROPOSITION 11.16 *Les fonctions lipschtziennes et à support borné sont denses dans L_1 .*

La proposition 11.16 est un corollaire du théorème 11.12.

PREUVE. (Proposition 11.15) Dans la suite g est C -lipschtzienne à support compact inclus dans $[-M, M]$.

On définit un 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

$$\begin{aligned} & \int_{\mathbb{R}} |g * K_h(x) - g(x)| dx \\ & \leq \underbrace{\int_{\mathbb{R}} |g * (K_h - L_h)(x)| dx}_{(a)} + \\ & \quad \underbrace{\int_{\mathbb{R}} \left| g * L_h(x) - g(x) \int_{\mathbb{R}} L_h(y) dy \right| dx}_{(b)} + \\ & \quad \underbrace{\int_{\mathbb{R}} \left| g(x) \int_{\mathbb{R}} (K_h(y) - L_h(y)) dy \right| dx}_{(c)} . \end{aligned}$$

D'après l'inégalité de Young,

$$(a) \leq \|g\|_1 \times \|K_h - L_h\|_1 ,$$

alors que

$$\begin{aligned} (c) &\leq \int_{\mathbb{R}} |g(x)| \int_{\mathbb{R}} |K_h(y) - L_h(y)| dy dx \\ &\leq \int_{\mathbb{R}} |g(x)| dx \times \int_{\mathbb{R}} |K_h(y) - L_h(y)| dy \\ &= \|g\|_1 \times \|K_h - L_h\|_1 . \end{aligned}$$

Comme

$$\begin{aligned} \|K_h - L_h\|_1 &\leq \int_{\mathbb{R}} |K(x)| \mathbb{I}_{|x|>r} dx , \\ (a) + (c) &\leq 2 \|g\|_1 \times \int_{\mathbb{R}} |K(x)| \mathbb{I}_{|x|>r} dx . \end{aligned}$$

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.

$$\begin{aligned} (b) &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} (g(x-y) - g(x)) L_h(y) dy \right| dx \\ &\leq \int_{\mathbb{R}} \int_{-M-rh}^{M+rh} |(g(x-y) - g(x))| dx L_h(y) dy \\ &\leq \int_{-rh}^{rh} \int_{-M-rh}^{M+rh} C|y| dx |L_h(y)| dy \\ &\leq 2(M+rh)Crh \int_{-r}^r |L(y)| dy \\ &\leq 2(M+rh)Crh \int_{\mathbb{R}} |K(y)| dy . \end{aligned}$$

En combinant les majorations de (a), (b) et (c),

$$\limsup_{h \rightarrow 0} \int_{\mathbb{R}} |g * K_h(x) - g(x)| dx \leq 2 \|g\|_1 \times \int_{\mathbb{R}} |K(x)| \mathbb{I}_{|x|>r} dx .$$

En faisant tendre r vers ∞ , on conclut

$$\limsup_{h \rightarrow 0} \int_{\mathbb{R}} |g * K_h(x) - g(x)| dx = 0 .$$

□

L'étude de l'approximation des fonctions intégrables devient alors facile.

PREUVE.(Théorème 11.14) Dans la preuve f est intégrable, et g est Lipschtzienne à support compact.

$$\int_{\mathbb{R}} |f * K_h(x) - f(x)| dx \leq \underbrace{\int_{\mathbb{R}} |(f-g) * K_h(x)| dx}_{(a)} + \underbrace{\int_{\mathbb{R}} |g * K_h(x) - g(x)| dx}_{(b)} + \underbrace{\int_{\mathbb{R}} |g(x) - f(x)| dx}_{(c)} .$$

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 11.15, que

$$\limsup_{h \rightarrow 0} \int_{\mathbb{R}} |f * K_h(x) - f(x)| dx \leq (1 + \|K\|_1)\|g - f\|_1 .$$

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 11.17 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.

THÉORÈME 11.18 (CONSISTANCE UNIVERSELLE POUR LE RISQUE L_1) *Si le noyau $K \in L_1$ est un noyau positif avec $\int_{\mathbb{R}} K(x)dx = 1$ (K est une densité) et $\kappa := \int_{\mathbb{R}} K^2(x)dx < \infty$, si $h_n \rightarrow 0$ et $nh_n \rightarrow \infty$, alors pour toute densité f sur \mathbb{R} , en définissant $f_n := f * K_{h_n}$,*

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}} |\widehat{f}_n(x) - f(x)| dx \right] = 0.$$

L'argument reprend l'analyse déjà effectuée à propos de la méthode des fenêtres glissantes.

PREUVE.

$$\begin{aligned} & \mathbb{E} \left[\int_{\mathbb{R}} |\widehat{f}_n(x) - f(x)| dx \right] \\ & \leq \mathbb{E} \left[\int_{\mathbb{R}} |\widehat{f}_n(x) - f * K_{h_n}(x)| dx \right] + \mathbb{E} \left[\int_{\mathbb{R}} |f(x) - f * K_{h_n}(x)| dx \right] \\ & = \mathbb{E} \left[\int_{\mathbb{R}} |\widehat{f}_n(x) - f * K_{h_n}(x)| dx \right] + \int_{\mathbb{R}} |f(x) - f * K_{h_n}(x)| dx. \end{aligned}$$

Le théorème 11.14 garantit que le second terme tend vers 0 quand $h_n \rightarrow 0$.

Nous allons maintenant nous attacher à prouver que le premier terme tend vers 0 quand $nh_n \rightarrow \infty$ et $h_n \rightarrow 0$.

$$\begin{aligned} & \mathbb{E} \left[\int_{\mathbb{R}} |\widehat{f}_n(x) - f * K_{h_n}(x)| dx \right] \\ & = 2 \int_{\mathbb{R}} \mathbb{E} \left[\left(f * K_{h_n}(x) - \widehat{f}_n(x) \right)_+ \right] dx \\ & \leq 2 \int_{\mathbb{R}} \min \left(\mathbb{E} \left[\left(f * K_{h_n}(x) - \widehat{f}_n(x) \right)_+ \right], f * K_{h_n}(x) \right) dx \\ & \leq 2 \int_{\mathbb{R}} \min \left(\sqrt{\mathbb{E} \left[\left(f * K_{h_n}(x) - \widehat{f}_n(x) \right)^2 \right]}, f * K_{h_n}(x) \right) dx \\ & = 2 \int_{\mathbb{R}} \min \left(\sqrt{\text{var} \left(\widehat{f}_n(x) \right)}, f * K_{h_n}(x) \right) dx \\ & \leq 2 \int_{\mathbb{R}} \min \left(\sqrt{\text{var} \left(\widehat{f}_n(x) \right)}, f(x) \right) dx + 2 \int_{\mathbb{R}} |f(x) - f * K_{h_n}(x)| dx. \end{aligned}$$

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.

$$\begin{aligned} \text{var} \left(\widehat{f}_n(x) \right) & \leq \frac{1}{n} \int_{\mathbb{R}} K_{h_n}(y-x)^2 f(y) dy \\ & \leq \frac{1}{nh_n} \int_{\mathbb{R}} K_{h_n}^2(y-x) f(y) dy. \end{aligned}$$

Comme $\kappa = \int_{\mathbb{R}} K^2(y)dy$, K^2/κ est normalisé, et on peut lui appliquer le théorème 11.14

$$\lim_{h_n \searrow 0} \frac{1}{\kappa} \int_{\mathbb{R}} K_{h_n}^2(y-x) f(y) dy = f(x).$$

et même

$$\lim_{h_n \searrow 0} \int_{\mathbb{R}} \left| \frac{1}{\kappa} \int_{\mathbb{R}} K_{h_n}^2(y-x)f(y)dy - f(x) \right| dx = 0.$$

Si $nh_n \nearrow \infty$, on a donc en tout x , $\lim_n \text{var}(f_n(x)) = 0$. Ce qui suffit pour conclure. □

11.7 VITESSE DE CONVERGENCE

Le résultat de consistance universelle n'est malheureusement pas accompagné d'une garantie d'uniformité.

THÉORÈME 11.19 (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 ,*

$$\sup_{f: \|f\|_1=1, f \geq 0} \inf_{h > 0} \mathbb{E} \left[\int_{\mathbb{R}} |K_h * P_n(x) - f(x)| dx \right] = 2.$$

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 11.20 (CLASSE DE DENSITÉ W)

$$W := \left\{ f : f, f' \text{ absolument continues et } \int |f''| < \infty \text{ et } \int \sqrt{f^*} < \infty \right\}$$

REMARQUE 11.21 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 :

$$\int_{\mathbb{R}} \sqrt{f(x)} dx \leq \sqrt{|I|},$$

et donc

$$\int_{\mathbb{R}} \sqrt{f(x)} dx \leq \sqrt{s(f)}.$$

Si on note $c(f) := \int_{\mathbb{R}} |f''(x)| dx$, pour $c > 0$, on définit $W_c \subset W$ par

$$W_c := \{f : f \in W, \quad s(f)^2 \times c(f) < c\}.$$

Deux paramètres importants permettent de quantifier les performances d'un noyau.

$$\text{sd}(K) := \int_{\mathbb{R}} x^2 K(x) dx \quad R(K) := \int_{\mathbb{R}} K(x)^2 dx.$$

THÉORÈME 11.22 *Soit K un noyau symétrique, borné à support inclus dans $[-1, 1]$ tel que $\int_{-1}^1 K(y) dy = 1$ avec $\text{sd}(K) < \infty$ 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*

$$\mathbb{E} \left[\int_{\mathbb{R}} |f(x) - \hat{f}_n(x)| dx \right] \leq \kappa(K) \frac{c^{1/5}}{n^{2/5}}.$$

On peut choisir $\kappa(K) \leq 2(4 \text{sd}(K) R(K)^2)^{1/5}$.

Dans la preuve du théorème 11.22, nous utiliserons la notion suivante.

DÉFINITION 11.23 (NOYAU ASSOCIÉ) Etant donné un noyau intégrable K , le noyau L associé à K est défini par

$$L(x) = \begin{cases} \int_x^\infty (y-x)K(y)dy & \text{pour } x \geq 0 \\ L(-x) & \text{pour } x < 0. \end{cases}$$

EXEMPLE 11.24 Pour le noyau rectangulaire $\frac{1}{2}\mathbb{I}_{[-1,1]}$, le noyau associé est $\frac{1}{4}(1-|x|)_+^2$.

PROPOSITION 11.25 Si K est un noyau intégrable et L son noyau associé, alors

- i) L est pair (symétrique) ;
- ii) $\int_{\mathbb{R}} L(x)dx = \int_0^\infty x^2 K(x)dx$;
- iii) Si K est pair, $\int_{\mathbb{R}} L(x)dx = \frac{1}{2} \int_{\mathbb{R}} x^2 K(x)dx$;
- iv) Si K est pair, $\int_{\mathbb{R}} |L(x)|dx \leq \frac{1}{2} \int_{\mathbb{R}} x^2 |K(x)|dx$;

PREUVE. La parité est inscrite dans la définition.

On a

$$\begin{aligned} \int_{\mathbb{R}} L(x)dx &= 2 \int_0^\infty \int_0^\infty \mathbb{I}_{y>x}(y-x)K(y)dydx && \text{parité de } L \\ &= \int_0^\infty \int_0^y 2(y-x)dxK(y)dy && \text{Fubini} \\ &= \int_0^\infty y^2 K(y)dy. \end{aligned}$$

Si K est pair,

$$\int_0^\infty y^2 K(y)dy = \frac{1}{2} \int_{\mathbb{R}} y^2 K(y)dy,$$

ce qui établit la deuxième assertion.

Si K est pair, en invoquant encore la parité de L , puis le théorème de Fubini,

$$\begin{aligned} &\int_{\mathbb{R}} |L(x)|dx \\ &= 2 \int_0^\infty \left| \int_0^\infty \mathbb{I}_{y>x}(y-x)K(y)dy \right| dx \\ &\leq 2 \int_0^\infty \int_0^\infty |\mathbb{I}_{y>x}(y-x)K(y)| dydx \\ &= \int_0^\infty \int_0^y 2(y-x)dx |K(y)|dy \\ &= \int_0^\infty y^2 |K(y)|dy \\ &= \frac{1}{2} \int_{\mathbb{R}} y^2 |K(y)|dy. \end{aligned}$$

Ceci établit la troisième assertion. □

PREUVE.(Théorème 11.22)

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 $\int_{\mathbb{R}} |f(x) - K_h * f(x)| dx$.

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 $x > 0$

$$\begin{aligned} (L)_h(x) &= \int_x^\infty (y-x) \frac{1}{h} K\left(\frac{y}{h}\right) dy \\ &= \int_x^\infty \frac{y-x}{h} K\left(\frac{y}{h}\right) dy \\ &= h^2 \times \frac{1}{h} \int_{x/h}^\infty (y-x/h) K(y) dy \\ &= h^2 \times \frac{1}{h} L\left(\frac{x}{h}\right) \\ &= h^2 L_h(x) . \end{aligned}$$

Comme on suppose K symétrique et $\int_{-1}^1 K(y) dy = 1$,

$$\begin{aligned} K_h * f(x) - f(x) &= \int_{\mathbb{R}} (f(x-y) - f(x)) K_h(y) dy \\ &= \int_{\mathbb{R}} (f(x+y) - f(x)) K_h(y) dy . \end{aligned}$$

Comme f et f' sont supposées absolument continues, pour presque tout couple (x, y) ,

$$f(x+y) - f(x) = y f'(x) + \int_x^{x+y} (x+y-v) f''(v) dv$$

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 entraîne $\int_{\mathbb{R}} y K(y) dy = 0$).

$$\begin{aligned} &K_h * f(x) - f(x) \\ &= \int_{\mathbb{R}} \int_x^{x+y} (x+y-v) f''(v) dv K_h(y) dy \\ &= \int_x^\infty \int_{v-x}^\infty (x+y-v) f''(v) K_h(y) dy dv - \int_{-\infty}^x \int_{-\infty}^{v-x} (x+y-v) f''(v) K_h(y) dy dv \\ &= \int_x^\infty f''(v) \int_{v-x}^\infty (x+y-v) K_h(y) dy dv - \int_{-\infty}^x f''(v) \int_{-\infty}^{v-x} (x+y-v) K_h(y) dy dv \\ &= \int_x^\infty f''(v) (L)_h(v-x) dv + \int_{-\infty}^x f''(v) (L)_h(v-x) dv \\ &= \int_{\mathbb{R}} f''(v) (L)_h(x-v) dv \\ &= h^2 f'' * L_h(x) . \end{aligned}$$

En intégrant et en invoquant l'inégalité de Young et la proposition 11.25, on obtient une majoration du biais qui fait intervenir largeur de bande, régularité de la densité, et moments du noyau K :

$$\begin{aligned} \int_{\mathbb{R}} |K_h * f(x) - f(x)| dx &\leq h^2 \int_{\mathbb{R}} |f'' * L_h(x)| dx \\ &\leq h^2 \int_{\mathbb{R}} |f''| dx \times \int_{\mathbb{R}} |L_h(x)| dx \\ &\leq \frac{h^2}{2} \int_{\mathbb{R}} |f''| dx \times \int_{\mathbb{R}} x^2 |K(x)| dx \\ &= \frac{h^2}{2} \times c(f) \times \text{sd}(K) . \end{aligned}$$

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 Hölder,

$$\begin{aligned} \mathbb{E} \left[\left| \widehat{f}_n(x) - f * K_h(x) \right| \right] &\leq \left(\frac{\text{var}(K_h(x - X))}{n} \right)^{1/2} \\ &\leq \left(\frac{\mathbb{E} [K_h(x - X)^2]}{n} \right)^{1/2} \\ &= \left(\frac{K_h^2 * f(x)}{nh} \right)^{1/2} \end{aligned}$$

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 :

$$\begin{aligned} \int_{\mathbb{R}} \left(\frac{K_h^2 * f(x)}{nh} \right)^{1/2} dx &= |J| \times \int_J \frac{1}{|J|} \left(\frac{K_h^2 * f(x)}{nh} \right)^{1/2} dx \\ &\leq \sqrt{|J|} \times \left(\int_J \frac{K_h^2 * f(x)}{nh} dx \right)^{1/2}. \end{aligned}$$

En revenant à notre objectif initial,

$$\begin{aligned} \mathbb{E} \left[\int_{\mathbb{R}} \left| \widehat{f}_n(x) - f * K_h(x) \right| dx \right] &= \int_{\mathbb{R}} \mathbb{E} \left[\left| \widehat{f}_n(x) - f * K_h(x) \right| \right] dx \\ &\leq \int_{\mathbb{R}} \left(\frac{K_h^2 * f(x)}{nh} \right)^{1/2} dx \\ &\leq \frac{\sqrt{s(f) + 2h}}{\sqrt{nh}} \sqrt{\int_{\mathbb{R}} K_h^2 * f(x) dx} \\ &\leq \frac{\sqrt{s(f) + 2h}}{\sqrt{nh}} \sqrt{\|K_h^2\|_1 \times \|f\|_1} \\ &\leq \frac{\sqrt{s(f) + 2h}}{\sqrt{nh}} \sqrt{R(K)}. \end{aligned}$$

En combinant la majoration du biais et celle des fluctuations, on obtient :

$$\mathbb{E} \left[\int_{\mathbb{R}} \left| f(x) - \widehat{f}_n(x) \right| dx \right] \leq \frac{\sqrt{s(f) + 2h}}{\sqrt{nh}} \sqrt{R(K)} + \frac{h^2}{2} c(f) \times \text{sd}(K)$$

On impose $2h \leq s(f)$. On peut optimiser $\frac{\sqrt{2s(f)}}{\sqrt{nh}} \sqrt{R(K)} + \frac{h^2}{2} c(f) \times \text{sd}(K)$ en h . On se contente d'équilibrer les deux termes en choisissant

$$h = \left(\frac{8s(f)R(K)}{nc(f)^2 \text{sd}(K)^2} \right)^{1/5}$$

Pour n assez grand, la contrainte $2h \leq s(f)$ est respectée. On obtient

$$\mathbb{E} \left[\int_{\mathbb{R}} \left| f(x) - \widehat{f}_n(x) \right| dx \right] \leq 2 \frac{(s(f)^2 c(f))^{1/5}}{n^{2/5}} (4R(K)^2 \text{sd}(K))^{1/5} \leq 2 \frac{c^{1/5}}{n^{2/5}} (4R(K)^2 \text{sd}(K))^{1/5}.$$

□

REMARQUE 11.26 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 équilibre du biais et de la variance.

Le théorème 11.22 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.
- 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.

Il y a donc un (assez) long chemin à parcourir pour transformer un résultat comme le théorème 11.22 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 conduit à des estimées sensiblement différentes.

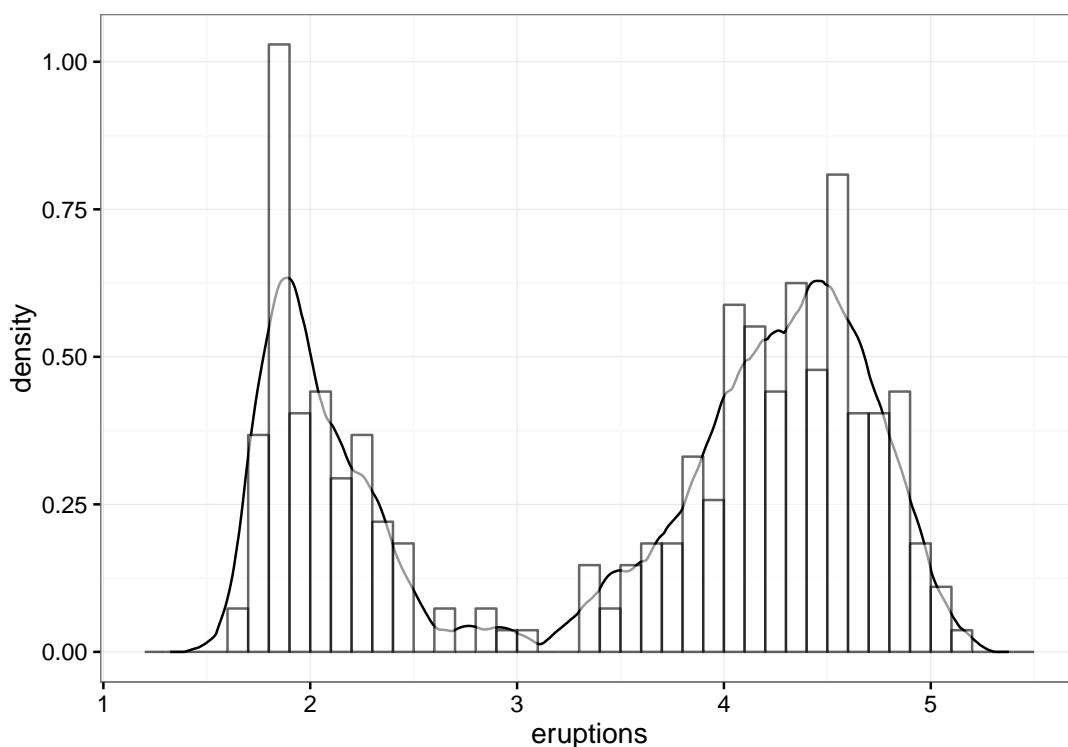


FIG. 11.1 : Estimation de densité sur les données `faithful` à l'aide de noyaux d'Epanechnikov en sélectionnant la largeur de bande avec la méthode de Sheather-Jones

```
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")
```

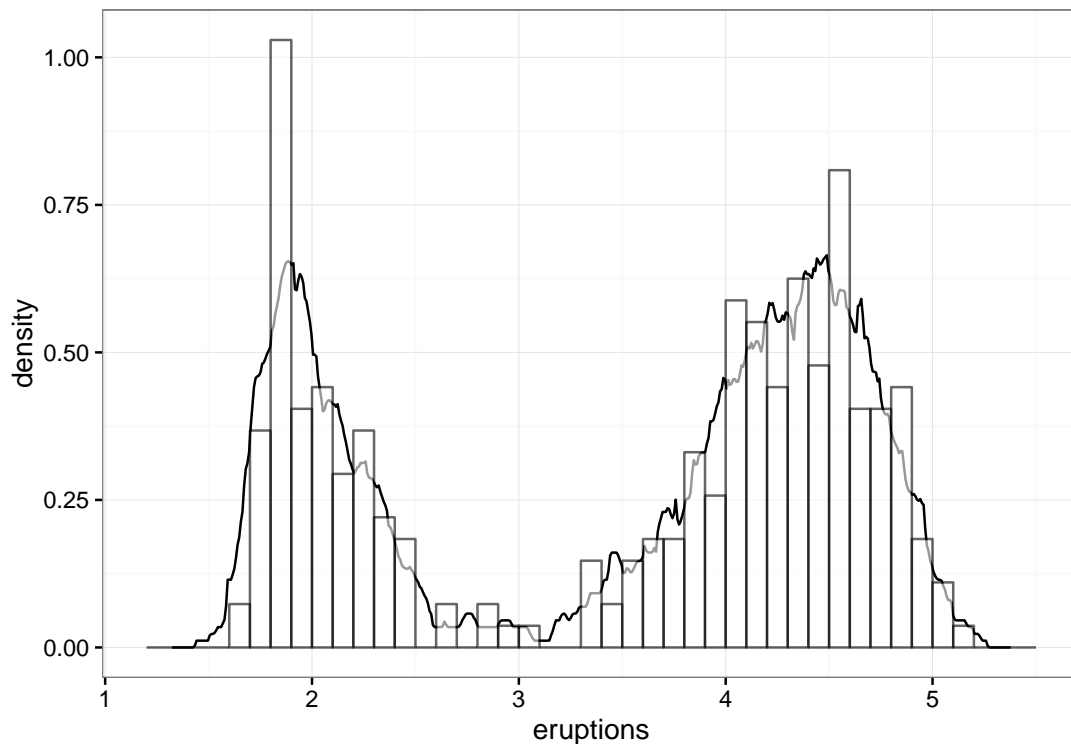


FIG. 11.2 : Estimation de densité sur les données *faithful* à l'aide de noyaux rectangulaires en sélectionnant la largeur de bande avec la méthode de Sheather-Jones

11.8 REMARQUES BIBLIOGRAPHIQUES

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.

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 $\mathcal{M}_{n,n}$*

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

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

$$\vec{v} := \left[\begin{array}{c} -\mathbf{A}^{-1}\mathbf{B}^t\vec{u} \\ \vec{u} \end{array} \right]$$

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

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

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 :

$$\mathbf{K} = \left[\begin{array}{c|c} \mathbf{L}_1 & 0 \\ \hline \mathbf{B}(\mathbf{L}_1^t)^{-1} & \mathbf{L}_2 \end{array} \right] \times \left[\begin{array}{c|c} \mathbf{L}_1^t & \mathbf{L}_1^{-1}\mathbf{B}^t \\ \hline 0 & \mathbf{L}_2^t \end{array} \right].$$

□

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 , $\mathbf{U}^t\mathbf{U} = \mathbf{I}_n$ et $\mathbf{V}^t\mathbf{V} = \mathbf{I}_p$.
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

$$\|\mathbf{A}\|_{\text{op}} = \sup \{ u^t \mathbf{A} v : u \in \mathbb{R}^n, v \in \mathbb{R}^p, \|u\| \leq 1, \|v\| \leq 1 \}.$$

PREUVE. On note $\sigma_1 := \|\mathbf{Z}\|_{\text{op}}$. En dimension finie, les boules unité sont compactes, il existe donc $\hat{u} \in \mathbb{R}^n$ et $\hat{v} \in \mathbb{R}^p$ de normes 1, tels que $\sigma_1 = \hat{u}^t \mathbf{Z} \hat{v}$. Soit $\mathbf{A} \in \mathcal{M}_{n,n-1}$, $\mathbf{B} \in \mathcal{M}_{p,p-1}$ des matrices telles que $[\hat{u} \mid \mathbf{A}]$ et $[\hat{v} \mid \mathbf{B}]$ soient *orthogonales* de dimensions $n \times n$ et $p \times p$

On décompose les matrices en blocs :

$$\begin{bmatrix} \hat{u}^t \\ \mathbf{A}^t \end{bmatrix} \times \mathbf{Z} \times [\hat{v} \mid \mathbf{B}] = \begin{bmatrix} \sigma_1 & w^t \\ \mathbf{0} & \mathbf{A}^t \mathbf{Z} \mathbf{B} \end{bmatrix} =: \mathbf{Y}.$$

En effet $\mathbf{A}^t \mathbf{Z} \hat{v} = \mathbf{0}$, car $\mathbf{Z} \hat{v}$ est colinéaire avec \hat{u} et les colonnes de \mathbf{A} sont orthogonales à \hat{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,

$$\|\mathbf{Y}\|_{\text{op}} \geq \frac{\left\| \mathbf{Y} \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|}{\left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|} \geq \frac{\sigma_1^2 + w^t w}{\sqrt{\sigma_1^2 + w^t w}} = \sqrt{\sigma_1^2 + w^t w}.$$

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.

$$\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}'^t \end{bmatrix} \times \begin{bmatrix} \hat{u}^t \\ \mathbf{A}^t \end{bmatrix} \times \mathbf{Z} \times \begin{bmatrix} \hat{v} & \mathbf{B} \end{bmatrix} \times \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}' \end{bmatrix} = \begin{bmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}' \end{bmatrix}$$

Les matrices

$$\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}'^t \end{bmatrix} \times \begin{bmatrix} \hat{u}^t \\ \mathbf{A}^t \end{bmatrix} \quad \text{et} \quad \begin{bmatrix} \hat{v} & \mathbf{B} \end{bmatrix} \times \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}' \end{bmatrix}$$

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 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 symétriques 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, \dots, e_p of \mathbb{R}^p et une suite décroissante de réels $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ 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 $\lambda_1 \geq \dots \geq \lambda_n$.

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 appelée le *spectre* de la matrice. \mathbf{K} est symétrique non négative (resp. positive) ssi ses valeurs propres sont non-négatives (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

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\text{HS}} = \text{Trace}(\mathbf{A} \times \mathbf{B}^t)$$

(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, \dots, u_n et v_1, \dots, 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.

$$\begin{aligned} \langle u_i v_j^t, u_k v_\ell^t \rangle_{\text{HS}} &= \text{Trace}(u_i v_j^t v_\ell u_k^t) \\ &= \langle v_j, v_\ell \rangle \text{Trace}(u_i u_k^t) \\ &= \langle v_j, v_\ell \rangle \times \langle u_i, u_k \rangle \end{aligned}$$

□

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} ,

$$\mathbf{Z} = \sum_{i=1}^p \sigma_i u_i v_i^t \quad \text{avec} \quad \sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_p = 0.$$

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 $\sum_{i \leq p} \mu_{i,i} u_i v_i^t$ 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 $k+1 \leq i \leq p$. \square

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 $\mathbf{W}y = 0$.

$$\|\mathbf{Z}y\|^2 = \sum_{i=1}^{k+1} \sigma_i^2 \langle v_i, y \rangle^2 \geq \sigma_{k+1}^2$$

which proves that $\|\mathbf{Z} - \mathbf{W}\|_2 \geq \sigma_{k+1}$. \square

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 moindres carrés ordinaires sans supposer que \mathbf{A} est de plein rang. Le problème de minimisation n'a plus de solution unique. On peut privilégier 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 définition impliquent que pour toute pseudo-inverse \mathbf{B} , $\mathbf{A}\mathbf{B}$ est la projection orthogonale sur l'espace engendré par les colonnes de \mathbf{A} , et $\mathbf{B}\mathbf{A}$ est la projection orthogonale sur l'espace engendré par les colonnes de \mathbf{B} . \square

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 $\mathbf{X}\mathbf{X}^+$ représente la projection orthogonale sur le sous-espace engendré par les colonnes de \mathbf{X} , $\hat{\theta} = \mathbf{X}^+y$ 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 $\hat{\theta} + \eta$ où η est orthogonal au sous-espace engendré par les colonnes de \mathbf{V} ($\hat{\theta}$ appartient au sous-espace engendré par les colonnes de \mathbf{V} et on le suppose non-nul). La norme euclidienne de $\hat{\theta} + \eta$ est minimisée en choisissant $\eta = 0$. \square

REMARQUE A.15 Si $\mathbf{X}^t \times \mathbf{X}$ est inversible alors

$$(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t = \mathbf{X}^+.$$

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. ISBN : 0-8018-5413-X ; 0-8018-5414-8. MR : 1417720, 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 CALCUL DE DENSITÉ IMAGE

THÉORÈME B.1

B.2 CONVERGENCES

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 B.2 (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 $\epsilon > 0$

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

quand n tend vers $+\infty$.

THÉORÈME B.3 (LOI FORTE DES GRANDS NOMBRES) Si X_1, \dots, X_n, \dots 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

$$\lim_{n \rightarrow \infty} \bar{X}_n = \mu \quad \text{avec} \quad \bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i.$$

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 B.4 (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

$$X_n \rightsquigarrow X \quad \text{ou} \quad X_n \rightsquigarrow \mathcal{L}$$

(\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.

THÉORÈME B.5 (PORTE-MANTEAU) 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 tout sous-ensemble fermé F de \mathbb{R}^k , $\limsup P_n\{F\} \leq P\{F\}$.
- iv) 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 $\mathbb{E}_P[f]$.
- v) ...

Le théorème suivant est un corollaire immédiat du théorème de continuité de Lévy.

THÉORÈME B.6 (CRAMER-WOLD) La loi d'un vecteur aléatoire $(X_1, \dots, X_n)^t$ 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 $(t_1, \dots, t_n)^t$ parcourt \mathbb{R}^n .

B.3 MÉTRISATIONS DE LA CONVERGENCE EN LOI

DÉFINITION B.7 (DISTANCE DE KOLMOGOROV-LÉVY)

$$d_K(W, Z) = \sup_x |\mathbb{P}\{W \leq x\} - \mathbb{P}\{Z \leq x\}|.$$

THÉORÈME B.8

DÉFINITION B.9 (DISTANCE DE TRANSPORT)

$$d_M(W, Z) = \sup_{h \text{ 1-Lipschitz}} |\mathbb{E}h(W) - \mathbb{E}h(Z)|.$$

PROPOSITION B.10

$$d_K(W, Z) \leq \sqrt{2C d_M(W, Z)}.$$

PREUVE. □

THÉORÈME B.11

PREUVE. □

B.4 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 distribués.

THÉORÈME B.12 (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 $\sigma_n^2 = \sum_{i \leq n} \mathbb{E}X_{i,n}^2$),*

$$\frac{1}{\sigma_n^2} \sum_{i \leq n} \mathbb{E}[X_{i,n}^2 \mathbb{I}_{|X_{i,n}| > \epsilon \sigma_n}] \xrightarrow{n \rightarrow \infty} 0$$

alors

$$\frac{1}{\sigma_n} \left(\sum_{i \leq n} X_{i,n} \right) \rightsquigarrow \mathcal{N}(0, 1).$$

THÉORÈME B.13 (INÉGALITÉ DE BERRY-ESSEEN) *Si X_1, \dots, X_n, \dots 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 $\mathcal{N}(0, 1)$.

Nous donnons une démonstration combinée des deux théorèmes.

Sans perdre en généralité, nous supposons que $\sum_{i=1}^n \mathbb{E}X_{i,n}^2 = 1$ pour tout n . On note $S_n = \sum_{i=1}^n X_{i,n}$. On suppose la condition de Lindeberg vérifiée :

$$\forall \epsilon > 0, \quad \lim_{n \rightarrow \infty} \sum_{i \leq n} \mathbb{E}[X_{i,n}^2 \mathbb{I}_{|X_{i,n}| > \epsilon}] = 0$$

On cherche à établir

$$\sup_{x \in \mathbb{R}} |\mathbb{P}\{S_n \leq x\} - \Phi(x)| \leq C \sum_{i=1}^n \mathbb{E}|X_{i,n}|^3$$

pour une constante universelle C .

B.5 MÉTHODE DELTA

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 $(X_n Y_n)_n$, on ne peut pas affirmer à coup sûr que $X_n Y_n \rightsquigarrow XY$. 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 B.14 (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 B.15 (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 $\mathbb{R}^{k''}$

$$g(X_n, Y_n) \rightsquigarrow g(X, y).$$

Le lemme de Slutsky est un élément de la preuve du théorème suivant appelé *méthode delta*.

THÉORÈME B.16 (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) $a_n (f(X_n) - f(\vec{x})) \rightsquigarrow Df X$.

La notion de famille de lois de probabilités *tendue* est définie par le critère suivant.

DÉFINITION B.17 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

$$\forall P \in \mathcal{C}, \quad P\{K(\epsilon)\} \geq 1 - \epsilon.$$

Le critère de Prokhorov relie tension et convergence étroite (en distribution).

THÉORÈME B.18 *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.*

B.6 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 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.

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.

C.1 THÉORÈMES D'INVERSION

THÉORÈME C.1 (THÉORÈME D'INVERSION LOCALE) *Soit E un espace euclidien, $U \subseteq E$ ouvert, $f : U \rightarrow E$ 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 C.2 (THÉORÈME D'INVERSION GLOBALE) *Soit E un espace euclidien, $U \subseteq E$ ouvert, $f : U \rightarrow E$ 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 $f(U)$)*

C.2 CONVEXITÉ