Statistiques

C. Levrard pour la dernière couche,
basée sur les polys de S. Gaiffas et S. Boucheron

Chapitre 1

Modèles statistiques

1.1 Différence de point de vue entre probas et stats

Supposons qu'on s'intéresse au nombre de naissances X en 2022 à Cherbourg.

Point de vue probabiliste

Le probabiliste n'observe pas X, mais va chercher à le modéliser. Si Cherbourg contient N femmes (ou foyers), à partir du taux de fécondité de l'année 2021, il va bricoler un θ_0 et supposer que $X \sim \mathcal{B}(N, \theta_0)$. A partir des quantiles de cette loi, il pourra conseiller la mairie sur le besoin de places en crèche pour l'année 2022. Dans un sens le probabiliste se fiche d'observer X, son modèle lui permet de donner des conseils à la mairie en amont.

Point de vue statisticien

Le statisticien se place à la fin de l'année 2022, il peut donc *observer* le nombre de naissances (disons x_{obs}). Il va reprendre le modèle du probabiliste en supposant que $X \sim \mathcal{B}(N, \theta)$, mais cette fois ci sans supposer que l'on connaît la valeur de θ. Le but du jeu est alors d'approcher un θ à partir de l'observation x_{obs}, que le probabiliste pourra utiliser pour ses prévisions de naissance pour l'année 2023. Par exemple, le statisticien peut affirmer

$$\theta \approx x_{obs}/N,$$

c'est à dire refiler la valeur x_{obs}/N au probabiliste pour ses prévisions ultérieures. Un des buts de la statistique va être de déterminer à quel point cette approximation est pertinente.

En résumé :

— Le probabiliste connaît le modèle génératif et prédit des choses sur les observations à venir.

— Le statisticien observe des *données*, et essaie de dire des choses sur le processus qui les a généré.

Nécessité d'un modèle

Le modèle du probabiliste sert en statistiques à "généraliser" les observations. Dans l'exemple du nombre de naissances en 2022, le modèle $\mathcal{B}(N, \theta)$ permet à partir de l'observation des naissances en 2022 de dire des choses pour l'année 2023 via le paramètre θ que l'on suppose commun aux deux années.

Sans le modèle ,rien n'empêche d'affirmer que X a été tirée selon $\delta_{x_{obs}}$, ce qui correspondrait parfaitement à l'observation en 2022 mais serait de peu d'intérêt pour faire des prévisions pour l'année 2023

Dessin : proba, on connaît le processus génératif, prédire des trucs intéressant sur ce qui va arriver. Stats, c'est arrivé, dire des trucs intéressants sur ce qui à généré.

1.2 Modèle statistique

Definition 1.1

Un modèle statistique est un triplet $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ composé des éléments suivants :

- $\mathcal{X}$: espace d'observation (dans lequel vont vivre nos observations),
- $\mathcal{A}$: tribu sur $\mathcal{X}$, pour donner du sens à ce qui est observable ou non,
- $(P_\theta)_{\theta\in\Theta}$: famille de lois sur $\mathcal{X}$ indexée par $\theta \in \Theta$ (Θ est appelé *espace des paramètres*).

Il est implicitement supposé qu'on observe X v.a. de loi P_θ, pour un θ **inconnu**, et qu'on cherche à estimer $q(\theta)$ à partir de ces observations.

Remarque 1.2. Par convention, on confondra souvent X v.a. de loi P_θ et $x = X(\omega)$ (observation). Cette convention se justifie en statistique "mathématique" par le fait qu'on s'intéresse aux observations uniquement au travers de leur loi pour essayer d'attraper θ. Si on avait à manipuler des "vraies" données, on distinguerait x_{obs} (réalité) de X (qui nous sert à construire une procédure de statistique inférentielle en général).

Remarque 1.3. On peut faire plus général et distinguer espace d'observation et espace probabilisé de départ. Dans ce sens, un modèle est plutôt $(\Omega, \mathcal{F}, \mathcal{X}, X, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, où on spécifie la v.a. de loi P_θ, c-à-d l'observation comme fonction de $\omega \in \Omega$.

Vu qu'on s'intéresse peu aux observations en tant que telles, mais bien plutôt à leurs lois, on peut se passer de $(\Omega, \mathcal{F}, X)$ la plupart du temps.

Remarque 1.4. Par convention, dans les modèles "classiques" où l'espace d'observation et la tribu qui y est associée sont évidents, on peut omettre de les mentionner. Par exemple le modèle $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\theta, \sigma_0^2)^{\otimes n})_{\theta\in\mathbb{R}})$ (tirage de n Guassiennes indépendantes de variance connue et de moyenne inconnue) est souvent abrégé en $(\mathcal{N}(\theta, \sigma_0^2)^{\otimes n})_{\theta\in\mathbb{R}}$.

Remarque 1.5. Si $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ est un modèle, et $g : \Theta \mapsto \Theta$ une bijection, alors $(\mathcal{X}, \mathcal{A}, (P_{g(\theta)})_{\theta\in\Theta})$ en est un autre, qui est équivalent d'un point de vue statistique. Par exemple, dans l'exemple du nombre de naissances à Cherbourg, on peut paramétrer le nombre de naissances X par $\mathcal{B}(N, \theta)$, $\mathcal{B}(N, 1-\theta)$, $\mathcal{B}\left(N, \frac{2}{1+\theta} - 1\right)$. La quantité intéressante (taux de fécondité "annuel" par foyer) sera alors θ dans le premier cas et $q(\theta)$ dans les deux autres cas.

Si on anticipe légèrement, chercher à estimer $q(\theta)$ dans le modèle P_θ revient à chercher à estimer $q \circ g(\theta)$ dans le modèle $P_{g(\theta)}$.

Traditionnellement, on distingue deux grands types de modèles.

Modèle paramétrique

C'est le cas où Θ (l'ensemble des paramètres) est un sous-ensemble de $\mathbb{R}^d$, ou plus généralement est de dimension finie (dépend donc de la définition de "dimension" en toute généralité).

Exemple 1.6 *: Naissances à Cherbourg en* 2022 *(suite).*

Si on observe, pour chacun des N foyers cherbourgeois, la variable Y_i qui vaut 1 si un enfant apparaît dans l'année, 0 sinon. En supposant que les $(Y_i)_{i=1,\dots,N}$ sont indépendants, on se retrouve avec le modèle statistique suivant.

— $\mathcal{X} = \{0,1\}^N$,

— $\mathcal{A} = \mathcal{P}\left(\{0,1\}^N\right)$ (parties de $\{0,1\}^N$),

— $\Theta =]0,1[$. Pour $\theta \in \Theta$, $P_\theta = \mathcal{B}(\theta)^{\otimes N}$.

Exemple 1.7 *: Naissances à Cherbourg en* 2022 *(fin) ?.*

On peut reprendre l'exemple précédent en supposant que l'on observe le nombre total de naissances en 2022 (noté X). Le modèle devient alors

— $\mathcal{X} = [\![0, N]\!]$,

— $\mathcal{A} = \mathcal{P}\left([\![0, N]\!]\right)$,

— $\Theta =]0,1[$. Pour $\theta \in \Theta$, $P_\theta = \mathcal{B}(N, \theta)$.

On imagine bien que, si le but est d'estimer θ (probabilité pour un foyer d'avoir un enfant en 2022), savoir quel foyer a eu un enfant n'est pas une information importante au regard du nombre d'enfants observé. En d'autres termes, le deuxième modèle, bien que contenant moins d'information que le premier (il en décrit une fonction) en général, contient tout autant d'information sur θ que le premier. On formalisera cette intuition plus tard via la notion d'exhaustivité.

Exemple 1.8 *: Taille et poids dans une tranche d'âge.* On observe, sur n individus d'une tranche d'âge, les vecteurs (T_i, P_i) (taille et poids). En supposant que ces observations sont i.i.d. de loi celle d'un vecteur Gaussien, on a le modèle :

— $\mathcal{X} = (\mathbb{R}^2)^n$, $\mathcal{A} = \mathcal{B}((\mathbb{R}^2)^n)$,

— $\Theta = \mathbb{R}^2 \times S_{2,+}(\mathbb{R})$. Pour $\theta = (\mu, \Sigma)$, $P_\theta = \mathcal{N}(\mu, \Sigma)^{\otimes n}$.

Dans cet exemple θ contient beaucoup d'information diverses concernant le processus générant taille et poids au sein d'une classe d'âge. Si on s'intéresse aux tailles et poids moyens sur l'ensemble de la classe d'âge, on s'intéresse en fait à $q_1(\theta) = \mu$. Si on s'intéresse, au sein d'une tranche d'âge, à la dépendance entre poids et taille, on s'intéresse à $q_2(\theta) = \Sigma$.

Modèle non paramétrique

En toute logique, c'est le cas où l'espace des paramètres n'est pas de dimension finie.

Exemple 1.9 *: Modélisation de la "dépendance" entre taille et poids dans une tranche d'âge.*

Supposons que l'on connaisse la distribution marginale des poids et tailles dans une tranche d'âge (P_1 et P_2, par exemple $\mathcal{N}(\mu_{P/T}, \sigma^2_{P/T})$, et que l'on observe les poids et taille de n individus tirés au hasard dans la tranche d'âge. Le modèle devient :

— $\mathcal{X} = (\mathbb{R}^2)^n$, $\mathcal{A} = \mathcal{B}((\mathbb{R}^2)^n)$,

— $\Theta = \{P \in \mathcal{P}(\mathbb{R}^2) \mid e_j \# P = P_j, j = 1, 2\}$. Pour $\theta \in \Theta$, $P_\theta = \theta^{\otimes n}$.

Remarque : On peut aussi paramétrer la structure de dépendance par une *copule*, c'est à dire une fonction $C : [0,1]^2 \to [0,1]$ telle que $F(t_1, t_2) = C(F_1(t_1), F_2(t_2))$, où F est la fonction de répartition (multivariée) de θ, F_j est la fonction de répartition (univariée) de P_j. L'ensemble des fonctions copules permettant de relier P_1 et P_2 est bien de dimension non finie.

Exemple 1.10 *: Modélisation de la pollution atmosphérique à Paris.*

Si on dispose des mesures quotidiennes de concentration en particules fines à Paris, que l'on suppose les mesures indépendantes et de même loi, et que cette loi commune à une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, le modèle statistique correspondant est

— $\mathcal{X} = R^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$,

— $\Theta = \{\text{densité sur } \mathbb{R}\}$. Pour $\theta \in \Theta$, $P_\theta = (\theta \mathrm{d}\lambda)^{\otimes n}$.

Si maintenant Paris est modélisé disons par un carré $[0,T]^2$, et que pour chaque journée i à 12h on dispose du relevé de concentration en particules fine à chaque endroit $X^{(i)} : [0,T]^2 \mapsto \mathbb{R}^+$. Une modélisation fruste peut être de considérer que les relevés de concentration journalier sont indépendants et de même loi que celle de $X : t \mapsto f(t) + \varepsilon_t$, où f est continue et ε_t est un bruit Gaussien de fonction de covariance connue (disons Ornstein-Uhlenbeck par exemple). Dans ce cas le modèle est

— $\mathcal{X} = \mathcal{C}([O,T]^2, \mathbb{R})^n$, $\mathcal{A} = \mathcal{B}(\mathcal{C}([O,T]^2, \mathbb{R}^+))^{\otimes n}$ (plus petite tribu qui rend les applications coordonnées mesurables),

— $\Theta = \mathcal{C}([O,T]^2, \mathbb{R}^+)$. Pour $\theta \in \Theta$, $P_\theta = (\mathcal{L}(\theta + \varepsilon))^{\otimes n}$.

En pratique, les mesures observées ne sont pas de type fonctionnel, mais plutôt sur un grillage de $[O,T]^2$. Le modèle de bruit Gaussien dans ce cas nous ramène à un modèle Gaussien paramétrique (à structure de covariance connue) de très grande dimension (nombre de pixels).

La plupart du temps on considèrera des modèles correspondant à la réalisation de n variables aléatoires indépendantes et de même loi (i.i.d.). On parle alors de *n-échantillon*. Cela correspond à des modèles de type $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta \in \Theta})$. Dans ce type de modèle, on notera $X_{1:n}$ un vecteur aléatoire de loi $P_\theta^{\otimes n}$.

Dans un modèle (identifiable), la statistique inférentielle (classique) permet de faire trois choses :

1. Trouver une valeur approchée du paramètre θ caché (estimation ponctuelle).
2. Donner une zone de Θ dans laquelle le vrai paramètre θ a des chances de se trouver (intervalle de confiance).
3. Répondre à des questions binaires sur θ (test). Par exemple "θ est-il positif ?".

1.3 Estimation (ponctuelle)

Un peu de vocabulaire pour commencer.

Definition 1.11 : Statistique

Pour un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, une *statistique* est une fonction mesurable sur $(\mathcal{X}, \mathcal{A})$.

En somme, une statistique est une fonction des observations. Le point important est qu'une statistique ne peut pas prendre θ en argument. Ses valeurs ne doivent dépendre du paramètre θ qu'au travers de P_θ. Lorsque cela a du sens, pour une statistique f, on notera

$$E_\theta(f(X)) = \int_{\mathcal{X}} f(u)P_\theta(du),$$

correspondant à l'espérance de $f(X)$ lorsque $X \sim P_\theta$. Si le but est de deviner la valeur de θ à partir des observations, il est naturel de considérer des statistiques à valeurs dans Θ. C'est précisément la définition d'un estimateur.

Definition 1.12 : Estimateur

Dans le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, un estimateur de $q(\theta)$, où $q : \Theta \to \mathcal{Y}$ est juste une statistique à valeur dans $\mathcal{Y}$.

En statistique paramétrique on prendra toujours $\mathcal{Y} = \mathbb{R}^k$. La distinction entre statistique et estimateur est purement affaire de convention : un estimateur est une statistique qui a un but (estimer $q(\theta)$). La notion de statistique en général sera utile dans le cadre des modèles qui comportent une statistique exhaustive.

Exemple 1.13.

— Dans le modèle $(\mathcal{N}(\theta, \sigma_0^2)^{\otimes n})_{\theta\in\mathbb{R}}$, un estimateur standard de θ est la moyenne empirique

$$T(X_{1:n}) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

— Dans le modèle $(\mathcal{U}(]0, \theta[)^{\otimes n})_{\theta>0}$ (tirage de n lois uniformes i.i.d.), deux estimateurs raisonnables de θ sont

$$T_1(X_{1:n}) = 2 * \bar{X}_n,$$
$$T_2(X_{1:n}) = \max_{i=1\dots n} X_i.$$

On peut aussi trouver les notations $\hat{\theta}$, $\hat{\theta}_n$ pour désigner un estimateur de θ (le n rappelle qu'il est basé sur un n-échantillon). Par convention le chapeau est réservé aux statistiques/estimateurs (observables à partir des données).

Risque quadratique

Une manière d'évaluer la qualité d'estimation ponctuelle est de considérer le *risque quadratique* de l'estimateur T.

Definition 1.14

Dans le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, le *risque quadratique* de T est

$$R_T(\theta) = E_\theta\|T(X) - q(\theta)\|^2.$$

Si $E_\theta\|T(X)\|^2 = +\infty$, le risque quadratique de T est aussi infini. Ce n'est pas la seule manière d'évaluer la qualité d'un estimateur, on verra plus finement les manières de comparer les estimateurs dans le chapitre bayésien.

Décomposition biais/variance : On peut décomposer le risque quadratique de la manière suivante :

$$R_T(\theta) = \|E_\theta(T(X)) - q(\theta)\|^2 + \mathrm{Var}_\theta(T(X)),$$

où

- le terme $E_\theta(T(X)) - q(\theta)$ est appelé *biais* de l'estimateur T,
- $Var_\theta(T(X)) = E_\theta\left[\|T(X) - E_\theta(T(X))\|^2\right]$ est la variance de l'estimateur T sous P_θ.

Un estimateur T tel que

$$\forall\theta \in \Theta \quad E_\theta(T(X)) = q(\theta)$$

est dit *non-biaisé*, son risque quadratique se résume alors à sa variance.

Exemple 1.15. Dans le modèle $(\mathcal{U}(]0,\theta[)^{\otimes n})_{\theta\in\Theta}$,

- l'estimateur $T_1(X_{1:n}) = 2\bar{X}_n$ est sans biais, son risque quadratique est

$$R_{T_1}(\theta) = \frac{4}{n}\mathrm{Var}(\mathcal{U}(]0,\theta[)) = \frac{4\theta^2}{n}\mathrm{Var}(\mathcal{U}(]0,1[) = \frac{\theta^2}{3n}.$$

- l'estimateur $T_2(X_{1:n})$ est biaisé. Un calcul simple montre que T_2 a pour densité $nt^{n-1}\theta^{-n}$, on en déduit
  - biais : $\theta - \frac{n\theta}{n+1} = \frac{\theta}{n+1}$,
  - risque quadratique : $R_{T_2}(\theta) = \frac{\theta^2}{(n+1)(n+2)}$.

Dans cet exemple le risque de T_2 est sensiblement meilleur que celui de T_1. De manière générale, l'absence de biais est une propriété souhaitable mais ne garantit pas l'optimalité.

Comportements asymptotiques souhaitables

Lorsque l'on parle de "convergence" d'estimateurs, on se place dans le cas où on observe un n-échantillon i.i.d.. On prendra donc le modèle $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta}$.

Un estimateur étant une fonction mesurable des observations (et cet espace d'observation variant avec n), il serait plus précis de parler de suite d'estimateurs. Encore un abus de langage et une convention.

Une propriété minimale des estimateurs est que lorsque l'information disponible (n, taille d'échantillon) croît, l'estimateur converge vers la valeur souhaitée.

Definition 1.16 : Consistance

Un(e) (suite d') estimateur(s) T (de $q(\theta) \in \mathbb{R}^k$) est dit *consistant* si

$$\forall \theta \in \Theta \quad T(X_{1:n}) \xrightarrow[n\to+\infty]{\mathbb{P}} q(\theta).$$

Lorsque la convergence a lieu p.s. on parle de consistance forte.

Recettes de consistance : Prouver une consistance peut se faire avec la loi des grands nombres où en utilisant la convergence du risque quadratique vers 0.

Exemple 1.17. En reprenant le modèle $(\mathcal{U}(]0,\theta[^{\otimes n})_{\theta>0}$, avec les deux estimateurs T_1 et T_2.

— Comme $E_\theta|X_1| < +\infty$, la loi des grands nombres donne $\bar{X}_n \to_{n\to+\infty} \theta/2$ p.s., on en déduit que T_1 est (fortement consistant).

— Comme $R_{T_2}(\theta) \to_{n\to\infty} 0$, T_2 converge vers θ dans $L_2(P_\theta)$, donc en proba (P_θ), il est lui aussi consistant. On peut montrer la forte consistance en utilisant Borel-Cantelli (exercice).

Definition 1.18 : Normalité asymptotique

Dans le cas où $q(\theta) \in \mathbb{R}$, un estimateur T est dit *asymptotiquement normal en* θ s'il existe une suite r_n positive et $\sigma_\theta^2 > 0$ tels que

$$r_n(T(X_{1:n}) - q(\theta)) \rightsquigarrow_{n\to+\infty} \mathcal{N}(O, \sigma_\theta^2).$$

La normalité asymptotique en θ est la convergence en loi de l'estimateur renormalisé vers une loi normale non dégénérée. La normalité asymptotique désigne la même propriété lorsqu'elle est valide pour tout $\theta \in \Theta$. Enfin on peut étendre la définition en dimension supérieure en requérant une matrice de covariance non-nulle.

La normalité asymptotique n'est pas intéressante en elle-même, l'idée est de chercher le comportement asymptotique de la statistique recentrée pour pouvoir en déduire ultérieurement des garanties en terme de risque asymptotique ou d'intervalle de confiance. Le théorème central limite indique que le comportement asymptotique normal est relativement fréquent.

Exemple 1.19. On reprend notre modèle favori $(\mathcal{U}(]0,\theta[)^{\otimes n}$, et nos deux estimateurs.

— Comme $E_\theta X_1^2 < +\infty$, le théorème central limite donne

$$\sqrt{n}\,(T_1(X_{1:n}) - \theta) \rightsquigarrow \mathcal{N}\left(0, \frac{\theta^2}{3}\right).$$

— Pour T_2, un calcul rapide donne, pour $t > 0$,

$$P_\theta(n(\theta - T_2(X_{1:n})) > t) = \left(1 - \frac{t}{n\theta}\right)^n \mathbb{1}_{t\le n\theta} \longrightarrow_{n\to\infty} \exp(-t/\theta),$$

et donc $n(\theta - T_2(X_{1:n})) \rightsquigarrow \mathcal{E}(\theta^{-1})$. Pas de normalité asymptotique donc, mais on a quand même un comportement asymptotique.

Recettes pour normalité asymptotique : La plupart des normalités asymptotiques se prouvent à l'aide du théorème central limite et de deux outils : le lemme de Slutsky et la Δ-méthode.

Théorème 1.20 : Lemme de Slutsky

> *Soient X_n et Y_n deux suites de vecteurs aléatoires convergeant en loi respectivement vers X et y (vecteur aléatoire constant valant y p.s.). Alors $Y_n \xrightarrow{\mathbb{P}} y$ et $(X_n, Y_n) \rightsquigarrow (X, y)$.*

En particulier le lemme de Slutsky autorise certaines opérations sur les limites en loi. Par exemple $X_n \rightsquigarrow \mathcal{N}(0, \sigma^2)$ et $\hat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$ implique $X_n/\sigma \rightsquigarrow \mathcal{N}(0, 1)$, ce qui sera assez utile pour les intervalles de confiance. Une conséquence du lemme de Slutsky est la "Méthode Δ", permettant de transférer la propriété de normalité asymptotique via fonctionnelle différentiable.

Théorème 1.21 : Δ-méthode

> *Soit $(X_n)_{n\geq 1}$ une suite de variable aléatoires, et $(r_n)_{n\geq 1}$ suite de réels positifs tendant vers $+\infty$ tels que $r_n(X_n - x) \rightsquigarrow X$, pour un $x \in \mathbb{R}$ et X une variable aléatoire sur $\mathbb{R}$. Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction différentiable en x, alors*
>
> $$r_n(g(X_n) - g(x)) \rightsquigarrow_{n\to+\infty} g'(x)X.$$

Proof of Theorem 1.21. Comme $r_n \to +\infty$, une première application du Lemme de Slutsky à $(r_n^{-1}, r_n(X_n - x))$ permet de montrer $X_n \xrightarrow{\mathbb{P}} x$. On peut alors déduire de la différentiabilité de g en x que

$$\frac{g(X_n) - g(x)}{X_n - x} \xrightarrow{\mathbb{P}} g'(x).$$

Le Lemme de Slutsky garantit alors que $(r_n(X_n - x), (g(X_n) - g(x))/(X_n - x)) \rightsquigarrow (X, g'(x))$, et par continuité du produit $r_n(g(X_n) - g(x)) \rightsquigarrow g'(x)X$.

□

Exemple 1.22. Dans le modèle $(\mathcal{E}(\theta)^{\otimes n})_{\theta>0}$ (observations de n v.a. exponentielles de paramètres θ indépendantes) où on cherche à estimer θ. On peut partir du théorème central limite

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\theta}\right) \rightsquigarrow \mathcal{N}\left(0, \frac{1}{\theta^2}\right),$$

et appliquer la méthode Δ avec la fonction $u \mapsto 1/u$ pour montrer que l'estimateur $T(X_{1:n}) = \bar{X}_n^{-1}$ vérifie une normalité asymptotique

$$\sqrt{n}\left(T(X_{1:n}) - \theta\right) \rightsquigarrow -\theta^2\mathcal{N}\left(0, \frac{1}{\theta^2}\right) = \mathcal{N}(0, \theta^2).$$

1.4 Intervalle de confiance

À partir d'un estimateur T de $q(\theta)$, le but est de quantifier l'incertitude liée à cette estimation. Plus précisément on va bâtir à partir de T des régions de $\mathbb{R}^k$ dans lesquelles le vrai paramètre $q(\theta)$ devrait se trouver, avec forte probabilité.

Pour simplifier un peu, on suppose que $\Theta \subset \mathbb{R}$ et $q(\theta) = \theta$

1.4.1 Non asymptotique

Definition 1.23 : Intervalle de niveau de confiance $1-\alpha$

Soit $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ un modèle et $\alpha \in]0,1[$. Un intervalle de confiance (par défaut non asymptotique) de niveau $1-\alpha$ pour θ est un couple de statistiques (T^-, T^+) tel que

$$\forall \theta \in \Theta \quad P_\theta\left(T^- \leq \theta \leq T^+\right) \geq 1-\alpha.$$

Lorsque l'inégalité est une égalité on parle de niveau de confiance *exact*. Prendre $T^- = -\infty$ et $T^+ = +\infty$ garantit toujours un niveau $1-\alpha$, le but implicite est de trouver des intervalles de confiance les plus petits possibles. En ce sens, des intervalles de confiances "croissants" en fonction du niveau de confiance sont naturels.

Dans le cas où Θ n'est pas un sous-ensemble de $\mathbb{R}$, on peut définir plus généralement des régions de confiances (plus nécessairement des intervalles) comme des sous-ensembles aléatoires de Θ, ce qui nécessite d'équiper Θ avec une tribu et de vérifier certaines hypothèses de mesurabilité.

Recettes pour les IC non asymptotiques : Il y a deux manières de faire, à partir d'un estimateur T de θ :

1. soit on connaît la loi d'une *quantité pivotale* (usuellement de type $(T-\theta)/a$, idéalement ne dépendant pas de θ) et on peut en inférer un intervalle de confiance (cas idéal),
2. soit on passe par des inégalités de concentration.

Exemple 1.24. Dans le modèle $\mathcal{U}(]0,\theta[)^{\otimes n}$, avec $T_2(X_{1:n}) = \max_{i=1,\dots,n} X_i$, on peut remarque que la loi de $\frac{T_2}{\theta}$ ne dépend pas de θ, $\frac{T_2}{\theta}$ va jouer le rôle de quantité pivotale.

On sait que, pour $t \in]0,1[$,

$$P_\theta\left(t \leq \frac{T_2}{\theta} \leq 1\right) = 1 - t^n.$$

Pour $\alpha \in]0,1[$, en prenant $t_\alpha = \alpha^{\frac{1}{n}}$, on en déduit

$$\forall \theta > 0 \quad P_\theta\left(\alpha^{\frac{1}{n}} \leq \frac{T_2}{\theta} \leq 1\right) = 1-\alpha,$$

et donc que $\left[T_2, T_2\alpha^{-\frac{1}{n}}\right]$ est un intervalle de confiance pour θ au niveau de confiance α.

On rappelle les inégalités de concentration "classiques" suivantes :

Lemme 1.25 : Markov et Bienaymé-Tchebychev

1. ***(Markov)*** *: Si* $\mathbb{E}|X| < +\infty$*, alors, pour tout* $t \in \mathbb{R}$*,*

$$t\mathbb{P}(X \geq t) \leq \mathbb{E}(X\mathbb{1}_{X\geq t}) \leq \mathbb{E}(|X|).$$

2. ***(Bienaymé-Tchebychev)*** *: Si* $\mathbb{E}(X^2) < +\infty$*, alors, pour tout* $t > 0$*,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Dans le cas particulier où X est une somme de variables indépendantes bornées (ce qui arrive souvent dans un cadre statistique), l'inégalité de Hoeffding est un outil très utile.

THÉORÈME 1.26 : INÉGALITÉ D'HOEFFDING

Soient $X_1, \dots, X_n$ des variables indépendantes telles que, pour tout $i \in [\![1, n]\!]$, $a_i \leq X_i \leq b_i$ p.s., pour $a_i, b_i \in \mathbb{R}$. En notant $S = \sum_{i=1}^n X_i$, on a, pout tout $t \geq 0$,

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

Cette inégalité peut se généraliser aux martingales à accroissements bornés (appelée alors inégalité d'Azuma-Hoeffding), et peut aussi se voir comme une inégalité de déviation sous-Gaussienne (la partie en $\exp(-ct^2)$ est une décroissance de queue de type gaussienne en effet). Avant de passer à la preuve regardons un exemple d'application.

Exemple 1.27 *: Taux d'éclosion des oeufs de pingouins.*

On collecte n oeufs de pingouins fécondés, et on note X_i la variable qui vaut 1 si l'oeuf i éclôt, 0 sinon. On peut modéliser cette expérience par un modèle $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathcal{B}(\theta)^{\otimes n}_{\theta \in]0,1[})$. Le paramètre d'intérêt est alors θ, taux d'éclosion "théorique" de ces oeufs.

Un estimateur sans biais de θ est donné par $T(X_{1:n}) = \bar{X}_n$ (moyenne empirique), qui sous P_θ a pour loi $\mathcal{B}(n, \theta)/n$.

Soit $\alpha > 0$, on cherche un IC de niveau de confiance $1 - \alpha$ pour θ. Bienaymé Tchebychev donne

$$P_\theta\left(|T - \theta| \geq t\right) \leq \frac{\theta(1-\theta)}{nt^2} \leq \frac{1}{4nt^2}.$$

On en déduit que $[T \pm t_\alpha^{BT}]$ est un IC au niveau de confiance $1 - \alpha$ pour θ, avec $t_\alpha^{BT} = \frac{1}{2\sqrt{n\alpha}}$.

En utilisant l'inégalité de Hoeffding, on obtient

$$P_\theta\left(|T - \theta| \geq t\right) \leq 2e^{-2nt^2},$$

et alors $[T \pm t_\alpha^H]$ est un autre intervalle de confiance de niveau $1 - \alpha$, avec $t_\alpha^H = \sqrt{\frac{\log(2/\alpha)}{2n}}$.

Pour comparer ces deux intervalles, on peut regarder, à n fixé, la longueur de ces intervalles lorsque α tend vers 0 (confiance de plus en plus grande). Dans le premier cas, on obtient une longueur en $1/\sqrt{n\alpha}$, dans le deuxième cas, une longueur en $\sqrt{2\log(1/\alpha)/n}$, qui est négligeable devant la première.

Une manière plus pragmatique de comparer : supposons que l'on fixe le niveau de confiance à 90% (soit $\alpha = 0.1$), et que l'on veuille une précision (longueur d'intervalle de 2%). Les tailles minimales d'échantillon pour atteindre ces objectifs sont alors de l'ordre

— **Bienaymé-Tchebychev** : $n \geq 25000$,

— **Hoeffding** : $n \geq 15000$.

Preuve du Théorème 1.26. Avant toute chose on remarque que l'on peut supposer les X_i centrés quitte à considérer $\tilde{X}_i = X_i - \mathbb{E}(X_i)$ (ça ne change rien aux longueurs des intervalles, ni à $S - \mathbb{E}(S)$). La preuve de ce résultat combine la méthode de Chernoff et le Lemme d'Hoeffding. Pour le volet méthode de Chernoff : en appliquant l'inégalité de Markov à la variable $e^{\lambda S}$, pour un $\lambda \geq 0$, on obtient, pout tout $t > 0$,

$$\begin{aligned} e^{\lambda t}\mathbb{P}\left(e^{\lambda S} \geq e^{\lambda t}\right) &\leq \mathbb{E}\left(e^{\lambda S}\right) \\ &\leq \prod_{i=1}^{n} \mathbb{E}e^{\lambda X_i}, \end{aligned}$$

par indépendance des X_i. Il reste alors à contrôler les transformées de Laplace $\mathbb{E}e^{\lambda X_i}$, ce que l'on fait généralement en considérant

$$\psi_{X_i}(\lambda) = \log(\mathbb{E}(e^{\lambda X_i})),$$

plus facile à manipuler. C'est l'objet du Lemme de Hoeffding (que l'on prouvera à la fin).

Lemme 1.28 : Lemme de Hoeffding

> *Si X est une variable aléatoire centrée prenant ses valeurs dans $[a, b]$, alors*
>
> $$\psi_X(\lambda) \leq \frac{\lambda^2(b-a)^2}{8},$$
>
> *pour tout $\lambda \geq 0$.*

En utilisant ce lemme dans la méthode de Chernoff, on obtient

$$\mathbb{P}(S \geq t) = \mathbb{P}\left(e^{\lambda S} \geq e^{\lambda t}\right) \leq e^{-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2},$$

et en optimisant le λ dans le terme à droite donne le résultat pour $\lambda = \frac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}$. □

Preuve du Lemme 1.28. En notant P la loi de la variable X, on peut réécrire

$$\psi_X(\lambda) = \log\left(\int e^{\lambda x} P(dx)\right)$$

Comme $e^{\lambda x} \leq e^{\lambda b}$ P-p.s., on peut intervertir dérivation et intégration, ce qui donne

$$\begin{aligned} \psi'_X(\lambda) &= \frac{\int x e^{\lambda x} P(dx)}{\int e^{\lambda x} P(dx)} \\ \psi''_X(\lambda) &= \frac{\int x^2 e^{\lambda x} P(dx) - \left(\int x e^{\lambda x} P(dx)\right)^2}{\left(\int e^{\lambda x} P(dx)\right)^2}. \end{aligned}$$

L'astuce consiste à reconnaître respectivement une espérance et une variance. Si on pose $Q_\lambda(dx) = e^{\lambda x}/(\int e^{\lambda x} P(dx))P(dx)$, c'est à dire Q_λ ayant pour densité $e^{\lambda x}/(\int e^{\lambda x} P(dx))$ par rapport à P, on retombe sur une mesure de probabilité, et on a alors

$$\begin{aligned} \psi'_X(\lambda) &= \mathbb{E}(Y) \\ \psi''_X(\lambda) &= \mathrm{Var}(Y), \end{aligned}$$

où $Y \sim Q_\lambda$. Comme X prend ses valeurs dans $[a,b]$, Y aussi, et on peut brutalement majorer sa variance par

$$\mathrm{Var}(Y) \leq \mathbb{E}\left(Y - \frac{a+b}{2}\right)^2 \leq \frac{(b-a)^2}{4}.$$

Par ailleurs, $Q_0 = P$, donc $\psi'_X(0) = \mathbb{E}(X) = 0$. On en déduit alors

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

De manière similaire, $\psi_X(0) = 0$, et donc

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

□

1.4.2 Asymptotique

Nous somme toujours dans le cadre simple $q(\theta) = \theta \in \mathbb{R}$.

Definition 1.29 : Intervalle de niveau de confiance asymptotique $1-\alpha$

Dans un modèle i.i.d. $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta})$ et pour $\alpha \in]0,1[$, un intervalle de confiance *asymptotique* de niveau $1-\alpha$ pour θ est un couple de statistiques (T_n^-, T_n^+) tel que

$$\forall \theta \in \Theta \quad \lim_{n\to+\infty} P_\theta^{\otimes n}\left(T_n^- \leq \theta \leq T_n^+\right) \geq 1-\alpha.$$

On peut là aussi étendre cette notion au delà de $\mathbb{R}$, et pour des suites de lois P_θ^n plus générales que le cadre i.i.d.. L'intérêt des intervalles de confiance asymptotique est de pouvoir baser des intervalles de confiance sur la loi (asymptotique), ce qui est souvent plus précis que via des inégalités de concentration.

Recettes pour les ICA : Classiquement on construit des intervalles de confiance asymptotique à partir de convergence en lois (TCL par exemple), à renfort parfois de méthode Δ.

Exemple 1.30 *: Taux d'éclosion des oeufs de pingouin, suite.* En reprenant l'exemple des oeufs de pingouins, le Théorème Central Limite donne

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

Par continuité de la loi limite, on en déduit que

$$\lim_{n\to+\infty} P_\theta\left(\theta \in \left[T \pm \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} q_{\alpha/2}\right]\right) = 1-\alpha,$$

où q_u désigne le quantile d'ordre u d'une loi $\mathcal{N}(0,1)$. Comme le terme en $\theta(1-\theta)$ dépend évidemment de θ, l'intervalle ci-dessus n'est pas un intervalle de confiance. On peut y remédier de plusieurs manières.

1. En utilisant le lemme de Slutsky : la loi des grands nombres donne $T \xrightarrow{\mathbb{P}} \theta$, donc
$$\sqrt{n}\frac{(T-\theta)}{\sqrt{T(1-T)}} \rightsquigarrow \mathcal{N}(0,1),$$
le terme de gauche est alors *quantité pivotale* asymptotique. Cela fournit l'intervalle asymptotique de niveau de confiance $1-\alpha$ suivant : $[T \pm \frac{\sqrt{T(1-T)}}{\sqrt{n}} q_{\alpha/2}]$.
2. En utilisant la méthode Δ : si on arrive à trouver une fonction G différentiable telle que $G'(\theta) = (\theta(1-\theta))^{-\frac{1}{2}}$, alors la méthode Δ donne
$$\sqrt{n}(G(T) - G(\theta)) \rightsquigarrow \mathcal{N}(0,1).$$
L'intervalle de confiance asymptotique correspondant serait alors $G^{-1}\left([G(T) \pm \frac{q_{\alpha/2}}{\sqrt{n}}]\right)$ (dont on n'est même pas sûrs que ce soit un intervalle).
3. En majorant brutalement le terme de variance : comme $\theta(1-\theta) \leq 1/4$, on en déduit
$$\lim_{n\to+\infty} P_\theta\left(\theta \in \left[T \pm \frac{1}{2\sqrt{n}} q_{\alpha/2}\right]\right) \geq 1-\alpha,$$
et donc $\left[T \pm \frac{1}{2\sqrt{n}} q_{\alpha/2}\right]$ est un ICA de niveau $1-\alpha$.

Ces trois méthodes permettent de couvrir beaucoup de situation, leur pertinence relève du cas par cas. Pour comparer le troisième intervalle avec les intervalles obtenus de manière non-asymptotique, si on fixe $\alpha = 0.1$ et que l'on veut une précision de 2%, la taille d'échantillon requise est $n \geq 6700$ (mieux que via Hoeffding donc).

Dans le cadre particulier d'approximation d'une loi binomiale par une loi normale, si $n(\theta) \geq 5$ et $n(1-\theta) \geq 5$, alors les quantiles de $\sqrt{n}(\theta - T)/\sqrt{\theta(1-\theta)}$ et ceux de sa limite coïncident jusqu'au 2eme chiffre après la virgule (inclus).

En pratique, l'utilisation d'un ICA dans ce modèle binomial est considérée comme valide dès lors que $n(\theta \wedge (1-\theta)) \geq 5$.

L'utilisation pratique des intervalles de confiance asymptotique est tributaire de la rapidité de la convergence en loi qui la sous-tend. Pour certains modèles (comme celui vu en exemple) des conventions existent. Dans un cadre plus général, on peut relier intervalles de confiance asymptotiques et non-asymptotiques en utilisant des résultats de type Berry-Esséen.

Théorème 1.31 : Théorème de Berry-Esséen

Soit $X_1, \ldots, X_n$ une suite de variables i.i.d. telles que $\mathbb{E}(X_1) = 0$, $\mathrm{Var}(X_1) = \sigma^2$, et $\mathbb{E}(|X_1|^3) = \kappa < +\infty$. Si on note, pour $t \in \mathbb{R}$,
$$F_n(t) = \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\bar{X}_n \leq t\right)$$
la "vraie" fonction de répartition, et par Φ la fonction de répartition d'une loi $\mathcal{N}(0,1)$ (celle de la loi limite donc), on a
$$\sup_{t\in\mathbb{R}} |F_n(t) - \Phi(t)| \leq \frac{c\kappa}{\sigma^3\sqrt{n}},$$
où c est une constante numérique (à ce jour valant 0.4748).

Pour conclure cette partie, on comprend maintenant l'importance énorme des quantiles de la loi normale standard dans toutes les applications des statistiques (et du fameux 1.96). En pratique, ces quantiles sont tabulés avec une précision suffisante (avec des tables "manuscrites" ou logiciels). Une méthode plus matheuse d'encadrer ces quantiles se base sur le mini-résultat suivant

Lemme 1.32 : Encadrement des quantiles d'une loi normale standard

Pour $x \in \mathbb{R}$, on note $\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ (densité d'une $\mathcal{N}(0,1)$), et

$$\bar{\Phi}(x) = \int_x^{+\infty} \phi(t)dt,$$

aussi appelée fonction de survie (de la loi $\mathcal{N}(0,1)$). On a alors, pour $x \geq 1$,

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

Démonstration. C'est essentiellement de l'intégration par parties. D'une part, on a

$$\begin{aligned}\bar{\Phi}(x) &= \int_x^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt \\ &= \left[\frac{-e^{-t^2/2}}{\sqrt{2\pi}t}\right]_x^{+\infty} - \int_x^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}t^2}dt \\ &\leq \frac{\phi(x)}{x}.\end{aligned}$$

D'autre part, une inégalité de Markov donne

$$\bar{\Phi}(x) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda x}\right) \leq e^{-\lambda x + \frac{\lambda^2}{2}},$$

en choisissant $\lambda = x$ on obtient l'autre majoration de $\bar{\Phi}(x)$.

Pour la minoration, reprenons la précédente IPP, et remarquons que

$$\begin{aligned}\int_x^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}t^2}dt &= \left[\frac{-e^{-t^2/2}}{\sqrt{2\pi}t^3}\right]_x^{+\infty} - \int_x^{+\infty} \frac{3e^{-t^2/2}}{\sqrt{2\pi}t^4}dt \\ &\leq \frac{\phi(x)}{x^3}.\end{aligned}$$

□

1.5 Tests

Le point de vue des tests est moins naturel que celui de l'estimation. Plutôt que d'estimer θ, on cherche plutôt à répondre à une question binaire dessus. Dans le cadre de l'estimation du taux d'éclosion des oeufs de pingouins, on s'intéresse à la question "est-il plus grand que 1/2" plutôt qu'à son estimation.

Evidemment, si on peut donner des garanties en estimation (via des intervalles de confiance par exemple), on pourra donner des garanties en test et la réciproque est fausse. En ce sens, tester est plus "facile" qu'estimer.

Formellement, une question binaire sur $q(\theta)$ revient à choisir deux sous ensemble Θ_0 et Θ_1 de Θ. On parle alors *d'hypothèses*

$$\begin{aligned} H_0 &: \theta \in \Theta_0, \\ H_1 &: \theta \in \Theta_1. \end{aligned}$$

Par convention H_0 est appelée *hypothèse nulle*, et H_1 *hypothèse alternative*. On verra plus bas que ces deux rôles ne sont pas symétriques.

Tester revient donc à estimer $g : \theta \mapsto \mathbb{1}_{\theta \in \Theta_1}$, dès lors un test est juste un estimateur de cette quantité.

Definition 1.33 : Test

Dans un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$, un test T est une fonction mesurable de $\mathcal{X}$ dans $\{0, 1\}$.

Par convention toujours, lorsque la sortie du test est 0, on dit qu'on *accepte* (sous-entendu l'hypothèse nulle), tandis que $T = 1$ correspond au *rejet* de cette hypothèse.

Comme pour les estimateurs, il s'agit maintenant de mesurer la qualité d'un test, au vu des deux hypothèses que l'on cherche à discriminer. Une approche naturelle est d'équiper $\{0, 1\}$ avec une distance, par exemple $\mathbb{1}_{y \neq y'}$, et à calculer le risque d'un test défini par

$$R_T(\theta) = E_\theta \left(\mathbb{1}_{T(X) \neq g(\theta)} \right) = P_\theta(T(X) \neq g(\theta)),$$

qui quantifie la probabilité que notre test se trompe sous P_θ. Plutôt que de regarder $\sup_\theta R_T(\theta)$ (qui donnerait la même importance aux erreurs sous H_0 et sous H_1), on distingue les erreurs maximales sous les deux hypothèses.

Definition 1.34 : Erreurs de première et seconde espèce, puissance

Dans un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$, pour un test T et deux hypothèses Θ_0 et Θ_1, on définit

- *l'erreur de première espèce* de T : $\sup_{\theta \in \Theta_0} P_\theta(T = 1)$ (probabilité max de rejeter à tort),
- *l'erreur de seconde espèce* de T : $\sup_{\theta \in \Theta_1} P_\theta(T = 0)$.

On parle aussi de *puissance* (minimale) d'un test :

$$\inf_{\theta \in \Theta_1} P_\theta(T = 1),$$

qui est juste 1 moins l'erreur de seconde espèce.

La plupart du temps, un type d'erreur est plus "grave" que l'autre. Dans le cadre d'un test qui prend en entrée divers paramètres d'un patient (par exemple le résultat d'un sondage nasal) et dont la sortie est 0 (patient sain) ou 1 (patient malade), les faux négatifs (patients malades dont le test indique la santé) sont plus préoccupants que les faux négatifs (patients sains que le test détecte comme malade). Par convention, un *test de niveau* α est un test qui contrôle l'erreur la plus grave.

Definition 1.35 : Test de niveau α

Dans un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, pour deux hypothèses Θ_0 et Θ_1, et un paramètre $\alpha \in [0,1]$, un test T est dit de niveau α si son erreur de première espèce est majorée par α, c-à-d

$$\sup_{\theta\in\Theta_0} P_\theta(T=1) \leq \alpha.$$

Le fait qu'un test soit de niveau α ne dépend que de son comportement sous H_0, le fait qu'il arrive à correctement détecter H_1 est de ce point de vue secondaire. Evidemment, le but va être, sous une contrainte de niveau α, de trouver des tests les plus puissants possibles. Comme exemples limites de tests de niveau α aveugles à H_1, on peut citer le test nul $T \equiv 0$, qui est de niveau 0 (mais aussi de puissance nulle), ou alors un test purement aléatoire, de loi $\mathcal{B}(\alpha)$ indépendante des observations (qui de niveau α et de puissance α).

La plupart des tests utilisés sont calibrés par leur niveau. On se rend compte alors que la seule certitude que l'on peut avoir à l'issue d'un tel test est dans le cas d'un rejet de H_0 (lorsque $T=1$) : dans ce cas la probabilité de faire une erreur est majorée par α. En revanche, il n'y a aucune garantie sur l'erreur faite lorsque l'on accepte H_0.

Le choix (dissymétrique) des hypothèses H_0 et H_1 est alors crucial en pratique. De manière informelle, il faut mettre en H_0 le contraire de ce que l'on cherche à prouver.

Exemple 1.36 *: Oeufs de pingouins suite.* Dans le modèle où on observe l'éclosion ou non de n oeufs de pingoins, on peut se poser la question de savoir si la fonte des glaces à un effet sur le taux d'éclosion θ, en supposant que le taux d'éclosion normal est $\theta = 1/2$.

Si vous êtes dans la pétrochimie et cherchez à prouver que non, le taux d'éclosion reste normal, il vous faudra prendre $H_0 : \theta < 1/2$ et $H_1 : \theta = 1/2$.

À l'inverse, si vous êtes plutôt écologiste et cherchez à prouver que ce taux d'éclosion a diminué, il vous faudra prendre $H_0 : \theta = 1/2$ et $H_1 : \theta < 1/2$.

Recettes de construction de test : On verra dans les chapitres suivants des tests classiques, déjà construits et répondants à certains modèles et situations. Si vous devez construire "manuellement" un test de niveau α pour les hypothèses Θ_0 et Θ_1 dans le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}$, on procède généralement comme suit :

1. Choisir une statistique S censée bouger suivant les deux hypothèses (contre ex : $P_\theta = \mathcal{N}(\theta,1)^{\otimes n}$, $S = \sum_i (X_i - \bar{X}_n)^2$).
2. Choisir a priori la forme d'une *région de rejet* R_α en fonction de l'alternative : $S \in R_\alpha$ correspondra à la valeur 1 du test, c-a-d on posera $T = \mathbb{1}_{S\in R_\alpha}$. (ex dans le même cas : si $H_1 : \theta \geq 10$ et $H_0 : \theta < 10$, $S = \bar{X}_n$, on prendra naturellement R_α du type $[t_\alpha, +\infty[$.
3. Calibration de R_α de manière à avoir

$$\sup_{\theta\in\Theta_0} P_\theta(S \in R_\alpha) \leq \alpha,$$

idéalement avec égalité. Cela se fait souvent à l'aide d'une *quantité pivotale* (comme en IC), c'est à dire en trouvant une transformation g_θ de S dont

la loi ne dépend pas de θ : on calibre alors $\sup_{\theta\in\Theta_0}\mathbb{P}(g_\theta(S)\in g_\theta(R_\alpha))\leq\alpha$ (calibration en $g_\theta(R_\alpha)$, la loi de $g_\theta(S)$ étant fixe). On peut aussi utiliser de la domination stochastique.

Exemple 1.37 *: Oeufs de pingouins, fin.* Dans le modèle d'éclosion de n oeufs donné par $(\{0,1\}^n,\mathcal{P}(\{0,1\}^n),\mathcal{B}(\theta)^{\otimes n}_{\theta\in]0,1[})$.

Point de vue écolo : On cherche à prouver que la fonte des glaces a réduit le taux déclosion des oeufs par rapport à la normale 1/2. On a alors

$$\begin{array}{rcl} H_0 & : & \theta=1/2,\\ H_1 & : & \theta<1/2.\end{array}$$

On peut prendre comme statistique de test $S=\sum_{i=1}^n X_i$. Sous H_1 on s'attend à ce que S soit petite, donc on pose $R_\alpha=[0,t_\alpha]$. Comme

$$P_{\theta_0}(S\in R_\alpha)=\mathbb{P}\left(\mathcal{B}(n,1/2)\leq t_\alpha\right),$$

si on prend comme $t_\alpha+1$ le quantile d'ordre α d'une $\mathcal{B}(n,1/2)$, $T=\mathbb{1}_{S\leq t_\alpha}$ est un test de niveau α.

Point de vue pétrolier : On cherche à prouver que le taux d'éclosion des oeufs est resté à la normale. Cela correspond au choix d'hypothèses

$$\begin{array}{rcl} H_0 & : & \theta<1/2,\\ H_1 & : & \theta=1/2.\end{array}$$

La statistique de test reste la même, en revanche on prendra plutôt comme région de rejet $R_\alpha=[t_\alpha,n]$. Il faut maintenant calibrer t_α de sorte que

$$\sup_{\theta<1/2}P_\theta(S\geq t_\alpha)\leq\alpha.$$

Il s'agit maintenant de calculer $\sup_{\theta<1/2}P_\theta(S\geq t_\alpha)$. Vu que sous P_θ, $S\sim\mathcal{B}(n,\theta)$, on s'attend à ce que $\theta\mapsto P_\theta(S\geq t_\alpha)$ soit croissante en θ. On peut le prouver par le calcul, on peut aussi utiliser un argument de couplage.

L'idée est la suivante : si on pose $U_1,\dots,U_n\sim\mathcal{U}(]0,1[)$ i.i.d., et que l'on définit $Y_\theta=|\{i\mid U_i\leq\theta\}|$, alors Y_θ a pour loi P_θ, et on a $Y_{\theta_1}\leq Y_{\theta_2}$ si $\theta_1\leq\theta_2$. On a alors, pour $\theta_1\leq\theta_2$,

$$P_{\theta_1}(S\geq t_\alpha)=\mathbb{P}\left(Y_{\theta_1}\geq t_\alpha\right)\leq\mathbb{P}\left(Y_{\theta_2}\geq t_\alpha\right)=P_{\theta_2}(S\geq t_\alpha).$$

On en déduit alors que

$$\sup_{\theta<1/2}P_\theta(S\geq t_\alpha)=P_{1/2}(S\geq t_\alpha),$$

donc en choisissant t_α le quantile d'ordre $1-\alpha$ d'une $\mathcal{B}(n,1/2)$, on a un test de niveau α. On peut remarquer alors que la puissance aussi est majorée par α (c'est souvent le cas lrosque les deux hypothèses sont contigües).

On peut généraliser la méthode utilisée pour calculer $\sup_{\theta\in\Theta_0}P_\theta(T=1)$ dans le cadre de la domination stochastique.

Proposition 1.38 : Domination stochastique

Soient P et Q deux mesures de proba sur $\mathbb{R}$. On dit que Q domine P (ou $P \preccurlyeq Q$) dès lors qu'un des points suivants est vérifié :

1. *Il existe $X \sim P$ et $Y \sim Q$ telles que $\mathbb{P}(X \leq Y) = 1$.*
2. *Pour tout $t \in \mathbb{R}$, $F_P(t) \geq F_Q(t)$.*
3. *Pour tout $t \in [0,1]$, $q_P(t) \leq q_Q(t)$ (où Q désigne la fonction quantile $\inf\{u \mid F(u) \geq t\}$).*
4. *Pour toute fonction bornée croissante f,*

$$\int f(u)P(du) \leq \int f(u)Q(du).$$

Si $\{P_\theta\}_{\theta\in\Theta_0}$ est stochastiquement ordonnée, on peut alors calculer $\sup_{\theta\in\Theta_0} P_\theta(T = 1)$ dans beaucoup de cas.

Démonstration. Le passage (3) $\Rightarrow$ (1) se fait avec un argument de couplage. En notant $U \sim \mathcal{U}(]0,1[)$ et $X = q_P(U)$, $Y = q_Q(U)$, on a, pour $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(X \leq t) &= \mathbb{P}\left(\inf\{u \mid F_P(u) \geq U\} \leq t\right) \\ &= \mathbb{P}\left(\{F_P(t) \geq U\}\right) \quad F \text{ croissante càd} \\ &= F_P(t).\end{aligned}$$

Donc $X \sim P$, et de la même manière $Y \sim Q$. $q_P \leq q_Q$ entraîne alors $\mathbb{P}(X \leq Y) = 1$.

Le passage (2) $\Rightarrow$ (3) est un classique de l'inverse généralisé d'une fonction de répartition (ou quantile) : comme $F_P \geq F_Q$, on a pour tout $t \in [0,1]$, $\{u \mid F_Q(u) \geq t\} \subset \{u \mid F_P(u) \geq t\}$, et alors $q_p \leq q_Q$.

Le passage (4) $\Rightarrow$ (2) consiste à prendre des f du type $\mathbb{1}_{]-\infty,t]}$, pour $t \in \mathbb{R}$.

Enfin, le passage (1) $\Rightarrow$ (4) est trivial : si f est bornée, alors $\mathbb{E}(f(X))$ et $\mathbb{E}(f(Y))$ existent. Comme f est croissante, on a $f(X) \leq f(Y)$ p.s., et donc

$$\int f(u)P(du) = \mathbb{E}(f(X)) \leq \mathbb{E}(f(Y)) = \int f(u)Q(du).$$

□

En pratique, lorsque vous lancez un test à partir d'un jeu de données via `R` par exemple, l'issue de la procédure est une p-valeur, à partir de laquelle vous allez décider de rejeter (ou accepter). Mathématiquement, une p-valeur est une statistique un peu particulière :

Definition 1.39 : p-valeur

Dans un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}$, pour un problème de test d'hypothèse nulle Θ_0, une *p-valeur* est une fonction $p : \mathcal{X} \to [0,1]$ telle que

$$\forall \theta \in \Theta_0 \quad \forall u \in [0,1] \quad P_\theta\left(q(X) \leq u\right) \leq u.$$

C'est la définition "hard" de la p-valeur (celle notamment utilisée en tests multiples), on verra juste après comment associer une p-valeur à une famille croissante de

régions de rejet. À l'inverse, si on dispose d'une p-valeur, on dispose immédiatement d'un test de niveau α.

LEMME 1.40

Si p est une p-valeur, et $\alpha \in [0,1]$, alors

$$T : X \mapsto \mathbb{1}_{p(X)\leq\alpha}$$

est un test de niveau α.

Ce résultat un peu évident fournit la manière prudente d'interpréter une p-valeur : en la comparant à un seuil de test défini auparavant.

Inversement, à partir d'une famille croissante de régions de rejet on peut construire une p-valeur.

LEMME 1.41

Si $(R_\alpha)_{\alpha\in[0,1]}$ est une famille croissante de régions de rejet associées à une famille de tests de niveau α, alors

$$\hat{\alpha} : X \mapsto \inf\{\alpha > 0 \mid X \in R_\alpha\}$$

est une p-valeur.

Démonstration. Soit $\theta \in \Theta_0$, $u \in [0,1]$ et $\varepsilon > 0$ assez petit, alors

$$P_\theta(\hat{\alpha}(X) \leq u) \leq P_\theta(X \in R_{u+\varepsilon}) \leq u + \varepsilon.$$

En faisant tendre ε vers 0 on peut conclure. □

Cet autre résultat donne la deuxième manière d'interpréter une p-valeur : comme le plus petit niveau d'un test qui conduirait au rejet de H_0 en se basant sur les observations. De là à interpréter une p-valeur comme un indice de plausibilité d'une observation sous H_0, ou une probabilité de rejeter H_0 à tort en se basant sur une observation, il n'y a qu'un pas.

Exemple 1.42. Si on reprend l'exemple des oeufs de pingouins, du point de vue écolo, avec $n = 1000$, en observant S, alors

1. $[0, S]$ est la plus petite région de rejet contenant S,
2. elle correspond au niveau de test $F_0(S)$, où F_0 est la fonction de répartition d'une $\mathcal{B}(1000, 1/2)$.

On en déduit que $F_0(S)$ est une p-valeur.

Application : si on observe $S_{obs} = 460$ oeufs éclos, la p-valeur correspondante est de $0,6\%$. Le point de vue écolo semble alors fondé.

Tests dans un cadre asymptotique

Comme pour les intervalles de confiance, il existe des notions de niveau asymptotique et de consistance pour les tests. On se place ici dans le modèle i.i.d. $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta})$. La notion de consistance est la celle usuelle pour l'estimation de $\mathbb{1}_{\theta\in\Theta_0}$ (c-à-d on veut

que $T \xrightarrow{\mathbb{P}} \mathbb{1}_{\theta\in\Theta_0}$ pour tout θ). La notion de niveau asymptotique est un peu similaire à celle du niveau de confiance asymptotique pour les IC.

Definition 1.43

Pour une hypothèse nulle $\Theta_0 \subset \Theta$ et $\alpha \in [0,1]$, un test T est dit de niveau *asymptotique* α si

$$\sup_{\theta\in\Theta_0} \lim \sup_{n\to+\infty} P_\theta^{\otimes n}(T(X_{1:n}) = 1) \leq \alpha.$$

La recette de construction d'un test de niveau asymptotique α est la même que dans le cadre général, avec une étape de convergence en loi la plupart du temps.

Exemple 1.44 *: Oeufs de pingouins encore.* Dans le modèle d'éclosion des oeufs de pingouins, testé d'un point de vue pétrolier, on cherche toujours un test de la forme $S \geq t_\alpha$, mais on va calibrer le seuil de manière asymptotique.

Pour ce faire, on se souvient que, pout $t \geq 0$

$$\sup_{\theta<1/2} P_\theta(S \geq t) \leq P_{1/2}(S \geq t).$$

Or, sous $P_{1/2}$, $\frac{2(S-n/2)}{\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$. Si on note $q_{1-\alpha}$ le $1-\alpha$-quantile d'une $\mathcal{N}(0,1)$, on a

$$\lim_{n\to\infty} P_{1/2}\left(\frac{2(S-n/2)}{\sqrt{n}} \geq q_{1-\alpha}\right) = \alpha,$$

et donc $T = \mathbb{1}_{S\geq n/2+\sqrt{n}q_{1-\alpha}/2}$ est un test de niveau asymptotique α.

1.6 Méthodes classiques d'estimation

Moments, M-estimation, logV. Recette G M-estimation concentration convexité.

Chapitre 2

Modèle linéaire Gaussien

Les modèles linéaires Gaussiens sont d'une part assez utilisés en pratique (en biologie, linguistique, etc.), et servent souvent de modèles limites via le théorème central limite. Cela vaut donc la peine d'y consacrer un peu de temps.

2.1 Vecteurs Gaussiens

2.1.1 Rappels sur la loi normale

Commençons par des rappels standards sur les lois normales.

Definition 2.1 : Loi normale (ou Gaussienne)

Soit X une variable aléatoire réelle, d'espérance μ et variance σ^2. X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$ si l'une des propriétés équivalentes suivantes est vérifiée :

1. X admet pour densité $f(t) = \frac{1}{\sqrt{2\sigma\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$,
2. pour tout $t \in \mathbb{R}$,

$$\phi_X(t) = \mathbb{E}(e^{itX}) = e^{it\mu} e^{-\frac{\sigma^2 t^2}{2}}.$$

Les propriétés de bases des lois normales sont les suivantes.

Proposition 2.2 : Lois Normales-Propriétés

1. *Si $X \sim \mathcal{N}(\mu, \sigma^2)$, et $(a, b) \in \mathbb{R}^2$, alors $aX + b \sim \mathcal{N}(a\mu + b, b^2\sigma^2)$.*
2. *Si $X \sim \mathcal{N}(\mu, \sigma^2)$, $Y \sim \mathcal{N}(\nu, \tau^2)$ et $X \perp\!\!\!\perp Y$, alors $X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2)$.*

On peut aussi caractériser les lois Gaussiennes via l'identité de Stein. Ces propriétés sont suffisantes pour la suite.

2.1.2 Vecteurs Gaussiens, définitions et propriétés

Definition 2.3 : Vecteur Gaussien

> Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$. X est un vecteur Gaussien si et seulement si pour tout $t \in \mathbb{R}^d$ $\langle t, x\rangle$ suit une loi Gaussienne.

Remarquons que si X est un vecteur Gaussien, alors $\mathbb{E}(\|X\|^2 = \sum_{j=1}^d \mathbb{E}(\langle e_j, X\rangle^2) < +\infty$. En particulier, $\mu = \mathbb{E}(X)$ et $\Sigma = \mathrm{Cov}(X)$ sont bien définis, avec

$$\Sigma_{i,j} = \mathbb{E}((X_i - \mu_i)(X_j - \mu_j)),$$

Σ étant appelée *matrice de covariance* (et est toujours positive). Le résultat suivant montre que le veteur moyenne et la matrice de covaraince caractérisent entièrement la loi d'un vecteur Gaussien.

Lemme 2.4 : Fonction caractéristique d'un vecteur Gaussien

> *Soit X un vecteur Gaussien de moyenne μ et matrice de covariance Σ. Alors, pour tout $u \in \mathbb{R}^d$,*
>
> $$\phi_X(u) = \mathbb{E}(e^{i\langle u, X\rangle}) = e^{i\langle u,\mu\rangle - \frac{u^T\Sigma u}{2}}.$$

La preuve vient du fait que $\langle u, X\rangle$ est Gaussienne (définition), de moyenne $\langle u, \mu\rangle$ et variance $u^T\Sigma u$ (calcul). La fonction caractéristique caractérisant la loi (encore un Théorème de Lévy), un vecteur Gaussien est entièrement caractérisé en loi par son vecteur moyenne et sa matrice de covariance. On notera alors une telle loi $\mathcal{N}(\mu, \Sigma)$.

On peut en déduire les propriétés utiles suivantes :

Proposition 2.5 : Vecteurs Gaussiens-Propriétés

> 1. *Si $X \sim \mathcal{N}(\mu, \Sigma)$, $A \in M_{k\times d}(\mathbb{R})$ et $b \in \mathbb{R}^k$, alors $AX + b \sim \mathcal{N}(A\mu, A\Sigma A^T)$.*
> 2. *Si $X \sim \mathcal{N}(\mu, \Sigma)$, $X' \sim \mathcal{N}(\mu', \Sigma')$ et $X \perp\!\!\!\perp X'$, alors $X + X' \sim \mathcal{N}(\mu + \mu', \Sigma + \Sigma')$.*
> 3. *Si Σ est définie, alors $X \sim \mathcal{N}(\mu, \Sigma)$ admet une densité sur $\mathbb{R}^d$, donnée par*
>
> $$f(x) = \frac{1}{(2\pi \mathrm{Det}(\Sigma))^{\frac{d}{2}}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(X-\mu)\right].$$

Dans le cas où Σ n'est pas inversible, la loi de X est supportée par l'espace vectoriel engendré par les vecteurs propres de Σ. De manière générale, on pourra toujours décomposer X en $\mu + AY$, avec Y vecteur Gaussien standard (à composantes i.i.d. $\mathcal{N}(0, 1)$) à valeurs dans $\mathbb{R}^k$, $k \leq d$ et $A : \mathbb{R}^k \to \mathbb{R}^d$ linéaire de rang k.

2.1.3 Indépendances et conditionnements

L'idée fondamentale de cette section est que, puor un vecteur Gaussien, décorrélation et indépendance coïncident. On peut donc lire la structure de dépendance directement sur la matrice de covariance.

Lemme 2.6

Soit $X = (X_{1:k}^T, X_{k+1:d}^T)^T$ un vecteur Gaussien ayant une matrice de covariance de type

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{kk} & 0 \\ \hline 0 & \Sigma_{k'k'} \end{array}\right),$$

avec $k + k' = d$. Alors $X_{1:k}$ et $X_{k+1:d}$ sont indépendants.

La preuve se fait en regardant la fonction caractéristique. L'hypothèse que X soit un vecteur Gaussien est capitale (il est assez facile de construire deux Gaussiennes décorrélées mais non indépendantes).

On peut en déduire les espérances (et lois) conditionnelles de coordonnées sachant les autres.

Proposition 2.7 : Espérances conditionnelles Gaussiennes

Soit $X = (X_{1:k}^T, X_{k+1:d}^T)^T$ un vecteur Gaussien de moyenne μ et ayant pour matrice de covariance

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{kk} & B^T \\ \hline B & \Sigma_{k'k'} \end{array}\right),$$

avec Σ_{kk} définie. On a alors

$$\mathbb{E}\left(X_{k+1:d} \mid X_{1:k}\right) = \mu_{k+1:d} + B(\Sigma_{kk})^{-1}\left(X_{1:k} - \mu_{1:k}\right).$$

De plus, $(X_{k+1:d} - \mathbb{E}\left(X_{k+1:d} \mid X_{1:k}\right) \perp\!\!\!\perp X_{1:k})$.

Démonstration. En notant E le terme de droite, on remarque que $(X_{1:k}^T, (X_{k+1:d} - E)^T)^T$ est un vecteur Gaussien (fonction affine de vecteur Gaussien). Le bloc de sa matrice de covariance en bas à gauche s'écrit

$$\begin{aligned}\mathbb{E}\left((X_{k+1:d} - E)(X_{1:k} - \mu_{1:k})^T\right) &= B - \mathbb{E}((E - \mu_{k+1:d})(X_{1:k} - \mu_{1:k})^T) \\ &= B - B = 0.\end{aligned}$$

On en déduit que $(X_{k+1:d} - E) \perp\!\!\!\perp X_{1:k}$, et E étant $X_{1:k}$-mesurable que $E = \mathbb{E}\left(X_{k+1:d} \mid X_{1:k}\right)$. □

En d'autres termes, l'espérance conditionnelle d'un bout de vecteur Gaussien sachant le reste est sa projection au sens $L_2(\Omega, \mathbb{P})$ sur l'espace affine engendré par le reste.

2.1.4 Théorème(s) de Cochran

La version clés en main du paragraphe précédent est le Théorème de Cochran. Auparavant, quelques rappels de lois usuelles classiques sur $]0; +\infty[$.

Definition 2.8 : Loi Gamma - Définition et propriétés

Une variable aléatoire X suit une loi $\gamma(a,b)$ si elle a pour densité

$$x^{a-1}e^{-bx}\frac{b^a}{\Gamma(a)}\mathbb{1}_{x>0}.$$

Par ailleurs, les propriétés suivantes sont vérifiées.

1. $\mathbb{E}\gamma(a,b) = \frac{a}{b}$, $\mathrm{Var}(\gamma(a,b)) = \frac{a}{b^2}$.
2. Si $X \sim \gamma(a,b)$, $\phi_X(t) = \left(\frac{b}{b-it}\right)^a$.
3. Si $X \sim \gamma(a,b)$ et $Y \sim \gamma(a',b)$, $X \perp\!\!\!\perp Y$, alors $X+Y \sim \gamma(a+a',b)$.
4. Si $X \sim \gamma(a,b)$ et $\lambda > 0$, $\lambda X \sim \gamma(a, b/\lambda)$.

Un cas particulier de loi Gamma est la loi du Chi-deux.

Definition 2.9 : Loi du Chi-Deux- Définition et propriétés

Soit $k \in \mathbb{N}^*$. La loi du Chi-deux à k degrés de libertés, notée $\chi^2(k)$ est la loi $\gamma(\frac{k}{2}, \frac{1}{2})$.

Elle vérifie les propriétés suivantes :

1. Si $X_1, \dots, X_k$ sont les $\mathcal{N}(0,1)$ indépendantes, alors $\sum_{i=1}^k X_i^2 \sim \chi^2(k)$.
2. Si $X \sim \chi^2(k_1)$, $Y \sim \chi^2(k_2)$, $X \perp\!\!\!\perp Y$, alors $X+Y \sim \chi^2(k_1+k_2)$.

On aura aussi besoin de la loi du χ^2 décentrée, définie comme suit.

Definition 2.10 : Loi du χ^2 décentrée

Soit $X \sim \mathcal{N}(\mu, I_d)$, $\mu \neq 0$. Alors la loi de $\|X\|^2$ ne dépend que de $\|\mu\|$, et est appelée *loi du χ^2 décentrée* $\chi^2(d, \|\mu\|)$. En particulier

$$\|X\|^2 \sim \chi^2(d, \|\mu\|) \sim (\|\mu\| + Z_1)^2 + \sum_{j=2}^{d} Z_j^2,$$

où $Z_1, \dots, Z_d$ sont i.i.d. $\mathcal{N}(0,1)$.

Démonstration. Si P est une matrice orthonormale, de premier vecteur $\frac{\mu}{\|\mu\|}$, en notant $Y = P^T X$, on a $\|Y\| = \|X\|$, donc $\|X\|^2$ et $\|Y\|^2$ ont même loi, donnée par l'expression à droite. □

Le(s) théorème(s) de Cochran peuvent alors s'énoncer comme suit.

Théorème 2.11 : Cochran version centrée

Si $X \sim \mathcal{N}(0, \sigma^2 I_d)$, et $E_1, \dots, E_k$ forment une décomposition de $\mathbb{R}^d$ en sous-espaces orthogonaux, alors $(\pi_{E_1}X, \dots, \pi_{E_k}X)$ sont indépendants, et, pour tout $j \in [\![1,k]\!]$, $\|\pi_{E_j}X\|^2 \sim \sigma^2\chi^2(\dim(E_j))$.

Démonstration. Notons π_j les matrices de projections, et $Z = ((\pi_1 X)^T, \dots, (\pi_k X)^T)^T$. Z est un vecteur Gaussien (fonction linéaire d'un vecteur Gaussien). Par ailleurs, si $i \neq j$, $\mathbb{E}(\pi_i X (\pi_j X)^T) = \sigma^2 \pi_i \pi_j^T = 0$. On en déduit que $(\pi_{E_1} X, \dots, \pi_{E_k} X)$ sont indépendants.

Par ailleurs, en notant U_j une base de vecteurs orthonormés de E_j, on a $\pi_j X = U_j U_j^T X$, donc $\pi_j X \sim \mathcal{N}(0, \sigma^2 U_j I_{p_j} U_j^T)$, avec $p_j = \dim(E_j)$. On en déduit que $\pi_j X \sim U_j (Z_1, \dots, Z_{p_j})^T$, avec des Z_i i.i.d. $\mathcal{N}(0, \sigma^2)$. Comme $\|U_j Z\|^2 = \|Z\|^2 \sim \sigma^2 \chi^2(p_j)$, on en déduit le résultat. □

Théorème 2.12 : Cochran version décentrée

Soit $\mathbb{X} \sim \mathcal{N}(\mu, \sigma^2 I_d)$, et $E_1, \dots, E_k$ une décomposition de $\mathbb{R}^d$ en sous-espaces orthogonaux. Alors $(\pi_{E_1} X, \dots, \pi_{E_k} X)$ sont indépendants, et, pour tout $j \in [\![1, k]\!]$, $\|\pi_{E_j} X\|^2 \sim \sigma^2 \chi^2(\dim(E_j), \|\pi_{E_j} \mu\|)$

Démonstration. La preuve concernant l'indépendance est exactement la même que dans le cas centré. Soit $j \in [\![1, k]\!]$, et U_j une BON de E_j. On a $\pi_{E_j} X = U_j U_j^T X \sim \mathcal{N}(U_j (U_j^T \mu), \sigma^2 U_j I_{p_j} U_j^T)$, donc $\pi_j X \sim U_j Z$, où $Z \sim \mathcal{N}(U_j^t \mu, \sigma^2 I_k)$. En particulier $\|\pi_j X\|^2 = \|Z\|^2 \sim \sigma^2 \chi^2(p_j, \|U_j^T \mu\|) = \sigma^2 \chi^2(p_j, \|\pi_{E_j} \mu\|)$. □

Une application directe (on en verra d'autres) est donnée par le test du Chi-deux d'appartenance à un sous-espace.

σ^2 connu, test du Chi-deux d'appartenance à un sous-espace

Dans le modèle $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), (\mathcal{N}(u, \sigma^2 I_d)_{\mu,\sigma})$ où on connaît σ, on cherche à tester $H_0 : \mu \in V$ contre $H_1 : \mu \notin V$, où V est un sous-espace vectoriel de $\mathbb{R}^d$ de dimension k.

Comme souvent, l'astuce consiste à trouver une quantité pivotale sous H_0 dont on connaît la loi. Ici, sous H_0, $X - \pi_V X \sim \mathcal{N}(0, \sigma^2 \pi_{V^\perp} \pi_{V^\perp}^T)$. On en déduit alors que, sous H_0

$$S(X) = \frac{\|X - \pi_V X\|^2}{\sigma^2} \sim \chi^2(d - k).$$

Sous H_1, on s'attend à ce que $S(X)$ soit grand. On prend donc comme test

$$T = \mathbb{1}_{S \geq q_{1-\alpha}},$$

où $q_{1-\alpha}$ est le $1-\alpha$-quantile d'un $\chi^2(n-k)$, ce qui donne bien un test de niveau α.

2.2 Une application asymptotique : tests du Chi-deux d'adéquation et d'homogénéité

Les applications en statistiques asymptotique du formalisme "vecteur Gaussien" se basent sur le théorème central limite "vectoriel".

Théorème 2.13 : TCL vectoriel

Soient $X_1, \dots, X_n$ des vecteurs aléatoires i.i.d. tels que $\mathbb{E}(\|X_1\|^2) < +\infty$. On a

$$\sqrt{n}\left(\bar{X}_n - \mathbb{E}(X_1)\right) \rightsquigarrow \mathcal{N}(0_k, \Sigma),$$

où $\Sigma = \mathrm{Cov}(X_1)$.

2.2.1 Test du chi-deux d'adéquation (ou d'ajustement)

On se place dans le cadre où on observe n tirages i.i.d. d'une variable X d'une distribution pouvant prendre au plus k valeurs. On peut modéliser cette situation via

$$([\![1,k]\!]^n, \mathcal{P}([\![1,k]\!]^n), (P_{\mathbf{p}})_{\mathbf{p}\in S_k}^{\otimes n}),$$

où $S_k = \{(p_1, \dots, p_k) \mid \sum_{j=1}^k p_j = 1\}$.

Le test du χ^2 d'adéquation consiste à tester $H_0 : \mathbf{p} = \mathbf{p}_0$ contre $H_1 : \mathbf{p} \neq \mathbf{p}_0$, pour une valeur de $\mathbf{p}_0$ fixée au préalable. Par exemple un casino souhaitant déterminer si un dé est bien équilibré prendra $\mathbf{p}_0 = (1/6, \dots, 1/6)$.

Si on note $N_j = \sum_{i=1}^n \mathbb{1}_{X_i=j}$, le vecteur $N = (N_1, \dots, N_k)$ suit une loi multinomiale $\mathcal{M}(n, \mathbf{p})$, avec

$$\mathbb{P}(N = (n_1, \dots, n_k)) = \frac{n!}{n_1! \dots n_k!} \prod_{j=1}^k p_j^{n_j},$$

avec $\sum_{j=1}^k n_j = n$.

On peut montrer que le vecteur N, convenablement normalisé, converge en loi. On peut commencer par remarquer que $N = \sum_{i=1}^n Z_i$, où $Z_i = (\mathbb{1}_{X_i=1}, \dots, \mathbb{1}_{X_i=k})$, les Z_i étant i.i.d., de vecteur moyenne $\mathbf{p}$, et de matrice de covariance $\Sigma = \mathrm{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$.

Comme $\mathbb{E}(\|Z_1\|^2) = 1$, le théorème central limite s'applique, et on obtient alors

$$\sqrt{n}(\frac{N}{n} - \mathbf{p}) \rightsquigarrow \mathcal{N}(0_k, \Sigma).$$

En pratique, l'approximation normale est loisible dès lors que pour tout $j = 1, \dots, k$, $np_j \geq 5$. Sous H_0, on a alors $\sqrt{n}(\frac{N}{n} - \mathbf{p}_0) \approx \mathcal{N}(0, \Sigma_0)$ dès que $n\mathbf{p}_0 \geq (5, \dots, 5)$. Il reste à trouver une quantité pivotale. Pour cela, il suffit de remarquer que, sous H_0,

$$\mathrm{Diag}(1/\sqrt{\mathbf{p}_0})\sqrt{n}(\frac{N}{n} - \mathbf{p}_0) \rightsquigarrow \mathcal{N}(0_k, I_k - \sqrt{\mathbf{p}_0}\sqrt{\mathbf{p}_0}^T),$$

en supposant que pour tout j, $p_{j,0} \neq 0$. Enfin, on remarque que $\pi_0 = I_k - \sqrt{\mathbf{p}_0}\sqrt{\mathbf{p}_0}^T$ est une matrice de projection (sur l'orthogonal à $\sqrt{\mathbf{p}_0}$). Le théorème de Cochran implique alors que

$$\left\|\mathrm{Diag}(1/\sqrt{\mathbf{p}_0})\sqrt{n}(\frac{N}{n} - \mathbf{p}_0)\right\|^2 \rightsquigarrow \|\pi_0 \mathcal{N}(0_k, I_k)\|^2 = \chi^2(k-1).$$

Un test de niveau α pour $H_0 : \mathbf{p} = \mathbf{p}_0$ est donc $\mathbb{1}_{S \geq q_{1-\alpha}}$, où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une $\chi^2(k-1)$, et S est la statistique de test

$$S = \sum_{j=1}^k \frac{(N_j - np_{j,0})^2}{np_{j,0}}.$$

2.2.2 Un test du chi-deux d'homogénéité

On suppose cette fois-ci que l'on observe $X_1, \dots, X_n$ i.i.d. à valeurs dans $[\![1, k]\!]$, et $Y_1, \dots, Y_m$ i.i.d. à valeurs dans le même espace. La question que l'on se pose est celle de l'égalité en loi des Y et des X. Par exemple, on peut vouloir comparer la répartition de certains caractères entre deux populations.

Une manière de s'en sortir est de passer par un test du χ^2 d'indépendance (plus tard). On peut aussi passer par un test du χ^2 d'appartenance à un sous-espace. Pour se faire, on dissocie les deux espaces d'arrivée pour obtenir $[\![1, 2k]\!]$, et on regroupe les observations en supposant que l'on observe $\tilde{X}_1, \dots, \tilde{X}_{n+m}$ i.i.d. à valeurs dans $[\![1, 2k]\!]$, où les n premières observations correspondent aux X_i, les m dernières aux $Y_i + k$.

Avec ce bidouillage, on a, pour $1 \leq j \leq k$, $\mathbb{P}(\tilde{X}_i = j) = \frac{n}{n+m}\mathbb{P}(X_1 = j)$, et $\mathbb{P}(\tilde{X}_i = j) = \frac{m}{n+m}\mathbb{P}(Y_1 = j)$ si $k+1 \leq j \leq 2k$. Si on note $Z_i = (\mathbb{1}_{\tilde{X}_i = j})_{j=1,\dots,2k}$, Z_i suit une loi multinomiale $\mathcal{M}(n+m, (\frac{n}{n+m}\mathbf{p}_X, \frac{m}{m+n}\mathbf{p}_Y))$. On supposera dans la suite que toutes les proportions sont non nulles.

En notant les proportions $\alpha_n = \frac{n}{n+m}$, $\alpha_m = \frac{m}{m+n}$ et $\mathbf{p}_{\tilde{X}}$ le vecteur moyenne de $\tilde{X}$, on a, sous H_0,

$$A\mathbf{p}_{\tilde{X}} = 0_k,$$

où A est la matrice $(\mathrm{Diag}(\alpha_n^{-1}, k)|\mathrm{Diag}(-\alpha_m^{-1}, k)) \in M_{k,2k}$. Toujours sous H_0, on en déduit alors

$$A(Z_1 - \mathbf{p}_{\tilde{X}}) = AZ_1.$$

Une application du théorème central limite donne alors

$$\sqrt{n+m}A\bar{Z}_{n+m} \rightsquigarrow \mathcal{N}(0_k, A\Sigma A^T),$$

avec $\Sigma = \mathrm{Diag}(\mathbf{p}_{\tilde{X}}) - \mathbf{p}_{\tilde{X}}\mathbf{p}_{\tilde{X}}^T$. Comme $A\mathbf{p}_{\tilde{X}} = 0_k$, un calcul matriciel donne

$$A\Sigma A^T = A\mathrm{Diag}(\mathbf{p}_{\tilde{X}})A^T = (\alpha_n^{-1} + \alpha_m^{-1})\mathrm{Diag}(\mathbf{p}) \in M_{k,k},$$

où $\mathbf{p} = \mathbf{p}_X = \mathbf{p}_Y$ (rappelons que l'on travaille sous H_0).

On en déduit alors que, sous H_0,

$$\mathrm{Diag}(\mathbf{p})^{-\frac{1}{2}} \frac{A\bar{Z}_{n+m}}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \rightsquigarrow \mathcal{N}(0, I_k).$$

On peut estimer $\mathrm{Diag}(\mathbf{p})$ via $\frac{1}{n+m}(N_{j,X} + N_{j,Y})_{j=1,\dots,k}$, où $N_{j,X} = \sum_{i=1}^n \mathbb{1}_{X_i=j}$ et $N_{j,Y} = \sum_{i=1}^m \mathbb{1}_{Y_i=j}$. La convergence en proba étant assurée par la loi des grands nombres, le lemme de Slustsky donne alors en prenant la norme au carré

$$\sum_{j=1}^k \frac{nm\left(\frac{N_{j,X}}{n} - \frac{N_{j,Y}}{m}\right)^2}{N_{j,X} + N_{j,Y}} \rightsquigarrow \chi^2(k),$$

sous H_0. Ce dont on peut déduire un test de niveau α.

Attention : Ce n'est pas la forme classique du test du χ^2 d'homogénéité. On verra plus tard une autre version plus "morale".

2.3 Régression linéaire homoscédastique à design fixe

Le contexte de la régression est le suivant : on observe (X_i, Y_i), les $X_i \in \mathbb{R}^k$ étant appelés *variables prédictives* ou *paramètres*, la variable à prédire étant $Y \in \mathbb{R}$. Le but est de construire un *régresseur*, c'est à dire une fonction $\hat{f}_n : \mathbb{R}^k \to \mathbb{R}$ de manière à pouvoir prédire une valeur de y étant donné un nouveau paramètre x. La *régression linéaire* consiste juste à chercher un prédicteur affine en les paramètres, c'est à dire de la forme $x \mapsto \langle \theta, x \rangle + b$.

Exemple 2.14. Dans un champ de maïs où n plants ont reçu différentes doses d'engrais X_i, on peut essayer d'expliquer la hauteur du plant Y_i par une fonction affine de la dose d'engrais reçue, c'est à dire via $aX_i + b$, où a et$b$ sont à déterminer pour prévoir les futures hauteurs de plan.

On peut se débarasser du côté affine dès maintenant : si, on transforme le paramètre $X_i \in \mathbb{R}^k$ en le paramètre $\tilde{X}_i = (X_i^T, 1)^T \in \mathbb{R}^{k+1}$, une fonction affine des X_i sera une fonction linéaire des $\tilde{X}_i$. Poser un modèle linéaire (Gaussien) revient juste à supposer que la variable à prédire s'exprime comme fonction linéaire des prédicteurs, à erreur additive (Gaussienne) i.i.d. près.

Definition 2.15 : Modèle linéaire (Gaussien) homoscédastique

Un *modèle linéaire homoscédastique* consiste à supposer que la loi des observations (X_i, Y_i) est de la forme

$$Y = X\theta + \varepsilon,$$

où $Y = (Y_1, \ldots, Y_n)^T$, X est la matrice $n \times k$ dont la $i - ème$ ligne vaut X_i^T, $\theta \in \mathbb{R}^k$, et $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$, les ε_i étant i.i.d. d'espérance nulle et de variance σ^2 inconnue a priori.

Si de plus les ε_i sont supposées Gaussiennes, on parle de modèle linéaire *Gaussien*.

Remarque : Le côté homoscédastique vient du fait que l'on a supposé les variances des erreurs ε_i homogènes. Si ce n'est pas le cas (on n'est plus dans un cadre i.i.d. mais seulement indépendant alors), on parle de modèle *hétéroscédastique*.
Remarque 2 : Dans toute la suite on supposera que les X_i sont *fixes*, on parle alors de régression à design fixe. Si les X_i sont eux-mêmes considérés comme des réalisations d'une variable aléatoire X, on parle alors de design aléatoire. Beaucoup de résultat en design fixe peuvent s'étendre au designe aléatoire en conditionnant par $X_1, \ldots, X_n$. Pour alléger un peu le formalisme on ne traitera ici que le design fixe.

2.3.1 Modèle linéaire général - Moindre carrés

Dans le modèle linéaire général $Y = X\theta + \varepsilon$, on peut vérifier que

$$\theta = \arg\min_{u \in \mathbb{R}^k} E_\theta \|Y - Xu\|^2 := \arg\min_{u \in \mathbb{R}^k} R(u),$$

où l'espérance porte sur les Y_i. La fonction R est couramment appelée *risque en prédiction* : on se donne un nouvel échantillon Y' de même loi que Y et on regarde le risque quadratique moyen encouru à prédire Y' par Xu. L'estimateur par moindre carré consiste à supposer $R_n(u) = \|Y - Xu\|^2 \approx E_\theta\|Y' - Xu\|^2 = R(u)$, et donc à chercher

$$\hat{\theta}_{LS} \in \arg\min_u \|Y - Xu\|^2 := \arg\min_u R_n(u).$$

Un calcul simple montre que

$$\nabla_\theta R_n = 2(X^T Xu - X^T Y),$$

donc $\hat{\theta}_{LS}$ est solution de

$$X^T Xu = X^T Y.$$

Cela revient à demander $Xu = \pi_{V(X)}(Y)$, où $V(X)$ est le s-e-v. de $\mathbb{R}^n$ engendré par les colonnes de X. Cette équation admet une unique solution dans le cas où $X^T X$ est inversible. Dans le cas général, on a un s-e-v- de solutions.

Definition 2.16 : Estimateur des moindres carrés

Dans le modèle linéaire, si la matrice de précision $X^T X$ est inversible, l'estimateur des moindres carrés est défini par

$$\hat{\theta}_{LS} = (X^T X)^{-1} X^T Y.$$

À partir de maintenant, sauf précision contraire on supposera que la matrice de précision $X^T X$ est inversible. En particulier, cela implique que le nombre de paramètres est plus petit que le nombre d'observations, $k \leq n$.

Proposition 2.17 : Propriétés de l'estimateur des moindres carrés

Dans un modèle linéaire homoscédastique, l'estimateur $\hat{\theta}_{LS}$ vérifie les propriétés suivantes :

1. $E_\theta(\hat{\theta}_{LS}) = \theta$,
2. $\mathrm{Cov}_\theta(\hat{\theta}_{LS}) = \sigma^2 (X^T X)^{-1}$.

On peut en déduire les bornes de risque suivantes :

— $E_\theta(\|\hat{\theta}_{LS} - \theta\|^2) = \sigma^2 \mathrm{Tr}((X^T X)^{-1})$.

— $E_\theta(R(\hat{\theta}_{LS}) - R(\theta)) = k\sigma^2$,

Démonstration. Dans un modèle linéaire homoscédastique, on peut écrire

$$\begin{aligned}\hat{\theta}_{LS} &= (X^T X)^{-1} X^T (X\theta + \varepsilon) \\ &= \theta + (X^T X)^{-1} X^T \varepsilon.\end{aligned}$$

Comme $\mathbb{E}(\varepsilon) = 0$, on en déduit $E_\theta(\hat{\theta}_{LS}) = \theta$. Par ailleurs,

$$\begin{aligned}\mathrm{Cov}_\theta(\hat{\theta}_{LS}) &= E_\theta(\hat{\theta}_{LS} - \theta)(\hat{\theta}_{LS} - \theta)^T \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}.\end{aligned}$$

Le troisième point découle du second :

$$E_\theta(\|\hat{\theta}_{LS} - \theta\|^2) = \mathrm{Tr}(\mathrm{Cov}_\theta(\hat{\theta}_{LS})) = \sigma^2 \mathrm{Tr}((X^T X)^{-1}).$$

Pour le troisième point, il faut remarquer que, pour $u \in \mathbb{R}^k$,

$$\begin{aligned} R(u) - R(\theta) =& \mathbb{E}\left[\|X(\theta - u) + \varepsilon'\|^2 - \|\varepsilon'\|^2\right] \\ &= \|X(\theta - u)\|^2. \end{aligned}$$

On en déduit alors que $R(\hat{\theta}_{LS}) - R(\theta) = \|X(\theta - \hat{\theta}_{LS})\|^2$ (qui est aléatoire en Y), et donc

$$\begin{aligned} E_\theta\left(R(\hat{\theta}_{LS}) - R(\theta)\right) &= \mathbb{E}\left\|X(X^T X)^{-1} X^T \varepsilon\right\|^2 \\ &= \mathbb{E}\left(\varepsilon^T X (X^T X)^{-1} X^T \varepsilon\right) \\ &= \sum_{i=1}^{n} (X(X^T X)^{-1} X^T)_{i,i} \sigma^2 \\ &= k\sigma^2, \end{aligned}$$

où on a abondamment utilisé le fait que $X(X^T X)^{-1} X^T$ est la matrice de projection orthogonale sur l'espace $V(X)$, qui est donc de trace k. □

Valeur ajustées, résidus et leviers

Une fois notre estimateur par moindre carrés $\hat{\theta}_{LS}$ construit, on peut définir deux quantités intrinsèquement intéressantes :

— les *valeurs ajustées* (ou prédites) : $\hat{Y} = X\hat{\theta}_{LS}$, c'est à dire les valeurs de Y prédites par le modèle,

— et les *résidus*, $\hat{\varepsilon} = Y - \hat{Y}$, c'est à dire le vecteur des écarts entre valeurs prédites et valeurs observées.

Si on note H la matrice de projection sur $V(X)$, c'est à dire $H = X(X^T X)^{-1} X^T$, on peut réécrire

$$\begin{aligned} \hat{Y} &= HY, \\ \hat{\varepsilon} &= (I_n - H)(X\theta + \varepsilon) = (I_n - H)\varepsilon. \end{aligned}$$

Les résidus peuvent être utilisés pour fournir un estimateur non-biaisé de σ : en effet, on a

$$\mathbb{E}(\|\hat{\varepsilon}\|^2) = \sigma^2 \mathrm{Tr}(I_n - H) = \sigma^2(n - k),$$

car $I_n - H$ est la matrice de projection sur $V(X)^\perp$. On en déduit que

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n - k}$$

est un estimateur non biaisé de σ^2.

En pratique, la validation d'un modèle se fait en regardant le comportement des résidus, qui doit être proche du comportement théorique attendu : si on suppose que $\varepsilon \sim \mathcal{N}(0, I_n)$ (modèle linéaire Gaussien), on s'attend à des résidus de loi $\mathcal{N}(0_n, I_n - H)$.

Une dernière précision concernant la matrice H : elle est souvent appelée en anglais *hat matrix* ou *matrice d'influence*, et ses coefficients diagonaux *leverages* ou *leviers*. L'idée est que ces coefficients diagonaux caractérisent l'influence d'une observation Y_i sur le modèle ajusté via sa prédiction $\hat{Y}_i$. En effet

$$\frac{\partial \hat{Y}_i}{\partial Y_i} = H_{i,i} \in [0,1].$$

Une observation "influente" correspond à un levier proche de 1 : si on perturbe cette observation on perturbe beaucoup la prédiction, comparativement aux observations à levier faible. Une autre manière de voir est de remarquer que $\mathrm{Var}(\hat{\varepsilon}_i) = (1 - H_{i,i})$: une observation de levier fort conduit à un résidu "faible", le modèle ajuste particulièrement cette observation (toujours comparativement aux observations de levier faible).

Exemple 2.18 *: Explication du terme levier.* Dans le modèle le plus simple où $X_i \in \mathbb{R}$, le modèle linéaire s'écrit $Y = \theta X$, où $\theta \in \mathbb{R}$. On a alors $\hat{\theta}_{LS} = \sum_{i=1}^n Y_i X_i / \sum_{i=1}^n X_i^2$. L'effet de Y_i sur $\hat{\theta}_{LS}$ et $\hat{Y}_i$ est proportionnel à $X_i/\|X\|^2$ (resp. $X_i^2/\|X\|^2$). Les observations les plus influentes sont donc celles correspondants aux valeurs de $|X_i|$ les plus éloignées. (FAIRE DESSIN).

Concluons sur deux remarque : en pratique les leviers peuvent être utilisés pour détecter des observations anormalement influentes sur le modèle (outliers). Cela a du sens théoriquement si on considère le cas d'un design aléatoire : une observation X_i loin de la distribution "centrale" de X aura un effet énorme sur le modèle, et peut correspondre en pratique toujours à une erreur de mesure sur les paramètres X_i.

Enfin, les leviers ont une importance théorique dans le cadre du design aléatoire, voir par exemple (Mourtada (2020), *Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices*) ou le poly de l'année dernière.

2.3.2 Modèle linéaire Gaussien

Dans le modèle linéaire général, le modèle au sens statistique pouvait s'écrire

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\bigotimes_{i=1}^n P_{\langle \theta, X_i \rangle + \varepsilon_i})_{\varepsilon, \theta}),$$

où $P_{\langle \theta, X_i \rangle + \varepsilon_i}$ désigne la loi de cette variable, étant donnée celle de l'erreur ε_i. Ce modèle est donc paramétré par θ, mais aussi par la loi commune des erreurs individuelles ε_i, que l'on suppose juste de variance σ^2. En ce sens c'est un modèle non paramétrique.

Si on suppose de plus $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, le modèle linéaire devient Gaussien et s'écrit

$$(\mathcal{N}(X\theta, \sigma^2 I_n))_{\theta \in \mathbb{R}^k},$$

et on retombe sur un modèle paramétrique, où on observe un vecteur Gaussien de moyene $X\theta$ et covariance $\sigma^2 I_n$, $k+1$ paramètres à estimer donc.

Les résultats supplémentaires par rapport au chapitre précédents concernent les lois des estimateurs précédemment déterminés, desquelles vont découler des régions de confiance et tests sur θ.

Estimation de θ et σ dans le cas Gaussien

Théorème 2.19 : Lois des estimateurs dans le modèle linéaire Gaussien

Dans le modèle linéaire Gaussien homoscédastique, les estimateurs $\hat{\theta}_{LS} = (X^TX)^{-1}X^TY$ et résidus $\hat{\varepsilon} = Y - X\hat{\theta}_{LS}$ vérifient

- $\hat{\theta}_{LS} \sim \mathcal{N}(\theta, \sigma^2(X^TX)^{-1})$,
- $\hat{\varepsilon} \sim \mathcal{N}(0_n, \sigma^2(I_n - H))$,
- $\hat{\theta}_{LS} \perp\!\!\!\perp \hat{\varepsilon}$.

En particulier, $\hat{\sigma}^2 = \frac{1}{n-k}\|\hat{\varepsilon}\|^2 \sim \frac{\sigma^2}{n-k}\chi^2(n-k)$, et $\hat{\sigma}^2 \perp\!\!\!\perp \hat{\theta}_{LS}$.

Démonstration. On rappelle que H est la matrice de projection orthogonale sur $V(X)$. Le théorème de Cochran indique alors que HY et $(I_n - H)Y = \hat{\varepsilon}$ sont indépendants, et $\|\hat{\varepsilon}\|^2 \sim \sigma^2\chi^2(n-k)$. On a immédiatement

$$\hat{\varepsilon} \sim \mathcal{N}((I_n - H)(X\theta), \sigma^2(I_n - H)) = \mathcal{N}(0_n, \sigma^2(I_n - H))$$
$$\hat{\sigma}^2 \sim \sigma^2\frac{\chi^2(n-k)}{n-k}.$$

Par ailleurs, $HY = X\hat{\theta}_{LS} \sim \mathcal{N}(H(X\theta), \sigma^2 H) = \mathcal{N}(X\theta, \sigma^2 H)$. On en déduit

$$\hat{\theta}_{LS} = (X^TX)^{-1}X^T(HY) \sim \mathcal{N}(\theta, \sigma^2(X^TX)^{-1}X^THX(X^TX)^{-1}) = \mathcal{N}(\theta, \sigma^2(X^TX)^{-1}),$$

et $\hat{\theta}_{LS} \perp\!\!\!\perp \varepsilon$, $\hat{\theta}_{LS} \perp\!\!\!\perp \hat{\sigma}^2$. □

Ellipsoïde de confiance, σ^2 connu

Dans le cas où σ^2 est connu, on peut bâtir une région de confiance pour θ à partir de la quantité pivotale

$$\frac{(X^TX)^{\frac{1}{2}}(\hat{\theta}_{LS} - \theta)}{\sigma} \sim \mathcal{N}(0_k, I_k).$$

En effet, si $q_{k,1-\alpha}$ désigne le $1-\alpha$-quantile d'une $\chi^2(k)$, alors

$$\left\{u \mid \|(X^TX)^{\frac{1}{2}}(\hat{\theta}_{LS} - u)\|^2 \leq \sigma^2 q_{k,1-\alpha}\right\} = \left\{u \mid \|X(\hat{\theta}_{LS} - u)\|^2 \leq \sigma^2 q_{k,1-\alpha}\right\}$$

est une région (elliptique) de niveau de confiance $1-\alpha$.

Dans le cas particulier unidimensionnel $Y_i \sim \mathcal{N}(\theta X_i, \sigma^2)$, on peut partir de

$$\frac{\|X\|}{\sigma}(\hat{\theta}_{LS} - \theta) \sim \mathcal{N}(0,1)$$

pour obtenir un intervalle de niveau de confiance $1-\alpha$:

$$\left[\hat{\theta}_{LS} \pm \frac{\sigma}{\|X\|}q_{1-\frac{\alpha}{2}}\right],$$

où $q_{1-\frac{\alpha}{2}}$ est cette fois-ci le quantile d'ordre $1-\frac{\alpha}{2}$ d'une Gaussienne standard.

Dans le cadre multidimensionnel, on peut obtenir des intervalles de confiance sur les coordonnées θ_j obtenus de la même manière. Bien que l'on puisse en déduire des rectangles de confiance pour θ, ce n'est pas l'approche standard (celle des ellipsoïdes de confiance).

Ellipsoïde de confiance, σ^2 inconnu, lois de Fisher et Student

Dans le cas où σ^2 est inconnu, on peut partir de la quantité pivotale suivante

$$\frac{\|X(\hat{\theta}_{LS} - \theta)\|^2/k}{\hat{\sigma}^2} \sim \mathcal{F}(k, n-k),$$

où $\mathcal{F}(p,q)$ est la loi de $(Z_1/p)/(Z_2/q)$, Z_1 et Z_2 étant indépendantes, de loi respectives $\chi^2(p)$ et $\chi^2(q)$. Cette loi est appelée *loi de Fisher* à(p,q) degrés de liberté, et est une loi standard (au sens où on connaît ses quantiles, ou tout du moins on sait les approcher).

Un ellipsoïde de confiance sur θ est alors donné par

$$\left\{u \mid \|X(\hat{\theta}_{LS} - u)\|^2 \leq k\hat{\sigma}^2 q_{k,n-k,1-\alpha}\right\},$$

où $q_{k,n-k,1-\alpha}$ est le $1-\alpha$-quantile d'une loi $\mathcal{F}(k, n-k)$. Comme dans le cas où σ^2 est connu, la forme de l'ellipse dépendra essentiellement des valeurs propres de la matrice X^TX.

Dans le cas unidimensionnel $Y_i \sim \mathcal{N}(\theta X_i, \sigma^2)$, on peut partir d'une autre quantité pivotale

$$\frac{\|X\|(\hat{\theta}_{LS} - \theta)}{\hat{\sigma}} \sim \mathcal{T}(n-1),$$

où $\mathcal{T}(p)$ est la loi de N/Z, avec $N \perp\!\!\!\perp Z$, N et Z étant de lois respectives $\mathcal{N}(0,1)$ et $\sqrt{\chi^2(p)/p}$. Une telle loi est appelée *loi de Student* à p degrés de libertés, et est aussi standard. Un intervalle de confiance est alors donné par

$$\left[\hat{\theta}_{LS} \pm \frac{\hat{\sigma}}{\|X\|} q_{n-1,1-\alpha/2}\right],$$

où $q_{n-1,1-\alpha/2}$ est le $1-\alpha/2$ quantile d'une $\mathcal{T}(n-1)$. Dans le cas multidimensionnel, on peut en déduire un intervalle de confiance sur θ_j :

$$\left[\hat{\theta}_{LS,j} \pm \hat{\sigma}\sqrt{(X^TX)^{-1}_{j,j}} q_{n-k,1-\alpha/2}\right],$$

$q_{n-k,1-\alpha/2}$ étant alors le $1-\alpha/2$-quantile d'une loi $\mathcal{T}(n-k)$.

Plus généralement, on peut bâtir des intervalles de confiance de type Student pour toute quantité de la forme $\langle a, \theta\rangle$, et des ellipsoïdes de confiance de type Fisher pour toute quantité de la forme $A\theta$, où $A \in M_{p,k}$.

Le problème est plus simple si on veut donner un intervalle de confiance sur σ^2 : vu que

$$\hat{\sigma}^2 \sim \frac{\chi^2(n-k)}{n-k},$$

un intervalle de niveau de confiance $1-\alpha$ sur σ^2 est donné par

$$\left[\hat{\sigma}^2 - \frac{q_{\beta_1,n-k}}{n-k}, \hat{\sigma}^2 + \frac{q_{\beta_2,n-k}}{n-k}\right],$$

où $q_{\beta,n-k}$ désigne le β-quantile d'une $\chi^2(n-k)$, et $\beta_1 < \beta_2$ vérifiant $\beta_1 + 1 - \beta_2 = \alpha$ (les cas extrêmes étant donnés par $\beta_1 = 0$ où $\beta_2 = 1$).

Concluons cette partie en remarquant qu'il est possible de construire des tests sur $A\theta$ et σ^2 en utilisant les mêmes recettes.

Intervalle de confiance en prédiction

Rappelons que le but initial était, pour une nouvelle valeur de paramètre X_{new}, de prédire la valeur $Y_{new} = \langle \theta, X_{new} \rangle + \varepsilon_{new}$ qui lui est associée. Une fois le modèle ajusté sur $(X_1, Y_1), \ldots, (X_n, Y_n)$, cela se fait via $\hat{Y}_{new} = \left\langle \hat{\theta}_{LS}, X_{new} \right\rangle$.

En remarquant que (X_{new}, Y_{new}) est indépendant de $\hat{\theta}_{LS}$, on a

$$\begin{aligned} Y_{new} - \hat{Y}_{new} &= \left\langle (\theta - \hat{\theta}_{LS}), X_{new} \right\rangle + \varepsilon_{new} \\ &\sim \mathcal{N}(0, \sigma^2 (X_{new}^T (X^T X)^{-1} X_{new} + 1)). \end{aligned}$$

Comme $\hat{\sigma}^2$ est indépendante de $\hat{\theta}_{LS}$ (et donc de $\hat{Y}_{new}$), on en déduit que

$$\frac{Y_{new} - \hat{Y}_{new}}{\hat{\sigma}\sqrt{(X_{new}^T (X^T X)^{-1} X_{new} + 1)}} \sim \mathcal{T}(n-k).$$

Un intervalle de confiance sur Y_{new} est alors donné par

$$\left[\hat{Y}_{new} \pm \hat{\sigma}\sqrt{(X_{new}^T (X^T X)^{-1} X_{new} + 1)} q_{1-\alpha/2, n-k} \right],$$

où $q_{1-\alpha/2,n-k}$ est le $1-\alpha/2$ quantile d'une loi $\mathcal{T}(n-k)$.

Test de Fisher d'appartenance à un sous-espace (ANOVA)

En pratique, dans le modèle linéaire Gaussien,

$$Y_i = \sum_{j=1}^{k} \theta_j X_{i,j} + \varepsilon_i,$$

une question qui revient souvent est "le j-ème paramètre est il vraiment important" ?

Par exemple, dans le modèle où $X_{i,1}$ représente la dose d'engrais reçu par le plant de maïs i, une question intéressante peut être de déterminer si cette dose d'engrais a bien une influence sur la hauteur de plant Y_i. Dans le modèle, cette question revient à déterminer si $\theta_1 = 0$ ou non.

De manière plus générale, on peut considérer le test d'hypothèses

$$\begin{cases} H_0 & : \quad \theta \in W_0, \\ H_1 & : \quad \theta \notin W_0, \end{cases}$$

où W_0 est un sous-espace vectoriel de dimension $k_0 < k$ de $\mathbb{R}^k$. On peut penser par exemple à

$$W_0 = \{u \in \mathbb{R}^k \mid u_1 = u_2 = \ldots u_{k-k_0} = 0\},$$

qui est de dimension k_0 et revient à tester si les $k-k_0$ premières variables explicatives du modèle on une influence.

En notant $V_0 = X(W_0)$ (image de W_0 par X), on a $\dim(W_0) = \dim(V_0) = k_0$ (X étant de rang k, elle est injective). Un problème équivalent est alors de tester si $X\theta \in V_0 \lhd V(X)$, avec pour alternative $X\theta \in V(X) \lhd \mathbb{R}^n$.

Le principe est alors de remarquer que, sous H_0, $X\theta - \pi_{V_0}(X\theta) = 0$. On a alors

$$X\hat{\theta}_{LS} - \pi_{V_0}(X\hat{\theta}_{LS}) \sim \mathcal{N}(0_n, \sigma^2 \pi_{V_0^\perp} \pi_{V(X)} \pi_{V_0^\perp}) = \mathcal{N}(0_n, \sigma^2 \pi_{V_0^\perp \cap V(X)}).$$

Le paramètre σ^2 étant inconnu, on peut l'estimer par $\hat{\sigma}^2$, qui est indépendant de $X\hat{\theta}_{LS}$. On en déduit alors que, sous H_0,

$$S(Y) = \frac{\|\pi_{V(X)\cap V_0^\perp}(Y)\|^2/(k-k_0)}{\|\pi_{V(X)^\perp}(Y)\|^2/(n-k)} = \frac{\|\pi_{V_0^\perp}(X\hat{\theta}_{LS})\|^2/(k-k_0)}{\hat{\sigma}^2} \sim \mathcal{F}(k-k_0, n-k).$$

Le test de Fisher associé, aussi appelé ANOVA (analyse de variance), est

$$\mathbb{1}_{S(Y)\geq q_{1-\alpha,k-k_0,n-k}},$$

où $q_{1-\alpha,k-k_0,n-k}$ est le quantile d'ordre $1-\alpha$ d'une loi $\mathcal{F}(k-k_0, n-k)$, et est bien de niveau α.

L'idée derrière le terme ANOVA est que l'on peut décomposer la "variance globale", ou "inertie" de Y, définie par $I_{tot}(Y) = \|Y\|^2$, via

$$I_{tot}(Y) = \|\pi_{V_0}(Y)\|^2 + \|\pi_{V_0^\perp\cap V(X)}(Y)\|^2 + \|\pi_{V(X)^\perp}(Y)\|^2,$$

les trois termes étant indépendants par ailleurs dans le modèle Gaussien. Le dernier terme peut s'interpréter comme une variance résiduelle (la partie non expliquée par le modèle linéaire), la somme des deux premiers comme la variance expliquée par le modèle "total" $V(X)$ que l'on suppose être juste.

Cette variance expliquée par le modèle total peut elle même se décomposer comme la somme de la variance expliquée par le sous-modèle V_0 et d'une partie expliquée par la partie de V orthogonale à V_0. Une Anova consiste alors à comparer la variance expliquée par V sans V_0 à la variance résiduelle, qui sous H_0 doivent être de même nature (aux dimensions près).

Puissance du test de Fisher

Comme pour la loi du χ^2, on peut définir la loi de Fisher décentrée.

Definition 2.20 : Loi de Fisher décentrée

Z suit une loi de Fisher décentrée $\mathcal{F}(p, q, \|\mu\|)$ si

$$Z \sim \frac{Y_1/p}{Y_2/q},$$

où $Y_1 \sim \chi^2(p, \|\mu\|)$, $Y_2 \sim \chi^2(q)$, $Y_1 \perp\!\!\!\perp Y_2$.

Si $\theta \notin W_0$, en notant $\mu = \pi_{V(X)\cap V_0^\perp}(X\theta)$, on a

$$S(Y) = \frac{\|\pi_{V(X)\cap V_0^\perp}(Y)\|^2/(k-k_0)}{\|\pi_{V(X)^\perp}(Y)\|^2/(n-k)} \underset{P_\theta}{\sim} \mathcal{F}(k-k_0, n-k, \|\mu\|).$$

On a alors, en notant $q_{1-\alpha,k-k_0,n-k}$ le $1-\alpha$ quantile d'une loi $\mathcal{F}(k-k_0, n-k)$,

$$\begin{aligned} P_\theta\left(S(Y)=1\right) &= \mathbb{P}\left(\mathcal{F}(k-k_0, n-k, \|\mu\|) \geq q_{1-\alpha,k-k_0,n-k}\right) \\ &\underset{\|\mu\|\to 0}{\longrightarrow} \mathbb{P}\left(\mathcal{F}(k-k_0, n-k) \geq q_{1-\alpha,k-k_0,n-k}\right) \\ &= \alpha, \end{aligned}$$

car $\mathcal{F}(k-k_0, n-k, \|\mu\|) \rightsquigarrow \mathcal{F}(k-k_0, n-k)$ lorsque $\|mu\| \to 0$, et $\mathcal{F}(k-k_0, n-k)$ est diffuse. En remarquant que si $\|\mu\| \leq \|\mu'\|$, $\mathcal{F}(k-k_0, n-k, \|\mu\|) \preccurlyeq \mathcal{F}(k-k_0, n-k, \|\mu'\|)$, on en déduit que la puissance minimale pour $H_1 : \theta \notin W_0$ vaut α.

On peut aussi en déduire que si H_1 est de la forme $\mathrm{d}(\theta, W_0) \geq t$, on a $\inf_{\mathrm{d}(\theta,W_0)\geq t} \mathrm{d}(X\theta, V_0) = u = \sqrt{\lambda_{\min}}t$, où $\lambda_{\min} = \lambda_{\min}(X^TX)$. La puissance minimale sur H_1 est alors

$$\mathbb{P}\left(\mathcal{F}(k-k_0, n-k, u) \geq q_{1-\alpha,k-k_0,n-k}\right).$$

Comme $\mathcal{F}(k-k_0, n-k, u) \rightsquigarrow \chi^2(k-k_0, u)$ lorsque $n \to +\infty$ (qui est diffuse), et $q_{1-\alpha,k-k_0,n-k} \to q_{1-\alpha,k-k_0}$ (quantile d'un $\chi^2(k-k_0)$, on aura

$$\inf_{\mathrm{d}(\theta,W_0)\geq t} P_\theta\left(S(Y)=1\right) \underset{n\to+\infty}{\longrightarrow} 1$$

dès lors que $\lambda_{\min}(X^TX) \to_{n\to+\infty} +\infty$.

Exemple 2.21 *: Test d'égalité de k moyennes.*

Dans le cas où on observe la même variable Y (par exemple la durée de vie) sur n individus divisés en k sous-catégories (par exemple en fonction de la catégorie socio-professionnelle) équilibrées (d'effectif n/k), on peut se demander si la moyenne de Y est la même catégorie par catégorie, pour éventuellement ajuster une réforme des retraites.

D'un point de vue stats, ce problème se traduit via le modèle Gaussien

$$Y = X\theta + \varepsilon,$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$,

$$X = \begin{pmatrix} \mathbf{1}_{n/k} & 0 & \dots & 0 \\ 0 & \mathbf{1}_{n/k} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{1}_{n/k} \end{pmatrix},$$

et $\theta = (\theta_1, \dots, \theta_k)^T$ représente le vecteur des moyennes conditionnelles de Y_i. On teste alors

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_k,$$

contre le contraire. On a $W_0 = \mathrm{Vect}(\mathbf{1}_k)$, $V_0 = \mathrm{Vect}(\mathbf{1}_n)$.

En notant $\bar{Y_j} = \frac{\sum_{i=(j-1)n/k+1}^{jn/k} Y_i}{n/k}$ la moyenne observée pour la condition j, on a

$$\|\pi_{V(X)\cap V_0^\perp}(Y)\|^2 = \frac{n}{k}\sum_{j=1}^{k}(\bar{Y_j} - \bar{Y_n})^2,$$

$$\|\pi_{V(X)^\perp}(Y)\|^2 = \sum_{j=1}^{k}\sum_{i=(j-1)n/k+1}^{jn/k}(Y_i - \bar{Y_j})^2.$$

La première quantité peut s'interpréter comme une *variance inter classes*, la seconde quantité comme une *variance intra-classes* et on peut déduire un test T de niveau α basé sur le ratio de ces deux variances (ce qui est une autre manière d'expliquer le terme ANalysis Of Variances).

Une dernière remarque : dans ce cas $X^TX = \frac{n}{k}I_k$, et $\lambda_{\min}(X^TX) = \frac{n}{k}$. On en déduit alors que si $\theta \notin H_0$,

$$P_\theta(T(X)=1) \underset{n\to+\infty}{\longrightarrow} 1,$$

ou bien encore que ce test sera consistant dès lors que l'on prend pour H_1 une hypothèse de type "il existe deux moyennes conditionnelles séparées d'au moins u".

Chapitre 3

Maximum de vraisemblance

3.1 Méthodes d'estimations classiques

Jusque ici nous avons vu des méthodes d'estimation ad-hoc suivant les modèles (moyennes dans le modèle Gaussien, moindres carrés pour le modèle linéaire, max dans un modèle uniforme, etc.). Ces méthodes et les manières de les traiter théoriquement font partie de méthodes générales relativement classiques en statistiques (M-estimation et moments), qu'on va maintenant expliciter.
Remarque : En pratique, le "fait maison" quand il est possible est souvent l'approche la plus pertinente. Cette généralisation est surtout intéressante d'un point de vue théorique.

3.1.1 Méthode des moments

Dans un modèle où on observe $X_1, \dots, X_n$ i.i.d. de loi commune P_θ, où $\theta \in \Theta \subset \mathbb{R}^k$, une approche "naturelle" consiste à trouver un système de k équations linéaires déterminant entièrement θ, via des fonctions tests $f_1, \dots, f_k$, c'est à dire en écrivant

$$E_\theta(f_j(X))_{j=1,\dots,k} = \psi(\theta). \tag{3.1}$$

Si $\psi : \Theta \to \mathbb{R}^k$ est inversible, θ est déterminé par ces k équations, c'est à dire

$$\theta = \psi^{-1}(E_\theta(\mathbf{f}(X))),$$

où $\mathbf{f} = (f_1, \dots, f_k)^T$. Un exemple simple est donné par le modèle Gaussien où on observe n variables $\mathcal{N}(\mu, \sigma^2)$ i.i.d. : en prenant $f_1(x) = x$ et $f_2(x) = x^2$, on a

$$\begin{aligned} E_\theta(f_1(X)) &= \mu \\ E_\theta(f_2(X)) &= \mu^2 + \sigma^2. \end{aligned}$$

On peut alors définir $\psi^{-1}(x, y) = (x, y - x^2)$ qui permet alors de retrouver θ à partir de $(E_\theta(f_1(X)), E_\theta(f_2(X))$.

Une fois ce système théorique d'équation trouvé, la procédure d'estimation par méthode des moments consiste à remplacer $E_\theta(\mathbf{f})$ par sa version empirique (basée sur échantillon) $E_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^n \mathbf{f}(X_i)$. On définit alors l'estimateur des moments par

$$\hat{\theta}_{Moments} = \psi^{-1}(E_n(\mathbf{f})).$$

Toujours dans l'exemple Gaussien, cela revient à prendre

$$\hat{\mu}_{Moments} = \bar{X}_n,$$
$$\hat{\sigma}^2_{Moments} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

On retrouve alors les estimateurs par moindres carrés dans le modèle linéaire gaussien correspondant, à un facteur $(n/n-1)$ près pour la variance.

Un des avantages de la méthode de cette méthode est qu'elle fournit des estimateurs qui suivent naturellement des lois normales, sous certaines hypothèses.

THÉORÈME 3.1 : NORMALITÉ ASYMPTOTIQUE DES ESTIMATEURS PAR MOMENTS

Soit $\theta \in \mathring{\Theta}$. Si $\psi : \theta \mapsto E_\theta(\mathbf{f})$ est localement $\mathcal{C}^1$ autour de θ, avec $D_\theta\psi$ inversible, et $E_\theta\|\mathbf{f}(X)\|^2 < +\infty$, alors

$$\sqrt{n}\left(\hat{\theta}_{Moments} - \theta\right) \rightsquigarrow \mathcal{N}(0_k, (D_\theta\psi)^{-1}\mathrm{Cov}_\theta(\mathbf{f}(X))((D_\theta\psi)^{-1})^T).$$

Démonstration. C'est une conséquence de la méthode Δ. On commence par se donner U voisinage ouvert de θ dans Θ et V voisinage ouvert de $E_\theta(f(X))$ dans $\mathbb{R}^k$ tels que ψ soit un $\mathcal{C}^1$-difféomorphisme entre U et V. On pose ensuite $Z_n = \mathbb{1}_{E_n(\mathbf{f})\in V}$. Comme $E_n(\mathbf{f}) \xrightarrow{\mathbb{P}} E_\theta(\mathbf{f})$ (loi des grands nombres), on a $Z_n \xrightarrow{\mathbb{P}} 1$. Lorsque $Z_n = 1$, $\hat{\theta}_{Moments}$ est bien défini par $\psi^{-1}(E_n(\mathbf{f}))$. Lorsque $Z_n = 0$, prenons par convention $\hat{\theta}_{Moments} = 0$ (valeur arbitraire peu importante).

Comme $E_\theta\|\mathbf{f}(X)\|^2 < +\infty$, le théorème central limite donne

$$\sqrt{n}\left(E_n(\mathbf{f}) - \psi(\theta)\right) \rightsquigarrow \mathcal{N}(0_k, \mathrm{Cov}_\theta(\mathbf{f}(X))).$$

En utilisant le Lemme de Slutsky il vient

$$\sqrt{n}Z_n(\hat{\theta}_{Moments} - \theta) \rightsquigarrow (D_{\psi(\theta)}(\psi^{-1}))\mathcal{N}(0_k, \mathrm{Cov}_\theta(\mathbf{f}(X))).$$

On conclut avec Slutsky encore en utilisant $Z_n \xrightarrow{\mathbb{P}} 1$. □

En pratique, on prend souvent comme fonction f_j des fonctions puissance (de coordonnées éventuellement), d'où l'appellation "méthode des moments". Dans le cadre des modèles exponentiels, d'autres fonctions naturelles apparaissent. Enfin, dans un contexte de M-estimation, des fonctions test peuvent apparaître en prenant le gradient d'une fonction de risque.

3.1.2 M-estimation

On se place toujours dans un modèle i.i.d., où les lois individuelles des X_i suivent P_θ, pour $\theta \in \Theta \subset \mathbb{R}^k$. Le principe général de la M-estimation se base sur la donnée d'une *fonction de contraste* $\gamma : \Theta \times \mathcal{X} \to \mathbb{R}$ telle que

$$M : \begin{cases} \Theta & \to & \mathbb{R} \\ u & \mapsto & E_\theta(\gamma(u, X)), \end{cases} \tag{3.2}$$

soit minimale en θ. La fonction M est parfois appelée risque : le risque d'un estimateur $\hat{\theta}_n$ est alors $M(\hat{\theta}_n) = \mathbb{E}\left(\gamma(\hat{\theta}_n, X_{new}) \mid X_{1:n}\right)$, c'est à dire ce que l'on paye en moyenne en prédiction, lorsque une nouvelle donnée X_{new} arrive de loi P_θ.

L'idée est alors de trouver un estimateur $\hat{\theta}_{M-est}$ non pas en minimisant $M(\theta)$ (inaccessible) mais plutôt sa version empirique $M_n(\theta) = E_n(\gamma(u, .))$, c'est à dire en choisissant

$$\hat{\theta}_{M-est} \in \arg\min_{\Theta} E_n(\gamma(u, .)) = \arg\min_{\Theta} M_n(u).$$

Remarque : Il se peut qu'un tel estimateur ne soit pas défini (minimiseur en dehors des frontières ou à l'infini). On verra un exemple d'un tel cas pathologique dans la section maximum de vraisemblance. Sous certaines conditions (convexité du risque, cible à l'intérieur par exemple), on peut énoncer des résultats d'existence généraux.

Le point de vue M-estimation permet de retrouver les estimateurs des moindre carrés encore : par exemple, dans le modèle $(\mathcal{P}(\theta)_{\theta>0}$, pour la fonction de contraste $\gamma(u, x) = (u-x)^2$, on a bien $\theta \in \arg\min_{]0,+\infty[} M(u)$. Le M-estimateur correspondant est alors $\bar{X}_n$.

On peut aussi récupérer des estimateurs plus généraux, comme la médiane empirique : dans un modèle $(\theta + \frac{1}{2}x^{-2}\mathbb{1}_{|x|>1}dx)_{\theta\in\mathbb{R}}$, on a $\theta \in \arg\min M(u)$, avec $\gamma(u, x) = |u-x|$, le M-estimateur correspondant est la médiane empirique de l'échantillon.

On remarque que d'un point de vue empirique, peu importe le modèle, les estimateurs de médiane et moyenne empirique peuvent toujours être définis par ce biais. Les M-estimateurs sont par nature plus du domaine de la statistique *prédictive* qu'inférentielle (on se fiche du vrai θ, on veut un estimateur performant en prédiction, sur de nouvelles données).

Dernière remarque : il existe des passerelles entre méthode des moments et M-estimation. Dans les conditions où on peut intervertir intégrale et dérivation, la solution de $\arg\min_u [E_\theta \langle u, \mathbf{f}(X)\rangle + b(u)]$ vérifie $E_\theta(\mathbf{f}(X)) = \nabla_\theta b$, et on se retrouve avec un estimateur par méthode des moments. Inversement, si on prend comme fonction de contraste $\gamma(u, x) = \|\mathbf{f}(x) - E_u(\mathbf{f}(x))\|^2$, le M-estimateur correspondant permet de retomber sur l'estimateur par moments.

Recette pour les M-estimateurs

Les garanties que l'on peut obtenir sur les M-estimateurs dépendent généralement de la conjonction des deux ingrédients suivants :

1. un résultat de concentration **uniforme** de type $\sup_\theta |M_n(\theta) - M(\theta)|$, permettant de borner $M(\hat{\theta}_{M-est})$,
2. un résultat structurel sur la fonctionnelle M permettant de relier $M(\hat{\theta}_{M-est}) - M(\theta)$ à $\|\hat{\theta}_{M-est} - \theta\|$ (type convexité locale).

Plutôt que d'énoncer un résultat général, regardons sur un exemple.

Exemple 3.2 *: Retour sur le modèle linéaire homoscédastique.* Dans le modèle linéaire homoscédastique $Y = X\theta + \varepsilon$, où les ε_i sont i.i.d. de moyenne nulle et de variance σ^2, on peut considérer l'estimateur par moindre carrés comme un M-estimateur : en effet

$$\theta \in \arg\min_u M(u),$$
$$\hat{\theta}_{LS} \in \arg\min M_n(u),$$

où $M(u) = E_\theta\|Y' - Xu\|^2$ (où Y' est une copie de Y indépendante, c'est un risque en *prédiction*), et $M_n(u) = \|Y - Xu\|^2$ (risque observé).

On peut alors écrire

$$\begin{aligned} M(\hat{\theta}_{LS}) - M(\theta) &= (M - M_n)(\hat{\theta}_{LS}) - (M - M_n)(\theta) + M_n(\hat{\theta}_{LS}) - M_n(\theta) \\ &\leq (M - M_n)(\hat{\theta}_{LS}) - (M - M_n)(\theta), \end{aligned}$$

car $\hat{\theta}_{LS}$ minimise M_n. Si on était dans un cadre i.i.d. on pourrait s'en sortir en majorant $\sup_u(M - M_n)(u)$ (arguments de concentration uniforme). Ici, on peut ruser un peu. En effet,

$$\begin{aligned} M_n(\hat{\theta}_{LS}) - M_n(\theta) &= \sum_{i=1}^n (\left\langle \theta - \hat{\theta}_{LS}, x_i \right\rangle + \varepsilon_i)^2 - \varepsilon_i^2 \\ &= \sum_{i=1}^n \left\langle \theta - \hat{\theta}_{LS}, x_i \right\rangle^2 + 2\sum_{i=1}^n \left\langle \theta - \hat{\theta}_{LS}, x_i \right\rangle \varepsilon_i \\ &= \|X(\theta - \hat{\theta}_{LS})\|^2 + 2\varepsilon^T X(\hat{\theta}_{LS} - \theta). \end{aligned}$$

En se souvenant que $M(u) - M(\theta) = \|X(u - \theta)\|^2$, on en déduit

$$\begin{aligned} M(\hat{\theta}_{LS}) - M(\theta) &\leq \left|2\varepsilon^T(X(\hat{\theta}_{LS} - \theta)\right| \\ &\leq 2\|(X(\hat{\theta}_{LS} - \theta)\|\|\pi_{V(X)}(\varepsilon)\| \\ &\leq 2\sqrt{M(\hat{\theta}_{LS}) - M(\theta)}\|\pi_{V(X)}(\varepsilon)\|. \end{aligned}$$

On en déduit alors

$$M(\hat{\theta}_{LS}) - M(\theta) \leq 4\|\pi_{V(X)}(\varepsilon)\|^2.$$

On n'a pas utilisé la formule exacte pour $\hat{\theta}_{LS}$, en particulier, si $\text{rang}(X) \geq n$, on aura quand même la borne

$$M(\hat{\theta}_{LS}) - M(\theta) \leq 4\|\varepsilon\|^2,$$

dont on peut déduire

$$E_\theta\left(M(\hat{\theta}_{LS}) - M(\theta)\right) \leq 4n\sigma^2,$$

et éventuellement une borne avec grande proba sur $M(\hat{\theta}_{LS}) - M(\theta)$. Dans le cas classique $\text{rang}(X) = k < n$, on retombe sur

$$E_\theta\left(M(\hat{\theta}_{LS}) - M(\theta)\right) \leq 4k\sigma^2,$$

et les bornes en proba correspondantes. Cette borne est toutefois moins bonne que celle obtenue avec l'espression directe de $\hat{\theta}_{LS}$ (d'un facteur 4). De manière générale, on a toujours intérêt à considérer l'expression exacte de l'estimateur si disponible.
Remarque : On peut aussi trouver une expression "exacte" de $\hat{\theta}_{LS}$ dans le cadre $\text{rang}(X) \geq n$ (à coups de pseudo-inverses), conduisant à la même borne sans le facteur 4.

Si maintenant on veut relier $M(\hat{\theta}_{LS}) - M(\theta) = \|X(\hat{\theta}_{LS} - \theta)\|^2$, on doit obligatoirement supposer X^TX inversible : en effet, sinon pour $u \in (\text{im}(X^TX))^\perp$, on a

$\|Xu\|^2 = 0$ et $\|u\|$ potentiellement > 0. Une minoration de type $\|u\| \leq \|Xu\|$ semble alors impossible. Par ailleur, dans ce cas, θ n'est pas le seul minimiseur de M, et c'est aussi un cas où le modèle n'est pas identifiable (plusieurs θ donnent la même loi d'observation). Chercher à estimer θ est sans espoir.

On se place donc dans le cas (X^TX) inversible. Pour relier $M(u)-M(\theta)$ à $\|u-\theta\|$ en toute généralité, on peut écrire

$$M(u) - M(\theta) \geq \lambda_{\min}\|u-\theta\|^2,$$

où $\lambda_{\min}$ est la plus petite valeur propre de X^TX. On a alors

$$\|\hat{\theta}_{LS} - \theta\|^2 \leq \frac{4\|\pi_{V(X)}(\varepsilon)\|^2}{\lambda_{\min}},$$

ce qui est moins bon que l'expression trouvée en explicitant $\hat{\theta}_{LS}$, mais donne les bons ordres de grandeur toutefois.

Exemple 3.3 *: Régression linéaire par moindres écarts absolus.* On considère maintenant le modèle $Y_i = \theta X_i + e_i$, où $\theta \in]0,1[$, les (X_i, e_i) sont i.i.d., avec $X_i \perp\!\!\!\perp e_i$, $X_i \sim \mathcal{U}(]0,1[)$, et les e_i sont des erreurs symétriques sur $]-1,1[$, de densité $f \geq c > 0$. En particulier, $Y_i \in]-2,2[$. Dans ce modèle on observe $(X_i, Y_i)_{i=1,\dots,n}$. Le but est de faire de la régression, c'est à dire d'essayer de prévoir Y par une fonction de la forme uX.

On considère la fonction de contraste

$$\gamma : \begin{cases}]0,1[\times]0,1[\times]-2,2[& \to & \mathbb{R}^+ \\ (u,x,y) & \mapsto & |(y-ux)|. \end{cases}$$

L'objectif est alors de minimiser $E_\theta(|Y_{new} - uX_{new}|) = M(u)$. Le risque empirique associé est alors

$$M_n(u) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - uX_i|.$$

Comme M_n est convexe, elle admet un minimiseur $\hat{\theta}_{M-est}$. Une expression exacte de $\hat{\theta}_{M-est}$ est assez pénible à obtenir (via appartenance de 0 au sous-gradient), il n'y a même pas unicité de $\hat{\theta}_{M-est}$ en général.

Pour essayer de dire des choses sut $\hat{\theta}_{M-est}$, on part comme dans l'exemple précédent :

$$\begin{aligned} M(\hat{\theta}_{M-est}) - M(\theta) &= (M-M_n)(\hat{\theta}_{M-est}) - (M-M_n)(\theta) + M_n(\hat{\theta}_{M-est}) - M_n(\theta) \\ &\leq (M-M_n)(\hat{\theta}_{M-est}) - (M-M_n)(\theta). \end{aligned}$$

Cette fois-ci on va majorer brutalement les déviations par $\sup_u (M-M_n)(u)$, ce qui donne

$$M(\hat{\theta}_{M-est}) - M(\theta) \leq 2 \sup_{u\in]0,1[} |(M_n - M)(u)|.$$

Reste à majorer ce sup de déviations. A ce niveau, une manière pédestre de faire est de se ramener à un sup sur un ensemble fini, ce qui est possible en gros si la fonction

de contraste est suffisamment régulière en u. C'est le cas ici : on a, pour tout u,v dans $]0,1[$, et (x,y),

$$|\gamma(u,x,y) - \gamma(v,x,y)| \leq |u-v|.$$

On peut alors en déduire que pour tout u,v dans $]0,1[$,

$$|(M-M_n)(u) - (M-M_n)(v)| \leq 2|u-v|.$$

Par conséquent, si $\varepsilon > 0$, en notant C_ε un grillage régulier de $]0,1[$ de pas ε (comptant au plus $\lceil \frac{1}{\varepsilon} \rceil$ éléments), on a

$$\sup_{u\in]0,1[} |(M_n - M)(u)| \leq \sup_{u\in C_\varepsilon} |(M_n - M)(u)| + 2\varepsilon.$$

On peut maintenant travailler ponctuellement : pour un u fixé, l'inégalité de Hoeffding donne,

$$\mathbb{P}\left(|(M-M_n)(u)| \geq 2\sqrt{\frac{2x}{n}} \right) \leq 2e^{-x}.$$

On en déduit alors que, pour $x > 0$,

$$\mathbb{P}\left(\sup_{u\in C_\varepsilon} |(M-M_n)(u)| \geq 2\sqrt{\frac{2x}{n}} \right) \leq 2\lceil \frac{1}{\varepsilon} \rceil e^{-x},$$

en appliquant une borne d'union. On peut réécrire cette borne en

$$\mathbb{P}\left(\sup_{u\in C_\varepsilon} |(M-M_n)(u)| \geq 2\sqrt{\frac{2(x+\log\left(2\lceil \frac{1}{\varepsilon} \rceil\right)}{n}} \right) \leq e^{-x},$$

On en déduit alors que, avec probabilité plus grande que $1-e^{-x}$,

$$\sup_{u\in]0,1[} |(M_n - M)(u)| \leq 2\sqrt{\frac{2(x+\log\left(2\lceil \frac{1}{\varepsilon} \rceil\right)}{n}} + 2\varepsilon.$$

En choisissant $\varepsilon = 1/\sqrt{n}$ pour équilibrer les deux termes, et pour $n \geq 16$ (assez grand quoi), on peut mettre cette borne sous la forme

$$\sup_{u\in]0,1[} |(M_n - M)(u)| \leq 4\sqrt{\frac{\log(n)}{n}} + \frac{2\sqrt{2x}}{\sqrt{n}},$$

avec probabilité plus grande que $1-e^{-x}$, ce dont on déduit

$$\mathbb{P}\left(M(\hat{\theta}_{M-est}) - M(\theta) \geq 8\sqrt{\frac{\log(n)}{n}} + \frac{4\sqrt{2x}}{\sqrt{n}} \right) \leq e^{-x},$$

et on a donc une borne en déviation sur la performance de $\hat{\theta}_{M-est}$ en termes du risque M. On peut éventuellement la convertir en borne en espérance en utilisant $E(Y) = \int_0^+ \infty \mathbb{P}(Y \geq t)dt$ (pour une variable Y positive), ce qui donne

$$\mathbb{E}\left(M(\hat{\theta}_{M-est}) - M(\theta) \right) \leq C\sqrt{\frac{\log(n)}{n}},$$

pour une constante numérique C. Les bornes en déviations sont cependant plus utiles, notamment pour construire des intervalles de confiance. On peut remarquer que toute cette partie marche dès lors que le bruit e_i est borné. Les propriétés de centrage n'interviennent pas.

Si on veut maintenant des garanties sur $|\hat{\theta}_{M-est} - \theta|$, il faut relier $M(\hat{\theta}_{M-est}) - M(\theta)$ à $|\hat{\theta}_{M-est} - \theta|$ (et c'est là que l'hypothèse de centrage du bruit intervient). Pour cela, on écrit, pour $u, \theta \in [0,1]$,

$$\begin{aligned} M(u) - M(\theta) &= \mathbb{E}\left(|(\theta-u)X + e| - |e|\right) \\ &= \int_0^1 dx \int_{-1}^1 dv (|(\theta-u)x + v| - |v|) f(v). \end{aligned}$$

Pour un x tel que $(\theta - u)x > 0$,

$$\begin{aligned} \int_{-1}^1 dv(|(\theta-u)x+v| - |v|)f(v) &= \int_{-1}^{-(\theta-u)x} -(\theta-u)x dv + \int_0^1 (\theta-u)x f(v) dv \\ &\quad + \int_{-(\theta-u)x}^0 ((\theta-u)x + 2v) f(v) dv \\ &= \int_{-1}^0 -(\theta-u)x f(v)dv + \int_0^1 (\theta-u)x f(v) dv \\ &\quad + 2\int_{-(\theta-u)x}^0 ((\theta-u)x + v) f(v) dv \\ &= 2\int_{-(\theta-u)x}^0 1((\theta-u)x + v) f(v) dv \\ &\geq 2c \int_0^{(\theta-u)x} ((\theta-u)x - v) dv \\ \geq c((\theta-u)x)^2. \end{aligned}$$

Par symétrie, on a, pour tout $x \in]0,1[$,

$$\int_{-1}^1 dv(|(\theta-u)x+v| - |v|)f(v) \geq c((\theta-u)x)^2.$$

On en déduit alors

$$\begin{aligned} M(u) - M(\theta) &\geq \int_0^1 c((\theta-u)x)^2 dx \\ &\geq c\frac{(u-\theta)^2}{3}. \end{aligned}$$

On peut alors conclure qu'avec probabilité plus grande que $1 - e^{-x}$,

$$|\hat{\theta}_{M-est} - \theta|^2 \leq \frac{24}{c}\sqrt{\frac{\log(n)}{n}} + \frac{12\sqrt{2x}}{c\sqrt{n}},$$

ce qui donne une vitesse de convergence de $\hat{\theta}_{M-est}$ vers θ de l'ordre de $n^{-1/4}$. Cette vitesse n'est pas forcément optimale : en rajoutant quelques hypothèses sur le bruit (par exemple Gaussien) et en utilisant des outils de concentration un peu plus fins on aurait plutôt une vitesse de l'ordre de $n^{-1/2}$.

Cette méthode de preuve (concentration uniforme + "convexité locale du risque") est assez générale pour obtenir des résultats non asymptotique. Insistons encore sur

le fait qu'elle conduit souvent à des résultats sous-optimaux comparativement à des méthodes qui exploitent un expression explicite de l'estimateur. Enfin, des résultats asymptotiques peuvent être obtenus en remplaçant l'étape de concentration uniforme par une étape de "convergence uniforme en loi". Dans les deux cas, il s'agit d'étudier les comportement de processus empirique (de manière non-asymptotique ou via leur convergence en loi). Pour une approche générale de ce domaine, on peut consulter le Wellner, Van der Waart, et/ou Massart.

3.2 Un peu plus sur les modèles

3.2.1 Propriétés usuelles et souhaitables des modèles

Pour faire du maximum de vraisemblance côté maths on a besoin de formaliser différents aspects des modèles rencontrés jusque ici. Une propriété fondamentale en statistique inférentielle (celle qui veut retrouver θ) est celle *d'identifiabilité*.

Definition 3.4 : Modèle Identifiable

> Un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ est *identifiable* si $\theta \mapsto P_\theta$ est injective.

L'identifiabilité du modèle est nécessaire si on veut pouvoir estimer θ à partir d'observations tirées suivant P_θ (stats inférentielles), mais ne l'est pas si l'objectif est plutôt d'approcher P_θ ou de minimiser un risque (en statistique prédictive par exemple).

Exemple 3.5. Pour l'exemple du nombre de naissances, le modèle $(\mathcal{B}(N, \theta))_{\theta\in]0,1[}$ est identifiable, car $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$.

Exemple 3.6. Un exemple classique de non-indentifiabilité est celui des modèles de mélange (utilisés en classification non-supervisée), par exemple, pour $(\theta_1, \theta_2) \in \mathbb{R}^2$, $P_{(\theta_1,\theta_2)} = \frac{1}{2}\mathcal{N}(\theta_1, \sigma_0^2) + \frac{1}{2}\mathcal{N}(\theta_2, \sigma_0^2)$. Comme, pour $\theta_1 \neq \theta_2$, $P_{(\theta_1,\theta_2)} = P_{(\theta_2,\theta_1)}$, le modèle n'est pas identifiable. D'un point de vue prédictif, estimer (θ_1, θ_2) à permutation près suffit donc on s'en fiche un peu. On peut aussi rendre le modèle identifiable en "fixant" l'ordre, c'est à dire en prenant $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 \mid \theta_1 \leq \theta_2\}$.

Exemple 3.7. Dernier exemple déjà rencontré de non-identifiabilité : le modèle linéaire Gaussien avec $(X^T X)$ non inversible. Dans ce cas, pour un $\theta \in \mathbb{R}^k$ donné, n'importe quel $\theta + u$ avec $u \in Ker(X)$ (tel que $X\theta = X(\theta + u)$) donnera la même loi que P_θ.

Proposition 3.8 : Non-identifiable implique inconsistance

> *Dans un modèle $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta})$, aucun estimateur ne peut être consistant.*

Démonstration. Soit T un estimateur consistant, et $\theta_1 \neq \theta_2$ tels que $P_{\theta_1} = P_{\theta_2}$. On a alors

$$\lim_{n\to+\infty} P_{\theta_1}^{\otimes n}(T = \theta_1) = 1 = \lim_{n\to+\infty} P_{\theta_2}^{\otimes n}(T = \theta_1).$$

Donc $\lim_{n\to+\infty} P_{\theta_2}^{\otimes n}(T = \theta_2) = 0$, et T n'est pas consistant. □

La réciproque est vraie modulo le fait que l'on arrive à relier proximité entre P_θ et $\mathbb{P}'_\theta$ à $|\theta - \theta'$. Il existe donc plusieurs résultats, qui dépendent de la manière dont on mesure la proximité entre P_θ et $\mathbb{P}'_\theta$. Citons par exemple celui concernant la variation totale.

Definition 3.9 : Variation Totale

Si P et Q sont deux lois sur $\Omega, \mathcal{F}$, la *distance en variation totale* entre P et Q est définie par

$$\mathrm{d}_{TV}(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

On peut vérifier que cela définit bien une distance sur l'espace des mesures de probas sur $(\Omega, \mathcal{F})$. Si distance et variation totale et distance Euclidenne sur Θ sont compatibles, l'identifiabilité implique l'existence d'un estimateur consistant.

Théorème 3.10 : Identifiabilité implique consistance

Si le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$ est identifiable, $\Theta \subset \mathbb{R}^k$, et d_{TV} est équivalente à $\|.\|$ sur Θ, alors il existe un estimateur consistant.

Preuve : cf Ibragimov Theorem 4.1. □

La deuxième hypothèse courante que l'on peut faire sur un modèle est celle de la domination.

Definition 3.11 : Modèle dominé

Le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$ est *dominé* (par une mesure μ sigma finie sur $\mathcal{X}$) si $P_\theta << \mu$, pour tout $\theta \in \Theta$. Dans ce cas on note p_θ ou $p(. \mid \theta)$ la densité de P_θ par rapport à μ (unique μ-p.s).

C'est une hypothèse techniquement utile qui permet de travailler avec des densités plutôt qu'avec des distributions. Elle est nécessaire pour plusieurs méthodes d'estimation (maximum de vraisemblance en particuler), pour définir certaines distances entre distributiosn (type Kullback ou Hellinger), et est vérifiée dans la plupart des modèles classiques.

Exemple 3.12.

- Le modèle $\Big([0, +\infty[, \mathcal{B}([0, +\infty[), (\mathcal{E}(\theta))_{\theta > 0}\Big)$, correspondant à l'observation d'une variable aléatoire exponentielle X de paramètre $\theta > 0$ est dominé par λ_1 (mesure de Lebesgue sur $[0, +\infty[$.
- Le même modèle "infatué en 0", c'est à dire $(\pi\mathcal{E}(\theta) + (1 - \pi)\delta_0)_{\theta > 0, \pi \in [0,1]}$ reste dominé, par $\delta_0 + \lambda_1$, les $P_{\theta,\pi}$ ayant pour densité

$$p_{\theta,\pi}(x) = (1 - \pi)\mathbb{1}_{x=0} + \pi\theta \exp(-\theta x)\mathbb{1}_{x>0}.$$

- Le modèle arbitrairement censuré, correspondant à l'observation de $X \wedge t$, où $X \sim \mathcal{E}(\theta)$ et $t > 0$ est un paramètre du modèle n'est pas dominé : en effet une mesure dominante donnerait une masse positive à $\{t\}$, pour tout $t > 0$, difficile dans ces conditions d'être σ-finie.

Dans les modèles dominés, on peut aller un peu plus loin dans les résultats négatifs que l'inconsistance si non-identifiabilité, en raisonnant sur "l'information sur θ contenue dans le modèle", ou aussi *information de Fisher*.

Definition 3.13 : Information de Fisher

Dans un modèle dominé, si $u \mapsto \ell_u(x) = \log(p_u(x))\,\mathbb{1}_{p_u(x)>0}$ est dérivable en $\theta \in \mathring{\Theta}$ μ-p.s., de dérivée $\dot{\ell}_\theta \in L_2(P_\theta)$, on définit *l'information de Fisher* en θ par

$$I(\theta) = E_\theta\left(\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^T\right).$$

La fonction ℓ_u est appelée *fonction score*.

L'interprétation de l'information de Fisher comme l'information contenue sur θ par le modèle est le point de vue historique de Fisher. La notion d'information est à prendre au sens "information en terme de précision pour estimer θ au sens de la norme au carré" (on a vu ou verra qu'il existe différentes manières d'évaluer la précision d'un estimateur).

Intuitivement, si p_u reste constante dans un voisinage de θ, alors, les lois des observations seront les mêmes dans un voisinage de θ et on ne pourra pas estimer θ de manière consistante. Inversement, si p_u varie beaucoup en θ, il semble plus facile de discriminer entre θ et un point de son voisinage sur la base des lois des observations P_u.

Formellement, c'est la Hessienne de la fonction de risque en distance de Kullback (permettant donc de relier précision Kullback induite par M-estimation max de vraisemblance) à la précision Euclidienne. Plus généralement, l'information de Fisher donne une précision maximale en termes de variance pour les estimateurs non biaisés.

Théorème 3.14 : Borne de Cramer-Rao

*Dans un modèle dominé où on peut définir une information de Fisher en $\theta \in \mathring{\Theta}$, soit T un estimateur **sans biais** localement autour de θ.*

Si on peut écrire

$$\nabla_\theta\left(\int_{\mathcal{X}} p_\theta(x)\mu(dx)\right) = \int \dot{p}_\theta(x)\mu(dx)(=0),$$
$$D_\theta\left(\int_{\mathcal{X}} T(x)p_\theta(x)\mu(dx)\right) = \int T(x)\dot{p}_\theta(x)^T\mu(dx)(=I_k),$$

alors

$$\mathrm{Cov}_\theta(T) \succcurlyeq I(\theta)^{-1}.$$

*En particulier, si $\Theta \subset \mathbb{R}$ et T est un estimateur **sans biais** de θ,*

$$E_\theta(T(X)-\theta)^2 \geq I(\theta)^{-1}.$$

Démonstration. On donne la preuve dans le cas $\Theta \subset \mathbb{R}$ (on peut étendre au cas multivarié en regardant des $\langle\theta, v\rangle$). Soit donc T un estimateur non-biaisé autour de

$\theta \in \mathring{\Theta}$. On a alors, pour tout u autour de θ,

$$E_u(T(X)) = \int_{\mathbb{R}} (T(x)) p_u(x) \mu(dx) = u.$$

En dérivant et intervertissant, on obtient

$$\begin{aligned} 1 &= \int_{\mathbb{R}} T(x) \dot{p}_\theta(x) \mu(dx) \\ &= \int_{\mathbb{R}} T(x) \dot{\ell}_\theta(x) p_\theta(x) \mu(dx) \\ &= E_\theta \left[\dot{\ell}_\theta(X) T(X) \right]. \end{aligned}$$

Par ailleurs,

$$E_\theta \left[\theta \dot{\ell}_\theta(X) \right] = \theta \frac{d}{d\theta} \int_{\mathbb{R}} \dot{p}_\theta(x) \mu(dx) = 0.$$

On en déduit

$$E_\theta \left[(T(X) - \theta) \dot{\ell}_\theta(X) \right] = 1.$$

Si $\mathrm{Var}_\theta(T) = +\infty$ la borne est triviale. Dans le cas où $\mathrm{Var}_\theta(T) < +\infty$, l'inégalité de Cauchy-Schwarz donne

$$1 \leq \sqrt{E_\theta (T(X) - \theta)^2} \sqrt{I(\theta)},$$

soit le résultat voulu. □

Remarque : Dans le cas réel, si $I(\theta) = 0$, alors $\dot{\ell}_\theta = 0$ P_θ-p.s. et donc, pour tout estimateur T non biaisé autour de θ, $E_\theta \left[(T(X) - \theta) \dot{\ell}_\theta(X) \right] = 0$. La preuve de Cramer-Rao restant vraie dès lors qu'on a un estimateur non-biaisé autour de θ, cela prouve qu'on ne peut avoir dans ce cas d'estimateur non-biaisé autour de θ. Plus généralement, les modèles où $I(\theta)$ est non-inversible sont considérés comme pathologiques ou mal conçus (par exemple non-identifiables), on fera souvent l'hypothèse implicite que l'information de Fisher est inversible. Si cela chagrine les esprits les plus pointilleux, la borne $I(\theta)\mathrm{Var}_\theta(T) \geq 1$ reste toujours valide pour les estimateurs sans biais.

Remarque importante : Si on note $I(\theta)$ l'information de Fisher d'un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta}$ et $I_n(\theta)$ l'information de Fisher du modèle $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta \in \Theta}$ (correspondant à l'observation de n variables i.i.d. tirées suivant le premier modèle), on a alors

$$I_n(\theta) = nI(\theta).$$

En particulier, la borne de Cramer-Rao s'écrit alors

$$\mathrm{Var}_\theta(T) \geq \frac{1}{nI(\theta)},$$

pour les estimateurs non-biaisés correspondant.

Les estimateurs non-biaisés atteignant la borne de Cramer-Rao sont appelés estimateurs *efficaces*. Dans certais cas favorables des estimateurs efficaces peuvent être construits facilement (famille exponentielle par exemple).

Exemple 3.15 *: Efficacité des moindre carrés dans le modèle linéaire Gaussien.*

Dans le modèle linéaire Gaussien $Y = X\theta + \varepsilon$, où pour simplifier on considère $\varepsilon \sim \mathcal{N}(0, I_n)$, on peut choisir comme mesure dominante $\mathcal{L}_n$, et densité

$$p_\theta(y_{1:n}) = \frac{1}{(\sqrt{2\pi})^n} \prod_{i=1}^{n} e^{-(y_i - \langle X_i, \theta\rangle)^2/2}.$$

La fonction score s'écrit alors

$$\begin{aligned}\ell_\theta(y_{1:n}) &= -\frac{1}{2}\sum_{i=1}^{n}(y_i - \langle X_i, \theta\rangle)^2 - \frac{n}{2}\log(2\pi) \\ &= -\frac{1}{2}\|y - X\theta\|^2 - \frac{n}{2}\log(2\pi).\end{aligned}$$

On en déduit

$$\nabla_\theta \ell_u(y_{1:n}) = X^T(y - X\theta),$$

et alors

$$\begin{aligned}I(\theta) &= E_\theta\left(\nabla_\theta \ell_u(Y)\nabla_\theta \ell_u(Y)^T\right) \\ &= \mathbb{E}\left[X^T\varepsilon\varepsilon^T X\right] \\ &= X^T X.\end{aligned}$$

La borne de Cramer-Rao implique alors que tout estimateur sans biais de θ doit vérifier

$$\mathrm{Cov}_\theta(T) \succcurlyeq (X^T X)^{-1}.$$

Comme $\mathrm{Cov}_\theta(\hat{\theta}_{LS}) = (X^T X)^{-1}$, $\hat{\theta}_{LS}$ est efficace.

Dans d'autre cas la borne de Cramer-Rao ne peut être atteinte. La notion d'efficacité est tombée en désuétude essentiellement pour la raison suivante : dans beaucoup de cas on peut construire des estimateurs **biaisés** plus performants que des estimateurs efficaces en terme d'écart quadratique. Dans le modèle Gaussien linéaire le phénomène de Stein nous fournira un tel estimateur (cf chapitre sur le bayésien). Sans chercher très loin, on peut regarder l'exemple suivant.

Exemple 3.16 *: Loi binomiale.* Dans le modèle $\mathcal{B}(n, \theta)_{\theta\in]0,1[}$, une densité par rapport à la mesure de comptage est donnée par

$$p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

On peut en déduire la fonction de score $\ell_\theta(x) = x\log(\theta) + (n-x)\log(1-\theta) + \log\left(\binom{n}{x}\right)$ correspondante, de dérivée

$$\dot{\ell}_\theta(x) = \frac{x - n\theta}{\theta(1-\theta)}.$$

L'information de Fisher s'écrit alors

$$I(\theta) = \frac{1}{(\theta(1-\theta))^2}\mathrm{Var}_\theta(X) = \frac{n}{\theta(1-\theta)}.$$

L'estimateur $T(X) = \frac{X}{n}$ (sans biais) vérifie bien

$$E_\theta \left(T(X) - \theta\right)^2 = \frac{\theta(1-\theta)}{n} = I(\theta)^{-1},$$

et est donc efficace.

Regardons maintenant l'estimateur $S(X) = \frac{X+\sqrt{n}/2}{n+\sqrt{n}}$. Il est biaisé, mais on peut écrire

$$\begin{aligned} E_\theta \left(S(X) - \theta\right)^2 &= \left(E_\theta(S(X) - \theta)\right)^2 + \mathrm{Var}_\theta(S(X)) \\ &= \frac{1}{(n+\sqrt{n})^2}\left[\left(\frac{\sqrt{n}}{2} - \sqrt{n}\theta\right)^2 + n\theta(1-\theta)\right] \\ &= \frac{1}{4(1+\sqrt{n})^2}. \end{aligned}$$

On a alors $E_\theta \left(S(X) - \theta\right)^2 < E_\theta \left(T(X) - \theta\right)^2$ si θ est suffisamment proche de $1/2$. En anticipant un peu, le pire des risques de S est meilleur que celui de T, c'est à dire

$$\frac{1}{4(1+\sqrt{n})^2} = \sup_{\theta \in]0,1[} E_\theta \left(S(X) - \theta\right)^2 < \sup_{\theta \in]0,1[} E_\theta \left(T(X) - \theta\right)^2 = \frac{1}{4n}.$$

On prouvera par la suite que S est optimal au sens du pire des risques (on appelle ça minimax).

En somme, l'efficacité des estimateurs est une propriété sympathique (notamment des estimateurs du maximum de vraisemblance dans les modèles exponentiels), mais n'est qu'une notion d'optimalité parmi d'autres. En particulier, elle ne garantit en rien la minimaxité (qui est une propriété un peu plus "moderne" que l'on définira dans le chapitre bayésien).

En revanche, l'information de Fisher reste une quantité fondamentale, notamment dans l'étude du comportement limite des estimateurs du maximum de vraisemblance.

3.2.2 Exhaustivité

Dans le modèle des naissances où on observe $X_{1:n} \sim \mathcal{B}(\theta)^{\otimes n}$, l'intuition laisse à penser que toute l'information sur θ apportée par $X_{1:n}$ est portée par $\sum_{i=1}^n X_i$. En d'autre termes, une fois $\sum_{i=1}^n X_i$ observée, l'observation résiduelle de $X_{1:n}$ n'apporte pas plus d'information sur θ. On peut formaliser cette intuition via la notion *d'exhaustivité*.

Definition 3.17

Soit $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ un modèle statistique, et soit $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ une statistique. T est dite exhaustive pour ce modèle si et seulement si pour toute fonction $\phi : \mathcal{X} \to \mathbb{R}^+$ mesurable il existe une fonction $g_\phi : \mathcal{Y} \to [0, +\infty]$ telle que, pour tout $\theta \in \Theta$,

$$E_\theta \left(\phi(X)|T(X)\right) = g_\phi(T(X)) \quad P_\theta - p.s..$$

En particulier, $E_\theta \left(\phi(X)|T(X)\right)$ ne dépend pas de θ.

Cette définition générale sera surtout utile dans un cadre Bayésien. Dans le cadre des modèles dominés, la caractérisation suivante des statistiques exhaustives est souvent prise comme définition.

PROPOSITION 3.18

Soit $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ un modèle statistique dominé par une mesure μ sigma-finie. Soit $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ une statistique. Alors T est exhaustive si et seulement si il existe une version $p(x, \theta)$ de $\frac{dP_\theta}{d\mu}$ satisfaisant

$$p(x, \theta) = g(x)h(T(x), \theta),$$

pour tout $\theta \in \Theta$.

Démonstration. On commence par le sens indirect, en utilisant la remarque suivante : si μ est une mesure σ-finie, alors il existe $\tilde{\mu}$ mesure de probabilité telle que $\mu << \tilde{\mu}$. On peut alors supposer sans perte de généralité que

$$\frac{dP_\theta}{d\mu}(x) = g(x)h(T(x), \theta),$$

où μ est une mesure de probabilité. Soit alors $\phi : \mathcal{X} \to \mathbb{R}^+$ mesurable, $\psi : \mathcal{Y} \to \mathbb{R}^+$ mesurable, et $\theta \in \Theta$. On a alors

$$\begin{aligned} E_\theta\left(\phi(X)\psi(T(X))\right) &= E_\mu(\phi(X)\psi(T(X))g(X)h(T(X), \theta)) \\ &= E_\mu\left(\psi(T(X))h(T(X), \theta)E_\mu\left(g(X)\phi(X) \mid T(X)\right)\right). \end{aligned}$$

En notant $g_1(T(X)) = E_\mu\left(g(X)\phi(X) \mid T(X)\right)$, $g_2(T(X)) = E_\mu(g(X) \mid T(X))$. On peut voir avec des fonctions tests que $g_2(x) = 0$ implique $g_1(x) = 0$ μ-p.s., et donc écrire

$$\begin{aligned} E_\theta\left(\phi(X)\psi(T(X))\right) &= E_\mu\left(\psi(T(X))h(T(X), \theta)E_\mu(g(X) \mid T(X))\frac{g_1(T(X))}{g_2(T(X)}\right) \\ &= E_\mu\left(\psi(T(X))h(T(X), \theta)g(X)\frac{g_1(T(X))}{g_2(T(X)}\right) \\ &= E_\theta\left(\psi(T(X))g_\phi(T(X))\right), \end{aligned}$$

avec

$$g_\phi(T(X)) = \frac{g_1(T(X))}{g_2(T(X))} = \frac{E_\mu\left(\phi(X)g(X) \mid T(X)\right)}{E_\mu\left(g(X) \mid T(X)\right)}.$$

Le sens direct est plus compliqué. On admettra le lemme suivant.

LEMME 3.19 : THÉORÈME A.78, SCHERVISH 95

Si $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ est un modèle dominé par une mesure sigma-finie μ, alors il existe $(\theta_i)_{i=1,\dots,+\infty}$ et $(c_i)_{i=1,+\infty}$ tels que

1. *$\forall i \geq 1$ $c_i \geq 0$, $\sum_{i=1}^{+\infty} c_i = 1$.*
2. *En notant $\nu = \sum_{i=1}^{\infty} c_i P_{\theta_i}$, on a*

$$\forall\theta \in \Theta \quad P_\theta << \nu << \mu.$$

Soit $\phi : \mathcal{X} \to \mathbb{R}^+$ une fonctions mesurable. On a (en prenant ν comme dans le lemme)

$$\begin{aligned}
E_\theta(\phi(X)) &= E_\theta\left(E_\theta(\phi(X) \mid T(X))\right) = E_\theta\left(g_\phi(T(X))\right) \\
&= E_\nu\left(\frac{dP_\theta}{d_\nu}(X) g_\phi(T(X))\right) \\
&= E_\nu\left(E_\nu\left(\frac{dP_\theta}{d_\nu}(X) \mid T(X)\right) g_\phi(T(X))\right) \\
&= \sum_i c_i E_{\theta_i}\left(E_\nu\left(\frac{dP_\theta}{d_\nu}(X) \mid T(X)\right) g_\phi(T(X))\right) \\
&= \sum_i c_i E_{\theta_i}\left(E_\nu\left(\frac{dP_\theta}{d_\nu}(X) \mid T(X)\right) E_{\theta_i}(\phi(X) \mid T(X)))\right) \\
&= \sum_i c_i E_{\theta_i}\left(E_\nu\left(\frac{dP_\theta}{d_\nu}(X) \mid T(X)\right) \phi(X))\right) \\
&= E_\nu\left(E_\nu\left(\frac{dP_\theta}{d_\nu}(X) \mid T(X)\right) \phi(X)\right).
\end{aligned}$$

En notant $h(T(X), \theta) = E_\nu\left(\frac{dP_\theta}{d_\nu}(X) \mid T(X)\right)$, on a

$$\begin{aligned}
E_\theta(\phi(X)) &= E_\nu(h(T(X), \theta)\phi(X)) \\
&= E_\mu\left(\frac{d\nu}{d\mu}(X) h(T(X), \theta)\phi(X)\right).
\end{aligned}$$

En posant $g(x) = \frac{d\nu}{d\mu}(x)$, on a le résultat. □

L'intérêt majeur des statistiques exhaustives est qu'elles permettent souvent de simplifier les modèles.

PROPOSITION 3.20

Soit $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ un modèle, et S une statistique exhaustique pour θ. Alors, pour tout estimateur T il existe un estimateur $\tilde{T}(S(X))$ vérifiant

$$\forall \theta \in \Theta E_\theta\left\|\theta - \tilde{T}(S(X))\right\|^2 \leq E_\theta\|\theta - T(X)\|^2.$$

Démonstration. En définissant $\tilde{T}(S(X)) = E_\theta\left(T(X) \mid S(X)\right)$, on définit bien un estimateur car cette espérance conditionnelle ne dépend pas de θ. $\tilde{T}$ vérifie, pour tout $\theta \in \Theta$,

$$\begin{aligned}
E_\theta\left\|\theta - \tilde{T}(S(X))\right\|^2 &= E_\theta\left\|\theta - E_\theta(T(X) \mid S(X))\right\|^2 \\
&\leq E_\theta\left[E_\theta\left\|\theta - T(X)\right\|^2 \mid S(X)\right] \\
&= E_\theta\left\|T(X) - \theta\right\|^2.
\end{aligned}$$

□

En d'autres termes, si le but est d'approcher θ (ou $q(\theta)$) en terme de norme Euclidienne ou de toute autre fonction convexe, on peut se restreindre aux estimateurs de la forme $\tilde{T}(S(X))$, c'est-à-dire changer de modèle et regarder $(\mathcal{Y}, \mathcal{B}, S\#(P_\theta)_{\theta\in\Theta})$.

Exemple 3.21.

Dans le modèle d'éclosion des oeufs de pingouins $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathcal{B}(\theta)^{\otimes n})$, une mesure dominante est la mesure de comptage sur $\{0,\}^n$, pour laquelle les densités se mettent sous la forme

$$\begin{aligned} p_\theta(x_{1:n}) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

On en déduit que $S(X) = \sum_{i=1}^n X_i$ est exhaustive pour ce modèle. Si le but est d'estimer θ en distance Euclidienne, on peut se ramener au modèle $(\mathcal{B}(n,\theta))_{\theta\in]0,1[}$, plus commode à manipuler (personne n'a envie de manipuler des n-uplets quand il peut s'en passer).

Trouver une statistique exhaustive permet donc de s'épargner du calcul en réduisant le modèle, ce sera particulièrement vrai dans le chapitre sur le bayésien. Une autre manière de dire qu'on ne perd rien à changer de modèle (pour l'estimation en norme Euclidienne) peut se formuler en termes d'information de Fisher : on ne perd pas d'information de Fisher à changer de modèle, cela caractérise même les statistiques exhaustives (cf Ibragimov Théorème 7.2).

3.2.3 Modèles exponentiels

On peut définir les modèles exponentiels comme suit.

Definition 3.22 : Modèle exponentiel sous forme canonique

Soit μ mesure σ-finie sur $\mathcal{X}$, et $T_1, \ldots, T_k$ des variables aléatoires réelles. On définit

$$\Theta_{dom} = \left\{\theta \in \mathbb{R}^k \mid Z(\theta) = \int_{\mathcal{X}} e^{\langle\theta, T(x)\rangle}\mu(dx) < +\infty\right\},$$

avec $T = (t_1, \ldots, T_k)$. Le modèle exponentiel sous forme canonique associé à T et μ est alors

$$(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}),$$

où $\Theta = \mathring{\Theta}_{dom}$, $P_\theta = e^{\langle\theta,T(x)\rangle - \log(Z(\theta))}\mu$. Θ_{dom} est appelé *domaine* du modèle, Z sa fonction de partition.

Un modèle exponentiel est donc par définition dominé, et admet T pour statistique exhaustive. Beaucoup de modèles classiques se mettent sous cette forme. À partir d'ici on parlera de modèles exponentiels avec la convention implicite qu'ils sont sous forme canonique.

Exemple 3.23 *: Modèles exponentiels classiques.*

— Le modèle $\mathcal{N}(\mu, \sigma^2)_{\mu\in\mathbb{R},\sigma^2>0}$ peut se mettre sous la forme exponentielle. En prenant pour mesure dominante la mesure de Lebesgue sur $\mathbb{R}$, on écrit

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - Z_1(\mu,\sigma^2)} \\ &= e^{\langle\theta, T(x)\rangle - \log(Z(\theta))}, \end{aligned}$$

avec $T(x) = (x, -x^2)^T$, $\theta = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)^T$, et

$$Z(\theta_1, \theta_2) = \exp\left(\frac{\theta_1^2}{2\theta_2}\right)\sqrt{\frac{\pi}{\theta_2}}.$$

— Pour le modèle $\gamma(a, b)_{a,b>0}$, avec pour mesure dominante Lebesgue sur $]0, +\infty[$, on écrit

$$\begin{aligned} p_{a,b}(x) &= x^{a-1}e^{-bx}\frac{b^a}{\Gamma(a)} = e^{-bx+(a-1)\log(x)-Z_1(a,b)} \\ &= e^{\langle\theta, T(x)\rangle - \log(Z(\theta))}, \end{aligned}$$

avec $T(x) = (\log(x), -x)^T$, $\theta = (a-1, b)^T$, et

$$Z(\theta) = \frac{\Gamma(\theta_1 + 1)}{\theta_2^{\theta_1+1}}.$$

La plupart du temps il s'agit donc de reparamétrer les modèles classiques.

Remarque : Si on a un modèle exponentiel défini par μ et T, on peut définir un modèle exponentiel équivalent sur $\mathcal{Y} = T(\mathcal{X})$ par rapport à la mesure $\nu = T\#\mu$ via les $T\#P_\theta$. En effet, si $\phi : \mathcal{Y} \to \mathbb{R}^+$ est mesurable

$$\begin{aligned} E_{T\#P_\theta}(\phi(y)) &= E_\theta(\phi(T(x)) \\ &= E_\mu(\phi(T(x))e^{\langle\theta, T(x)\rangle - \log(Z(\theta))}) \\ &= E_\nu(\phi(y)e^{\langle\theta, y\rangle - \log(Z(\theta))}). \end{aligned}$$

Ce qui montre alors que les $T\#P_\theta$ ont pour densité $e^{\langle\theta,y\rangle-\log(Z(\theta))}$ par rapport à ν. Ce modèle est équivalent au sens où il fournit des estimateurs au moins aussi bon que le premier au sens de l'écart en norme Euclidienne, on verra aussi par la suite qu'il conduit aux mêmes estimateurs du maximum de vraisemblance (et donc aux mêmes propriétés asymptotiques les concernant).

Les modèles exponentiels ont les propriétés particulières suivantes.

Théorème 3.24 : Convexité des domaines et fonctions de partition

Soit μ mesure σ-finie sur $\mathcal{X}$, et $T : \mathcal{X} \to \mathbb{R}^k$ mesurable. Alors, s'il n'est pas vide,

$$\Theta_{dom} = \left\{\theta \in \mathbb{R}^k \mid Z(\theta) := \int_{\mathcal{X}} e^{\langle T(x),\theta\rangle}\mu(dx) < +\infty\right\}$$

est convexe. De plus $\theta \mapsto \log(Z(\theta))$ est convexe sur Θ_{dom}

Démonstration. La preuve est basée sur l'inégalité de Hölder. Soient θ_1, θ_2 dans Θ_{dom}, et $\lambda \in [0,1]$, on a

$$\begin{aligned}\int_{\mathcal{X}} e^{\langle T(x), \lambda\theta_1 + (1-\lambda)\theta_2\rangle} &= \int_{\mathcal{X}} \left[e^{\langle T(x),\theta_1\rangle}\right]^{\lambda} \left[e^{\langle T(x),\theta_2\rangle}\right]^{(1-\lambda)} \mu(dx) \\ &\leq \left[\int_{\mathcal{X}} e^{\langle T(x),\theta_1\rangle}\mu(dx)\right]^{\lambda} \left[\int_{\mathcal{X}} e^{\langle T(x),\theta_2\rangle}\mu(dx)\right]^{1-\lambda}.\end{aligned}$$

On en déduit que Θ_{dom} et $\log(Z)$ sont convexes. □

On peut aussi caractériser l'identifiabilité des modèles exponentiels en termes de structure de covariance de T.

Definition 3.25 : Modèle exponentiel minimal

Un modèle exponentiel est dit *minimal* si, pour tout hyperplan affine $H \subset \mathbb{R}^k$ et $\theta \in \Theta$,

$$P_\theta\left(T(X) \in H\right) < 1.$$

En d'autres termes, un modèle est minimal si les $T\#P_\theta$ ne peuvent être reparamétrés avec un degré de liberté en moins. Un contre-exemple classique est le cas où $T_i = T_j$, pour un $i \neq j$. Dans ce cas, le modèle n'est évidemment pas identifiable : échanger les coordonnées i et j de θ donnent la même distribution. On peut généraliser ce phénomène.

Théorème 3.26 : Identifiabilité des modèles exponentiels

Un modèle exponentiel (sous forme canonique) est identifiable si et seulement si il est minimal.

Démonstration. Supposons qu'il existe $\theta_0 \in \Theta$, $v \in \mathbb{R}^k$ et $c \in \mathbb{R}$ tels que

$$P_{\theta_0}\left\{\langle v, T(X)\rangle = c\right\} = 1.$$

En prenant $\theta_t = \theta_0 + tv \in \Theta$ pour t assez petit (on rappelle que $\Theta = \overset{\circ}{\Theta}_{dom}$), et comme $p_{\theta_0} > 0$, on a $P_{\theta_t} << \mu << P_{\theta_0}$, et

$$\frac{dP_{\theta_t}}{dP_{\theta_0}} = \frac{p_{\theta_t}}{p_{\theta_0}} = e^{\langle T(x), tv\rangle - \log\left(\frac{Z(\theta_t)}{Z(\theta_0)}\right)} = e^{tc - \log\left(\frac{Z(\theta_t)}{Z(\theta_0)}\right)},$$

la dernière égalité étant au sens P_{θ_0} presque sûr. On en déduit que $\frac{dP_{\theta_t}}{dP_{\theta_0}}$ est constante donc vaut 1 P_{θ_0}-p.s., et donc que $P_{\theta_t} = P_{\theta_0}$ (donc non-identifiabilité). □

Moments et information de Fisher des modèles exponentiels

Les modèles exponentiels correspondent à des lois à queues relativement faibles (sous-exponentielles). En particulier, cela implique l'existence de moments de n'importe quel ordre pour la statistique exhaustive T.

Proposition 3.27 : Existence des moments de T

Dans un modèle exponentiel sous forme canonique, pour tout $\theta \in \Theta(= \mathring{\Theta}_{dom})$,

$$u \mapsto E_\theta \left(e^{u\|T(X)\|} \right)$$

est bien définie autour de 0. En particulier, pour tous $j \in [\![1, k]\!]$ et $p \geq 1$,

$$E_\theta \left| T_j(X) \right|^p < +\infty.$$

Démonstration. Sans perte de généralité on regarde le cas $k = 1$. Comme $\theta \in \Theta$, $\theta \pm h \in \Theta$ pour h assez petit. On en déduit que, pour $|u| < h$,

$$E_\theta(e^{uT(X)}) = E_\mu \left(e^{\langle T(X), \theta+u \rangle - \log(Z(\theta+u)) + \log\left(\frac{Z(\theta+u)}{Z(\theta)}\right)} \right) = \frac{Z(\theta+u)}{Z(\theta)} < +\infty.$$

En particulier, si $0 \leq u < h$,

$$E_\theta(e^{u|T(X)|}) \leq E_\theta \left(e^{-uT(X)} + e^{uT(X)} \right) < +\infty,$$

et on en déduit que $E_\theta |T(X)|^p < +\infty$ pour tout $p \geq 0$ (décomposition en série puis Fubini). □

Il ressort de cette preuve que les moments de T est la fonction de partition Z sont liés. De fait, on peut établir les identités suivantes.

Théorème 3.28 : Régularité de Z et moments de T

Dans un modèle exponentiel sous forme canonique, $\theta \mapsto Z(\theta)$ est $\mathcal{C}^\infty$ sur Θ. De plus, pour tout $\theta \in \Theta$,

$$\begin{array}{rcl} \nabla_\theta(\log(Z)) & = & E_\theta(T(X)), \\ H_\theta(\log(Z)) & = & \mathrm{Cov}_\theta(T(X)), \end{array}$$

où H désigne la matrice Hessienne. Enfin, l'information de Fisher de ce modèle existe, et vérifie

$$I(\theta) = H_\theta(Z) = \mathrm{Cov}_\theta(T(X)).$$

Démonstration. C'est un jeu d'interversion entre intégrale et dérivée. Pour $h \in \mathbb{R}^k$ suffisamment petit tel que $\theta + h \in \Theta$, on a

$$\begin{aligned} \frac{Z(\theta+h) - Z(\theta)}{\|h\|} &= E_\mu \left(e^{\langle \theta, T(X) \rangle} \frac{e^{\langle h, T(X) \rangle} - 1}{\|h\|} \right) \\ &= Z(\theta) E_\theta \left(\frac{e^{\langle h, T(X) \rangle} - 1}{\|h\|} \right). \end{aligned}$$

Par ailleurs, pour $x \in \mathbb{R}^k$, $\|h\| \leq h_0$ tels que $E_\theta(e^{2h_0\|T(X)\|}) < +\infty$, on a

$$\frac{e^{\langle h, T(x) \rangle} - 1}{\|h\|} \leq \|T(x)\| e^{h_0\|T(x)\|} \in L_1(P_\theta),$$

d'après la proposition précédente. On en déduit que l'interversion dérivation/intégrale est possible, et

$$\nabla_\theta Z = Z(\theta) E_\theta(T(X)).$$

Pour la dérivée seconde, on peut jouer au même jeu en considérant

$$\begin{aligned}\frac{E_{\theta+h}(T(X)) - E_\theta(T(X))}{\|h\|} &= E_\mu\left[T(X)e^{\langle\theta,T(X)\rangle}\frac{1}{\|h\|}\left(e^{\langle h,T(X)\rangle - \log(Z(\theta+h))} - e^{-\log(Z(\theta))}\right)\right] \\ &= E_\theta\left[T(X)\frac{1}{\|h\|}\left(e^{\langle h,T(X)\rangle - \log(Z(\theta+h))} - e^{-\log(Z(\theta))}\right)\right]\end{aligned}$$

permettant de justifier l'interversion dérivation/intégrale suivante :

$$\begin{aligned}D_\theta\left[E_\mu\left(T(X)e^{\langle T(X),u\rangle - \log(Z(u))}\right)\right] &= E_\mu\left(T(X)e^{\langle T(X),\theta\rangle - \log(Z(\theta))}T(X)^T\right) \\ &\quad - E_\mu\left(T(X)e^{\langle T(X),\theta\rangle - \log(Z(\theta))}\right)\nabla_\theta \log(Z)^T \\ &= E_\theta(T(X)T(X)^T) - E_\theta(T(X))(E_\theta(T(X)))^T \\ &= \mathrm{Cov}_\theta(T(X)).\end{aligned}$$

En particulier, la fonction score $\ell_\theta(x) = \langle\theta, T(x)\rangle - \log(Z(\theta))$ est bien différentiable et de différentielle dans $L_2(P_\theta)$. L'information de Fisher de ce modèle est alors

$$I(\theta) = E_\theta\left((T(X) - \nabla_\theta Z)(T(X) - \nabla_\theta Z)^T\right) = \mathrm{Cov}_\theta(T(X)).$$

□

Une conséquence immédiate de ce résultat est que l'identifiabilité des modèles exponentiels peut directement se déduire de la fonction de partition.

Corollaire 3.29 : Identifiabilité des modèles exponentiels (bis)

Un modèle exponentiel sous forme canonique est identifiable si et seulement si

$$H_\theta(\log(Z)) \succ 0.$$

En particulier, un modèle exponentiel sous forme canonique est identifiable si et seulement si

$$S : \begin{cases}\Theta & \to & S(\Theta) \\ \theta & \mapsto & \nabla_\theta(\log(Z))\end{cases}$$

est un C^∞-difféomorphisme.

Démonstration. On vérifie facilement que $H_\theta(\log(Z)) \succ 0$ équivaut à la minimalité du modèle exponentiel, lui-même équivalent à l'identifiabilité du modèle.

La seconde partie est une application du Théorème d'inversion globale. Il reste néanmoins à vérifier que l'identifiabilité du modèle implique l'injectivité (globale)

de S. Pour cela, supposons qu'il existe $\theta_1 \neq \theta_2 \in \Theta$ tels que $S(\theta_1) = S(\theta_2)$. Comme Θ est convexe, on peut définir

$$f : \begin{cases} [0,1] & \to & \mathbb{R} \\ t & \mapsto & \langle \theta_2 - \theta_1, S(t\theta_2 + (1-t)\theta_1) \rangle \end{cases},$$

qui est $\mathcal{C}^\infty$ sur $]0,1[$. Comme $f(0) = f(1)$, on déduit l'existence d'un $t_0 \in]0,1[$ tel qe

$$0 = f'(t_0) = \left\langle \theta_2 - \theta_1, (D_{\theta_{t_0}} S)(\theta_2 - \theta_1) \right\rangle,$$

où $\theta_{t_0} = t_0\theta_2 + (1-t_0)\theta_1 \in \Theta$. On en déduit que $H_{\theta_{t_0}}(\log(Z))$ n'est pas inversible, et donc le modèle est non identifiable. □

Le dernier point sera particulièrement utile dans la partie qui suit (estimation de θ). Concluons par un exemple.

Exemple 3.30 *: Lois gamma.* On a vu que le modèle $\gamma(a,b)_{a,b>0}$ pouvait se mettre sous la forme exponentielle avec $T(x) = (\log(x), -x)^T$, $\theta = (a-1, b)^T$ et $Z(\theta) = \frac{\Gamma(\theta_1+1)}{\theta_2^{\theta_1+1}}$. On peut alors écrire

$$\nabla_\theta(\log(Z)) = \left(\frac{\Gamma'(\theta_1)}{\Gamma(\theta_1} - \log(\theta_2), \frac{\theta_1+1}{\theta_2}\right)^T = \left(\psi(\theta_1) - \log(\theta_2), \frac{\theta_1+1}{\theta_2}\right)^T,$$

où ψ désigne la fonction digamma (croissante sur $]-1, +\infty[$). L'information de Fisher (ou matrice Hessienne de $\log(Z)$) est donnée par

$$H_\theta(\log(Z)) = \begin{pmatrix} \psi'(\theta_1) & -\frac{1}{\theta_2} \\ \frac{1}{\theta_2} & -\frac{\theta_1+1}{\theta_2^2} \end{pmatrix},$$

de déterminant $-\frac{(\theta_1+1)}{\theta_2^2}\psi'(\theta_1) + \frac{1}{\theta_2^2} < 0$. Comme le modèle est identifiable par ailleurs (deux lois $\gamma(a,b)$ de paramètres différents auront des moyennes et/ou variances différentes), on peut en déduire l'inégalité

$$(\theta_1+1)\psi'(\theta_1) < 1,$$

pour $\theta_1 > -1$.

3.3 Maximum de vraisemblance dans les modèles exponentiels

Dans toute cette partie on se donnera un modèle dominé correspondant à l'observation de $X_1, \ldots, X_n$ i.i.d. tirés suivant $P_\theta \sim p_\theta \mu$, pour $\theta \in \Theta$.

3.3.1 Principe de la maximisation de la vraisemblance

Le principe d'estimation par maximum de vraisemblance est une méthode de M-estimation valable uniquement dans les modèles dominés. Le critère empirique à maximiser est

$$M_n(u) = p_u^{\otimes n}(X_{1:n}) = \prod_{i=1}^n p_u(X_i) := V_n(u).$$

Moralement, cela correspond à chercher le paramètre θ pour lequel les observations sont le plus "vraisemblables", au sens des densités par rapport à la mesure commune μ (FAIRE DESSIN).

On sent bien que définir $\hat{\theta} \in \arg\max_u M_n(u)$ va poser quelques soucis :

1. Premièrement, les densités ne sont définies que μ-p.s.. Si on les modifie, on peut changer l'estimateur de maximum de vraisemblance (avec probabilité tendant vers 0 toutefois. Par exemple, dans un modèle Gaussien, on peut prendre pour densité $p_\mu(x) = (\sqrt{2\pi})^{-1} e^{-(x-\mu)^2} + 10 * \mathbb{1}_{x=3\mu}$. Pour une observation, on aura toujours $\hat{\theta} = 3\mu$. Pour plus de deux observations le cas pathologique disparaît avec probabilité 1. En toute généralité, il serait plus correct de parler "d'un estimateur du maximum de vraisemblance". En pratique, si parmi les versions des densités on en trouve une régulière, on la choisit par convention tacite.
2. Deuxièmement, même avec une convention de choix de densité, il se peut que l'estimateur du maximum de vraisemblance n'existe pas, cf exemple ci-dessous.

Exemple 3.31 *: Lois géométriques.* Pour le modèle de lois géométriques paramétré par $\theta > 0$, de densité

$$p_\theta(x) = \theta^{x-1}(1-\theta) = e^{(x-1)\log(\theta)+\log(1-\theta)},$$

la vraisemblance en $x_{1:n}$ s'écrit

$$V_n(\theta) = p_\theta^{\otimes n}(x_{1:n}) = \exp\left(\log(\theta)\sum_{i=1}^n (x_i - 1) + n\log(1-\theta)\right).$$

On en déduit que dans le cas où $x_{1:n} = (1, \dots, 1)$, V_n n'atteint pas son maximum sur $]0, +\infty[$, et l'EMV n'est pas défini. La probabilité qu'un tel évènement arrive sous P_θ est $(1-\theta)^n \to 0$.

Cet exemple est révélateur de deux phénomènes : premièrement que le maximum de vraisemblance va être défini si les observations ne tombent pas toutes à la frontière de leur domaine, dans le cadre des modèles exponentiels (les lois géométriques en sont un). Deuxièmement, toujours dans ce cadre exponentiel, la probabilité que les observations tombent à la frontière va toujours tendre vers 0 (dans les cas non dégénérés).

Bref, quand on définit un estimateur du maximum de vraisemblance, il convient d'être précautionneux. Dans l'exemple précédent, pour maximiser la vraisemblance, nous sommes passés par une exponentiation (pratique pour traiter des produits de densités). Cette approche revient à maximiser la *log-vraisemblance*

$$\ell_n(u) = \log(V_n(u)) = \sum_{i=1}^n \log(p_u(X_i)),$$

pouvant prendre comme valeur $-\infty$ néanmoins. Ces deux problèmes de maximisation sont équivalents. Néanmoins, l'expression du critère théorique associé à la log-vraisemblance permet de prouver que θ est bien un maximiseur du critère idéal

$$M(u) = E_\theta \log(p_u(X)),$$

sous certaines conditions d'intégrabilité. Pour prouver cela, on introduit la *divergence de Kullback* (ou entropie relative) entre deux distributions.

Definition 3.32 : Divergence de Kullback

Soient P et Q deux distributions dominées par μ, de densité respectives p et q. La *divergence de Kullback* entre P et Q, $\mathrm{d}_{KL}(P\|Q)$ est définie par

$$\mathrm{d}_{KL}(P\|Q) = \begin{cases} +\infty & \text{si} \quad P \not<< Q \\ \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\mu(dx) & \text{sinon} \quad \text{(et si cette intégrale existe).} \end{cases}$$

Par ailleurs, on a toujours $\mathrm{d}_{KL}(P\|Q) \geq 0$, avec égalité si et seulement si $P = Q$.

Démonstration. Si Q ne domine pas P, $\mathrm{d}_{KL}(P\|Q)$ est bien strictement positive. Sinon, $p(x)/q(x)$ est bien défini Q-p.s., et on a

$$\begin{aligned} \mathrm{d}_{KL}(P\|Q) &= E_Q\left(\frac{p}{q}(X)\log\left(\frac{p}{q}(X)\right)\right) \\ &= E_Q\left(\Phi\left(\frac{p}{q}(X)\right)\right), \end{aligned}$$

où $\phi(u) = u\log(u)$ pour $u > 0$. Comme Φ est strictement convexe, on a

$$\mathrm{d}_{KL}(P\|Q) = E_Q\left(\Phi\left(\frac{p}{q}(X)\right)\right) \geq \Phi\left(E_Q\left(\frac{p}{q}(X)\right)\right) = \Phi(1) = 0,$$

avec égalité si et seulement si $\frac{p}{q}(X)$ est constant Q-p.s., la seule constante possible étant 1, ce qui équivaut à $p = q$ Q-p.s., équivalent à $P = Q$. □

Remarque : Même lorsque $P << Q$ on peut avoir $\mathrm{d}_{KL}(P\|Q) = +\infty$. L'existence de l'intégrale se montre de manière standard ($L_1(\mu)$ ou positivité/négativité de l'intégrande).

Il existe des liens explicite entre divergence de Kullback et information de Fisher dans les modèles suffisamment réguliers, on illustrera ce phénomène dans les modèles exponentiels. Pour ce qui nous intéresse présentement, on peut relier maximisation de la log-vraisemblance idéale et divergence de Kullback.

Théorème 3.33 : Pertinence théorique de la maximisation de vraisemblance

Pour un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}$ dominé par μ, on suppose que $E_\theta|\log(p_u(X)| < +\infty$ pour tout $u \in \Theta$. On a alors, pour tout $u \in \Theta$,

$$M(u) = E_\theta(\log(p_u(X))) = -\mathrm{d}_{KL}(P_\theta\|P_u) + H(P_\theta),$$

où $H(P_\theta) = E_\theta(-\log(p_\theta(X))) \in]0, +\infty[$ est parfois appelée entropie de P_θ.
Par conséquent, θ est l'unique maximiseur de M.

Remarque : On peut généraliser parfois pour les cas où $E_\theta(\log(p_u)(X))$ sont définies et valent $-\infty$ (cas discret par exemple).

Démonstration. Commençons par remarque que les conditions d'intégrabilité demandées impliquent $P_\theta << P_u$. En effet, si ce n'était pas le cas, on aurait A tel que $\int_A p(x)\mu(dx) > 0$ et $q \equiv 0$ sur A μ-p.s.. On aurait alors

$$\int |p(x)\log(q(x))|\mu(dx) \geq \int_A |p(x)\log(q(x))|\mu(dx) = +\infty.$$

Comme $P_\theta << P_u$, on peut écrire

$$M(\theta) - M(u) = E_\theta(\log(p_\theta(X)) - \log(p_u(X))) = E_\mu\left[p_\theta(X)\log\left(\frac{p_\theta(X)}{p_u(X)}\right)\right] = \mathrm{d}_{KL}(P_\theta \| P_u),$$

cette dernière quantité étant bien définie (intégrande dans $L_1(\mu)$). □

3.3.2 Le cas des modèles exponentiels

On peut énoncer des résultats généraux sur des estimateurs du maximum de vraisemblance approchés (un élément quelconque dont associé à une valeur quasiment maximale à disons $1/n$ près), dans les modèles suffisamment réguliers pour que M admette une décomposition de Taylor à l'ordre 2 autour de θ (cela rejoint les principes généraux de preuves pour la M-estimation dont on a parlé précédemment). Ce genre de résultat peut par exemple se trouver dans le Théorème 5.23 du van der Waart.

Dans le modèle exponentiel, les choses sont plus simples : on peut parler du vrai minimiseur de la log-vraisemblance avec grande proba, et on peut se passer de la théorie des processus empiriques pour donner des résultats.

Commençons par le premier point.

Théorème 3.34 : Existence du maximum de vraisemblance

Dans un modèle exponentiel minimal sous forme canonique, désignons par $\Lambda = S(\Theta)$ (ouvert et convexe).

Si $(T(X_1), \dots, T(X_n))$, est tel que $\bar{T}_n \in \Lambda$, alors ℓ_n (fonction de log-vraisemblance) est strictement concave et atteint un unique maximum dans Θ défini par

$$\hat{\theta}_{EMV} = S^{-1}(\bar{T}_n),$$

et coïncide donc avec l'estimateur des moments associé à T.

Démonstration. Pour alléger les notations, notons $\psi = \log(Z)$. Commençons par rappeler que $\ell_n(\theta) = n\left(\left\langle \bar{T}_n, \theta\right\rangle - \psi(\theta)\right)$. Comme le modèle est minimal, ψ est strictement convexe (Corollaire 3.29). De plus, $\bar{T}_n \in Lambda$ implique l'existence de $\hat{\theta}_n \in \Theta$ tel que $\nabla_{\hat{\theta}_n}\psi = \bar{T}_n$. On va montrer que ce $\hat{\theta}_n$ est l'unique maximum ℓ_n sur $\bar{\Theta}$.

Comme $\hat{\theta}_n \in \Theta$, il existe $r > 0$ tel que $\mathrm{B}(\hat{\theta}_n, r) \subset \Theta$. Si $\hat{\theta}_n \neq u \in \mathrm{B}(\hat{\theta}_n, r)$, on a

$$\begin{aligned}\ell_n(\hat{\theta}_n) - \ell_n(u) &= n\left[\psi(u) - \psi(\hat{\theta}_n) - \left\langle \nabla_{\hat{\theta}_n}\psi, u - \hat{\theta}_n\right\rangle\right] \\ &> 0,\end{aligned}$$

par stricte convexité de ψ. Notons maintenant

$$\kappa = \min_{u \in \mathrm{S}(\hat{\theta}_n, r)} \psi(u) - \psi(\hat{\theta}_n) - \left\langle \nabla_{\hat{\theta}_n}\psi, u - \hat{\theta}_n\right\rangle > 0,$$

par continuité de ψ, et prenons $u \in \Theta$ tel que $\|u - \hat{\theta}_n\| \geq r$. On a alors

$$\psi(u) - \psi(\hat{\theta}_n) - \left\langle \nabla_{\hat{\theta}_n} \psi, u - \hat{\theta}_n \right\rangle = \psi\left(\hat{\theta}_n + \frac{r(u-\hat{\theta}_n)}{\|u-\hat{\theta}_n\|}\frac{\|u-\hat{\theta}_n\|}{r}\right) - \psi(\hat{\theta}_n) - \frac{\|u-\hat{\theta}_n\|}{r}\left\langle \nabla_{\hat{\theta}_n}\psi, \frac{r}{\|u-\hat{\theta}_n\|}(u-\hat{\theta}_n)\right\rangle.$$

Comme

$$\frac{1}{\|u-\hat{\theta}_n\|}\left(\psi\left(\hat{\theta}_n + \frac{r(u-\hat{\theta}_n)}{\|u-\hat{\theta}_n\|}\frac{\|u-\hat{\theta}_n\|}{r}\right) - \psi(\hat{\theta}_n)\right) \geq \frac{1}{r}\left(\psi\left(\hat{\theta}_n + \frac{r(u-\hat{\theta}_n)}{\|u-\hat{\theta}_n\|}\right) - \psi(\hat{\theta}_n)\right),$$

on en déduit

$$\begin{aligned}\psi(u) - \psi(\hat{\theta}_n) - \left\langle \nabla_{\hat{\theta}_n}\psi, u-\hat{\theta}_n\right\rangle &\geq \frac{\|u-\hat{\theta}_n\|}{r}\left[\psi\left(\hat{\theta}_n + \frac{r(u-\hat{\theta}_n)}{\|u-\hat{\theta}_n\|}\right) - \psi(\hat{\theta}_n) - \left\langle \nabla_{\hat{\theta}_n}\psi, \frac{r}{\|u-\hat{\theta}_n\|}(u-\hat{\theta}_n)\right\rangle\right] \\ &\geq \frac{\kappa\|u-\hat{\theta}_n\|}{r} \geq \kappa.\end{aligned}$$

On en déduit

$$\sup_{u \notin \mathrm{B}(\hat{\theta}_n, r)} \ell_n(\hat{\theta}_n) - \ell_n(u) \geq \kappa.$$

Comme ℓ_n admet son unique maximum sur $\mathrm{B}(\hat{\theta}_n, r)$ en $\hat{\theta}_n$, on en déduit que ℓ_n admet son unique maximum sur $\bar{\Theta}$ en $\hat{\theta}_n$ (en utilisant Fatou pour les points du bord et à l'infini). □

L'estimateur du maximum de vraisemblance est donc bien défini de manière unique dès lors que $\bar{T}_n \in \Lambda$. Il peut toutefois être bien défini en dehors de ce cas (cf poly de l'année dernière), mais son interprétation comme un estimateur par moments est tributaire de cette condition.

Exemple 3.35 *: Loi Binomiale.* Dans le modèle $(\mathcal{B}(n,p))_{p\in]0,1[}$, si on prend pour mesure dominante sur $[\![0,n]\!]$

$$\mu(\{k\}) = \binom{n}{k},$$

on peut se ramener à un modèle exponentiel de statistique exhaustive X, paramètre $\theta = \log\left(\frac{p}{1-p}\right) \in \mathbb{R}$, et fonction de partition

$$Z(\theta) = (1-p)^{-n} = \left(1+e^{\theta}\right)^n.$$

Le domaine $\Theta_{dom} = \Theta = \mathbb{R}$ et est bien ouvert. De plus,

$$(\log(Z))'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}}$$

induit un C^∞-difféo entre Θ et $]0,n[$. On aurait pu s'en rendre compte plus rapidement en remarquant que $E_\theta X = np = ne^\theta/(1+e^\theta)$. La log-vraisemblance s'écrit

$$\ell_n(\theta) = X\theta - n\log\left(1+e^\theta\right).$$

Lorsque $X \in]0,n[= \Lambda$, l'estimateur du maximum de vraisemblance est bien défini par

$$\hat{\theta}_{EMV} = \log\left(\frac{X}{n-X}\right).$$

Cela correspond aux situations où il correspond à l'estimateur par moments. Lorsque $X = n$ ou $X = 0$, l'estimateur du maximum de vraisemblance n'existe pas (moralement correspond à $+\infty$, $-\infty$). C'est un cas où bonne définition de l'EMV et équivalence avec la méthode des moments coïncident.

Exemple 3.36 *: Modèle Géométrique.* On regarde le modèle $\mathcal{G}(p)_{p>0}^{\otimes n}$, correspondant à n observations d'un modèle exponentiel i.i.d., de statistique exhaustive $X-1$, paramètre $\theta = \log(p)$, fonction de partition

$$\log(Z(\theta)) = -\log(1-p) = -\log(1-e^\theta).$$

Les domaines sont $\Theta_{dom} =]-\infty, 0]$ et $\Theta =]-\infty, 0[$. Comme $E_\theta(T) = \frac{1}{p} = e^{-\theta}$, on en déduit que $\Lambda =]1, +\infty[$.

Lorsque $\bar{X}_n > 1$, (donc dans Λ), l'estimateur du maximum de vraisemblance est bien défini par correspondance avec l'estimateur des moments :

$$\hat{\theta}_{EMV} = -\log((\bar{X}_n - 1).$$

Lorsque $\bar{X}_n = 1$ (correspondant à $X_1 = X_2 = \dots X_n = 1$), la log-vraisemblance s'écrit

$$\ell_n(\theta) = n\log(1-e^\theta),$$

qui atteint son maximum en $\theta = 0 \in \Theta_{dom} \setminus \Theta$. Dans ce cas l'estimateur par max de vraisemblance reste défini (mais ne coïncide pas avec l'estimateur par moment "classique").

On remarque que $\mathbb{P}\left(\hat{\theta}_{EMV} \neq \hat{\theta}_{moments}\right) = p^n = e^{\theta n}$. En toute généralité on aura toujours $\mathbb{1}_{\hat{\theta}_{EMV} = \hat{\theta}_{moments}} \to 1$ en probabilité, mais la vitesse se traite au cas par cas.

Comme, pour tout θ, $\bar{T}_n \to E_\theta(T(X)) \in \Lambda$ en P_θ-probabilité, on en déduit le résultat de convergence suivant pour le maximum de vraisemblance.

Théorème 3.37

Dans un modèle exponentiel minimal sous forme canonique, soit $\theta \in \Theta$. On a alors

$$\sqrt{n}(\hat{\theta}_{EMV} - \theta)\mathbb{1}_{\bar{T}_n \in \Lambda} \rightsquigarrow \mathcal{N}(0_k, I(\theta)^{-1}).$$

Démonstration. On commence par vérifier que $S : \theta \mapsto E_\theta(T(X)) = \nabla_\theta Z$ satisfait les hypothèses du Théorème 3.1 :

1. Z est un $\mathcal{C}^\infty$-difféomorphisme,
2. $D_\theta Z = I(\theta)$ est inversible,
3. $E_\theta(\|T(X)\|^2) < +\infty$.

On en déduit alors que

$$\sqrt{n}(S^{-1}(\bar{T}_n) - \theta)\mathbb{1}_{\bar{T}_n \in \Lambda} \rightsquigarrow \mathcal{N}(0_k, I(\theta)^{-1}\mathrm{Cov}_\theta(T(X))I(\theta)^{-1}).$$

En remarquant que $\mathrm{Cov}_\theta(T(X)) = I(\theta)$, et $S^{-1}(\bar{T}_n)\mathbb{1}_{\bar{T}_n \in \Lambda} = \hat{\theta}_{EMV}\mathbb{1}_{\bar{T}_n \in \Lambda}$, on en déduit le résultat. □

On peut en déduire une région de confiance sur θ : si $q_{1-\alpha,k}$ est le $1-\alpha$ quantile d'une distribution $\chi^2(k)$,

$$\left\{ u \mid \left\| \sqrt{nI(u)}(\hat{\theta}_{EMV} - u) \right\|^2 \leq q_{1-\alpha,k} \right\}$$

est une région de niveau de confiance $1-\alpha$. On peut construire de manière plus explicite une ellipsoïde de confiance en remarquant que $I(\hat{\theta}_{EMV})$ convergence en P_θ probabilité vers $I(\theta)$.

$$\left\{ u \mid \left\| \sqrt{nI(\hat{\theta}_{EMV})}(\hat{\theta}_{EMV} - u) \right\|^2 \leq q_{1-\alpha,k}. \right\}$$

Pour ces deux régions de confiance, on a omis le $\mathbb{1}_{\bar{T}_n \in \Lambda}$ (qui devrait y être pour assurer l'existence de $\hat{\theta}_{EMV}$.

Remarque : La "variance asymptotique" de $\hat{\theta}_{EMV}$ collant avec la borne de Cramer-Rao, et le modèle $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta \in \Theta})$ vérifiant les hypothèses du Théorème 3.14, un tel estimateur est dit *asymptotiquement efficace*.

Remarque 2 : On peut déduire de tout ce qui précède des résultats pours les modèles exponentiels sous forme non-canonique mais qui restent identifiables, en somme des modèles exponentiels minimaux reparamétrés. Si $\psi : U\Theta$ est un C^1-difféomorphisme (entre deux ouverts), et qu'on définit P_u par $P_{\psi(u)}$, alors

1. $\hat{u}_{EMV}$ est bien définit lorsque $\hat{\theta}_{EMV}$ l'est, et $\hat{u}_{EMV} = \psi^{-1}(\hat{\theta}_{EMV})$.
2. la méthode Δ donne

$$\sqrt{n}(\hat{u}_{EMV} - u) \rightsquigarrow \mathcal{N}(0_k, (D\psi)_u^{-1} I(\theta)^{-1}((D\psi)_u^{-1})^T).$$

On peut aussi vérifier que $(D\psi)_u^{-1} I(\theta)^{-1}((D\psi)_u^{-1})^T)$ correspond à $I(u)^{-1}$ (où $I(u)$ est l'information de Fisher associée à cette nouvelle paramétrisation). En effet, on peut écrire

$$\begin{aligned}
\ell(u) &= \langle T(X), \psi(u) \rangle - \log(Z(\psi(u)) \\
\dot{\ell}(u) &= ((D\psi)_u)^T \left(T(X) - S(\psi(u)) \right) = ((D\psi)_u)^T \left(T(X) - E_u(T(X)) \right),
\end{aligned}$$

de telle sorte que

$$I(u) = ((D\psi)_u)^T \mathrm{Cov}_u(T(X))(D\psi)_u) = ((D\psi)_u)^T I(\theta)(D\psi)_u).$$

Concluons cette partie avec un résultat de culture générale : on peut prouver que $\sqrt{n}(\hat{\theta}_{EMV} - \theta) \rightsquigarrow \mathcal{N}(0_k, I(\theta)^{-1})$ dès lors que le modèle est suffisamment régulier (voir par exemple le Théorème 3.3.15 du Dacunha Castelle Duflo, T2).

Mentionner Wald et Wilks ?

3.4 Tests basés sur maximum de vraisemblance

3.4.1 Test du rapport de vraisemblance

Dans un modèle $(\mathcal{X}, \mathcal{A}, (P_\theta \sim p_\theta \mu)_{\theta \in \Theta})$ dominé par μ, et deux hypothèses $\Theta_0, \Theta_1 \subset \Theta$, une manière standard de construire un test est de considérer le *rapport de vraisemblance* (basé sur un n-échantillon)

$$RV_{H_1,H_0}(x_{1:n}) = \frac{\sup_{\Theta_1} p_\theta^{\otimes n}(x_{1:n})}{\sup_{\Theta_0} p_\theta^{\otimes n}(x_{1:n})},$$

si dénominateur non-nul. En toute généralité c'est une heuristique qui donne la statistique sur laquelle baser le test, l'idée étant que sous H_1 on s'attend à observer de grandes valeurs de RV_{H_1,H_0}. Le test du rapport de vraisemblance est alors

$$T_{RV}(X_{1:n}) = \mathbb{1}_{RV_{H_1,H_0}(X_{1:n}) > t_\alpha},$$

où α est à a calibrer sous H_0. Lorsque l'on peut, on utilisera plutôt le log-ratio

$$\log(RV_{H_1,H_0}(x_{1:n})) = \sup_{\Theta_1} \ell_n(\theta) - \sup_{\Theta_0} \ell_n(\theta),$$

ou n'importe quelle fonction g telle que $RV_{H_1,H_0}(X_{1:n}) > t \Leftrightarrow g(X_{1:n}) \in A_t$, si le terme de droite est plus facilement manipulable. Les tests de rapport de vraisemblance ont quelques propriétés sympathiques dans des situations précises (deux points et modèles réguliers). En dehors de ces cadres, des tests "maisons" peuvent être préférables.

Hypothèses à deux points : théorie de Neyman-Pearson

On s'intérese ici au cas particulier où $\Theta_0 = \{\theta_0\}$, et $\Theta_1 = \{\theta_1\}$. En notant p_0 et p_1 les densités correspondantes, le test du rapport de vraisemblance est

$$T(X) = \mathbb{1}_{\left\{\frac{p_1}{p_0}(X) > t_\alpha\right\}},$$

où $t_\alpha = \sup\left\{t \mid P_0\left(\frac{p_1}{p_0}(X) \leq t\right) \leq 1-\alpha\right\}$. On aura toujours $P_0\left(\frac{p_1}{p_0}(X) > t_\alpha\right) \leq \alpha$ (continuité à droite de la fonction de répartition de $\frac{p_1}{p_0}(X)$ sous P_0). Lorsque ce test est de niveau exact α, on peut montrer le résultat d'optimalité suivant :

Théorème 3.38 : Lemme de Neyman-Pearson

Si le test du rapport de vraisemblance T_{RV} est de niveau exact α, alors, pour tout autre test T de niveau α, on a

$$\beta(T) \leq \beta(T_0),$$

où β est la puissance $\beta(T) = P_1(T = 1)$.

Par ailleurs, si T est un test de niveau exact α et de même puissance que T_{RV}, alors T_{RV} et T_0 sont égaux μ p.p..

En d'autres termes, lorsqu'ils sont de niveau exact, les tests du rapport de vraisemblance sont UPP ("uniformément les plus puissants"), dans le cas d'hypothèses à deux points, et cela peut être légèrement généralisé au cas des vraisemblances monotones.

Démonstration. D'après les hypothèses on a t_α tel que $P_0\left(\frac{p_1}{p_0}(X) > t_\alpha\right) = \alpha$. L'astuce consiste à regarder la quantité suivante

$$\int (T - T_{RV})(x)(p_1 - t_\alpha p_0)(x)\mu(dx).$$

Comme $p_1(x) > t_\alpha p_0(x) \Rightarrow T_{RV}(x) = 1$, on a

$$\begin{aligned} 0 &\geq \int (T - T_{RV})(x)(p_1 - t_\alpha p_0)(x)\mu(dx) \\ &\geq \beta(T) - t_\alpha P_0(T = 1) - \beta(T_{RV}) + t_\alpha P_0(T_{RV} = 1) \\ &\geq (\beta(T) - \beta(T_{RV})) + t_\alpha(\alpha - P_0(T = 1)). \end{aligned}$$

Comme $P_0(T = 1) \leq \alpha$, on en déduit $\beta(T) \leq \beta(T_{RV})$.

Pour l'unicité, remarquons que $P_0\left(\frac{p_1(X)}{p_0(X)} \leq t_\alpha\right) = 1-\alpha$ implique que $P_0\left(\frac{p_1(X)}{p_0(X)} = t_\alpha\right) = 0$ (pas de saut de la fonction de répartition à gauche implique continuité). Cela implqie en particulier que $\mu\left(\{p_1(x) = t_\alpha p_0(x)\}\right) = 0$. L'égalité des niveaux et fonctions puissances donne

$$\int (T - T_{RV})(x)(p_1 - t_\alpha p_0)(x)\mu(dx) = 0,$$

et on rappelle que $(T - T_{RV})(x)(p_1 - t_\alpha p_0)(x) \leq 0$ μ-p.p.. On en déduit $(T - T_{RV})(x)(p_1 - t_\alpha p_0)(x) = 0$ μ-p.p., et d'après ce qui précède $(T - T_{RV})(x) = 0$ μ-p.p.. □

Dans le cas où le test du rapport de vraisemblance est de niveau $< \alpha$, le test du rapport de vraisemblance peut ne pas être le plus puissant parmi les tests de niveau α. D'après ce qui précéde on comprend que cela correspond aux situations où $P_0\left(\frac{p_1}{p_0}(X) = t_\alpha\right) > 0$. On peut alors constuire un test *randomisé* à partir du test du rapport de vraisemblance qui lui va rester optimal (au sens de la puissance).

THÉORÈME 3.39 : NEYMAN-PEARSON RANDOMISÉ

Si $P_0(T_{RV} = 1) = \alpha_- < \alpha$, on note alors $\alpha_+ = P_0\left(\frac{p_1}{p_0}(X) \geq t_\alpha\right) \geq \alpha$, et on construit le test randomisé $\tilde{T}_{RV}$ suivant.

- *Si $\frac{p_1}{p_0}(x) < t_\alpha$ ou $\frac{p_1}{p_0}(x) > t_\alpha$, $\tilde{T}_{RV}(x) = T_{RV}(x)$.*
- *Si $\frac{p_1}{p_0}(x) = t_\alpha$, $\tilde{T}_{RV}(x) \sim \mathcal{B}(\kappa)$ (on tire au hasard la décision), avec*

$$\kappa = \frac{\alpha - \alpha_-}{\alpha_+ - \alpha_-}$$

$\tilde{T}_{RV}$ est alors de niveau exact α, est UPP parmi les tests de niveau α, et si T est un autre test de niveau α de même puissance, on a $T = T_{RV}$ sur $\left\{\frac{p_1}{p_0}(x) \neq t_\alpha\right\}$.

On comprend bien qu'on ne peut plus avoir unicité au sens μ-p.p., car on peut construire plusieurs versions randomisées du test de rapport de vraisemblance avec des tirages différents de Bernoulli qui donneront des test UPP. En revanche tous ces test vont coïncider avec T_{RV} en dehors de $\left\{\frac{p_1}{p_0}(x) = t_\alpha\right\}$.

Démonstration. Commençons par vérifier que $\tilde{T}_{RV}$ est de niveau exact α. On remarque que $P_0\left(\frac{p_1}{p_0}(X) = t_\alpha\right) = \alpha_+ - \alpha_-$

$$\begin{aligned} P_0\left(\tilde{T}_{RV} = 1\right) &= P_0\left(\tilde{T}_{RV} = 1 \cap \frac{p_1}{p_0}(X) = t_\alpha\right) + P_0\left(\tilde{T}_{RV} = 1 \cap \frac{p_1}{p_0}(X) \neq t_\alpha\right) \\ &= \kappa(\alpha_+ - \alpha_-) + P_0(T_{RV} = 1) \\ &= \alpha. \end{aligned}$$

Si T est de niveau α, on peut écrire

$$\begin{aligned} E_1(\mathbb{1}_{\tilde{T}_{RV}=1} - \mathbb{1}_{T=1}) &= E_1(\tilde{T}_{RV} - T) \\ &= E_1((\tilde{T}_{RV} - T)\mathbb{1}_{p_0(X)\neq 0}) + E_1((\tilde{T}_{RV} - T)\mathbb{1}_{p_0(X)=0} \\ &\geq E_0\left((\tilde{T}_{RV} - T)\frac{p_1(X)}{p_0(X)}\right), \end{aligned}$$

car $p_0(x) = 0 \Rightarrow \tilde{T}_{RV} = 1$. On a alors

$$\begin{aligned} \beta(\tilde{T}_{RV}) - \beta(T) &\geq E_0\left((\tilde{T}_{RV} - T)\frac{p_1(X)}{p_0(X)}\right) \\ &\geq E_0\left((\tilde{T}_{RV} - T)\left(\frac{p_1(X)}{p_0(X)} - t_\alpha\right)\right) + t_\alpha E_0(\tilde{T}_{RV} - T) \\ &\geq E_0\left((\tilde{T}_{RV} - T)\left(\frac{p_1(X)}{p_0(X)} - t_\alpha\right)\right) \\ &\geq 0. \end{aligned}$$

Pour l'unicité, en remontant ces inégalités, si T est de niveau α et de même puissance que $\tilde{T}_{RV}$, alors T est de niveau *exact* α, et vérifie

$$(\tilde{T}_{RV}(x) - T(x))\mathbb{1}_{\frac{p_1}{p_0}(x)\neq t_\alpha} = 0$$

P_0 p.s., ainsi que

$$(\tilde{T}_{RV}(x) - T(x))\mathbb{1}_{p_0(x)=0} = 0$$

P_1 p.s.. On en déduit que $(\tilde{T}_{RV}(x) - T(x))\mathbb{1}_{\frac{p_1}{p_0}(x)\neq t_\alpha} = 0$ μ-p.s.. □

On peut montrer que le test du rapport de vraisemblance est consistant, en se basant sur le théorème 3.33, en supposant que $E_0\left|\log\left(\frac{p_1}{p_0}\right)\right|$ et $E_1\left|\log\left(\frac{p_1}{p_0}\right)\right|$ sont bien définis. On met le test du rapport de vraisemblance sous la forme

$$T_{RV}(x_{1:n}) = \mathbb{1}_{\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) > t_{\alpha,n}},$$

et on suppose pour simplifier que le test du rapport de vraisemblance est de niveau exact, c'est à dire

$$P_0\left(\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) > t_{\alpha,n}\right) = \alpha.$$

Dans le cas général, on peut montrer que $\tilde{T}_{RV} - T_{RV}$ converge en proba (P_0 et P_1) vers 0 (les atomes de $\frac{p_1^{\otimes n}}{p_0^{\otimes n}}(x_{1:n})$ vont être de masse tendant vers 0). Sous H_0,

on a $\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) \to -\mathrm{d}_{KL}(P_0\|P_1)$ en P_0-probabilité. On en déduit que $t_{\alpha,n} \to -\mathrm{d}_{KL}(P_0\|P_1)$. En effet, supposons par exemple $\limsup t_{\alpha,n} \geq -\mathrm{d}_{KL}(P_0\|P_1) + r$, où $r > 0$, on aurait alors

$$\begin{aligned}\alpha &= \limsup P_0\left(\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) > t_{\alpha,n}\right)\\ &\leq \limsup P_0\left(\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) > -\mathrm{d}_{KL}(P_0\|P_1) + r\right)\\ &= 0,\end{aligned}$$

d'où la contradiction (ça marche dans l'autre sens si on suppose $\liminf t_{\alpha,n} \leq -\mathrm{d}_{KL}(P_0\|P_1) - r$).

Maintenant, on a $\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) \to \mathrm{d}_{KL}(P_1\|P_0)$ en probabilité sous P_1. On en déduit

$$\liminf P_1\left(\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) > t_{\alpha,n}\right) \geq \liminf P_1(\left(\bar{\ell}_n(\theta_1) - \bar{\ell}_n(\theta_0) \geq 0\right) = 1,$$

d'où la consistance du test du rapport de vraisemblance. On peut caractériser plus finement le comportement asymptotique de la puissance, en passant par les distances de Hellinger et l'inégalité de Pinsker (cf poly de Stéphane B).

Hypothèses composites

Citons sans le démontrer un résultat classique sur le test ru rapport de vraisemblance dans des modèles réguliers (pour lesquels on peut intervertir comme on veut dérivation des logs et espérance). Pour de tels modèles on aura toujours

$$\sqrt{n}\left(\hat{\theta}_{EMV} - \theta\right) \rightsquigarrow \mathcal{N}(0_k, (I(\theta))^{-1}),$$

sous P_θ.

Théorème 3.40 : Test du RV dans les modèles réguliers

Si $(P_\theta)_{\theta\in\Theta}$ est un modèle régulier identifiable de dimension k, et Θ_0 est un sous-ensemble de Θ de dimension r, alors, pour tout $\theta \in H_0$,

$$\Lambda_n := 2\left(\sup_{\theta\in\Theta} \ell_n(\theta) - \sup_{\theta\in\Theta_0} \ell_n(\theta_0)\right) \rightsquigarrow \chi^2(k-r).$$

En particulier, le test du rapport de vraisemblance

$$\mathbb{1}_{\Lambda_n > q_{1-\alpha,k-r}}$$

est asymptotiquement de niveau α ($q_{1-\alpha,k-r}$ désignant le $1-\alpha$-quantile d'une loi $\chi^2(k-r)$).

On peut trouver une preuve de ce résultat dans le Van der Waart Théorème 16.7, on montrera une version légèrement plus faible plus bas.

On remarque que, plutôt que de considérer $\sup_{\theta\in\Theta_1} \ell_n(\theta)$, la statistique de test fait intervenir $\sup_{\theta\in\Theta} \ell_n(\theta)$, où moralement $\Theta = \Theta_1 \cup \Theta_0$. Θ_0 étant de dimension $r < k$, Θ_1 peut être considéré dense dans Θ, et cela ne fait aucune différence.

En ce qui concerne la puissance de ce test, pour $\theta \notin \Theta_0$, on aura consistance dès lors que $\mathrm{d}_{KL}(\theta, \Theta_0) > 0$ et $\mathrm{d}_{KL}(\Theta_0, \theta) > 0$ (séparation des hypothèses).

Des hypothèses séparées sont une condition nécessaire pour la consistance. Dans le cas contraire, on peut montrer que la puissance minimale sur H_1 est toujours majorée par α. Dans le cadre du théorème ci-desssus, les deux hypothèses n'étant pas séparées, la puissance uniforme vaudra toujours α asymptotiquement.

3.4.2 Tests d'hypothèses sur les paramètres

On se place dans un cadre où

$$\sqrt{n}\left(\hat{\theta}_{EMV} - \theta\right) \rightsquigarrow \mathcal{N}(0_k, (I(\theta))^{-1}), \tag{3.3}$$

pour tout $\theta \in \Theta$ (modèle régulier, exponentiel par exemple). On peut prouver des versions plus faibles du Théorème 3.40, quand le sous ensemble Θ_0 est donné par un système d'équations.

PROPOSITION 3.41 : TEST D'HYPOTHÈSES LINÉAIRES SUR θ

Dans un modèle identifiable où on a (3.3), $\Theta \subset \mathbb{R}^k$ est ouvert, et $C : \mathbb{R}^k \to \mathbb{R}^r$ est linéaire de rang $r < k$, on regarde les hypothèses

$$\begin{array}{rcl} H_0 & : & C\theta = c \\ H_1 & : & C\theta \neq c, \end{array}$$

où c est un vecteur constant de $\mathbb{R}^r$. Si $\hat{I}_n$ est un estimateur consistant de $I(\theta)^{-1}$ (que l'on suppose inversible), on introduit la statistique de Wald

$$W_n = n(C\hat{\theta}_{EMV} - c)^T (C\hat{I}_n C^T)^{-1} (C\hat{\theta}_{EMV} - c) \rightsquigarrow \chi^2(r),$$

pour tout $\theta \in \Theta_0$. Le test de Wald associé est

$$\mathbb{1}_{W_n > q_{1-\alpha,r}},$$

où $q_{1-\alpha,r}$ est le $1-\alpha$-quantile d'un $\chi^2(r)$, et est de niveau asymptotique α.

On peut montrer que ce test est équivalent à celui du maximum de vraisemblance (cf le poly de Stéphane B, ou plus généralement phénomène de Wilks). La preuve en est très simple.

Démonstration. Soit $\theta \in \Theta_0$. On a alors $\sqrt{n}C(\hat{\theta}_{EMV} - \theta) = \sqrt{n}(C\hat{\theta}_{EMV} - c) \rightsquigarrow \mathcal{N}(0_r, CI(\theta)^{-1}C^T)$, dont on peut déduire

$$\left\| \sqrt{n}\sqrt{(CI(\theta)^{-1}C^T)^{-1}}(C\hat{\theta}_{EMV} - c) \right\|^2 \rightsquigarrow \chi^2(r).$$

Comme $\hat{I}_n$ est un estimateur consistant de $I(\theta)^{-1}$, le lemme de Slutsky donne le résultat escompté. □

Dans le cas des modèles exponentiels, un estimateur de $I(\theta)^{-1}$ est donné par l'inverse de la matrice de covariance empirique

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(T(X_i) - \bar{T}_n)(T(X_i) - \bar{T}_n)^T.$$

On peut généraliser ce résultat pour des Θ_0 engendré par des contraintes non-linéaires.

THÉORÈME 3.42 : TEST DE WALD SUR θ

Dans un modèle identifiable où on a (3.3), $\Theta \subset \mathbb{R}^k$ est ouvert, et $\psi : \mathbb{R}^k \to \mathbb{R}^r$ est une submersion $\mathcal{C}^1$ de rang $r < k$. Pour les hypothèses

$$\begin{array}{rcl} H_0 & : & \psi(\theta) = 0 \\ H_1 & : & \psi(\theta) \neq 0, \end{array}$$

la statistique de Wald s'écrit

$$W_n = n\psi(\hat{\theta}_{EMV})^T \left[D_{\hat{\theta}_{EMV}} \psi \hat{I}_n (D_{\hat{\theta}_{EMV}} \psi)^T \right]^{-1} \psi(\hat{\theta}_{EMV}) \rightsquigarrow \chi^2(r),$$

pour tout $\theta \in H_0$, où $\hat{I}_n$ est un estimateur consistant de $I(\theta)^{-1}$ (supposée inversible). Le test de Wald associé est

$$\mathbb{1}_{W_n > q_{1-\alpha,r}},$$

où $q_{1-\alpha,r}$ est le $1-\alpha$-quantile d'un $\chi^2(r)$, et est de niveau asymptotique α.

On remarque qu'on peut choisir $\psi_c = \psi - c$ de telle sorte que le 0 n'est pas limitant dans cet énoncé. Le preuve reste très simple.

Démonstration. Soit $\theta \in \Theta_0$. On commence par remarquer que si (3.3) est vérifiée, alors $\hat{\theta}_{EMV}$ est un estimateur consistant de θ. On a alors $\sqrt{n}\psi(\hat{\theta}_{EMV}) = \sqrt{n}(\psi(\hat{\theta}_{EMV}) - \psi(\theta)) \rightsquigarrow \mathcal{N}(0_r, (D_\theta\psi) I(\theta)^{-1} (D_\theta\psi)^T)$, en utilisant la méthode Δ. On en déduit

$$\left\| \sqrt{n} \sqrt{[(D_\theta\psi) I(\theta)^{-1} (D_\theta\psi)^T]^{-1}} \psi(\hat{\theta}_{EMV}) \right\|^2 \rightsquigarrow \chi^2(r).$$

On peut conclure en invoquant le Lemme de Slutsky, en remarquant que $\hat{I}_n$ est un estimateur consistant de $I(\theta)^{-1}$, et comme ψ est $\mathcal{C}^1$, $(D_{\theta_{EMV}}\psi)$ est un estimateur consistant de $(D_\theta\psi)$. □

La morale de ces deux résultats est la même que dans le Théorème 3.40 : un test basé sur le max de vraisemblance (Wald ou rapport de vraisemblance) d'un sous-modèle de dimension r se comportera moralement et asymptotiquement comme un χ^2 du nombre de degrés de libertés restant, c'est à dire $k - r$.

3.4.3 Test du chi-deux d'indépendance

Concluons cette partie avec un exemple très utilisé en pratique de test d'appartenance à un sous-modèle. On suppose ici qu'on observe un n-échantillon (X_i, Y_i) à valeurs dans $[\![1, q]\!] \times [\![1, r]\!]$, dont la loi est déterminée par le vecteur a (avec $p_{j,k} = P_(X, Y)(\{j\} \times \{k\})$. Le test du χ^2 d'indépendance va chercher à déterminer si X et Y sont indépendantes, ce qui se traduit sur les paramètres par $p_{j,k} = p_{j,.} \times p_{.,k}$,

avec $p_{j,.} = \sum_k p_{j,k}$, $p_{.,k} = \sum_j p_{j,k}$. Cela revient à tester l'appartenance de p à un sous-modèle $\mathcal{P}_0$, formé des vecteurs dans le simplexe de $\mathbb{R}^{qr}$ tels que $\psi(p) = 0_{qr}$, où

$$\psi : \begin{cases} \mathbb{R}^{qr} & \to & \mathbb{R}^{qr} \\ p & \mapsto & (p_{j,k} - p_{j,.} \times p_{.,k})_{j,k}. \end{cases}$$

On pourrait peut-être s'en sortir avec une adaptation du test de Wald (cf Théorème 3.42), mais c'est un peu calculatoire, et en pratique on utilise plutôt la statistique de Pearson :

$$C_n(\hat{p}) = \sum_{j,k} \frac{(N_{j,k} - n\hat{p}_{j,k})^2}{n\hat{p}_{j,k}}, \tag{3.4}$$

où $N_{j,k} = \sum_i \mathbb{1}_{(X_i,Y_i)=\{(j,k)\}}$, $\hat{p}_{j,k}$ est l'estimateur du maximum de vraisemblance sous H_0 (ici l'indépendance). Un peu de calcul montre que

$$\hat{p}_{j,k} = \hat{p}_{j,.}\hat{p}_{.,k} = \frac{N_{j,.}N_{.,k}}{n^2}.$$

De manière générale, la statistique de Pearson se mettra toujours sous la forme

$$C_n(\hat{p}) = \sum_s \frac{(N_s - N_{theo,s})^2}{N_{theo,s}},$$

où N_s est l'effectif observé pour la case s, et $N_{theo,s}$ est l'effectif attendu sous H_0 pour cette case s (et donné formellement par maximisation de vraisemblance). Le fait que la maximisation de vraisemblance est correctement définie tient aussi au fait que $(N_s)_s$ suit une loi multinomiale paramétrée par p, qui fait partie de la famille exponentielle.

Pour bâtir un test sur la statistique de Pearson $C_n(\hat{p})$, il nous faut déterminer sa loi limite sous H_0.

Théorème 3.43

Si $p \in \mathcal{P}_0$ (c-à-d $X \perp\!\!\!\perp Y$), on a

$$C_n(\hat{p}) \rightsquigarrow \chi^2((q-1)(r-1)).$$

Ce résultat peut s'interpréter comme pour le Théorème 3.40 : la statistique de Pearson va converger vers un χ^2 de degrés de libertés le nombre de degrés de liberté résiduel, ici : $(qr-1)$ (degrés de liberté total) - $(q-1)+(r-1)$ (degrés de liberté de $\mathcal{P}_0$) $= (q-1)(r-1)$. Ce n'est pas un hasard : on peut montrer que statistique de Pearson et log-ratios de vraisemblances convergent en loi vers la même limite (cf Van der Waart ou poly de Stéphane B).

On peut toutefois prouver le Théorème 3.43 sans utiliser directement ce résultat, et en évitant les calcul liés à l'interprétation comme un test de Wald.

Démonstration. On va prouver légèrement plus fort : si $\mathcal{P}_0$ est une sous-variété ouverte (disons $\mathcal{C}^2$) de dimension s_0 de $\mathcal{P}$ (simplexe dans l'espace de dimension s), alors $C_n(\hat{p}) \rightsquigarrow \chi^2(s-1-s_0)$, sous $p \in \mathcal{P}_0$, où $\hat{p}$ est l'EMV sous $\mathcal{P}_0$.

L'astuce est de relier $C_n(\hat{p})$ à une projection. Pour ce faire, on introduit le

$$\bar{p} = \arg\min_{\mathcal{P}_0} C'_n(u) = \arg\min_{u\in\mathcal{P}_0} \sum_{j=1}^{s} \frac{(N_j - nu_j)^2}{np_j}.$$

On remarque que $Z_n = ((N_j - np_j)/\sqrt{np_j})_{j=1,\dots,s} \rightsquigarrow Z \sim \mathcal{N}(0_s, I_s - \sqrt{p}\sqrt{p}^T)$ (cf partie 2.2.1). Par ailleurs, en écrivant pour tout $u \in \mathcal{P}_0$,

$$\frac{1}{\sqrt{np_j}}(N_j - nu_j) = \frac{1}{\sqrt{np_j}}\left((N_j - np_j) + (n(p_j - u_j)\right) = (Z_n)_j - \frac{\sqrt{n}(u_j - p_j)}{\sqrt{p_j}},$$

on remarque que

$$\min_{u\in\mathcal{P}_0} \sum_{j=1}^{s} \frac{(N_j - nu_j)^2}{np_j} = \min_{h\in H_n} \|Z_n - h\|^2,$$

où $H_n = \left\{\sqrt{n}\mathrm{Diag}(1/\sqrt{p})(u-p) \mid u \in \mathcal{P}_0\right\}$. L'étape principale de la preuve est le Lemme suivant (on peut trouver une version plus proba dans Van Der Waart, Lemme 7.13).

Lemme 3.44

Sous les hypothèses du Théorème, on a

$$\min_{h\in H_n} \|Z_n - h\|^2 \rightsquigarrow \min_{h\in H} \|Z - h\|^2,$$

où $H = \mathrm{Diag}(1/\sqrt{p})T_p\mathcal{P}_0$.

Démonstration. On va admettre le résultat de géomérie suivant : comme $\mathcal{P}_0$ est $\mathcal{C}^2$, il existe $c > 0$ et $\kappa > 0$ tels que $\pi_{T_p} : B_{\mathcal{P}_0}(p, c) \to \pi_{T_p}(B_{\mathcal{P}_0}(p, c)) \supset \mathrm{B}_{T_p}(0, c/2)$ soit un C^1-difféomorphisme vérifiant $\|(u-p) - \pi_T(u)\| \leq \kappa\|u-p\|^2$ et $\|\pi_T(u)\| \geq \|u-p\|/2$.

On note $D = Diag(1/\sqrt{p})$. Tout d'abord, comme $\mathrm{d}(Z_n, H) \rightsquigarrow \mathrm{d}(Z, H)$ (continuité de $\mathrm{d}(., H)$), on en déduit qu'il suffit de montrer que $\mathrm{d}(Z_n, H_n) - \mathrm{d}(Z_n, H)$ tend en probabilité vers 0 (pour appliquer Slutsky).

Deuxième remarque : Z_n converge en loi, donc est tendue (ou $O_P(1)$), on peut alors se ramener à montrer que $\sup_{z\in\mathrm{B}(0,M)} |\mathrm{d}(z, H_n) - \mathrm{d}(z, H)|$ tend vers 0. Soit donc $z \in \mathrm{B}(0, M)$, et u_n tel que $\mathrm{d}(z, H_n) = \|z - D\sqrt{n}(u_n - p)\|$. Comme $0 \in H_n$, on a $\mathrm{d}(z, H_n) \leq M$ et alors $\|u_n - p\| \leq \frac{CM}{\sqrt{n}}$ (dans toute la suite C va être une constante positive ne dépendant pas de n, ni des points de base z ou h, une quantité vraiment constante quoi). Pour n assez grand (ne dépendant pas de z), on a alors $\|u_n - p\| \leq c$. En notant $v = \pi_{T_p}(u_n - p)$, on a

$$\|z - \sqrt{n}D(u_n - p)\| \geq \|z - \sqrt{n}Dv\| - \sqrt{n}\frac{CM^2}{n} \geq \mathrm{d}(z, H) - \frac{CM^2}{\sqrt{n}}.$$

Réciproquement, si $h = Dv \in H$ tel que $\mathrm{d}(z, H) = \|z - h\|$, on a $\|v\| \leq CM$ (uniformément en z). On a donc, pour n assez grand (ne dépendant pas de z) $\|v\|/\sqrt{n} \in B_{T_p}(0, c/2)$, et en notant $\psi = \pi_T^{-1}$, on a, en posant $u_n = \psi^{-1}(\|v\|/\sqrt{n})$,

$$\|\frac{v}{\sqrt{n}} - (u_n - p)\| \leq \frac{CM^2}{n}.$$

Dès lors, on peut écrire

$$\begin{aligned} \mathrm{d}(z,H) = \|z - Dv\| &= \left\| z - \sqrt{n} D \frac{v}{\sqrt{n}} \right\| \\ &\geq \left\| z - \sqrt{n} D(u_n - p) \right\| - \frac{CM^2}{\sqrt{n}}. \end{aligned}$$

On en déduit que, pour n assez grand, $\sup_{\|z\| \leq M} |\mathrm{d}(z, H_n) - \mathrm{d}(z, H)| \leq CM^2/\sqrt{n} \to 0$, et donc le résultat. □

On peut maintenant complètement caractériser la loi de $\min_{h \in H} \|Z - h\|^2$. On commence par remarquer que $Z \sim \pi_{\sqrt{p}^\perp} N$, où N est un vecteur Gaussien standard de $\mathbb{R}^s$. Puis, comme pour tout $v \in \mathcal{P}_0$, $(\mathrm{Diag}(1\sqrt{p})(a-p))^T \sqrt{p} = 0$, en passant à la limite on déduit que $H \perp \sqrt{p}$. On a donc

$$\min_{h \in H} \|Z - h\|^2 = \left\| \pi_{(H \oplus \sqrt{p})^\perp} N \right\|^2 \sim \chi^2(s - 1 - s_0).$$

Il reste à relier $C_n'(\bar{p})$ à $C_n(\hat{p})$. Pour faire plus rapide, on va introduire les notions de O_P et o_P.

Definition 3.45

Soit X_n une suite de vecteurs aléatoires. On définit

$$\begin{array}{rcl} X_n = O_P(1) & \text{ssi} & X_n \text{ est tendue,} \\ X_n = o_P(1) & \text{ssi} & X_n \text{ converge vers 0 en proba.} \end{array}$$

On peut alors, pour X_n et Y_n deux suites de vecteurs aléatoires, définir

$$\begin{array}{rcl} X_n = O_P(Y_n) & \text{ssi} & X_n = Z_n Y_n, \text{ avec } Z_n = O_P(1), \\ X_n = o_P(Y_n) & \text{ssi} & X_n = Z_n Y_n, \text{ avec } Z_n = o_P(1). \end{array}$$

Dans le cas de suites déterministes, on retombe sur les notions de o et O classiques. Par ailleurs, les O_P et o_P satisfont les mêmes règles que leurs pendants déterministes. Par exemple, on peut relier développement limité et O_P, o_P.

Lemme 3.46

Soit f fonction définie sur un voisinage de 0 et telle que $f(0) = 0$. Soit $X_n = o_P(1)$. On a alors

- *Si $f(x) = o(\|x\|^p)$, alors $f(X_n) = o_P(\|X_n\|^p)$.*
- *Si $f(x) = O(\|x\|^p)$, alors $f(X_n) = O_P(\|X_n\|^p)$.*

La preuve est immédiate en regardant $g(x) = f(x)/\|x\|^p$. Le formalisme o_P, O_P est assez utile pour montrer l'équivalence des limites en loi de deux processus de manière rapide. Par exemple, dans le cas de la méthode Δ, si on a $X_n - a = o_P(1)$, $r_n(X_n - a) \rightsquigarrow Z$, et g différentiable en a, on peut écrire

$$g(X_n) - g(a) - D_a g(X_n - a) = o_P(\|X_n - a\|),$$

et donc

$$r_n(g(X_n) - g(a) - (D_a g)(X_n - a)) = o_P(r_n \|X_n - a\|) = o_P(1),$$

car $r_n(X_n - a)$ converge en loi donc est tendue. On en déduit (en utilisant Slutsky) que $r_n(g(X_n) - g(a))$ et $(D_a g)(r_n(X_n - a))$ ont même limite en loi, c'est à dire $(D_a g)Z$.

Dans le cas qui nous intéresse, on va relier C'_n, C_n et $\Lambda_n(p) = 2(\sup_{q\in\mathcal{P}} \ell_n(q) - \ell_n(p))$ (statistique du rapport de vraisemblance, celle que $\hat{p}$ minimise sur $\mathcal{P}_0$), pour $\bar{p}$ et $\hat{p}$. Commençons par remarquer que $\hat{p} - p = O_P(1/\sqrt{n})$ (c'est le Théorème 3.37). Par ailleurs, on a

$$C'_n(\bar{p}) = \sum_{j=1}^{s} \frac{(N_j - n\bar{p}_j)^2}{np_j} \leq \sum_{j=1}^{s} \frac{(N_j - np_j)^2}{np_j}.$$

À j, fixé, on en déduit

$$\frac{n|p_j - \bar{p}_j|}{\sqrt{np_j}} \leq \frac{|N_j - np_j|}{\sqrt{np_j}} + \frac{|N_j - n\bar{p}_j|}{\sqrt{np_j}} \leq \sqrt{C'_n(\bar{p})} + \sqrt{C'_n(p)} \leq 2\sqrt{C'_n(p)} = O_P(1),$$

et donc $\bar{p} - p = O_P(1/\sqrt{n})$. Montrons maintenant l'équivalence asymptotique de C'_n, C_n et Λ_n pour une suite $\hat{q}$ telle que $\hat{q} - p = O_P(1/\sqrt{n})$.

On remarque que

$$\Lambda_n(\hat{q})/2 = \sum_{j=1}^{s} N_j \log\left(\frac{N_j/n}{\hat{q}_j}\right) = -\sum_{j=1}^{s} N_j \log\left(1 + \left(\frac{n\hat{q}_j}{N_j} - 1\right)\right).$$

Par ailleurs $N_j/n = p_j + O_P(1/\sqrt{n})$ et $\hat{q}_j = p_j + O_P(1/\sqrt{n})$. En notant $Z_n = \left(\frac{n\hat{q}_j}{N_j} - 1\right)$, on en déduit que $Z_n = O_P(1/\sqrt{n})$. Le développement limité de $\log(1+x)$ donne alors

$$\Lambda_n(\hat{q})/2 = -\sum_{j=1}^{s} N_j \left(\frac{n\hat{q}_j}{N_j} - 1\right) + \sum_{j=1}^{s} N_j \left(\frac{n\hat{q}_j}{N_j} - 1\right)^2 /2 + \sum_{j=1}^{s} N_j o_P(\|Z_n\|^2).$$

Comme $N_j = O_P(n)$, $N_j o_P(\|Z_n\|^2) = o_P(1)$. Par ailleurs, $\sum_{j=1}^{s}(n\hat{q}_j - N_j) = n - n = 0$. On en déduit

$$\Lambda_n(\hat{q}) = \sum_{j=1}^{s} \frac{(N_j - n\hat{q}_j)^2}{N_j} + o_P(1).$$

Comme $N_j/n = p_j + o_P(1) = \hat{q}_j + o_P(1)$, on en déduit

$$\Lambda_n(\hat{q}) = C'_n(\hat{q}) + o_P(1) = C_n(\hat{q}) + o_P(1).$$

On peut alors conclure via les inégalités

$$\begin{aligned} C_n(\hat{p}) &= \Lambda_n(\hat{p}) + o_P(1) \leq \Lambda_n(\bar{p}) + o_P(1) = C'_n(\bar{p}) + o_P(1), \\ C_n(\hat{p}) &= C'_n(\hat{p}) + o_P(1) \geq C'_n(\bar{p}) + o_P(1), \end{aligned}$$

et Slutsky.

Pour être complet, il faudrait vérifier que $\mathcal{P}_0$ est une variété $\mathcal{C}^2$ de dimension $(q-1)+(r-s)$, ce dont on se convainc assez facilement en utilisant des paramétrisations locales. □

Remarque : Cette méthode de preuve est assez fidèle à l'esprit de la preuve du résultat plus général (Théorème 3.40) : montrer que le critère $\Lambda_n/2$ est asymptotiquement équivalent au carré d'une projection d'un vecteur Gaussien standard sur un espace de dimension $(k-r)$ (l'orthogonal du sous espace des paramètres de dimension r). Les étapes en sont peu ou prou les mêmes, en passant cette fois-ci par un autre critère $(\hat{\theta}-\theta)^T I(\theta)(\hat{\theta}-\theta)^T$ (qui lui va être équivalent asymptotiquement à une projection). En somme, cela montre l'équivalence asymptotique des tests du rapport de vraisemblance et de Wald.
Remarque finale : En pratique, pour que l'approximation normale sous-jacente aux tests du Chi-deux soit considérée comme valide, on demande de vérfier que, pour tout $j \in [\![1, s]\!]$, $N_{theo,j} \geq 5$, c'est à dire que les effectifs théoriques attendus sous H_0 (par exemple sous l'hypothèse d'indépendance soient suffisamment grands).

3.5 Limitations de l'approche max de vrais

Les estimations et tests par maximum de vraisemblance sont importants d'un point de vue culturel : beaucoup de tests couramment utilisés se basent dessus, et certaines approches modernes (comme la régression logistique) y sont directement reliés. En revanche, ils souffrent de certains défauts structurels, dont on peut résumer l'esprit ainsi : à n'utiliser que lorsque l'on est sûr que la vraie loi sous-jacente est dans un modèle régulier !

Limitations calculatoires

Même dans le cas des modèles exponentiels, une formule close pour un estimateur du maximum de vraisemblance n'est pas toujours possible. Par exemple, dans le modèle $\gamma(a,b)$, l'estimateur du maximum de vraisemblance vérifie le système d'équations

$$\frac{\Gamma'(\hat{\theta}_1)}{\Gamma(\hat{\theta}_1)} - \log(\hat{\theta}_2) = \overline{\log(X)}_n$$
$$\frac{\hat{\theta}_1 + 1}{\hat{\theta}_2} = -\bar{X}_n,$$

qui n'admet pas de formule close. On peut s'en sortir néanmoins avec des méthodes de résolution approchée, ou des méthodes one-step.

Limitations conceptuelles

Comme pour les M-estimateurs en général, si $P \notin (P_\theta)_{\theta\in\Theta}$, on n'a aucune chance de construire un estimateur consistant. En effet, la cible est alors $\theta^* \in \arg\min M(\theta)$, et le M estimateur va alors converger vers θ^*.

Dans le cadre de la maximisation de la vraisemblance, on aura dans le meilleur des cas (si on peut toujours bien définir un EMV) $\hat{\theta} \to \theta^*$, où $\theta^* \in \arg\min_\theta \mathrm{d}_{KL}(P\|P_\theta)$, en d'autre terms on convergera vers une projection au sens de la divergence de Kullback (pour peu que cette dernière soit bien définie). En somme, dans les modèles mal spécifiés, si $\mathrm{d}_{KL}(P\|P_\Theta)$ est grand, les performances en estimation seront médiocres. On peut pallier ce problème en "robustifiant" légèrement les estimateurs par maximum de vraisemblance (tests entre boules de Hellinger, ρ-estimation par exemple, cf les travaux de Y. Baraud, L. Birgé, M. Sard).

On peut aussi remarquer que les résultats sur l'EMV dans les mpodèles réguliers sont de nature asymptotique, et, bien que la borne de Cramer Rao soit atteinte asymptotiquement, il peut arriver que les EMV ne soient pas optimaux par rapport à des estimateurs biaisés (on verra ça dans le chapitre suivant).

Enfin, dans des modèles non-réguliers, le principe d'estimation par maxium de vraisemblance peut donner un peu n'importe quoi. Par exemple dans le modèle $(\mathcal{U}(]0,\theta[)^{\otimes n})_{\theta>0}$, supposons que l'on souhaite tester

$$\begin{aligned} H_0 :&\theta = 1 \\ H_1 :&\theta > 1. \end{aligned}$$

La vraisemblance en θ s'écrit $V(\theta) = \frac{1}{\theta^n}\mathbb{1}_{M_n<\theta}$, où $M_n = \max_{i=1,\dots,n} X_i$. On a alors, $\sup_{\theta\in H_1}(V(\theta)) = \mathbb{1}_{M_n<1} + \frac{1}{(M_n)^n}\mathbb{1}_{M_n\geq 1}$, et $\sup_{\theta\in H_0}(V(\theta)) = \mathbb{1}_{M_n<1}$. Le rapport de vraisemblance s'écrit alors

$$RV_{H_0,H_1} = \mathbb{1}_{M_n<1} + (+\infty)\mathbb{1}_{M_n\geq 1},$$

et le test du rapport de vraisemblance (pour un niveau de confiance $(1-\alpha) > 0$) sera toujours $\mathbb{1}_{M_n>1}$ (d'erreur de première espèce 0).

Chapitre 4

Statistiques Bayésiennes

Le chapitre sur la Statistique Bayésienne permet d'aborder deux notions : celle de la comparaison entre estimateurs et l'optimalité au sens minimax, ainsi que le point de vue bayésien qui permet de définir une autre gamme d'estimateurs, structurellement biaisés mais pertinents dans certaines applications.

4.1 Comparaison entre estimateurs

Pour mesurer la qualité d'un estimateur T, on définit une notion de proximité dans l'espace des paramètres (fonction de perte), et on peut regarder la distance entre l'estimateur et le paramètre, en moyenne par exemple :

Definition 4.1 : Perte et Risque

Dans le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, on suppose que l'on cherche à estimer $q(\theta) \in \mathbb{R}^k$. On munit $\mathbb{R}^k$ d'une *fonction de perte* $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^+$. Le *risque* (ou plutôt *fonction de risque*) d'un estimateur T de $q(\theta)$ est

$$R_T : \begin{cases} \Theta & \to & [0, +\infty] \\ \theta & \mapsto & E_\theta(\ell(q(\theta), T(X)). \end{cases}$$

Exemple 4.2 *: Exemples classiques.*

— La perte quadratique est définie par $\ell(x, y) = \|x - y\|^2$. Le risque associé est appelé *risque quadratique*, donné par

$$R_T(\theta) = E_\theta(\|q(\theta) - T(X)\|^2).$$

— On peut définir plus généralement les pertes ℓ_p par $\ell(x, y) = \|x - y\|^p$. Les plus utilisées sont pour $p = 2$ et $p = 1$.

— Une perte utilisée en classification est la perte 0/1 associé à un problème d'estimation "binaire" ($q(\theta) \in \{0, 1\}$), définie par $\ell(q(\theta), y)) = \mathbb{1}_{q(\theta)\neq y}$. Pour cette perte, la fonction de risque d'un estimateur T est

$$R_T(\theta) = E_\theta(\mathbb{1}_{q(\theta)\neq T(X)}) = P_\theta(T(X) \neq q(\theta)),$$

c'est à dire la probabilité sous P_θ que notre estimateur se trompe.

— Excès de risque en M-estimation : dans un cadre de M-estimation, on suppose que $\theta = \arg\min_{u\in\Theta} M(u) = E_\theta\gamma(u, X)$. On peut alors définir comme fonction de perte $\ell(\theta, u) = E_\theta(\gamma(u, X) - E_\theta(\gamma(\theta, X)) = M(u) - M(\theta)$, et la fonction de risque associée est, pour un estimateur T,

$$R_T(\theta) = E_\theta\left(M(T(X)) - M(\theta)\right),$$

qui est souvent appelée excès de risque (de fait on appelle risque la fonction M dans ce cas). Les notions de risque en M-estimation et en comparaison d'estimateurs ne doivent pas être confondues : dans le premier cas elle sert à définir un estimateur, dans le deuxième cas à comparer deux estimateurs au sens général.

On va essayer de comparer des estimateurs sur la base de leurs fonctions de risque.

Definition 4.3 : Meilleur que

Soient T_1 et T_2 deux estimateurs de $q(\theta)$. On dit que T_1 est meilleur que T_2 si

$$R_{T_1}(\theta) \leq R_{T_2}(\theta) \quad \forall\theta \in \Theta.$$

On dit que T_1 est strictement meilleur que T_2 si T_1 est meilleur que T_2 et s'il existe $\theta_0 \in \Theta$ tel que $R_{T_1}(\theta_0) < R_{T_2}(\theta_0)$.

Remarque : La relation "est meilleur que" définie sur l'ensemble des estimateurs de $q(\theta)$ est une relation d'ordre partiel : en général deux estimateurs T_1 et T_2 ne sont pas comparables.

Exemple 4.4. On observe $X_1, \ldots, X_n$ i.i.d. de loi $\mathrm{B}(\theta)$, la loi de Bernoulli de paramètre $\theta \in [0, 1]$, et θ est inconnu. Le modèle statistique correspondant est $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ avec

$$\mathcal{X} = \{0, 1\}^n, \quad \mathcal{A} = \mathcal{P}\left(\{0, 1\}^n\right), \quad P_\theta = \mathcal{B}(\theta)^{\otimes n},\ \theta \in [0, 1].$$

On note $X_i : \{0, 1\}^n \to \{0, 1\}$ la i-ème application coordonnée, $1 \leq i \leq n$. La moyenne empirique $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur de θ et son risque quadratique s'écrit $R_{\bar{X}_n}(\theta) = \theta(1 - \theta)/n$. $T = 1_{\mathbb{X}}/2$ est un estimateur (constant) de θ de risque quadratique $R_T(\theta) = (\theta - 1/2)^2$. T et $\bar{X}_n$ ne sont pas comparables.

Cet exemple est généralisable : il n'existe pas de meilleur estimateur de $q(\theta)$ (si le modèle n'est pas pathologique).

Proposition 4.5

Supposons que $\ell(x, y) > 0$ pour tout $x, y \in \mathbb{R}^k$ tels que $x \neq y$ et que q n'est pas une fonction constante. Supposons de plus que pour tous $\theta, \theta' \in \Theta$ les probabilités P_θ et $P_{\theta'}$ ne sont pas étrangères. Alors il n'existe pas d'estimateur T^* de $q(\theta)$ tel que T^* soit meilleur que tout autre estimateur T de $q(\theta)$.

Démonstration. Supposons qu'un tel estimateur T^* existe. Alors pour tout $\theta \in \Theta$, T^* est meilleur que l'estimateur constant $q(\theta)1_{\mathbb{X}}$ donc

$$E_\theta\left(\ell(T^*, q(\theta))\right) \leq E_\theta\left(\ell(q(\theta), q(\theta))\right) = 0,$$

ce qui implique $P_\theta(T^* = q(\theta)) = 1$ vu l'hypothèse sur ℓ. Puisque q n'est pas une fonction constante, il existe $\theta_1, \theta_2 \in \Theta$ tels que $q(\theta_1) \neq q(\theta_2)$. On a donc

$$P_{\theta_1}(T^* = q(\theta_1)) = 1, \quad P_{\theta_2}(T^* = q(\theta_2)) = 1, \quad \{T^* = q(\theta_1)\} \cap \{T^* = q(\theta_2)\} = \emptyset$$

autrement dit P_{θ_1} et P_{θ_2} sont étrangères. □

Remarque : Dans le modèle de Bernoulli, $\bar{X}_n$ est un estimateur sans biais de θ. On peut démontrer que $\bar{X}_n$ est meilleur, pour le risque quadratique, que tout autre estimateur sans biais de θ (en utilisant la borne de Cramer Rao 3.14).

4.1.1 Admissibilité, minimaxité

Bien qu'on ne puisse pas en général comparer deux estimateurs, on peut trouver des estimateurs tels qu'il n'en existe pas de meilleur. De tels estimateurs sont dits admissibles.

Definition 4.6 : Admissibilité

> Un estimateur T de $q(\theta)$ est dit *admissible* s'il n'existe pas d'estimateur qui soit strictement meilleur que lui.

La notion d'admissibilité est un peu passée de mode, trouver des estimateurs admissibles ne fournit pas de garantie sur les vitesses de convergence. Pour ce dernier point, on regarder plutôt les risques maximaux sur Θ,

$$\bar{R}_T = \sup_{\theta \in \Theta} R_T(\theta).$$

Dès lors, on peut toujours comparer deux estimateurs sur la base de leur risque maximaux (on définit alors une relation d'ordre total). Des estimateurs optimaux au sens de cette relation sont dit *minimax*.

Definition 4.7 : Minimaxité

> Un estimateur T de $q(\theta)$ est dit *minimax* (sur Θ) si
>
> $$\bar{R}_T = \sup_{\theta \in \Theta} R_T(\theta) \leq \sup_{\theta \in \Theta} R_S(\theta) = \bar{R}_S \quad \text{pour tout estimateur } S.$$

Prouver qu'un estimateur est minimax se fait généralement en deux temps : dans un premier temps on calcule $\bar{R}_T$. Dans un deuxième temps on minore le risque minimax $\inf_S \bar{R}_S$ en utilisant des techniques bayésiennes. Au passage, cela permet de déterminer la vitesse d'estimation minimax sur la classe θ, définie par

$$v(\Theta) := \inf_S \bar{R}_S,$$

qui correspond au plus petit risque maximal possible encouru par un estimateur sur cette classe. Dans le cas d'un modèle correspondant à un n-échantillon, la dépendance en n de cette vitesse est cruciale.

Exemple 4.8 *: Modèle Gaussien standard.* Dans le cadre du modèle Gaussien $(\mathcal{N}(\theta,1)^{\otimes n})_{\theta\in\mathbb{R}}$, l'estimateur $\bar{X}_n$ satisfait

$$\sup_{\theta\in\mathbb{R}} R_{\bar{X}_n}(\theta) = \frac{1}{n},$$

pour le risque quadratique. La vitesse minimax sur cette classe,

$$v_n = \inf_T \sup_{\theta\in\mathbb{R}} E_\theta(\theta - T(X))^2$$

vérifie donc $v_n \leq 1/n$. On prouvera par la suite que $1/n$ est précisément la vitesse minimax sur cette classe.

En général, admissibilité et minimaxité n'ont rien à voir.

Exemple 4.9 *: Minimax n'implique pas admissible.*

- Exemple idiot : Dans le modèle $(\mathcal{U}(]0,\theta[)^{\otimes n})_{\theta>0}$, l'estimateur $M_n = \max_{i=1,\dots,n} X_i$ a pour risque quadratique $2\theta^2/((n+1)(n+2))$, et donc pour risque max $+\infty$. On peut prouver que le risque quadratique minimax sur cette classe vaut $+\infty$, M_n est donc minimax. En revanche, $\frac{n+2}{n+1}M_n$ est strictement meilleur que M_n.
- Exemple moins idiot : Phénomène de Stein.
  Dans le modèle $(\mathcal{N}(\theta,\sigma^2 I_k))_{\theta\in\mathbb{R}^k}$, où $\sigma^2 > 0$ est connu. L'estimateur X est minimax (on le prouvera). En revanche il n'est pas admissible : l'estimateur de Stein

  $$\hat{\theta}_{JS} = (1 - \sigma^2\frac{k-2}{\|X\|^2})X$$

  est strictement meilleur dès lors que $k \geq 3$.

Exemple 4.10 *: Admissible n'implique pas minimax.* Dans le modèle $(\mathcal{B}(n,\theta))_{\theta\in]0,1[}$, on considère l'estimateur constant $T \equiv \frac{1}{2}$. Soit S un estimateur meilleur que T. On a alors

$$R_S\left(\frac{1}{2}\right) \leq R_T\left(\frac{1}{2}\right) = 0,$$

au sens du risque quadratique, ce dont on déduit $P_{1/2}(S = 1/2) = 1$, et donc $S \equiv \frac{1}{2}$ $P_{1/2}$ p.s.. Comme, pour tout $A \subset [\![0,n]\!]$, $P_{1/2}(A) > 0$, on en déduit que $S \equiv \frac{1}{2}$ sur $\mathcal{X} = [\![0,n]\!]$, et donc T est admissible. De manière plus générale, si les P_θ ne sont pas étrangères deux à deux, les estimateurs constants seront admissibles.

S n'est pas minimax : $\bar{R}_S = \frac{1}{4}$, et $\bar{R}_X = \frac{1}{4n} < \frac{1}{4}$ si $n \geq 2$.

4.2 Le point de vue bayésien

4.2.1 Approche informelle

L'approche fréquentiste (ou minimax) peut sembler pessimiste : la qualité d'un estimateur T sera jugée dans le pire des cas possible via $\bar{R}_T$. Par exemple, si le but est d'estimer la durée de vie moyenne $\theta > 0$ d'un français né au 18eme siècle, l'approche fréquentiste se basera sur $\sup_{\theta>0} R_T(\theta)$, c'est à dire en considérant les cas extrêmes où cett edurée de vie moyenne est proche de 0 ou supérieure à 100.

L'approche bayésienne consiste à pallier cet extrêmisme, en intégrant une connaissance *a priori* sur le paramètre à estimer. L'intégration d'une connaissance a priori peut se faire via une restriction arbitraire du champ des paramètres (par exemple $]40, 100[$ au lieu de $]0, +\infty[$ dans l'exemple de la durée de vie moyenne). Elle peut aussi se faire plus généralement en considérant qu'il y a des zones de Θ plus probables que d'autres, ce que l'on peut modéliser par une loi *a priori* π sur Θ. Dès lors, plutôt que considérer un risque max $\bar{R}_T$ pour mesurer la qualité d'un estimateur, on regardera son *risque intégré* (ou moyen) par rapport à cette loi a priori,

$$\rho_T(\pi) = \int_\Theta R_T(\theta)\pi(d\theta),$$

pour peu que cette intégrale ait du sens.

Conceptuellement, l'approche bayésienne change la nature des observations. Dans le cadre classique on observe $X_1, \ldots, X_n$ i.i.d. suivant P_θ, où θ est fixe et inconnu. Dans le monde bayésien, la nature tire un $\tilde{\theta}$ au hasard via la loi π, puis $\tilde{X}_1, \ldots, \tilde{X}_n$ tels que $\tilde{X} \mid \tilde{\theta}, \ldots \tilde{X}_n \mid \tilde{\theta}$ sont i.i.d. suivant $P_{\tilde{\theta}}$, si cela a du sens (on le montrera formellement juste après). Les observations $\tilde{X}_1, \ldots, \tilde{X}_n$ ne sont alors plus i.i.d. : elles le sont conditionnellent à $\tilde{\theta}$, mais si on relâche ce conditionnement elles suivent une loi de mélange (par rapport à π de lois indépendantes). De fait, elles dépendent les unes des autres via leur paramètre $\tilde{\theta}$ commun qui est maintenant aussi aléatoire.

L'inférence bayésienne consiste alors à renverser le conditionnement : on connaît π (loi a priori) et le modèle, et on cherche à retrouver le $\tilde{\theta}$ qui a été tiré. Pour cela, on va chercher à caractériser la loi $\tilde{\theta} \mid \tilde{X}$, la loi du paramètre sachant les observations, qu'on appelera *loi a posteriori*. Par exemple, si on a mis une loi a priori uniforme sur $]40, 100[$ pour la durée de vie moyenne, on s'attend à ce que $\tilde{\theta} \mid \tilde{X}$ mette plus de masse dans la zone de $]40, 100[$ sur une zone autour de la durée de vie moyenne observée.

Pour distinguer du cas fréquentiste où θ et fixe et $X_1, \ldots, X_n$ sont i.i.d. suivant P_θ, on mettra des tilde partout : $\tilde{\theta}$ est maintenant une variable aléatoire dans l'espace des paramètres, et $\tilde{X}_1, \tilde{X}_n$ ne sont plus i.i.d. de loi P_θ (les $\tilde{X}_i \mid \tilde{\theta}$ oui). Il va donc falloir formaliser un peu le concept de "loi conditionnelle".

4.2.2 Formalisme adapté : lois conditionnelles

Commençons par un rappel sur l'espérance conditionnelle.

Definition 4.11 : Espérance conditionnelle

Soit Y une quantité aléatoire à valeurs dans $(E, \mathcal{E})$, sur $(\Omega, \mathcal{F}, \mathbb{P})$, et X une variable aléatoire positive (resp. L_1) sur ce même espace. Il existe alors $h : E \to [0, +\infty]$ (resp. $\mathbb{R}$) telle que, pour toute variable aléatoire Z $\sigma(Y)$-mesurable positive (resp. bornée),

$$\mathbb{E}(ZX) = \mathbb{E}(Zh(Y)). \tag{4.1}$$

$h(Y)$ est alors appelée espérance conditionnelle de X sachant Y, notée $\mathbb{E}(X \mid Y)$, et est unique $\mathbb{P}$ p.s..

On ne listera pas ici les propriétés classiques de l'espérance conditionnelle, pour lesquelles on pourra se référer au polycopié de F. Le Gall par exemple et que l'on

supposera connues. Pour une quantité aléatoire X quelconque, l'idée est maintenant que si, pour toute $\phi : \mathcal{X} \to \mathbb{R}$ mesurable positive ou L_1, on peut écrire $\mathbb{E}(\phi(X) \mid Y)$ comme l'intégrale de ϕ par rapport à une loi dépendant de Y, on pourra alors définir la loi de X sachant Y. Il faut d'abord formaliser cette histoire de loi dépendant de Y.

Definition 4.12 : Probabilité/noyau de transition

Soient $(\mathcal{X}, \mathcal{A})$ et $(\mathcal{Y}, \mathcal{B})$ deux espace mesurés. On appelle *probabilité de transition* de $(\mathcal{X}, \mathcal{A})$ vers $(\mathcal{Y}, \mathcal{B})$ une application

$$\nu : \mathcal{Y} \times \mathcal{A} \to [0, 1]$$

qui vérifie les propriétés suivantes :

1. Pour tout $y \in \mathcal{Y}$, $\nu(y, .)$ est une probabilité sur $(\mathcal{X}, \mathcal{A})$.
2. Pour tout $A \in \mathcal{A}$, $\nu(., A)$ est $\mathcal{B}$-mesurable.

Remarque : Vous avez probablement déjà rencontrés de tels objets en cours de proba, notamment via les chaînes de Markov. Dans ce cas, la probabilité de transition entre deux états est donnée par le noyau de la chaîne de Markov. D'où l'appellation alternative *noyau de transition*.

Exemple 4.13 *: Cas à densité.* Soit μ une mesure positive σ-finie sur $(\mathcal{X}, \mathcal{A})$ et soit $g : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^+$ une application $\mathcal{B} \otimes \mathcal{A}$-mesurable telle que

$$\forall y \in \mathcal{Y} \quad \int_{\mathcal{X}} g(y, x)\mu(dx) = 1.$$

Alors

$$\nu(y, A) = \int_A g(y, x)\mu(dx)$$

définit une probabilité de transition de $\mathcal{Y}$ dans $\mathcal{X}$. La propriété (2) de la définition découle en particulier du théorème de Fubini.

Le but maintenant va être d'exprimer $\mathbb{E}(\phi(X) \mid Y)$ comme l'intégrale de ϕ par une probabiltié de transition. Les propriétés ci-dessous vont être utile, notamment pour la $\sigma(Y)$-mesurabilité.

Proposition 4.14

1. *Si h est une fonction positive (ou bornée) sur $(\mathcal{X}, \mathcal{A})$ alors*

$$\phi(y) = \int_{\mathbb{X}} \nu(y, dx)h(x), \quad y \in \mathcal{Y}$$

est une fonction mesurable positive (resp. bornée) sur $\mathcal{Y}$.

2. *Si λ est une probabilité sur $(\mathcal{Y}, \mathcal{B})$ alors*

$$A \mapsto \int_{\mathbb{Y}} \lambda(dy)\nu(y, A)$$

est une probabilité sur $(\mathcal{X}, \mathcal{A})$.

Démonstration. 1. On le vérifie pour h étagée...

2. Définition d'une mesure de proba + Fubini

□

On peut maintenant définir proprement la notion de loi de X sachant Y.

DEFINITION 4.15

Soient X et Y deux variables aléatoires à valeurs respectivement dans $(\mathcal{X}, \mathcal{A})$ et dans $(\mathcal{Y}, \mathcal{B})$. On appelle version de la loi conditionnelle de X sachant Y toute probabilité de transition ν de $\mathcal{Y}$ dans $\mathcal{X}$ telle que, pour toute fonction h mesurable positive sur $(\mathcal{X}, \mathcal{A})$, on ait

$$E\left(h(X) \mid Y\right) = \int_{\mathcal{X}} \nu(Y, dx) h(x) \quad \text{p.s..}$$

Remarque : $\nu(y, dx)$ est parfois noté de manière abusive $P(X \in dx \mid Y = y)$.

Il reste à montrer que de tels noyaux existent et sont uniques p.s.. On commence par le dernier point.

PROPOSITION 4.16

Si $\mathcal{X}$ est un espace métrique séparable muni de sa tribu borélienne $\mathcal{A}$ et si ν et ν' sont deux versions de la loi conditionnelle de X sachant Y, alors $\nu(Y, \cdot) = \nu'(Y, \cdot)$ presque sûrement.

Démonstration. Pour $x \in \mathcal{X}$ et $r \geq 0$ on note $\mathrm{B}(x, r)$ la boule de centre x et de rayon r. Soit S un ensemble au plus dénombrable et dense dans $\mathcal{X}$. Soit $\mathcal{C}$ la classe des ensembles de la forme

$$\mathcal{X} \setminus \left(\bigcup_{i=1}^{N} \mathrm{B}(x_i, r_i) \right)$$

où $N \in \mathbb{N}$, $(x_1, \ldots, x_N) \in S^N$, $r_1, \ldots, r_n \in \mathbb{Q}_+$. $\mathcal{C}$ est dénombrable et engendre $\mathcal{A}$ puisque tout ouvert U de $\mathcal{X}$ peut s'écrire comme la réunion d'une famille dénombrable de boules de la forme $\mathrm{B}(x, r)$ avec $x \in S$ et $r \in \mathbb{Q}_+$. Posons

$$\Omega_0 = \left\{ \nu(Y, A) = \nu'(Y, A) \quad \forall A \in \mathcal{C} \right\}.$$

Pour tout $\omega \in \Omega_0$, $\nu(Y(\omega, \cdot)) = \nu'(Y(\omega), \cdot)$ car $\mathcal{C}$ est stable par intersection finie, contient $\mathcal{X}$ et engendre $\mathcal{A}$. De plus $P(\Omega_0) = 1$ (car $\Omega_0 = \bigcap_{A \in \mathcal{C}} \{\nu(Y, A) = \nu'(Y, A)\}$, intersection dénombrable d'évènements de probabilité 1 selon la Définition 4.15, avec $h = \mathbb{1}_A$). □

THÉORÈME 4.17

Si $\mathcal{X}$ est un espace métrique séparable complet (on dit que X est un espace Polonais) muni de sa tribu borélienne $\mathcal{A}$, alors il existe une version de la loi conditionnelle de X sachant Y.

Par exemple $\mathbb{R}^d$, $C([0, 1], \mathbb{R})$ muni de la norme du sup, $\mathbb{R}^{\mathbb{N}}$, sont polonais. On admet le résultat suivant.

LEMME 4.18

Soit $\mathcal{X}$ un espace Polonais muni de sa tribu borélienne $\mathcal{A}$. Il existe $\phi : \mathcal{X} \to [0,1]$ mesurable injective telle que $\phi(\mathcal{X})$ est un borélien.

Preuve du théorème. Il suffit de considérer le cas où $\mathcal{X} = [0,1]$ (utiliser le lemme). Pour chaque $t \in [0,1] \cap \mathbb{Q}$ il existe une fonction $F_t : \mathcal{Y} \mapsto \mathbb{R}$ $\mathcal{B}$-mesurable telle que

$$F_t(Y) = E\left(\mathbb{1}_{X \leq t} \mid Y\right) \quad \text{p.s..}$$

Notons

$$B = \{y \in \mathcal{Y} \mid F_s(y) \leq F_t(y) \quad \text{pour tous } s,t \in \mathbb{Q} \text{ tels que } 0 \leq s \leq t \leq 1\}$$

On a $B \in \mathcal{B}$, et $\mathbb{P}(Y \in B) = 1$) (intersection dénombrable d'ensembles de probabilité 1). Pour $y \in B$ et $t \in [0,1[$ on pose

$$G_1(y) = 1, \quad G_t(y) = \inf_{s \in \mathbb{Q}, s > t} F_s(y).$$

Pour chaque $y \in B$, $t \mapsto G_t(y)$ est une fonction de répartition. Il existe donc une probabilité $\nu(y, dx)$ sur $[0,1]$ telle que $\nu(y,[0,t]) = F_t(y)$ pour tout $t \in [0,1]$. Pour $y \in \mathcal{Y} \setminus B$, on pose $\nu(y,dx) = \delta_0(dx)$. Soit $\mathcal{C}$ l'ensemble des boréliens A tels que $y \mapsto \nu(y,A)$ est mesurable et $E\left(\mathbb{1}_A(X) \mid Y\right) = \nu(Y,A)$ p.s.. $\mathcal{C}$ est une classe monotone et contient la classe des intervalles de la forme $[0,t]$, pour $t \in [0,1] \cap \mathbb{Q}$, qui est stable par intersection finie et engendre la tribu borélienne de $[0,1]$. Par le théorème de la classe monotone $\mathcal{C}$ est la tribu borélienne de $[0,1]$. □

4.2.3 Risque intégré, loi a posteriori

Revenons maintenant à notre problème de modélisation bayésienne. Pour que tout ce qui a été décrit plus haut ait un sens, dorénavant on supposera tacitement que Θ est équipée d'une tribu $\mathcal{T}$ telle que $(\theta, A) \mapsto P_\theta(A)$ soit un noyau de transition. Dans toute la suite on se donnera aussi $(\tilde{\theta}, \tilde{X})$ de loi donnée par $\pi(d\theta)P_\theta(dx)$, (c'est à dire $\tilde{\theta} \sim \pi$, et $\tilde{X} \mid \tilde{\theta} \sim P_{\tilde{\theta}}$ p.s.). Sous ces conditions, on peut définir un risque intégré.

PROPOSITION 4.19

On suppose que $P_\theta(dx)$ est une probabilité de transition de $(\Theta, \mathcal{T})$ dans $(\mathcal{X}, \mathcal{A})$. On suppose aussi que $q : \Theta \to \mathbb{R}^k$ est borélienne et que $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k$ est borélienne. Alors pour tout estimateur T de $q(\theta)$, la fonction $R_T : \Theta \to [0,\infty]$ est borélienne.

En particulier, le risque intégré

$$\rho_T(\pi) = \int_\Theta R_T(\theta)\pi(d\theta)$$

est bien défini et à valeurs dans $[0,+\infty]$.

Démonstration. L'ensemble des $\Gamma \in \mathcal{T} \otimes \mathcal{A}$ tels que l'application

$$\theta \mapsto E_\theta\left(\mathbb{1}_\Gamma(\theta, X)\right)$$

soit $\mathcal{T}$-mesurable est une classe monotone. Il contient de plus la classe des pavés de la forme $B \times A$ avec $B \in \mathcal{T}$ et $A \in \mathcal{A}$, stable par intersection finie et qui contient $\Theta \times \mathcal{X}$. Le théorème de la classe monotone implique donc que pour tout $\Gamma \in \mathcal{T} \otimes \mathcal{A}$, $\theta \mapsto E_\theta\left(\mathbb{1}_\Gamma(\theta, X)\right)$ est $\mathcal{T}$-mesurable. On en déduit que, pour toute variable aléatoire $Z : \Theta \times \mathcal{X} \to \mathbb{R}^+$ étagée, l'application $\theta \mapsto E_\theta\left(Z(\theta, X)\right)$ est $\mathcal{T}$-mesurable. Pour toute variable aléatoire $Z : \Theta \times \mathcal{X} \to \mathbb{R}^+$ mesurable, il existe une suite (Z_k) de fonctions étagées mesurables telle que $Z = \lim_k \uparrow Z_k$, donc pour tout $\theta \in \Theta$ on a $E_\theta\left(Z(\theta, X)\right) = \lim_k \uparrow E_\theta\left(Z_k(\theta, X)\right)$, ce qui implique que $\theta \mapsto E_\theta\left(Z(\theta, X)\right)$ est $\mathcal{T}$-mesurable.

On conclut en remarquant que $R_T(\theta) = E_\theta\left(\ell(T, q(\theta))\right.$ et que $(\theta, x) \mapsto \ell(T(x), q(\theta))$ est $\mathcal{T} \otimes \mathcal{A}$-mesurable, □

On peut alors définir proprement le *risque bayésien*, et ce qu'est un *estimateur bayésien*.

Definition 4.20 : Risque et estimateurs bayésiens

Sous les hypothèses précédentes,

$$\rho(\pi) = \inf_T \rho_T(\pi)$$

est appelé *risque bayésien*. Un estimateur T^* tel que $\rho_{T^*}(\pi) = \rho(\pi)$ est appelé *estimateur bayésien*.

On remarque que les risques et estimateurs bayésiens dépendent fortement de la fonction de perte ℓ, ce que la notation ne traduit pas. Le but maintenant va être d'essayer de construire des estimateurs bayésiens de manière générale, pour cela on aura besoin de déterminer la loi de $\tilde{\theta} \mid \tilde{X}$, aussi appelée loi a posteriori.

Loi a posteriori

Definition 4.21 : Loi a posteriori

Sous les hypothèses faites en début de section, la loi de $\tilde{\theta} \mid \tilde{X}$ est bien définie $P_{\tilde{X}}$-p.s., et est appelée loi a posteriori. On la notera par convention $Q_{\tilde{X}}(d\theta)$.

Le calcul effectif de la loi a posteriori peut se faire de deux manières, suivant que le modèle est dominé ou non. Commençons par le cas dominé.

Théorème 4.22 : Formule de Bayes

Supposons que le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ est dominé par μ (σ-finie). Supposons qu'il existe $(\theta, x) \mapsto p_\theta(x)$ bi-mesurable telle que pour tout $\theta \in \Theta$ $p_\theta = \frac{dP_\theta}{d\mu}$ μ-presque partout. Alors $\int_\Theta p_\tau(\tilde{X})\pi(d\tau) > 0$ $P_{\tilde{X}}$-p.s. et la loi a posteriori s'écrit

$$Q_x(d\theta) = \frac{p_\theta(x)}{\int_\Theta p_\tau(x)\pi(d\tau)}\pi(d\theta),$$

pour tout $x \in \mathcal{X}$ tel que $\int_\Theta p_\tau(x)\pi(d\tau) > 0$. Autrement dit pour toute fonction $h : \Theta \to \mathbb{R}^+$ borélienne

$$\mathbb{E}\left(h(\tilde{\theta}) \mid \tilde{X}\right) = \frac{\int_\Theta h(\theta)p_\theta(\tilde{X})\pi(d\theta)}{\int_\Theta p_\theta(\tilde{X})\pi(d\theta)} \quad P_{\tilde{X}}\text{-p.s..}$$

On peut aussi trouver la notation $p(x|\theta)$ pour les densités du modèle.

Démonstration. On commence par remarquer que, si $\phi : \mathcal{X} \to \mathbb{R}^+$ est mesurable,

$$\begin{aligned}\mathbb{E}(\phi(\tilde{X})) &= \int_\Theta \int_\mathcal{X} \phi(x)p_\tau(x)\pi(d\tau)\mu(dx)\\ &= \int_\mathcal{X} \left(\int_\Theta p_\tau(x)\pi(d\tau)\right)\phi(x)\mu(dx),\end{aligned}$$

en utilisant Fubini. Notons $q : x \mapsto (\int_\Theta p_\tau(x)\pi(d\tau))$ (qui est bien mesurable). On remarque alors que $\tilde{X}$ a pour densité q par rapport à μ. Comme $\int_\mathcal{X} q(x)\mu(dx) < +\infty$, on a alors $q(x) < +\infty$ μ-p.s.. Par ailleurs

$$\mathbb{P}\left(q(\tilde{X}) = 0\right) = \int_\mathcal{X} q(x)\mathbb{1}_{q(x)=0}\mu(dx) = 0,$$

d'après ce qui précède. Donc $q(\tilde{X}) > 0$ $P_{\tilde{X}}$-p.s..

Pour trouver la loi de $\tilde{\theta} \mid \tilde{X}$, on se donne $\phi : \mathcal{X} \to \mathbb{R}^+$ et $h : \Theta \to \mathbb{R}^+$ mesurables. On a alors

$$\begin{aligned}\mathbb{E}(\phi(\tilde{X})h(\tilde{\theta})) &= \int_\Theta \pi(d\theta)\left(\int_\mathcal{X} h(\theta)\phi(x)p_\theta(x)\mu(dx)\right)\\ &= \int_{\{q(x)>0\}} \phi(x)q(x)\left(\int_\Theta h(\theta)\frac{p_\theta(x)}{q(x)}\pi(d\theta)\right)\mu(dx)\\ &= \mathbb{E}\left(\phi(\tilde{X})\frac{1}{q(\tilde{X})}\int_\Theta h(\theta)p_\theta(\tilde{X})\pi(d\theta)\right).\end{aligned}$$

On peut vérifier que $(x, A) \mapsto \frac{1}{q(x)}\int_A p_\theta(x)\pi(d\theta)$ est bien un noyau de transition de $(\mathcal{X}, \mathcal{A})$ vers $(\Theta, \mathcal{T})$. Et on a alors

$$\tilde{\theta} \mid \tilde{X} \sim Q_{\tilde{X}}(d\theta) = \frac{p_\theta(\tilde{X})}{\int_\Theta p_\tau(\tilde{X})\pi(d\tau)}\pi(d\theta),$$

$P_{\tilde{X}}$-p.s.. □

On peut remarquer que si le modèle est dominé, alors la loi de $\tilde{\theta} \mid \tilde{X}$ sera elle même dominée par la loi a priori π. En pratique, on ne calcule $q(\tilde{X})$ qu'en dernier

ressort : si l'expression $p_\theta(\tilde{X})\pi(d\theta)$ fait ressortir une dépendance en θ familière, on déduira $(q(\tilde{X}))$ de cela.

Exemple 4.23 *: Binomiale/Beta.* On se place dans le modèle $(\mathcal{B}(n,\theta))_{\theta\in]0,1[}$, avec pour loi a priori $\pi \sim \beta(a,b)$, c'est à dire à densité sur $]0,1[$

$$\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1},$$

avec $a,b>0$, et $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Ce modèle est dominé (par la mesure de comptage sur $[\![0,n]\!]$), on peut prendre comme densité bi-mesurable

$$p_\theta(x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

La formule de Bayes donne alors, pour un x tel que $P_{\tilde{X}}(\{x\})>0$,

$$Q_x(d\theta)=\theta^{x-1}(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1}\times N(x),$$

où N ne dépend pas de θ. Comme Q_x définit une loi de probabilité, on peut directement reconnaitre l'expression d'une loi $\beta(a+x,b+n-x)$, et éventuellement en déduire $N(x)=\frac{1}{B(a+x,b+n-x}$ (mais ce n'est pas vraiment intéressant). On en déduit

$$\tilde{\theta}\mid\tilde{X}\sim\beta(a+\tilde{X},b+n-\tilde{X}),$$

$P_{\tilde{X}}$ p.s..

Remarque : De tels cas de couples (loi a priori, loi du modèle) où la loi a posteriori appartient à la même famille que la loi a priori sont appelés *lois conjuguées*, et définissent le cadre "agréable" pour le bayésien (sinon il faut calculer $q(x)$).

Si le modèle n'est pas dominé, il faut se référer à l'esprit de la preuve, où on essaye de caractériser les espérances conditionnelles comme des intégrations par rapport à une loi. Pour ce faire, on partira de $\mathbb{E}(\phi(\tilde{X})h(\tilde{\theta})$, que l'on essayera de mettre sous la forme $\int_{\mathcal{X}}P_{\tilde{X}}(dx)\int_\Theta\nu(\tilde{X},d\theta)h(\theta)$, pour un noyau de transition ν. Les choix acceptables pour ϕ et h sont : (mesurables) positives, L_1, de type $\mathbb{1}_A$, $\mathbb{1}_B$, ou encore $\mathcal{C}^\infty$ à support compact (si on veut jouer avec des dérivations).

Exemple 4.24 *: Calcul de la loi a posteriori dans un modèle non dominé.* On observe $X=\theta\varepsilon$ où $\theta\in]0,+\infty[$ et ε est une variable aléatoire à valeurs dans $\{1,2\}$ telle que $P(\varepsilon=1)=P(\varepsilon=2)=1/2$. Le modèle correspondant s'écrit

$$\mathcal{X}=]0,+\infty[,\quad \mathcal{A}=\mathcal{B}(]0,+\infty[),\quad P_\theta=\frac{1}{2}\delta_\theta+\frac{1}{2}\delta_{2\theta},\ \theta\in]0,+\infty[.$$

On choisit comme loi a priori $\pi(d\theta)=f(\theta)d\theta$ où f est une densité de probabilité sur $]0,+\infty[$. On veut calculer une version de la loi conditionnelle de $\tilde{\theta}$ sachant $\tilde{X}$. Pour $h:]0,+\infty[\to\mathbb{R}^+$ et $g:]0,+\infty[\to\mathbb{R}^+$ deux fonctions mesurables, on a

$$\begin{aligned}
\mathbb{E}\left[h(\tilde{\theta})g(\tilde{X})\right] &= \int_0^\infty f(\theta)h(\theta)\frac{1}{2}\left(g(\theta)+g(2\theta)\right)d\theta\\
&= \frac{1}{2}\int_0^\infty f(\theta)h(\theta)g(\theta)d\theta+\frac{1}{2}\int_0^\infty f(\theta)h(\theta)g(2\theta)d\theta\\
&= \frac{1}{2}\int_0^\infty f(\theta)h(\theta)g(\theta)d\theta+\frac{1}{4}\int_0^\infty f(\theta/2)h(\theta/2)g(\theta)d\theta\\
&= \int_0^\infty\left(\frac{1}{2}f(\theta)h(\theta)+\frac{1}{4}f(\theta/2)h(\theta/2)\right)g(\theta)d\theta
\end{aligned}$$

Cela montre que $\tilde{X}$ admet comme densité $x \mapsto \frac{1}{2}f(x) + \frac{1}{4}f(x/2)$ (prendre $h = 1$) et que

$$\mathbb{E}\left[h(\tilde{\theta})g(\tilde{X})\right] = \mathbb{E}\left(\frac{\frac{1}{2}f(\tilde{X})h(\tilde{X}) + \frac{1}{4}f(\tilde{X}/2)h(\tilde{X}/2)}{\frac{1}{2}f(\tilde{X}) + \frac{1}{4}f(\tilde{X}/2)}g(\tilde{X})\right)$$

Donc, puisque g est arbitraire,

$$\mathbb{E}\left(h(\tilde{\theta}) \mid \tilde{X}\right) = \frac{f(\tilde{X})h(\tilde{X}) + \frac{1}{2}f(\tilde{X}/2)h(\tilde{X}/2)}{f(\tilde{X}) + \frac{1}{2}f(\tilde{X}/2)} \quad P_{\tilde{X}}\text{-p.s..}$$

La loi a posteriori est donc donnée par

$$Q_x(d\theta) = \frac{f(x)}{f(x) + \frac{1}{2}f(x/2)}\delta_x + \frac{\frac{1}{2}f(x/2)}{f(x) + \frac{1}{2}f(x/2)}\delta_{x/2}$$

pour tout $x \in]0, +\infty[$ tels que $f(x) + \frac{1}{2}f(x/2) > 0$ (qui correspond à un événement de probabilité 1 pour la loi de $\tilde{X}$).

Lois a posteriori, risque, et stats exhaustives

Les statistiques exhaustives, on l'a déjà vu permettent de "réduire le modèle" dans le cadre des moindres carrés : la Proposition 3.20 montre en effet que les estiamateurs optimaux au sens des moindres carrés sont à chercher parmi les fonctions de $S(X)$, ou S est exhaustive. Dans un cadre i.i.d., si par exemple $S(X) = \bar{X}_n$, cela permet de passer d'un modèle n-uplet à un modèle à une observation (et de s'épargner du calcul). On peut généraliser cela à n'importe quelle fonction de perte convexe.

Proposition 4.25

Soit $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ un modèle et S une statistique exhaustive pour θ. On se donne une fonction de perte $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^+$ convexe en la deuxième variable, et, pour un estimateur T de $q(\theta) \in \mathbb{R}^k$ la fonction de risque associée $R_T(\theta) = E_\theta(\ell(q(\theta), T(X)))$.

Si T est un estimateur de $q(\theta)$, alors il existe un estimateur $\tilde{T}(S(X))$ meilleur que T.

Démonstration. Comme dans la Proposition 3.20, $\tilde{T}(S(X)) = E_\theta(T(X) \mid S(X))$ est meilleur que T. □

Attention, la fonction de perte ne sera pas toujours convexe, par exemple pour les problèmes de tests où $\ell(q(\theta), T(X)) = \mathbb{1}_{T(X)\neq q(\theta)}$. Le résultat général va rester vrai si on considère les estimateurs randomisés (on l'a déjà plus ou moins montré avec le test de Neyman-Pearson randomisé).

Dans le cadre bayésien, peu importe la fonction de perte, la loi de $\tilde{\theta} \mid \tilde{X}$ sera toujours égale à celle de $\tilde{\theta} \mid S(\tilde{X})$, justifiant la réduction de modèle dans tous les cas de figure.

Proposition 4.26

Si $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ est un modèle, et $T : \mathcal{X} \to \mathcal{Y}$ est exhaustive pour ce modèle. Alors, pour toute loi a priori π sur Θ,

$$\tilde{\theta}|\tilde{X} \sim \tilde{\theta}|T(\tilde{X}).$$

Démonstration. Dans un cadre dominé par μ on peut trouver une fonction $p_\theta(x)$ bi-mesurable telle que $p_\theta = \frac{dP_\theta}{d\mu}$ μ-presque partout et $p_\theta(x)$ se met sous la forme

$$p_\theta(x) = g(x)h_\theta(T(x)),$$

où g est une fonction mesurable positive, et $h_\cdot(.)$ est bi-mesurable positive. Dans ce cas, le théorème de Bayes nous indique que la loi a posteriori est donnée par

$$Q_x(d\theta) = \frac{g(x)h_\theta(T(x))}{\int_\Theta g(x)h_\tau(T(x))\pi(d\tau)}\pi(d\theta) = \frac{h_\theta(T(x))}{\int_\Theta h_\tau(T(x))\pi(d\tau)}\pi(d\theta),$$

pour $x \in \mathcal{X}$ tel que $g(x)\int_\Theta h_\tau(x)\pi(d\tau) > 0$. D'un autre côté, si $f : \Theta \times \mathcal{Y} \to \mathbb{R}^+$ est mesurable, on peut écrire

$$\begin{aligned}\mathbb{E}(f(\tilde{\theta}, T(\tilde{X})) &= \mathbb{E}\left(\mathbb{E}\left(f(\tilde{\theta}, T(\tilde{X})) \mid \tilde{X}\right)\right)\\ &= \mathbb{E}\left(\int_\Theta \frac{h_\theta(T(\tilde{X})f(\theta, T(\tilde{X}))}{\int_\Theta h_\tau(T(\tilde{X}))\pi(d\tau)}\pi(d\theta)\right)\end{aligned}$$

Donc $\tilde{\theta} \mid T(\tilde{X}) \sim Q_{\tilde{X}}(d\theta)$ p.p..

Dans le cadre général, par définition de l'exhaustivité on a, pour toute fonction $\phi : \mathcal{X} \to \mathbb{R}^+$,

$$\mathbb{E}(\phi(\tilde{X}) \mid (T(\tilde{X}), \tilde{\theta})) = g_\phi(\tilde{X}) = \mathbb{E}(\phi(\tilde{X}) \mid T(\tilde{X})), \tag{4.2}$$

où $g_\phi : \mathcal{X} \to \mathbb{R}^+$ est mesurable. Sachant cela, si $f : \Theta \to \mathbb{R}^+$ et $h : \mathcal{X} \to \mathbb{R}^+$ sont deux fonctions mesurables, on a

$$\begin{aligned}\mathbb{E}(f(\tilde{\theta})h(\tilde{X})) &= \mathbb{E}\left[\mathbb{E}(h(\tilde{X}) \mid \tilde{\theta})f(\tilde{\theta})\right]\\ &= \mathbb{E}\left[\mathbb{E}(\mathbb{E}(h(\tilde{X}) \mid (\tilde{\theta}, T(\tilde{X})) \mid \tilde{\theta})f(\tilde{\theta})\right]\\ &= \mathbb{E}\left[\mathbb{E}(g_h(T(\tilde{X})) \mid \tilde{\theta})f(\tilde{\theta})\right]\\ &= \mathbb{E}(f(\tilde{\theta})g_h(T(\tilde{X}))\\ &= \mathbb{E}(\mathbb{E}(f(\tilde{\theta}) \mid T(\tilde{X}))g_h(T(\tilde{X}))\\ &= \mathbb{E}(\mathbb{E}(f(\tilde{\theta}) \mid T(\tilde{X}))h(\tilde{X})),\end{aligned}$$

d'après (4.2), avec $g_h(T(\tilde{X})) = \mathbb{E}(h(\tilde{X}) \mid T(\tilde{X}))$. On en déduit que $\mathbb{E}(f(\tilde{\theta}) \mid T(\tilde{X})) = \mathbb{E}(f(\tilde{\theta}) \mid \tilde{X})$, et donc que $\tilde{\theta} \mid \tilde{X} \sim \tilde{\theta} \mid T(\tilde{X})$ $\mathbb{P}$-p.s.. □

Remarque : On a aussi l'implication inverse : on peut montrer que si $\tilde{\theta} \mid \tilde{X} \sim \tilde{\theta} \sim T(\tilde{X})$, pour toute loi a priori sur Θ, alors T est exhaustive. Voyons sur un exemple l'utilité de l'exhaustivité.

Exemple 4.27 *: Echantillon Gaussien dans $\mathbb{R}^k$.*

On observe n vecteurs gaussiens à valeurs dans $\mathbb{R}^k$, i.i.d., tous de loi $\mathcal{N}(\theta, \sigma_0^2 I_k)$ où $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ est inconnu ($\sigma_0^2 > 0$ est connu). On prend comme loi à priori $\pi_v = \mathcal{N}(0_k, vI_k)$ où $v > 0$. On remarque qu'on peut écrire $\tilde{X}_i = \tilde{\theta} + \varepsilon_i$, $1 \leq i \leq n$, où les ε_i sont des vecteurs aléatoires i.i.d. de loi $\mathcal{N}(0_k, \sigma_0^2 I_k)$ et $\tilde{\theta}, \varepsilon_1, \ldots, \varepsilon_n$ sont indépendants.

On veut déterminer la loi a posteriori $\tilde{\theta} \mid \tilde{X}_1, \ldots, \tilde{X}_n$.

Approche brutale : Le modèle étant dominé par $\mathcal{L}_{kn}$, on peut passer par la formule de Bayes.

On peut aussi passer par une approche "vecteur Gaussien" : le vecteur

$$\left(\tilde{\theta}_1, \ldots, \tilde{\theta}_k, \tilde{X}_1^1, \ldots, \tilde{X}_1^k, \ldots, \tilde{X}_n^1, \ldots, \tilde{X}_n^k\right)$$

étant un vecteur gaussien, on sait que $\tilde{\theta} - \mathbb{E}(\tilde{\theta} \mid \tilde{X}_{1:n})$ est un vecteur Gaussien indépendant de $\tilde{X}_{1:n}$. Par ailleurs, on sait aussi que $\mathbb{E}(\tilde{\theta} \mid \tilde{X}_{1:n})$ est la projection de $\tilde{\theta}$ sur l'espace engendré par $\tilde{X}_{1:n}$ au sens $L_2(\mathbb{P})$.

Pour $\ell \in \{1, \ldots, k\}$ notons $\bar{\theta}_\ell$ la projection orthogonale dans $L^2(\mathbb{P})$ de $\tilde{\theta}_\ell$ sur le sous-espace vectoriel engendré par 1 et les $\tilde{X}_i^p$, $1 \leq i \leq n$, $1 \leq p \leq k$. $\bar{\theta}_\ell$ est de la forme

$$\bar{\theta}_\ell = b_\ell + \sum_{i=1}^{n} \sum_{k=1}^{p} B_i(\ell, p) \tilde{X}_i^p \quad \text{avec } b_\ell \in \mathbb{R},\ B_i(\ell, p) \in \mathbb{R}$$

et est caractérisé par

$$\mathbb{E}\left(\tilde{\theta}_\ell - \bar{\theta}_\ell\right) = 0, \quad \mathbb{E}\left((\tilde{\theta}_\ell - \bar{\theta}_\ell)\tilde{X}_i^p\right) = 0,\ 1 \leq i \leq n,\ 1 \leq p \leq k.$$

$\tilde{\theta} - \bar{\theta}$ étant un vecteur Gaussien indépendant de $(\tilde{X}_1, \ldots, \tilde{X}_n)$, on en déduit que $\tilde{\theta} \mid \tilde{X}_1, \ldots, \tilde{X}_n \sim \mathcal{N}\left(\bar{\theta}, K\right)$ où K est la matrice de covariance de $\tilde{\theta} - \bar{\theta}$.

Calculons $\mathbf{b} \in \mathbb{R}^k$ et les matrices B_i. On peut écrire

$$\bar{\theta} = \mathbf{b} + \sum_{i=1}^{n} B_i \tilde{X}_i = \mathbf{b} + \sum_{i=1}^{n} B_i \tilde{\theta} + \sum_{i=1}^{n} B_i \varepsilon_i,$$

et on a

$$\mathbb{E}\left(\tilde{\theta} - \bar{\theta}\right) = 0, \quad \mathbb{E}\left((\tilde{\theta} - \bar{\theta})\tilde{X}_i^T\right) = 0,\ 1 \leq i \leq n.$$

Tout d'abord, comme $\mathbb{E}(\tilde{\theta}) = 0_k$, on en déduit que $\mathbf{b} = 0_k$. Il vient

$$\tilde{\theta} - \bar{\theta} = \left(I_k - \sum_i B_i\right) \tilde{\theta} - \sum_i B_i \varepsilon_i.$$

Donc, pour $i \in \{1, \ldots, n\}$,

$$\left(\tilde{\theta} - \bar{\theta}\right) \tilde{X}_i^T = \left(\tilde{\theta} - \bar{\theta}\left(\tilde{\theta}^T + \varepsilon_i^T\right)\right) = \left(\left(I_k - \sum_j B_j\right) \tilde{\theta} - \sum_j B_j \varepsilon_j\right) \left(\tilde{\theta}^T + \varepsilon_i^T\right).$$

Comme $\mathbb{E}\left(\tilde{\theta}\varepsilon_j^T\right) = 0$ pour tout j et $\mathbb{E}\left(\varepsilon_i \varepsilon_j^T\right) = 0$ pour tous $j \neq i$, on obtient

$$\mathbb{E}\left(\ \left(\tilde{\theta} - \bar{\theta}\right) \tilde{X}_i^T\right) = \left(I_k - \sum_j B_j\right) \mathbb{E}\left(\tilde{\theta}\tilde{\theta}^T\right) - B_i \mathbb{E}\left(\varepsilon_i \varepsilon_i^T\right).$$

D'où

$$\left(I_k - \sum_j B_j\right) vI_k - B_i\sigma_0^2 I_k = 0.$$

En sommant par rapport à i on obtient

$$n\left(I_k - \sum_j B_j\right) vI_k - \sum_j B_j\sigma_0^2 I_k = 0,$$

soit

$$\sum_j B_j = \frac{nv}{nv+\sigma_0^2} I_k.$$

Donc pour tout i on a

$$\left(1 - \frac{nv}{nv+\sigma_0^2}\right) vI_k - B_i\sigma_0^2 I_k = 0,$$

$$B_i = \frac{v}{nv+\sigma_0^2} I_k.$$

Autrement dit

$$\bar{\theta} = \frac{nv}{nv+\sigma_0^2}\frac{1}{n}\sum_{i=1}^n \tilde{X}_i.$$

Calculons maintenant $K = \mathbb{E}\left(\left(\tilde{\theta}-\bar{\theta}\right)\left(\tilde{\theta}-\bar{\theta}\right)^T\right)$. Puisque

$$\tilde{\theta} - \bar{\theta} = \left(1 - \frac{nv}{nv+\sigma_0^2}\right)\tilde{\theta} + \frac{nv}{nv+\sigma_0^2}\frac{1}{n}\sum_{i=1}^n \varepsilon_i,$$

il vient

$$K = \left(1 - \frac{nv}{nv+\sigma_0^2}\right) vI_k + \left(\frac{nv}{nv+\sigma_0^2}\right)^2 \frac{\sigma_0^2}{n} I_k = \frac{\sigma_0^2 v}{nv+\sigma_0^2} I_k.$$

En conclusion, la loi a posteriori de $\tilde{\theta}$ sachant $\tilde{X}$ est

$$\mathcal{N}\left(\frac{nv}{nv+\sigma_0^2}\frac{1}{n}\sum_{i=1}^n \tilde{X}_i, \frac{\sigma_0^2 v}{nv+\sigma_0^2} I_k\right).$$

Approche à l'économie : On recherche une statistique exhaustive. Pour $x_{1:n} \in (\mathbb{R}^k)^n$, et $\theta \in \mathbb{R}^k$, P_θ a pour densité

$$\begin{aligned} p_\theta(x_{1:n}) &= N(x_{1:n}) \prod_{i=1}^n e^{-\frac{\|x_i-\theta\|^2}{2\sigma_0^2}} \\ &\propto e^{-\frac{n}{2\sigma_0^2}\|\theta\|^2 + \frac{1}{\sigma_0^2}\langle\theta, n\bar{x}_n\rangle}. \end{aligned}$$

On en déduit que $S(X_{1:n}) = \bar{X}_n$ est exhaustive, sous P_θ, $\bar{X}_n \sim \mathcal{N}\left(\theta, \frac{\sigma_0^2}{n} I_k\right)$, et on peut alors se ramener au modèle $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), \mathcal{N}(\theta, \frac{\sigma_0^2}{n} I_k))$ (soit une seule observation). Pour ce modèle, la formule de Bayes donne (en notant $\sigma_n^2 = \sigma_0^2/n$),

$$\begin{aligned} Q_x(d\theta) &\propto e^{-\frac{1}{2v}\|\theta\|^2 - \frac{1}{2\sigma_n^2}\|\theta-x\|^2} \\ &\propto e^{-\frac{1}{2}\left(\frac{1}{v}+\frac{1}{\sigma_n^2}\right)\left\|\theta - \frac{x}{\sigma_n^2}\left(\frac{1}{v}+\frac{1}{\sigma_n^2}\right)^{-1}\right\|^2}, \end{aligned}$$

pour $x \in \mathbb{R}^k$ (on a $q(x) > 0$ pour tout $x \in \mathbb{R}^k$, car $\tilde{X} \sim \mathcal{N}(0_k, (\sigma_n^2 + v)I_k)$). On en déduit que

$$\tilde{\theta} \mid \tilde{X}_{1:n} \sim \tilde{\theta} \mid \bar{\tilde{X}}_n \sim \mathcal{N}\left(\frac{v}{v+\sigma_n^2}\bar{\tilde{X}}_n, \frac{\sigma_n^2 v}{v+\sigma_n^2}\right),$$

soit la formule précédente.

4.3 Calcul d'estimateurs et risques bayésiens

Le but est de minimiser en T le risque intégré

$$\rho_(\pi) = \int_\Theta R_T(\theta)\pi(d\theta) = \mathbb{E}\ell(q(\tilde{\theta}), T(\tilde{X})).$$

Or, pour un estimateur T de $q(\theta)$, on peut écrire

$$\mathbb{E}\ell(q(\tilde{\theta}), T(\tilde{X})) = \mathbb{E}\left(\mathbb{E}\left(\ell(q(\tilde{\theta}), T(\tilde{X})) \mid \tilde{X}\right)\right),$$

la quantité $\mathbb{E}\left(\ell(q(\tilde{\theta}), T(\tilde{X})) \mid \tilde{X}\right)$ étant appelée *risque a posteriori* de T. Comme cette quantité ne dépend que de la loi a posteriori (que l'on connaît), on peut être tenté de la minimiser.

Théorème 4.28

Si $T(\tilde{X}) \in \arg\min_{y\in\mathbb{R}^k} \mathbb{E}\left(\ell(q(\tilde{\theta}), y) \mid \tilde{X}\right)$ $P_{\tilde{X}}$ p.s., alors T est bayésien. Si ce minimiseur est unique sur A tel que $P_{\tilde{X}}(A) > 0$, alors tout estimateur bayésien T' coïncide avec T sur A, $P_{\tilde{X}}$ p.s..

On remarque que ce Théorème comporte une hypothèse implicite : le minimum de $\mathbb{E}\left(\ell(q(\tilde{\theta}), y) \mid \tilde{X}\right)$ doit être atteint. La preuve est évidente.

Démonstration. Sous réserve que ce minimum existe, notons T un tel estimateur. Pour un autre candidat estimateur S, on a

$$\mathbb{E}\left(\ell(q(\tilde{\theta}), S(\tilde{X}))\right) = \mathbb{E}\left(\mathbb{E}\left(\ell(q(\tilde{\theta}), S(\tilde{X})) \mid \tilde{X}\right)\right) \geq \mathbb{E}\left(\mathbb{E}\left(\ell(q(\tilde{\theta}), T(\tilde{X})) \mid \tilde{X}\right)\right) = \rho_T(\pi).$$

Dans le cas où le minimiseur est unique sur A et T' est un autre estimateur bayésien, on a alors forcément $\{x \mid \int_\Theta \ell(q(\theta), T'(x))Q_x(d\theta) > \int_\Theta \ell(q(\theta), T(x))Q_x(d\theta)\} \cap A$ est de $P_{\tilde{X}}$ mesure nulle. Par unicité du minimum sur A, on en déduit $T'(\tilde{X}) = T(\tilde{X})$ (toujours $P_{\tilde{X}}$ p.s.) sur A. □

Dans la plupart des cas standard, ce minimum sera atteint (fonction de perte propre, espace de paramètres cible compact, etc.). On présente deux cas d'école.

4.3.1 Perte quadratique

Ici la perte considérée est $\ell(y_1, y_2) = \|y_1 - y_2\|^2$, pour $y_1, y_2 \in \mathbb{R}^k$. Dans ce cas, les estimateurs bayésiens prennent un forme simple : ce sont les moyennes a posteriori.

PROPOSITION 4.29

S'il existe un estimateur S tel que $\rho_S(\pi) < +\infty$, alors

$$T_b(\tilde{X}) = \mathbb{E}\left(q(\tilde{\theta}) \mid \tilde{X}\right)$$

est bien défini, bayésien, et tout estimateur bayésien coïncide avec lui $P_{\tilde{X}}$ p.s..

On peut remarquer que s'il n'existe aucun estimateur S tel que $\rho_S(\pi) < +\infty$, alors tout estimateur est bayésien (cas stupide).

Démonstration. Soit S un estimateur vérifiant $\rho_S(\pi) < +\infty$. Alors son risque a posteriori est fini presque-surement, c'est à dire

$$\mathbb{E}\left(\|S(\tilde{X}) - q(\tilde{\theta})\|^2 \mid \tilde{X}\right) < +\infty \quad P_{\tilde{X}} \text{ p.s..}$$

Or

$$\mathbb{E}\left(\|S(\tilde{X}) - q(\tilde{\theta})\|^2 \mid \tilde{X}\right) \geq \frac{1}{2}\mathbb{E}\left(\|q(\tilde{\theta})\|^2 \mid \tilde{X}\right) - \|S(\tilde{X})\|^2.$$

On en déduit $\mathbb{E}\left(\|q(\tilde{\theta})\|^2 \mid \tilde{X}\right) < +\infty$ $P_{\tilde{X}}$ p.s.. Donc $T_b(\tilde{X}) = \mathbb{E}(q(\tilde{\theta}) \mid \tilde{X})$ existe $P_{\tilde{X}}$ p.s., et est l'unique minimiseur de

$$\mathbb{E}\left(\|y - q(\tilde{\theta}\|^2 \mid \tilde{X}\right).$$

Le Théorème 4.28 s'applique donc. □

Exemple 4.30 *: Échantillon Gaussien (suite).* Dans le modèle $\mathcal{N}(\theta, \sigma_0^2 I_k)^{\otimes n}$ étudié précédemment, avec $\tilde{\theta} \sim \mathcal{N}(0_k, I_k)$, on a $E(\|\tilde{\theta}\|^2) = kv < +\infty$, et donc l'unique estimateur bayésien (au sens $P_{\tilde{X}}$) est

$$T_b(\tilde{X}) = \frac{nv}{nv + \sigma_0^2}\bar{\tilde{X}}_n.$$

Son risque bayésien est alors

$$\begin{aligned}\mathbb{E}\left\|T_b(\tilde{X}) - \tilde{\theta}\right\|^2 &= \mathbb{E}\left(\mathbb{E}\left(\left\|T_b(\tilde{X}) - \tilde{\theta}\right\|^2 \mid \tilde{X}\right)\right) \\ &= \frac{k\sigma_0^2 v}{nv + \sigma_0^2}.\end{aligned}$$

Cet exemple illustre bien deux concepts et techniques propres au bayésien. Premièrement, d'un point de vue intuitif, l'estimateur bayésien va intégrer une "connaissance a priori" sur le paramètre, ici on s'attend à ce que le paramètre soit proche de 0 a priori, plus ou moins fortement en fonction de v. L'estimateur bayésien va alors réaliser une interpolation entre l'estimateur fréquentiste $\bar{\tilde{X}}_n$ et l'information a priori, se traduisant en

$$T_b(\tilde{X}) = \alpha\bar{\tilde{X}}_n + (1 - \alpha) \times 0.$$

Le paramètre $\alpha = \frac{nv}{nv+\sigma_0^2}$ traduit un mélange entre poids de l'information a priori (v) et poids des observations (n).

— Si $v << \frac{\sigma_0^2}{n}$, l'information a priori sur θ l'emporte et T_b tend vers 0 (estimateur limite où il n'y a que de l'information a priori).

— Si $n >> \sigma_0^2/v$, alors ce sont les observations qui l'emportent sur l' a priori, l'estimateur limite est dans ce cas $\bar{\tilde{X}}_n$, l'estimateur fréquentiste.

D'un point de vue plus technique, le calcul du risque bayésien se fait généralement en calculant le plus facile entre $\mathbb{E}(\ell(q(\tilde{\theta}), T_b(\tilde{X})) \mid \tilde{X})$ et $\mathbb{E}(\ell(q(\tilde{\theta}), T_b(\tilde{X})) \mid \tilde{\theta})$, puis en intégrant (par rapport à la loi de $\tilde{X}$ ou celle de $\tilde{\theta}$). Dans le cas du risque quadratique, $\mathbb{E}(\ell(q(\tilde{\theta}), T_b(\tilde{X})) \mid \tilde{X})$ est la *variance a posteriori* et est souvent facilement calculable.

Concluons en remarquant qu'il n'est pas nécessaire que $\mathbb{E}(\|q(\tilde{\theta})\|^2$ soit fini pour que le risque bayésien quadratique le soit, par exemple dans le cas suivant.

Exemple 4.31. On observe $X = \theta + \xi$ où $\xi \sim \mathcal{N}(0,1)$ et $\theta \in \mathbb{R}$. On prend comme loi a priori la loi de Cauchy standard (de densité $\theta \mapsto \frac{1}{\pi(1+\theta^2)}$). On veut estimer θ. Le risque quadratique de l'estimateur $T(x) = x$ s'écrit $R_T(\theta) = 1$ pour tout $\theta \in \Theta$ et donc $\rho_T(\pi) = 1$. Par le théorème de Bayes on a

$$\mathbb{E}\left(\tilde{\theta} \mid \tilde{X}\right) = \frac{\int_{\mathbb{R}} \theta \exp\left(-(\tilde{X}-\theta)^2/2\right) \frac{1}{1+\theta^2} d\theta}{\int_{\mathbb{R}} \exp\left(-(\tilde{X}-\theta)^2/2\right) \frac{1}{1+\theta^2} d\theta} \quad P_{\tilde{X}}\text{-p.s.},$$

et par conséquent

$$S(x) = \frac{\int_{\mathbb{R}} \theta \exp\left(-(x-\theta)^2/2\right) \frac{1}{1+\theta^2} d\theta}{\int_{\mathbb{R}} \exp\left(-(x-\theta)^2/2\right) \frac{1}{1+\theta^2} d\theta}$$

est bayésien.

4.3.2 Test bayésien

On considère deux hypothèses statistiques $H_0 \subset \Theta$ et $H_1 \subset \Theta$ telles que

$$\Theta = H_0 \cup H_1, \quad H_0 \cap H_1 = \emptyset, \quad H_0 \neq= \emptyset, \quad H_1 \neq= \emptyset.$$

Le problème de test associé correspond à l'estimation de

$$q(\theta) = 1_{H_1}(\theta),$$

pour la fonction de perte (de classification)

$$\ell\left(T, q(\theta)\right) = \mathbb{1}_{T \neq q(\theta)}.$$

Le risque d'un test T (i.e. un estimateur de $q(\theta)$) s'écrit

$$R_T(\theta) = P_\theta\left(T \neq q(\theta)\right) = P_\theta\left(\theta \notin H_T\right).$$

Le risque intégré de T s'écrit

$$\rho_T(\pi) = \mathbb{P}\left(\tilde{\theta} \notin H_{T(\tilde{X})}\right) = 1 - \mathbb{P}\left(\tilde{\theta} \in H_{T(\tilde{X})}\right).$$

Ce cadre est différent de celui du problème de test classique : on ne cherche pas à contrôler l'erreur de première espèce mais bien la probabilité de se tromper au total. Les hypothèses redeviennent symétriques ici, c'est plutôt un problème de *classification*.

Dans notre cadre bayésien, on déduit le résultat suivant :

Proposition 4.32

T est bayésien si et seulement si

$$T(\tilde{X}) \in argmax_{t\in\{0,1\}}\mathbb{P}\left(\tilde{\theta} \in H_t \mid \tilde{X}\right) \quad P_{\tilde{X}}\text{-p.s.},$$

c'est à dire si $P_{\tilde{X}}$-p.s. on a

$$T(\tilde{X}) = 1 \quad sur \quad \left\{\mathbb{P}(\tilde{\theta} \in H_1 \mid \tilde{X}) > \mathbb{P}(\tilde{\theta} \in H_0 \mid \tilde{X})\right\},$$

$$T(\tilde{X}) = 0 \quad sur \quad \left\{\mathbb{P}(\tilde{\theta} \in H_1 \mid \tilde{X}) < \mathbb{P}(\tilde{\theta} \in H_0 \mid \tilde{X})\right\}.$$

Démonstration. On a

$$\arg\min_{y\in\{0,1\}} \mathbb{E}\left(\ell(q(\tilde{\theta}), y) \mid \tilde{X}\right) = \arg\min y \in \{0,1\}\mathbb{P}\left(\tilde{\theta} \notin H_y \mid \tilde{X}\right)$$

est bien défini $P_{\tilde{X}}$ presque sûrement par

$$\arg\max_{t\in\{0,1\}} \mathbb{P}\left(\tilde{\theta} \in H_t \mid \tilde{X}\right),$$

et est unique lorsque $\mathbb{P}\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right) \wedge \mathbb{P}\left(\tilde{\theta} \in H_1 \mid \tilde{X}\right) < \frac{1}{2}$. On peut alors appliquer le Théorème 4.28. □

Le test bayésien choisira donc l'hypothèse de masse la plus importante a posteriori (ce qui semble assez naturel). Dans le cas d'un modèle dominé on peut s'épargner le calcul des masses a posteriori.

Proposition 4.33

Supposons que le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ est dominé par μ (σ-finie). Supposons qu'il existe $(\theta, x) \mapsto p_\theta(x)$ bi-mesurable telle que pour tout $\theta \in \Theta$ $p_\theta = \frac{dP_\theta}{d\mu}$ μ-presque partout. Alors un test T de la forme

$$T(x) = \mathbb{1}_{\int_{H_1} p_\theta(x)\pi(d\theta) > \int_{H_0} p_\theta(x)\pi(d\theta)}, \quad pour\ x\ tel\ que\ q(x) > 0,$$

est bayésien.

Démonstration. Comme $\int_{H_0} Q_x(d\theta) < \int_{H_1} Q_x(d\theta)$ équivaut à

$$\int_{H_1} p_\theta(x)\pi(d\theta) > \int_{H_0} p_\theta(x)\pi(d\theta)$$

d'après la formule de Bayes, c'est évident. □

Exemple 4.34. On observe $X \sim \mathcal{N}(\theta, 1)$ où $\theta \in \mathbb{R}$. On considère les hypothèses $H_0 = \{0\}$ et $H_1 = \mathbb{R}^*$. Il est important de choisir une loi a priori π telle que $\pi(H_0) > 0$ sinon aurait $\mathbb{P}(\tilde{\theta} \in H_0 \mid \tilde{X}) = 0$ $\mathbb{P}$-p.s. et le test bayésien consisterait à toujours choisir H_1. On considère comme loi a priori

$$\pi = q\delta_0 + (1-q)\mathcal{N}(0, v),$$

où $q \in]0,1[$ et $v > 0$ (modèle Gaussien infatué en 0). Le modèle est dominé par la mesure de Lebesgue et on a $p_\theta(x) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}(x-\theta)^2\right)$. D'une part

$$\begin{aligned}
&\int_{H_1} p_\theta(x)\pi(d\theta) \\
&= (1-q)\int_{\mathbb{R}^*} \frac{1}{2\pi\sqrt{v}} \exp\left(-\frac{1}{2}(x-\theta)^2 - \frac{1}{2v}x^2\right) d\theta \\
&= \frac{1-q}{2\pi\sqrt{v}} \exp\left(-\frac{1}{2}x^2(1-1/(1/v+1))\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2}(1/v+1)\left(\theta - x/(1/v+1)\right)^2\right) d\theta \\
&= \frac{1-q}{\sqrt{2\pi}\sqrt{v+1}} \exp\left(-\frac{1}{2}x^2/(v+1)\right).
\end{aligned}$$

D'autre part

$$\int_{H_0} p_\theta(x)\pi(d\theta) = q\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2).$$

Donc

$$\begin{aligned}
&\int_{H_1} p_\theta(x)\pi(d\theta) > \int_{H_0} p_\theta(x)\pi(d\theta) \\
\Leftrightarrow \quad &\frac{1-q}{\sqrt{2\pi}\sqrt{v+1}} \exp\left(-\frac{1}{2}x^2/(v+1)\right) > q\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \\
\Leftrightarrow \quad &\frac{1}{2}x^2\frac{v}{v+1} > \log\left(\frac{q}{1-q}\right) + \frac{1}{2}\log(1+v).
\end{aligned}$$

En conclusion, si $\log\left(\frac{q}{1-q}\right) + \frac{1}{2}\log(1+v) < 0$ alors $T = 1_{\mathbb{R}}$ est bayésien et si $\log\left(\frac{q}{1-q}\right) + \frac{1}{2}\log(1+v) \geq 0$, le test

$$T(x) = \mathbb{1}_{|x| > 2\frac{v+1}{v}\left(\log\left(\frac{q}{1-q}\right) + \frac{1}{2}\log(1+v)\right)^{1/2}}$$

est bayésien.

4.4 Utilisation en théorie minimax

Même sans adhérer au paradigme bayésien, les estimateurs bayésiens peuvent être utiles. Quoique structurellement biaisés, dans certaines situations ils sont meilleurs que les estimateurs de type max de vraisemblance (cf exemple plus bas). Ils peuvent aussi être approchés en pratique via simulation, dans certains cas où les estimateurs "fréquentistes" ne sont pas calculables. Bref, c'est une gamme d'estimateurs relativement classiques qu'il faut avoir en tête.

Même dans un cadre fréquentiste, la théorie bayésienne présente un intérêt : celui de donner une borne inférieure sur les vitesses minimax. En effet, pour peu qu'elles ait du sens, on aura toujours la série d'inégalités suivantes

$$\inf_T \sup_\theta R_T(\theta) \geq \inf_T \rho_T(\pi) = \rho(\pi),$$

et ce pour n'importe quelle loi a priori. Le but va être alors de trouver des lois a priori "les moins favorables" pour attester de l'optimalité d'un estimateur donné, au sens minimax.

Théorème 4.35 : Hodges et Lehmann

Pour tout $k \in \mathbb{N}^*$, soit π_k une loi a priori et soit T_k un estimateur bayésien relativement à π_k. Si T^* est un estimateur tel que

$$\bar{R}_{T^*} \leq \limsup_k \rho_{T_k}(\pi_k),$$

alors T^* est minimax.

Démonstration. Soit T un estimateur de $q(\theta)$. On a

$$\bar{R}_T \geq \int_\Theta R_T(\theta)\pi_k(d\theta),$$

et $\int_\Theta R_T(\theta)\pi_k(d\theta) \geq \rho_{T_k}(\pi_k)$ car T_k est bayésien. Donc $\sup_{\theta\in\Theta} R_T(\theta) \geq \limsup_k \rho_{T_k}(\pi_k) \geq R_{T^*}(\theta_0)$ pour tout $\theta_0 \in \Theta$ par hypothèse. On en déduit que $\bar{R}_T \geq \bar{R}_{T^*}$. □

Exemple 4.36 *: Optimalité de l'EMV dans le modèle poissonien.*

On considère le modèle $\mathcal{P}(\theta)^{\otimes n}$, où $\theta > 0$. On vérifie que c'est un modèle exponentiel dominé par la mesure de comptage sur $\mathbb{N}^n$, de statistique exhaustive $S(X) = \sum_{i=1}^n X_i$. Calculons $\hat{\theta}_{EMV}$.

- Si $S(x) = \sum_{i=1}^n x_i > 0$, $\ell_n(\theta) \propto -n\theta + S(x)\log(\theta)$, et $\hat{\theta}_{EMV}(x) = \bar{x}_n$.
- Si $S(x) = 0$, alors $\ell_n(\theta) \propto -n\theta$, et, avec un léger abus, on note toujours $\hat{\theta}_{EMV} = \bar{x}_n$ (quitte à se placer sur $[0, +\infty[$, naturel en considérant $\mathcal{P}(0) \sim \delta_0$, et pas grave de toutes façons car $P_\theta(S(X) = 0) = e^{-n\theta}$ si $\theta > 0$).

Au sens du risque quadratique, on a

$$R_{\hat{\theta}_{EMV}}(\theta) = E_\theta\left(\bar{X}_n - \theta\right)^2 = \frac{\mathrm{Var}_\theta(X_1)}{n} = \frac{\theta}{n}.$$

On en déduit

$$\bar{R}_{\hat{\theta}_{EMV}} = +\infty.$$

Pour éviter ces $+\infty$, on peut considérer la fonction de perte un peu tordue

$$\ell(y, \theta) = \frac{(y-\theta)^2}{\theta},$$

et le risque L qui y est associé. Au sens de ce risque, on a

$$\bar{L}_{\hat{\theta}_{EMV}} = \frac{1}{n}.$$

On peut vérifier que l'information de Fisher de ce modèle vaut $I(\theta) = \frac{1}{\theta}$, et donc $I_n(\theta) = \frac{n}{\theta}$. $\hat{\theta}_{EMV}$ atteint donc la borne de Cramer-Rao (Théorème 3.14), et est donc optimal parmi les estimateurs sans biais (efficace).

Regardons maintenant l'optimalité de $\hat{\theta}_{EMV}$ du point de vue minimax. Pour cela, on se donne la loi a priori

$$\tilde{\theta} \sim \gamma(a, b),$$

avec $a, b > 0$. En se rappelant que S est exhaustive, on se ramène au modèle $\tilde{X} \mid \tilde{\theta} \sim \mathcal{P}(n\tilde{\theta})$. Le modèle étant dominé par la mesure de comptage, la formule de Bayes s'applique, et on a

$$Q_x(d\theta) \propto \theta^{a+x-1} e^{-(b+n)\theta},$$

pour $x \geq 0$. On en déduit que $\tilde{X} \mid \tilde{\theta} \sim \gamma(a + \tilde{X}, b + n)$. Un peu de calcul sur les loi γ avant de poursuivre. Si $a > 0$, $\mathbb{E}(\gamma(a, b))$ existe et vaut

$$\int_O^{+\infty} u^a e^{-bu} \frac{b^a}{\Gamma(a)} du = \frac{\Gamma(a+1)}{b^{a+1}} \frac{b^a}{\Gamma(a)} = \frac{a}{b}.$$

Si $a > 1$, $\mathbb{E}(\gamma(a, b)^{-1})$ existe, et vaut

$$\int_0^{+\infty} u^{a-2} e^{-bu} \frac{b^a}{\Gamma(a)} du = \frac{\Gamma(a-1)}{b^{a-1}} \frac{b^a}{\Gamma(a)} = \frac{b}{a-1}.$$

Par ailleurs, $\mathrm{Var}(\gamma(a, b) = \frac{a^2}{b}$. Pour le risque quadratique, $\mathbb{E}(\tilde{\theta}^2 \mid \tilde{X}) < +\infty$, les estimateurs bayésiens sont donc de la forme

$$T_{b,1}(\tilde{X}) = \mathbb{E}(\tilde{\theta} \mid \tilde{X}) = \frac{a + \tilde{X}}{b + n},$$

$P_{\tilde{X}}$-p.s. Plutôt que de calculer le risque a posteriori, on peut toujours minorer

$$R_{T_{b,1}}(\theta) \geq \mathrm{Var}_\theta(T_{b,1}) \geq \frac{n\theta}{(b+n)^2},$$

ce dont on déduit

$$\rho(\gamma(a, b)) \geq \frac{an}{b(b+n)^2} \underset{b \to 0}{\longrightarrow} +\infty.$$

Et $\hat{\theta}_{EMV}$ est minimax au sens quadratique (ce qui est un peu stupide).

Pour ce qui est du risque ajusté L, essayons de trouver un estimateur bayésien. On prend $a > 1$. Comme $E(\tilde{\theta}^{-1} \mid \tilde{X})) < +\infty$ dans ce cas, on peut écrire, pour $y \geq 0$,

$$\mathbb{E}\left(\ell(y, \tilde{\theta}) \mid \tilde{X}\right) = y^2 E(\tilde{\theta}^{-1} \mid \tilde{X})) - 2y + \mathbb{E}(\tilde{\theta} \mid \tilde{X}),$$

qui est minimal en

$$T_{b,2}(\tilde{X}) = \left(E(\tilde{\theta}^{-1} \mid \tilde{X}))\right)^{-1}$$

uniquement, $P_{\tilde{X}}$-p.s.. Le risque a posteriori de $T_{b,2}$ s'écrit

$$\begin{aligned}
\mathbb{E}\left(\ell(T_{b,2}(\tilde{X}), \tilde{\theta}) \mid \tilde{X}\right) &= \mathbb{E}(\tilde{\theta} \mid \tilde{X}) - \left(E(\tilde{\theta}^{-1} \mid \tilde{X}))\right)^{-1} \\
&= \frac{a + \tilde{X}}{b + n} - \frac{a + \tilde{X} - 1}{b + n} \\
&= \frac{1}{b + n},
\end{aligned}$$

$P_{\tilde{X}}$-p.s.. On en déduit

$$\rho(\gamma(a, b)) = \frac{1}{b + n} \underset{b \to 0}{\longrightarrow} \frac{1}{n} = \bar{L}_{\hat{\theta}_{EMV}},$$

et $\hat{\theta}_{EMV}$ est minimax au sens de la perte ajustée ℓ.

Exemple 4.37 *: Binomiale/Beta.* Dans le modèle $(\mathcal{B}(n,\theta))_{\theta\in]0,1[}$, dominé par la mesure de comptage et par ailleurs exponentiel, on a déjà vu que on pouvait définir

$$\hat{\theta}_{EMV} = \frac{X}{n},$$

quitte à sortir de Θ lorsque $X = 0$ ou $X = n$. Son risque quadratique s'écrit

$$R_{\hat{\theta}_{EMV}} = \mathrm{Var}_\theta\left(\frac{X}{n}\right) = \frac{\theta}{(1-\theta)}n.$$

On peut aussi vérifier que l'information de Fisher de ce modèle est donnée par $I(\theta) = \frac{n}{\theta(1-\theta)}$, et donc que $\hat{\theta}_{EMV}$ est efficace au sens de la borne de Cramer-Rao. Au sens minimax, on a

$$\bar{R}_{\hat{\theta}_{EMV}} = \frac{1}{4n}.$$

Dans un cadre bayésien, considérons la loi a priori donnée par $\pi \sim \beta(a,b)$, où $a, b > 0$. Un peu de calcul donne

$$\begin{aligned}\mathbb{E}\left(\beta(a,b)\right) &= \int_0^1 \frac{u^a(1-u)^{b-1}}{B(a,b)}du = \frac{B(a+1,b)}{B(a,b)}\\ &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a}{a+b}.\end{aligned}$$

Le modèle étant dominé, la formule de Bayes donne

$$Q_x(d\theta) \propto \theta^{a+x-1}(1-\theta)^{b+(n-x)-1},$$

pour $x \in \{0,\dots,n\}$. On en déduit

$$\tilde{\theta} \mid \tilde{X} \sim \beta(a+\tilde{X}, b+n-\tilde{X}).$$

Comme $\|\beta(a+\tilde{X}, b+n-\tilde{X})\|_\infty < +_i nfty$ $P_{\tilde{X}}$-p.s., les estimateurs bayésiens (au sens quadratique) sont donnés par

$$T_b(\tilde{X}) = \mathbb{E}\left(\tilde{\theta} \mid \tilde{X}\right) = \frac{a+\tilde{X}}{b+n+a},$$

$P_{\tilde{X}}$-p.s., ce dont on peut déduire, pour $\theta \in]0,1[$,

$$R_{T_b}(\theta) = \frac{(a-(a+b)\theta)^2}{(b+n)^2} + \frac{n\theta(1-\theta)}{(b+n+a)^2}.$$

En choisissant $a = b = \sqrt{n}/2$, il vient

$$R_{T_b}(\theta) = \frac{1}{4(1+\sqrt{n})^2},$$

ce dont on déduit deux choses :

— $T_b(X) = \frac{X+\sqrt{n}/2}{n+\sqrt{n}}$ est minimax, avec

$$\bar{R}_{T_b} = \frac{1}{4(1+\sqrt{n})^2}.$$

— Comme $\bar{R}_{\hat{\theta}_{EMV}} < \bar{R}_{T_b}$, $\hat{\theta}_{EMV}$ n'est pas minimax.

On remarque que le pire des cas pour l'estimateur $\hat{\theta}_{EMV}$ est θ autour de 1/2. L'estimateur bayésien T_b tend à corriger le tir en intégrant une information a priori de type "θ proche de 1/2", de manière à contrebalancer les mauvaises performances relatives de $\hat{\theta}_{EMV}$ dans cette zone. Les zones "θ proches du bord" donnent des risques négligeables devant la zone du milieu, donc au sens minimax on ne perd pas grand chose à détériorer les performances sur ces zones en intégrant de l'information a priori favorisant le milieu.

L'approche bayésienne permet donc d'évaluer les performances d'estimateurs d'un point de vue minimax. Dans les modèles simples comme ci-dessus, les risques bayésiens sont explicitement calculables. Dans des cas plus compliqués, diverses techniques de minoration des risques minimax existent (Lecam, Assouad, Fano-Birgé, etc.), toutes basées sur des approches bayésiennes néanmoins. Citons par exemple le Lemme de Le Cam.

Théorème 4.38 : Lemme de Le Cam

Soit $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}$ un modèle statistique, $q : \Theta \to E$ un paramètre à estimer dans un espace métrique (E, d), et $\ell(x, y) = \mathrm{d}(x, y)$ la fonction de perte associée à la métrique dans E. Soient $\theta_0 \neq \theta_1 \in \Theta$. On a alors

$$\inf_T \bar{R}_T \geq \frac{1}{2}\mathrm{d}(q(\theta_0), q(\theta_1))(1 - \mathrm{d}_{TV}(P_{\theta_0}, P_{\theta_1}).$$

Avant de prouver ce résultat, on remarque que le choix de θ_0 et θ_1 est libre. Pour se ramener à la borne la plus fine possible, il convient de choisir θ_0 et θ_1 de sorte que P_{θ_0} et P_{θ_1} soient les plus proches possibles en termes de distance en variation totale, et tels que
$d(q(\theta_0), q(\theta_1))$ soit le plus large possible. Par ailleurs, on peut assouplir un peu la condition sur ℓ : si la condition de symétrie n'est pas vérifiée,

$$\tilde{\ell}(q(\theta_1), q(\theta_2)) = \frac{1}{2}\left(\ell(q(\theta_1), q(\theta_2)) + \ell(q(\theta_2), q(\theta_1))\right),$$

est elle symétrique, de sorte que si ℓ satisfait les inégalités triangulaires, $\tilde{\ell}$ est une pseudo-distance. On peut en déduire une borne sur le risque minimax pour ℓ via une borne pour celui associé à $\tilde{\ell}$ (le résultat marche encore pour les pseudo-distances).

Démonstration. Toute l'idée est de se ramener à un problème de test bayésien entre les deux hypothèses $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$. On peut commencer par écrire, pour un estimateur T de $q(\theta)$,

$$\sup_{\theta\in\Theta} R_T(\theta) \geq \sup_{\theta\in\{\theta_0,\theta_1\}} R_T(\theta),$$

de sorte que l'on peut ramener le modèle à $\Theta = \{\theta_0, \theta_1\}$, qui est toujours dominé (par $(P_{\theta_0} + P_{\theta_1})/2$ par exemple). Prenons la loi a priori $\pi \sim \frac{1}{2}(\delta_{\theta_0} + \delta_{\theta_1})$. Supposons $\mathbb{P}(\tilde{\theta} \in H_1 \mid \tilde{X}) \geq 1/2$. On a alors, pour un estimateur T, en remarquant que

$\mathrm{d}(y, q(\theta_0)) + \mathrm{d}(y, q(\theta_1)) \geq \mathrm{d}(q(\theta_0), q(\theta_1))$,

$$\begin{aligned}
\mathbb{E}\left(\mathrm{d}(T(\tilde{X}, q(\tilde{\theta}) \mid \tilde{X}\right) &= \mathrm{d}(y, q(\theta_1))\mathbb{P}\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right) + \mathrm{d}(y, q(\theta_0))\mathbb{P}\left(\tilde{\theta} \in H_1 \mid \tilde{X}\right) \\
&\geq \left(\mathrm{d}(T(\tilde{X}), q(\theta_0)) + \mathrm{d}(T(\tilde{X}), q(\theta_1))\right)\mathbb{P}\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right) \\
&\quad + \mathrm{d}(T(\tilde{X}), q(\theta_0))\left(\mathbb{P}\left(\tilde{\theta} \in H_1 \mid \tilde{X}\right) - \mathbb{P}\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right)\right) \\
&\geq \mathrm{d}(q(\theta_0, q(\theta_1))\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right).
\end{aligned}$$

On en déduit

$$\rho(\pi) \geq \mathrm{d}(q(\theta_0, q(\theta_1))\mathbb{E}\left(\mathbb{P}\left(\tilde{\theta} \in H_1 \mid \tilde{X}\right) \wedge \mathbb{P}\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right)\right)$$

Par ailleurs, $\mathbb{E}\left(\mathbb{P}\left(\tilde{\theta} \in H_1 \mid \tilde{X}\right) \wedge \mathbb{P}\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right)\right)$ étant le risque bayésien du problème de test, on a

$$\begin{aligned}
\mathbb{E}\left(\mathbb{P}\left(\tilde{\theta} \in H_1 \mid \tilde{X}\right) \wedge \mathbb{P}\left(\tilde{\theta} \in H_0 \mid \tilde{X}\right)\right) &= \inf_T \frac{1}{2}\left(P_{\theta_0}(T=1) - P_{\theta_1}(T=0)\right) \\
&= \frac{1}{2}\left(1 - \sup_A |(P_{\theta_1} - P_{\theta_0})(A)|\right) \\
&\geq \frac{1}{2}(1 - \mathrm{d}_{TV}(P_{\theta_0}, P_{\theta_1})).
\end{aligned}$$

□

Cette inégalité est utilisée en pratique sur des n-échantillons. Pour cela, on a besoin de quelques outils techniques.

Lemme 4.39

Soient P_0 et P_1 dominées par μ, de densités respectives p_0 et p_1. On a alors

$$\mathrm{d}_{TV}(P_0, P_1) = \frac{1}{2}\int_{\mathcal{X}} |p_0(x) - p_1(x)|\mu(dx) = 1 - \int_{\mathcal{X}} p_0(x) \wedge p_1(x)\mu(dx).$$

On en déduit

$$\mathrm{d}_{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq n\mathrm{d}_{TV}(P, Q).$$

Démonstration. Sans perte de généralité on suppose que μ est de proba. Pour $A \subset \mathcal{A}$, on peut écrire

$$P_1(A) - P_0(A) = \int (p_1(x) - p_0(x))\mathbb{1}_A(x)\mu(dx),$$

qui est maximal pour $A = \{p_1 \geq p_0\}$. On a alors

$$\mathrm{d}_{TV}(P_0, P_1) = \int (p_1 - p_0)_+(x)\mu(dx).$$

Or

$$\begin{aligned}
\int (p_1 - p_0)_+(x)\mu(dx) + \int (p_1 - p_0)_-(x)\mu(dx) &= \int |p_1 - p_0|(x)\mu(dx) \\
\int (p_1 - p_0)_+(x)\mu(dx) &= \int (p_1 - p_0)_-(x)\mu(dx),
\end{aligned}$$

ce dont on déduit les deux égalités. L'inégalité vient de

$$\begin{aligned}\mathrm{d}_{TV}(P_0^{\otimes n}, P_1^{\otimes n}) &= 1 - \int (p_0(x) \wedge p_1(x))^n \mu(dx) \\ &\leq 1 - (\int p_0(x) \wedge p_1(x) \mu(dx))^n \\ &\leq 1 - (1 - \mathrm{d}_{TV}(P_0, P_1))^n \\ &\leq n \mathrm{d}_{TV}(P_0, P_1).\end{aligned}$$

□

On peut avoir un peu mieux que $n\mathrm{d}_{TV}(P_0, P_1)$ comme majorant. Pour cela on va devoir introduire une nouvelle distance, la *distance de Hellinger*.

Definition 4.40 : Distance de Hellinger

Soient P_0 et P_1 dominées par μ, de densités respectives p_0 et p_1. La distance de Hellinger $\mathrm{d}_H(P_0, P_1)$ est définie par

$$\mathrm{d}_H(P_0, P_1)^2 = \frac{1}{2} \int_{\mathcal{X}} (\sqrt{p_0} - \sqrt{p_1})(x)^2 \mu(dx) = 1 - \int_{\mathcal{X}} \sqrt{p_0 p_1(x)} \mu(dx),$$

le terme

$$\rho(P_0, P_1) = \int_{\mathcal{X}} \sqrt{p_0 p_1(x)} \mu(dx)$$

étant appelé *affinité de Hellinger*.

On a immédiatement que $\mathrm{d}_H^2(P_0, P_1) \leq \mathrm{d}_{TV}(P_0, P_1)$ (via $\sqrt{a} - \sqrt{b} \leq \sqrt{a-b}$). Par ailleurs, l'affinité de Hellinger passe bien au produit, c'est à dire qu'on a de manière évidente

$$\rho(P_0^{\otimes n}, P_1^{\otimes n}) = \rho(P_0, P_1)^n.$$

On peut relier distance de Hellinger et en variation totale pour les lois produit comme suit.

Théorème 4.41

Soient P_0 et P_1 dominées par μ, de densités respectives p_0 et p_1. On a alors

$$\mathrm{d}_{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq 1 - \frac{1}{4} \left(1 - \mathrm{d}_H^2(P_0, P_1)\right)^{2n}.$$

On en déduit

$$\mathrm{d}_{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq 1 - \frac{1}{4} \left(1 - \mathrm{d}_{TV}(P_0, P_1)\right)^{2n}$$

Démonstration. Partons de l'affinité $\rho(P_0^{\otimes n}, P_1^{\otimes n})$. En posant $A = \{p_1^n \geq p_0^n\}$, il

vient

$$\begin{aligned}\rho(P_0^{\otimes n}, P_1^{\otimes n}) &= \int \sqrt{p_0^n(x)p_1^n(x)}\mathbb{1}_A(x)\mu(dx) + \int \sqrt{p_0^n(x)p_1^n(x)}\mathbb{1}_{A^c}(x)\mu_n(dx)\\ &\leq \sqrt{\int p_0^n(x)\mathbb{1}_A(x)\mu_n(dx)}\sqrt{\int p_1^n(x)\mu_n(dx)} + \sqrt{\int p_1^n(x)\mathbb{1}_{A^c}(x)\mu_n(dx)}\sqrt{\int p_0^n(x)\mu_n(dx)}\\ &\leq 2\sqrt{\int p_0^n(x) \wedge p_1^n(x)\mu_n(dx)}\\ &\leq 2\sqrt{1 - \mathrm{d}_{TV}(P_0^{\otimes n}, P_1^{\otimes n})}.\end{aligned}$$

Par ailleurs,

$$\rho(P_0^{\otimes n}, P_1^{\otimes n}) = \rho(P_0, P_1)^n = (1 - \mathrm{d}_H^2(P_0, P_1))^n,$$

ce dont on déduit la première inégalité. La deuxième inégalité est immédiate via $\mathrm{d}_{TV}(P_0, P_1) \leq \mathrm{d}_H^2(P_0, P_1)$. □

Avec ces deux ingrédients (Lemme de Le Cam et Théorème 4.41), on peut retrouve beaucoup de bornes inférieures pour les risques minimax, du moins en terme de dépendance en n (taille de l'échantillon). On peut aussi s'en servir pour montrer la "nécessité" de se restreindre à des modèles pas trop gros (par exemple paramétriques) : du point de vue minimax, un modèle trop gros entraînera "l'inconsistance uniforme", c'est à dire un risque minimax ne convergeant pas vers 0 avec la taille de l'échantillon.

Proposition 4.42 : Inconsistance de l'estimation de support

Soit Θ l'ensemble des lois de probabilités sur $\mathbb{R}$, et, pour $\theta \in \Theta$, on pose $P_\theta = \theta$. On s'intéresse à l'estimation du support, c'est à dire $q(\theta) = \mathrm{Supp}(\theta)$, en distance de Hausdorff. On a alors

$$\lim \inf_{n\to+\infty} \inf_T \sup_\theta E_\theta^n \mathrm{d_H}(T, \mathrm{Supp}(\theta)) = +\infty.$$

Cela traduit le principe général suivant : le risque minimax étant un risque uniforme sur une classe, choisir une classe trop grosse ne permet pas d'avoir des garanties uniformes convergeant vers 0 avec la taille de l'échantillon (on retrouvera ce principe avec le No-Free Lunch Theorem).

Démonstration. On choisit $P_0 = \delta_0$, et $P_1 = \alpha\delta_0 + (1-\alpha)\delta_N$, où N est grand et $\alpha > 1/2$. On a $\mathrm{d_H}(\mathrm{Supp}(P_0), \mathrm{Supp}(P_1)) = N$. Par ailleurs, $\mathrm{d}_{TV}(P_0, P_1) = 1 - \alpha$. Du Lemme de Le Cam on déduit

$$\begin{aligned}\inf_T \sup_\theta E_\theta^n(\mathrm{d_H}(T, \mathrm{Supp}(\theta)) &\geq \frac{1}{2}N(1 - \mathrm{d}_{TV}(P_0^{\otimes n}, P_1^{\otimes n})\\ &\geq \frac{1}{2}N(1 - n\mathrm{d}_{TV}(P_0, P_1))\\ &\geq \frac{1}{2}N(1 - n(1-\alpha))\end{aligned}$$

en utilisant la borne sous-optimale du Lemme 4.39. En prenant $1 - \alpha = \frac{1}{2n}$, on obtient

$$\inf_T \sup_\theta E_\theta^n(\mathrm{d_H}(T, \mathrm{Supp}(\theta)) \geq \frac{N}{4},$$

d'où le résultat en faisant tendre N vers $+\infty$. □

4.4.1 L'estimateur par moindre carrés est minimax

On a vu dans le Chapitre 2.3.1 que, dans le modèle $Y = X\theta + \varepsilon$, avec (X^TX) inversible, paramétré par $\theta \in \mathbb{R}^k$ et $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ vecteur d'erreurs i.i.d. centrés et de variance σ^2, l'estimateur par moindre carrés $\hat{\theta}_{LS}$ vérifie

$$\sup_{\varepsilon_1,\theta} \|\hat{\theta}_{LS} - \theta\|^2 = \sigma^2 \mathrm{Tr}((X^TX)^{-1}),$$

où le supremum en ε_1 est pris sur l'ensemble des lois d'erreurs centrées et de variances σ^2 (c'est donc un modèle non paramétrique).

On va montrer que l'estimateur par moindre carrés est optimal sur cette classe au sens minimax, c'est à dire

$$\inf_T \sup_{\varepsilon_1,\theta} \|T(X) - \theta\|^2 \geq \sigma^2 \mathrm{Tr}((X^TX)^{-1}).$$

On peut se restreindre au modèle Gaussien : en effet

$$\inf_T \sup_{\varepsilon_1,\theta} \|T(X) - \theta\|^2 \geq \inf_T \sup_{\varepsilon_1 \sim \mathcal{N}(0,\sigma^2),\theta} \|T(X) - \theta\|^2.$$

Pour ce modèle $Y = X\theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, cherchons une statistique exhaustive. Le modèle est dominé par $\mathcal{L}_n$, de densité

$$\begin{aligned} p_\theta(y_{1:n}) &\propto e^{-\|y - X\theta\|^2/2} \\ &\propto e^{\langle \theta, X^Ty \rangle - \|\theta\|^2/2}. \end{aligned}$$

On en déduit que X^TY est exhaustive. Sous P_θ, on a

$$X^TY \sim \mathcal{N}\left(X^TX\theta, \sigma^2(X^TX)\right).$$

Comme X^TX est symétrique et inversible, on peut écrire $X^TX = UDU^T$, où U est orthogonale et D diagonale strictement positive. En notant $Q = D^{-\frac{1}{2}}U^T$, on a

$$QX^TY \sim \mathcal{N}\left(Q^{-1}\theta, \sigma^2 I_k\right).$$

En reparamétrant ce dernier modèle par $\tau = Q^{-1}\theta$, on se ramène au problème d'estimation de $Q\tau$ pour le risque quadratique dans le modèle Gaussien

$$\mathcal{N}\left(\tau, \sigma^2 I_k\right).$$

Posons comme loi a priori $\tilde{\tau} \sim \mathcal{N}(0_k, vI_k)$. On a vu précédemment que

$$\tilde{\tau} \mid \tilde{Y} \sim \mathcal{N}\left(\frac{v}{v+\sigma^2}\tilde{Y}, \frac{\sigma^2 v}{v+\sigma^2} I_k\right).$$

Comme $\mathbb{E}(\|\tilde{\tau}\|^2) < +\infty$, l'estimateur bayésien est

$$\begin{aligned} T_b(\tilde{Y}) &= \left(\mathbb{E}(q(\tilde{\tau}) \mid \tilde{Y}\right) \\ &= Q\mathbb{E}(\tilde{\tau} \mid \tilde{Y}) \\ &= \frac{v}{v+\sigma^2} Q\tilde{Y}. \end{aligned}$$

Le risque a posteriori s'écrit

$$\begin{aligned}\mathrm{Tr}(\mathrm{Var}(Q\tilde{\tau} \mid \tilde{Y})) &= \frac{\sigma^2 v}{v+\sigma^2}\mathrm{Tr}(QQ^T) \\ &= \frac{\sigma^2 v}{v+\sigma^2}\mathrm{Tr}(D^{-1}) \\ &= \frac{\sigma^2 v}{v+\sigma^2}\mathrm{Tr}((X^TX)^{-1}).\end{aligned}$$

On a alors

$$\lim_{v\to+\infty}\rho(\pi_v) = \lim_{v\to+\infty}\frac{\sigma^2 v}{v+\sigma^2}\mathrm{Tr}((X^TX)^{-1}) = \sigma^2\mathrm{Tr}((X^TX)^{-1}).$$

On en déduit que $\hat{\theta}_{LS}$ est minimax optimal sur le modèle $Y = X\theta + \varepsilon$, les ε_i étant i.i.d. de moyenne nulle et variance σ^2.

Compléments : côté computationnel friendly ?

Chapitre 5

Quelques enjeux de la statistique paramétrique moderne

Sans forcément parler de "big data", on va illustrer quelques problèmes rencontrés lorsque la dimension et/ou la taille d'échantillon est trop grande pour que l'estimation paramétrique classique fonctionne, et fournir quelques solutions théoriques. Pour rester simple, on se restreindra au cadre de la régression paramétrique, c'est à dire aux modèles

$$Y = X\theta + \varepsilon,$$

où $\theta \in \mathbb{R}^k$, $\varepsilon = (\varepsilon_i)_{i=1,\dots,n}^T$ i.i.d. centrés de variance σ^2. On s'intéressera aux pertes en *prédiction* : $\|X(\hat{\theta} - \theta)\|^2$ principalement (la perte $\|\hat{\theta} - \theta\|^2$ sera appelée *perte en estimation*). On peut aussi trouver l'appellation "Mean Squared Error" (MSE) pour le risque en prédiction. Pour être un peu plus précis $\|X(\hat{\theta} - \theta)\|^2$ correspond à l'excès de risque en prédiction (risque en $\hat{\theta}$ moins risque en θ).

5.1 Problèmes liés à la dimension

5.1.1 Estimation Ridge

En termes de prédiction, on a vu que le risque optimal pour le problème de régression était

$$\inf_{\hat{\theta}} \sup_{\theta} E_\theta \|X(\hat{\theta} - \theta)\|^2 = k\sigma^2,$$

dans le cas où $X^T X$ est inversible. Si $X^T X$ n'est plus supposée inversible (par exemple pour $k > n$, la borne inférieure ne bouge pas : on a vu qu'on pouvait ramener ce problème au problème d'estimation de la moyenne d'un vecteur Gaussien de dimension celle engendrée par les colonnes de X. On a alors immédiatement

$$\inf_{\hat{\theta}} \sup_{\theta} E_\theta \|X(\hat{\theta} - \theta)\|^2 \geq r\sigma^2,$$

où r est le rang de X. On peut trouver un estimateur de type moindre carrés qui atteint cette borne. Pour cela, on prend l'inverse de Moore-Penrose de $H = X^T X$, défini par

$$H^\dagger = Q\mathrm{Diag}(((\mu_i)^{-1}\mathbb{1}_{\mu_i \neq 0} + 0\mathbb{1}_{\mu_i = 0}))_{i=1,\dots,k} Q^T,$$

si $H = Q\text{Diag}((\mu_i)_{i=1,\dots,k})Q^T$ est la décomposition spectrale de H. On peut alors définir un estimateur par moindre carrés "minimal" par

$$\hat{\theta}_{LS} = H^\dagger X^T Y,$$

de telle sorte que $X\hat{\theta}_{LS}$ soit la projection orthogonale de Y sur l'espace engendré par les colonnes de X, $V(X)$. Par ailleurs, $\hat{\theta}_{LS}$ est le u de plus petite norme vérifiant $Xu = \pi_{V(X)}Y$. De manière immédiate, pour n'importe quel u vérifiant $Xu = \pi_{V(X)}Y$, on a, pour le risque en prédiction

$$E_\theta(\|X(u-\theta)\|^2) = r\sigma^2,$$

de sorte que la borne inférieure est atteinte. Si $r < k$ on a vu que vouloir estimer θ est chimérique.

Si on ne veut pas passer par l'inversion de Moore-Penrose, une solution (très utilisée en pratique) est de biaiser légèrement le problème, en introduisant un *terme de régularisation* au problème à minimiser. On va maintenant rechercher

$$\hat{\theta}_{ridge} \in \arg\min \|Y - Xu\|^2 + \lambda\|u\|^2,$$

c'est à dire un estimateur des moindres carrés *pénalisés* (ou régularisés). Si le terme de régularisation est en norme 2, de type $\lambda\|u\|^2$, on parle d'estimateur *ridge*.

L'estimateur ridge est déterminé de manière unique par l'annulation du gradient du risque empirique pénalisé (maintenant strictement convexe). De fait, en notant $H(\lambda) = H + \lambda I_k$, on

$$\nabla_u(\|Y - Xu\|^2 + \lambda\|u\|^2) = 2(H + \lambda I_k)u - 2X^T Y,$$

et donc

$$\hat{\theta}_{ridge} = H(\lambda)^{-1}X^T Y.$$

Remarque : En pratique, on peut éviter l'inversion de $H(\lambda)$ (coûteux si $k >> n$) et se ramener à une inversion de matrice du type $(I_n + \lambda^{-1}XX^T)^{-1})$ (via l'identité de Woodbury) : en effet

$$(X^T X + \lambda I_k)X^T = X^T(XX^T + \lambda I_n),$$

ce dont on peut déduire

$$X^T((XX^T + \lambda I_n)^{-1}) = X^T H(\lambda)^{-1},$$

en multipliant à droite et à gauche par les inverses. Cela justifie l'intérêt des méthodes ridge en grande dimension.

On peut aussi calculer explicitement son risque en prédiction. Dans toute la suite, on décomposera H suivant

$$H = Q\text{Diag}((\mu_i)_{i=1,\dots,k})Q^T. \tag{5.1}$$

Proposition 5.1

En notant $\beta_i = (Q^T\theta)_i$, pour $i \in \{1,\dots,k\}$, on a

$$E_\theta(\|X(\hat{\theta}_{ridge} - \theta)\|^2) = \lambda^2 \sum_{j=1}^{k} \frac{\mu_j \beta_j^2}{(\mu_j + \lambda)^2} + \sigma^2 \sum_{j=1}^{k} \frac{\mu_j^2}{(\mu_j + \lambda)^2}.$$

Démonstration. On commence par écrire

$$\begin{aligned} X(\hat{\theta}_{ridge} - \theta) &= XH(\lambda)^{-1}X^T(X\theta + \varepsilon) - X\theta \\ &= (XH(\lambda)^{-1}H - X)\theta + XH(\lambda)^{-1}X^T\varepsilon, \end{aligned}$$

de sorte qu'on puisse décomposer

$$E_\theta \|X(\hat{\theta}_{ridge} - \theta)\|^2 = \left\|(XH(\lambda)^{-1}H - X)\theta\right\|^2 + \mathbb{E}\left\|XH(\lambda)^{-1}X^T\varepsilon\right\|^2,$$

une décomposition biais/variance habituelle. Commençons par le terme de variance. On a

$$\begin{aligned} \mathbb{E}\left\|XH(\lambda)^{-1}X^T\varepsilon\right\|^2 &= \mathbb{E}\left(\varepsilon^T XH(\lambda)^{-1}HH(\lambda)^{-1}X^T\varepsilon\right) \\ &= \sigma^2\left(\mathrm{Tr}(XH(\lambda)^{-1}HH(\lambda)^{-1}X^T\right) \\ &= \sigma^2 \mathrm{Tr}((HH(\lambda)^{-1})^2). \end{aligned}$$

Or, si $H = QDQ^T$, $H(\lambda)^{-1} = QD(\lambda)^{-1}Q^T$ (en notant $D = \mathrm{Diag}((\mu_i)_{i=1,\dots n})$ et $D(\lambda) = D + \lambda I_k$). On en déduit

$$\begin{aligned} \mathrm{Tr}((HH(\lambda)^{-1})^2) &= \mathrm{Tr}(QD^2D(\lambda)^{-2}Q^T) \\ &= \sum_{j=1}^k \frac{\mu_j^2}{(\mu_j + \lambda)^2}. \end{aligned}$$

Passons au terme de biais. On a

$$(XH(\lambda)^{-1}H - X)\theta = X(I_k - \lambda H(\lambda)^{-1} - I_k)\theta = \lambda XH(\lambda)^{-1}\theta,$$

et donc

$$\begin{aligned} \left\|(XH(\lambda)^{-1}H - X)\theta\right\|^2 &= \lambda^2\theta^T H(\lambda)^{-1}HH(\lambda)^{-1}\theta \\ &= \lambda^2\theta^T QD(\lambda)^{-1}DD(\lambda)^{-1}Q^T\theta \\ &= \lambda^2 \sum_{j=1}^k \frac{\mu_j}{(\mu_j + \lambda)^2}\beta_j^2. \end{aligned}$$

□

On peut remarquer que le terme de variance est plus petit que celui associé à celui des moindres carrés. En effet

$$\sigma^2 \sum_{j=1}^k \frac{\mu_j^2}{(\mu_j + \lambda)^2} = \sigma^2 \sum_{j=1}^r \frac{\mu_j^2}{(\mu_j + \lambda)^2} < r\sigma^2,$$

si $\lambda > 0$. La quantité $\sum_{j=1}^r \frac{\mu_j}{(\mu_j + \lambda)}$ est parfois appelé *degré de liberté effectif* de l'estimateur ridge (plus petit donc que r, celui de l'estimateur moindre carrés). Le prix à payer pour une réduction de variance est un terme de biais. On peut toutefois montrer qu'il existe un λ pour lequel on a un gain strict.

Théorème 5.2

Il existe $\lambda > 0$ tel que

$$E_\theta \|X(\hat{\theta}_{ridge} - \theta)\|^2 < r\sigma^2.$$

Démonstration. Ce λ est à chercher parmi les *lambda* petits. On remarque que le terme de biais est $O(\lambda^2)$. Le terme de variance lui peut s'écrire

$$\begin{aligned}\sigma^2 \sum_{j=1}^{k} \frac{\mu_j^2}{(\mu_j+\lambda)^2} &= \sigma^2 \sum_{j=1}^{r} \frac{1}{(1+\lambda/\mu_j)^2} \\ &= \sigma^2 \sum_{j=1}^{r} (1 - \frac{2\lambda}{\mu_j} + O(\lambda^2)) \\ &= \sigma^2 \left[r - 2\lambda \left(\sum_{j=1}^{r} \frac{1}{\mu_j} \right) \right] + O(\lambda^2),\end{aligned}$$

On a alors

$$E_\theta \|X(\hat{\theta}_{ridge} - \theta)\|^2 - r\sigma^2 = -2\lambda\sigma^2 \left(\sum_{j=1}^{r} \frac{1}{\mu_j} \right) + O(\lambda^2) < 0$$

pour λ assez petit. □

Cela ne contredit en rien le fait que $\hat{\theta}_{LS}$ soit minimax : le λ du Théorème 5.2 dépend de σ^2, H et surtout θ. De fait, on ne sert pas de ce Théorème en pratique pour calibrer un λ, on utilise plutôt des techniques de cross-validation.

On peut interpréter ce phénomène de deux autres manières.

Point de vue optimisation sous contrainte

On peut remarquer que, si $\hat{c}_\lambda = \|\hat{\theta}_{ridge}\|$, alors

$$\hat{\theta}_{ridge} = \arg \min_{\|\theta\| \leq \hat{c}_\lambda} \|Y - X\theta\|^2.$$

De fait, résoudre un problème ridge est équivalent à résoudre un problème de moindre carrés sous contrainte en norme 2, avec un rayon $\hat{c}_\lambda$ décroissant en λ. Pour se convaincre que ce point de vue justifie la possible meilleure performance du ridge en prédiction, plaçons nous dans le cas (équivalent) où on veut estimer la moyenne θ (correspond à anciennement $QD^{\frac{1}{2}}\theta$) d'un vecteur $Y \in \mathbb{R}^r$ (anciennement $D^{-\frac{1}{2}}Q^T X^T Y$, cf section sur la minimaxité de l'estimateur par moindre carrés), dans le modèle $Y = \theta + \varepsilon$, les ε_i étant i.i.d. centrés de variance σ^2.

Si on regarde

$$\hat{\theta}_M \in \arg \min_{u \in \mathrm{B}(0,M)} \|Y - u\|^2 = \pi_{\mathrm{B}(0,M)}(Y),$$

on a, dès lors que $M \geq \|\theta\|$,

$$E_\theta \|\hat{\theta}_M - \theta\|^2 = E_\theta \|\pi_{\mathrm{B}(0,M)} Y - \theta\|^2 \ < E_\theta \|Y - \theta\|^2 \mathbb{1}_{Y \notin \mathrm{B}(0,M)} + E_\theta \|Y - \theta\|^2 \mathbb{1}_{Y \in \mathrm{B}(0,M)},$$

car $\mathrm{B}(0,M)$ est strictement convexe et $\theta \in \mathrm{B}(0,M)$. On a alors que $\hat{\theta}_M$ a un meilleur risque en θ que Y si $P_\theta(Y \notin \mathrm{B}(0,M)) > 0$. Si on prend M très grand (correspond à λ tend vers 0), on retombe sur l'estimateur par moindre carrés classique. L'idéal est de connaître $\|\theta\|$ à l'avance pour pouvoir gagner à coup sûr (réduit la variance, pas de biais).

D'une certaine manière, l'estimateur ridge incorpore une information a priori sur la position de θ (dans une boule) à un estimateur fréquentiste. On peut l'interpréter comme un estimateur bayésien.

Point de vue bayésien

Revenons au modèle $Y = X\theta + \varepsilon$, où cette fois-ci $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Prenons maintenant la loi a priori $\pi(d\theta) \sim \mathcal{N}(0, \frac{1}{\lambda})$. Le modèle étant dominé, la formule de Bayes s'applique et on retrouve

$$Q_y(d\theta) \propto e^{-\frac{1}{2}\left[\|y - X\theta\|^2 + \lambda\|\theta\|^2\right]} d\theta.$$

On reconnaît une loi normale a posteriori. Plutôt que de calculer ses paramètres, remarquons que

$$\hat{\theta}_{ridge} \in \arg\max p(\theta \mid \tilde{Y}),$$

$P_{\tilde{Y}}$ presque sûrement, où $p(\theta \mid \tilde{Y})$ représente la densité de la loi a posteriori par rapport à $\mathcal{L}_k$. Cette densité a posteriori étant celle d'une Gaussienne, on en déduit que

$$\hat{\theta}_{ridge} = \mathbb{E}\left(\tilde{\theta} \mid \tilde{Y}\right),$$

et donc que l'estimateur ridge correspond à un estimateur bayésien pour la loi a priori $\mathcal{N}(0, \frac{1}{\lambda})$. Cela fournit une autre indication concernant l'estimateur ridge : il sera d'autant plus efficace que $\|\theta\|$ est petit (pour lequel on pourra choisir un grand λ).

Point de vue M-estimation sous contraintes

On va essayer d'illustrer quantitativement le fait qu'un petit $\|\theta\|$ (information connue a priori) améliore significativement la prédiction. On regarde un problème équivalent au ridge

$$\hat{\theta}_M \in \arg \min_{u \in \mathrm{B}(0,M)} \|Y - Xu\|^2,$$

pour un M quelconque. On a déjà vu qu'on pouvait concevoir les moindres carrés comme un problème de M-estimation, la fonction de risque empirique $R_n(u) = \|Y - Xu\|^2$ visant à approcher le risque idéal $\|X(u-\theta)\|^2 + n\sigma^2$. Pour appliquer les recettes de M-estimation, il faut contrôler

$$\begin{aligned}
\Delta_n &= E_\theta \sup_{u \in \mathrm{B}(0,M)} R(u) - R_n(u) \\
&= \mathbb{E} \sup_{u \in \mathrm{B}(0,M)} \|X(\theta - u)\|^2 + n\sigma^2 - \|X(\theta - u)\|^2 - \|\varepsilon\|^2 - 2\langle \varepsilon, X(\theta - u)\rangle \\
&\leq \mathbb{E} \sup_{u \in \mathrm{B}(0,M)} \langle -2\varepsilon, X(\theta - u)\rangle = \mathbb{E}\left[\langle -2\varepsilon, X\theta\rangle + 2 \sup_{u \in \mathrm{B}(0,M)} \langle \varepsilon, Xu\rangle\right] \\
&\leq 2\mathbb{E}(\|\pi_{V(X)}(\varepsilon)\|) \sup_{u \in \mathrm{B}(0,M)} \|Xu\|,
\end{aligned}$$

en utilisant l'inégalité de Cauchy-Schwartz. En utilisant l'inégalité de Jensen,

$$\mathbb{E}(\|\pi_{V(X}(\varepsilon)\|) \leq \sqrt{\mathbb{E}\|\pi_{V(X}(\varepsilon)\|^2} \leq \sigma\sqrt{r}.$$

Par ailleurs, pour $u \in \mathrm{B}(0, M)$, on a $\|Xu\| \leq \sqrt{\mu_1} M$ (en rappelant que μ_1 est la plus grande valeur propre de $H = X^T X$). Mis bout à bout, on obtient

$$\Delta_n \leq 2M\sigma\sqrt{r\mu_1}.$$

Il reste à dérouler le fil de la M-estimation. On note

$$\theta_M^* \in \arg\min_{\mathrm{B}(0,M)} R(u) = \arg\min_{u\in\mathrm{B}(0,M)} \|X(\theta-u)\|^2,$$

c'est à dire un des meilleur u atteignable au sens du risque en prédiction sur $\mathrm{B}(0,M)$.

On peut écrire

$$\begin{aligned} E_\theta(\|X(\hat\theta_M-\theta)\|^2) &= E_\theta R(\hat\theta_M) - R(\theta) \qquad (5.2)\\ &= E_\theta(R_n(\hat\theta_M + (R-R_n)(\hat\theta_M)) - R(\theta)\\ &\le E_\theta(R_n(\theta_M^*)) - R(\theta) + \Delta_n\\ &\le \inf_{u\in\mathrm{B}(0,M)} \|X(\theta-u)\|^2 + 2M\sigma\sqrt{r\mu_1}. \qquad (5.3)\end{aligned}$$

Une telle inégalité est appelée *inégalité oracle* : elle compare le risque d'un estimateur au meilleur risque atteignable sur une classe (inconnu, sauf si on a un oracle sous la main).

Le premier terme

$$\inf_{u\in\mathrm{B}(0,M)} \|X(\theta-u)\|^2$$

est un terme de "biais du modèle", traduisant le fait que l'optimal atteignable θ_M^* par notre estimateur peut être différent de la cible θ. On parle alors d'*erreur d'approximation*.

Le deuxième terme

$$2M\sigma\sqrt{r\mu_1}$$

est un terme de variance (uniforme entre risque et risque empirique sur la classe considérée). Ce terme est aussi appelé *erreur d'estimation*.

Pour le choix optimal $M = \|\theta\|$ (inatteignable en pratique), on a

$$E_\theta(\|X(\hat\theta_M-\theta)\|^2) \le 2\theta\sigma\sqrt{r\mu_1},$$

ce dont on déduit que plus θ est petit, plus performante sera l'estimation sous contrainte (et donc le ridge).

Remarque : On peut déduire de ce qui précède une majoration de la vitesse minimax sur $\{\theta \in \mathrm{B}(0,M)\}$:

$$\sup_{\theta\in\mathrm{B}(0,M)} E_\theta(\|X(\hat\theta_M-\theta)\|^2) \le 2M\sigma\sqrt{r\mu_1},$$

passant en dessous de $r\sigma^2$ pour M assez petit. Le problème restreint est donc plus "facile" statistiquement que le problème non restreint dans ce cas. On peut remarquer aussi $\{\theta \in \mathrm{B}(0,M)\}$ est un peu trop restrictif : on pourrait se contenter de $\{\theta^* \in \mathrm{B}(0,M)\}$, où

$$\theta^* \in \arg\min_{Xu=X\theta} \|u\|^2,$$

soit le vecteur de plus petite norme donnant la même prédiction que θ.

Remarque 2 : Le choix du M en pratique se fait par validation croisée. Théoriquement parlant, on pourrait utiliser une stratégie par pénalisation, comme ce qu'on va voir juste après.

5.1.2 Parcimonie et sélection de modèle

On se place maintenant dans le modèle de régression linéaire Gaussien, c'est à dire

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Dans le modèle de régression, faire une hypothèse de parcimonie consiste à supposer que, parmi les k variables explicatives $X_{.,j}$, seulement s sont pertinentes, avec $s \leq k$. En d'autre termes, cela revient à supposer

$$\|\theta_0\| \leq s,$$

pù $\|\theta\|_0 = \sum_{j=1}^k \mathbb{1}_{\theta_j \neq 0} = |\text{Supp}(S)|$. Comme on l'a vu pour le ridge, même si un "vrai" θ ne satisfait pas une hypothèse de parcimonie, il peut être intéressant de trouver un estimateur "biaisé" parcimonieux de risque en prédiction potentiellement plus faible.

Si on connaissait le support S de θ, pour peu que $X\theta$ soit de rang s on pourrait utiliser une méthode par moindre carrés sur le sous-modèle $Y = X_S\theta_S + \varepsilon$ (θ_S étant le vecteur formé par les composantes non-nulles de θ, X_S la sous-matrice de X formé des s colonnes correspondantes), et récupérer une erreur en prédiction en $\sigma^2 s$, ainsi qu'une erreur en estimation en $\sigma^2 \text{Tr}((X_S^T X_S)^{-1})$. Tout le problème vient du fait qu'on ne connaît pas le support a priori.

Du point de vue estimation, on pourrait s'en sortir en testant successivement $X\theta \in V^{-j}(X)$, où $V^{-j}(X)$ est le sev engendré par les $X\theta$ pour les θ ayant une j-ème composante nulle. Une fois cela fait, cela fournirait un estimateur du support $\hat{S}$ (qui vaudrait S avec forte proba), puis effectuer une régression par moindre carrés avec contrainte de support.

Toutefois cette approche n'est pas toujours optimale en prédiction : on ne peut pas biaiser nos estimateurs en utilisant ceci. Si on s'intéresse à l'approche en prédiction uniquement (ce qu'on supposera à partir de maintenant), l'idéal serait de connaître, pour chaque support potentiel S,

$$R(\hat{\theta}_S) = \|X(\theta - \hat{\theta}_S)\|^2 + n\sigma^2,$$

où $\hat{\theta}_S$ est obtenu par moindre carrés sur les variables dans S, puis de comparer ces $R(\hat{\theta}_S)$ pour identifier un S^* optimal via

$$S^* \in \arg\min_{|S| \leq s} R(\hat{\theta}_S),$$

et l'estimateur $\hat{\theta}_{S^*}$ correspondant. Si θ est de support S de taille s, cela garantirait en particulier que $R(\hat{\theta}_{S^*}) \leq R(\hat{\theta}_S) \leq R(\theta) + s\sigma^2$, avec un potentiel gain en prédiction induit par le biais. Dans le cadre général, on aurait

$$R(\hat{\theta}_{S^*}) - R(\theta) \leq \inf_{|S| \leq s} R(\theta_S^*) - R(\theta) + |S|\sigma^2,$$

soit un gain potentiel en prédiction en introduisant un biais ($R(\theta_S^*)$ représente ici le risque en prédiction optimal parmi les estimateurs de support S).

On ne dispose pas de $R(\hat{\theta}_S)$ mais d'une approximation $R_n(\hat{\theta}_S)$: $R_n(\hat{\theta}_S) = \|Y - X_S\hat{\theta}_S\|^2$ vérifie

$$E_\theta(R_n(\hat{\theta}_S)) = R(\hat{\theta}_S) - r(S)\sigma^2,$$

où $r(S)$ est le rang de X_S. Le problème est que pour comparer différents supports, on a plutôt besoin de majorer

$$E_\theta(\sup_S R(\hat{\theta}_S) - R_n(\hat{\theta}_S)),$$

d'où l'idée de la *pénalisation* : on va essayer de trouver $pen(S)$ tel que, uniformément en S, on ait

$$\begin{aligned} R(\hat{\theta}_S) &\leq R_n(\hat{\theta}_S) + pen(S), \\ R_n(\theta^*_S) &\leq R(\theta^*_S) + pen(S), \end{aligned}$$

sur un évènement A que l'on espère de grosse proba. En effet, si on définit

$$\hat{S} = \arg\min_S R_n(\hat{\theta}_S) + pen(S),$$

sur l'évènement A on a immédiatement, pour n'importe quel concurrent S,

$$\begin{aligned} R(\hat{\theta}_{\hat{S}}) &\leq R_n(\hat{\theta}_{\hat{S}}) + pen(\hat{S}) \leq R_n(\hat{\theta}_S) + pen(S) \\ &\leq R_n(\theta^*_S) + pen(S) \leq R(\theta^*_S) + 2pen(S) \\ &\leq \inf_S R(\theta^*_S) + 2pen(S). \end{aligned}$$

De fait, on peut trouver une telle pénalité, mais la preuve est un peu plus complexe que l'heuristique décrite plus haut.

Théorème 5.3

Dans le modèle Gaussien homoscédastique, si on définit

$$\hat{S} \in \arg\min_{S \subset \{1,\dots,k\}} R_n(\hat{\theta}_S) + pen(S),$$

avec

$$pen(S) \geq 4\sigma^2 r(S) + 8\sigma^2 \left((|S| + 1) \log(k) \right),$$

on obtient, avec probabilité plus grande que $1 - 3e^{-x}$,

$$R(\hat{\theta}_{\hat{S}}) - R(\theta) \leq 2 \min_{S \subset \{1,\dots,k\}} [R(\theta^*_S) - R(\theta) + 2pen(S)] + 32\sigma^2 x.$$

En particulier

$$\mathbb{E}\left(R(\hat{\theta}_{\hat{S}}) - R(\theta) \right) \leq 2 \min_{S \subset \{1,\dots,k\}} [R(\theta^*_S) - R(\theta) + 2pen(S)] + 96\sigma^2.$$

Démonstration. On va utiliser un lemme de concentration uniforme, pas optimal sous bien des aspects.

Lemme 5.4

Soit $x > 0$. Alors, uniformément en $S \subset \{1, \dots, k\}$, on a

$$\|\varepsilon_{V(X_S)}\| \leq \sigma\sqrt{2r(S) + 3(x + (|S| + 1)\log(k))}$$
$$|2\langle \varepsilon, X(\theta - \theta^*_S)\rangle| \leq 2\sigma\|X(\theta - \theta^*_S)\|\sqrt{2(x + (|S| + 1)\log(k))},$$

avec probabilité plus grande que $1 - 3e^{-x}$.

Démonstration. Pour la première inégalité on se base sur une inégalité de concentration pour la loi du χ^2 : si $X \sim \chi^2(p)$, alors

$$\mathbb{P}\left(X \geq 2p + 3x\right) \leq e^{-x}.$$

On peut trouver une preuve de cette inégalité dans *Estimation of a quadratic functional*, B. Laurent et P. Massart, inégalité 4.3. Soit S un support fixé. Comme $\|\|\varepsilon_{V(X_S)}\|^2 \sim \sigma^2\chi^2(r(S))$, on a

$$\mathbb{P}\left(\|\varepsilon_{V(X_S)}\| \geq \sigma\sqrt{2r(S) + 3(x + (|S| + 1)\log(k))}\right) \leq k^{-(|S|+1)}e^{-x}.$$

Notons $\mathcal{S}_j = \{S \mid |S| = j\}$. On a

$$|\mathcal{S}_j| = \binom{k}{j} \leq k^j.$$

On en déduit

$$\mathbb{P}\left(\exists S \in \mathcal{S}_j \mid \|\varepsilon_{V(X_S)}\| \geq \sigma\sqrt{2r(S) + 3(x + (|S| + 1)\log(k))}\right) \leq k^j \times k^{-(j+1)}e^{-x} = k^{-1}e^{-x}.$$

En prenant une borne d'union en $j = 1, \dots, k$, on obtient

$$\mathbb{P}\left(\exists S \subset \{1, \dots, k\} \mid \|\varepsilon_{V(X_S)}\| \geq \sigma\sqrt{2r(S) + 3(x + (|S| + 1)\log(k))}\right) \leq kk^{-1}e^{-x} = e^{-x}.$$

Passons à la deuxième inégalité. On se base ici sur l'inégalité de concentration Gaussienne standard : si $N \sim \mathcal{N}(0, v)$, alors

$$\mathbb{P}(\left(|N| \geq \sqrt{2vx}\right) \leq 2e^{-x}.$$

Commençons par fixer S. Comme $\langle \varepsilon, X(\theta - \theta^*_S)\rangle \sim \mathcal{N}(0, \sigma^2\|X(\theta - \theta^*_S)\|^2)$, une application de l'inégalité de concentration Gaussienne standard donne

$$\mathbb{P}\left(|2\langle \varepsilon, X(\theta - \theta^*_S)\rangle| \geq 2\sigma\|X(\theta - \theta^*_S)\|\sqrt{2(x + (|S| + 1)\log(k))}\right) \leq 2k^{-(|S|+1)}e^{-x}.$$

En utilisant une borne d'union comme précédemment on en déduit le résultat. □

De ces inégalités, on déduit que sur un évènement A de probabilité supérieure à $1-3e^{-x}$, on a, uniformément en S, et en u de support S,

$$\begin{aligned}
&|(R-R_n)(u)-(R-R_n)(\theta)| = |2\langle\varepsilon, X(u-\theta)\rangle| \\
&\quad\leq 2\,|\langle\varepsilon, X(\theta_S^*-\theta)\rangle| + 2\,|\langle\varepsilon, X(u-\theta_S^*)\rangle| \\
&\quad\leq 2\sigma\|X(\theta-\theta_S^*)\|\sqrt{2(x+(|S|+1)\log(k))} + 2\left|\left\langle\pi_{V(X_S)}\varepsilon, X(u-\theta_S^*)\right\rangle\right| \\
&\quad\leq 2\sigma\|X(\theta-\theta_S^*)\|\sqrt{2(x+(|S|+1)\log(k))} + 2\|X(u-\theta_S^*)\|\|\pi_{V(X_S)}\varepsilon\| \\
&\leq \frac{1}{2}(\|X(\theta-\theta_S^*)\|^2+\|X(u-\theta_S^*)\|^2) + 2\sigma^2(x+(|S|+1)\log(k)) + 2\|\pi_{V(X_S)}\varepsilon\|^2 \\
&\leq \frac{1}{2}(R(u)-R(\theta)) + 4\sigma^2 r(S) + 8\sigma^2(x+(|S|+1)\log(k)) \\
&\leq \frac{1}{2}(R(u)-R(\theta)) + pen(S) + 8\sigma^2 x.
\end{aligned}$$

On en déduit alors, sur l'évènement A,

$$\begin{aligned}
R(\hat\theta_{\hat S})-R(\theta) &\leq R_n(\hat\theta_{\hat S})-R_n(\theta)+|(R-R_n)(\hat\theta_{\hat S})-(R-R_n)(\theta)| \\
&\leq \frac{1}{2}\left(R(\hat\theta_{\hat S})-R(\theta)\right) + R_n(\hat\theta_{\hat S}) + pen(\hat S) - R_n(\theta) + 8\sigma^2 x,
\end{aligned}$$

et donc, pour n'importe quel $S\subset\{1,\dots,k\}$,

$$\begin{aligned}
\frac{1}{2}\left(R(\hat\theta_{\hat S})-R(\theta)\right) &\leq R_n(\hat\theta_S)-R_n(\theta)+pen(S)+8\sigma^2x \\
&\leq R_n(\theta_S^*)-R_n(\theta)+pen(S)+8\sigma^2x \\
&\leq R(\theta_S^*)-R(\theta)+|(R-R_n)(\theta_S^*)-(R-R_n)(\theta)|+pen(S)+8\sigma^2x \\
&\leq R(\theta_S^*)-R(\theta)+2pen(S)+16\sigma^2x.
\end{aligned}$$

On en conclut que

$$R(\hat\theta_{\hat S})-R(\theta) \leq 2\min_{S\subset\{1,\dots,k\}}\left[R(\theta_S^*)-R(\theta)+2pen(S)\right]+32\sigma^2x.$$

□

En particulier, si $\|\theta\|_0 = s \geq 1$, le Théorème 5.3 implique que, en choisissant les pénalités minimales

$$\mathbb{E}(R(\hat\theta_{\hat S})-R(\theta)) \leq C\sigma^2 s\log(k).$$

L'hypothèse de parcimonie permet alors de réduire l'influence de la dimension : on passe de $r\sigma^2$ (donc potentiellement $k\sigma^2$ si X est de plein rang) à $C\sigma^2\log(k)$, voire moins si un estimateur biaisé de plus petit support fait mieux. Cette propriété est particulièrement intéressante en grande dimension.

Remarque : On a montré que si on pénalisait suffisamment fortement, on arrivait grosso modo à retrouver le support (voire à faire mieux) du θ sous-jacent. On peut montrer une réciproque : une sous-pénalisation (de l'ordre de $r(S)\sigma^2$ par exemple, correspondant aux espérances des déviations sur chaque sous-modèle, non-uniformément) mène à la sélection de beaucoup trop de variables avec grosse probabilité, et donc à un risque en $k\sigma^2$ (si X est de rang plein). On pourra par exemple

se référer à la Proposition 4.3 du *Concentration inequalities and Model Selection*, P. Massart.

Remarque 2 : On peut montrer que la vitesse minimax sur l'ensemble des θ à support de taille plus petite que s est de l'ordre de

$$cs\sigma^2 \log(k/s).$$

On peut trouver ce résultat dans *Minimax risk for sparse regression : Ultra-High dimensional phenomenons*, N. Verzelen. À un facteur $\log(s)$ près le Théorème 5.3 fournit donc les bons ordres de grandeur. Évidemment les constantes de ce théorème sont clairement sous-optimales, mais même en les optimisant on ne pourrait atteindre cette borne en toute généralité. De fait, les estimateurs optimaux sont plutôt obtenus par aggrégation d'estimateurs de faible dimension dans ce cas (voir par exemple *Sharp oracle inequalities for aggregation of affine estimators*, A. Dalayan et J. Salmon.

Remarque 3 : Plutôt que de parler de variables, on peut aussi parler de "prédicteur individuels". Si on a $f_1, \dots, f_k$ prédicteurs de base, avec des arguments X_i fixés, chercher la meilleure combinaison linéaire au sens des moindres carrés pour approcher $Y = (f(X_1), \dots, f(X_n))^T + \varepsilon$ parmi les

$$f_\theta(X_1, \dots, X_n) = \sum_{j=1}^{k} \theta_j (f_j(X_1), \dots, f_j(X_n))^T,$$

on peut réécrire le problèmes des moindre carrés (éventuellement pénalisés) sous la forme

$$\arg\min_\theta \|Y - X'\theta\|^2 + pen(\theta),$$

avec $X'_{i,j} = f_j(X_i)$. Dans ce domaine $\{f_1, \dots, f_k\}$ est appelé un dictionnaire, et retrouver un sous-ensemble parcimonieux de ce dictionnaire restant performant en prédiction est le domaine du *sparse dictionary learning*.

Lien avec le "seuillage dur"

Regardons de plus près les solutions "pratiques" de

$$\arg\min_S \|Y - X\hat{\theta}_S\|^2 + \lambda\sigma^2|S|,$$

pour un λ donné. Regardons le $\hat{S}$ sélectionné, de cardinal $\hat{s}$. On a alors, pour toute variable $j \in \hat{S}$,

$$\|Y - \pi_{V(X_{\hat{S}})}Y\|^2 + \sigma^2\lambda\hat{s} < \|Y - \pi_{V(X_{\hat{S}-j})}Y\|^2 + \sigma^2\lambda(\hat{s} - 1),$$

ce qui équivaut à

$$\|\pi_{V(X_{\hat{S}})}Y - \pi_{V(X_{\hat{S}-j})}Y\|^2 > \sigma^2\lambda.$$

En particulier, on doit avoir $X_{\hat{S}}$ de rang $\hat{s}$, et donc $\hat{\theta}_j \neq 0$ (on rappelle ici que par convention quand X n'est pas inversible on prend le θ de plus petite norme qui donne $\pi_{V(X)}Y = X\theta$). Par ailleurs, on peut remarquer que

$$\pi_{V(X_{\hat{S}})}Y = X_{\hat{S}}\hat{\theta}_{\hat{S}} = X_{\hat{S}-j}\hat{\theta}^{-j}_{\hat{S}} + X^j\hat{\theta}_j,$$

où X^j est la j-ème colonne de X. On en déduit

$$\|\pi_{V(X_{\hat{S}})}Y - \pi_{V(X_{\hat{S}-j})}Y\|^2 \leq \hat{\theta}_j^2 \|X^j\|^2,$$

et donc nécessairement

$$|\hat{\theta}_j| > \frac{\sigma\sqrt{\lambda}}{\|X^j\|}.$$

Dans le cas où X est formée de vecteurs orthonormés, cette condition nécessaire est suffisante, et la pénalisation ℓ_0 est équivalente à regarder $\hat{\theta}$ l'estimateur par moindre carrés sur $V(X)$, et à seuiller les coefficients, c'est à dire

$$\hat{\theta}_{\hat{S}} = \left(\hat{\theta}_j \mathbb{1}_{|\hat{\theta}_j| > \sigma\sqrt{\lambda}}\right)_{j=1,\dots,k}^T.$$

Concluons qu'en pratique la pénalisation ℓ_0 est utilisée soit dans un cadre orthonormé où elle se ramène à un seuillage (dans un cadre fonctionnel par exemple), soit via des approximations itératives type AIC ou BIC : on part du modèle complet $\|Y - X\hat{\theta}\|^2$ est on enlève une à une les variables qui peuvent l'être (au sens des moindres carrés pénalisés).

Lorsque k est vraiment grand, cette approche est impraticable, et on se tourne alors vers une relaxation convexe de ce problème.

5.1.3 LASSO

La sélection de modèle de la section précédente peut se réécrire sous la forme

$$\hat{\theta} \in \arg\min_u \|Y - Xu\|^2 + \lambda\|u\|_0,$$

pour un λ assez grand. Comme cette fonction de θ n'est pas convexe, on en est réduit à parcourir les sous-espaces à support fixés ou éventuellement à seuiller dans les cas favorables.

L'idée du Lasso est de prendre l'enveloppe convexe de $\theta \mapsto \|\theta\|_0$, qui est juste $\theta \mapsto \|\theta\|_1$ (faire un dessin en dimension 1 pour s'en convaincre). Pour rappel, l'enveloppe convexe de f est définie par $f_c(x) = \sup\{g(x) \mid g \text{ est convexe et } g \leq f\}$.

On va donc chercher une solution de

$$\arg\min_u \|Y - Xu\|^2 + \lambda\|u\|_1.$$

Cette fonction est propre et convexe, ses minimums sont donc nécessairement atteints. Soit $\hat{\theta}$ un tel minimum. On va regarder des conditions nécessaires coordonnée par coordonnée. On note e_j le vecteur $(0,\dots,0,1,0,\dots,0)^T$, avec le 1 à la j-ème place, et f la fonction à minimiser. Supposons $\hat{\theta}_j \neq 0$, et prenons $h \in \mathbb{R}$ petit. On a alors

$$f(\hat{\theta} + he_j) - f(\hat{\theta}) = -2\left\langle X^j, Y\right\rangle h + 2(X^T X\hat{\theta})_j h + \lambda h \mathrm{sg}(\hat{\theta}_j) + O(h^2) \geq 0,$$

où $\mathrm{sg}(x)$ représente le signe de x, à valeurs dans $\{-1, 0, 1\}$. On en déduit alors que nécessairement

$$(X^T X\hat{\theta})_j = \left\langle X^j, Y\right\rangle - \frac{\lambda}{2}\mathrm{sg}(\hat{\theta}_j). \tag{5.4}$$

Supposons maintenant que $\hat{\theta}_j = 0$ et prenons h positif. On a alors

$$f(\hat{\theta} + he_j) - f(\hat{\theta}) - 2\left\langle X^j, Y\right\rangle h + 2(X^T X\hat{\theta})_j h + \lambda h + O(h^2) \geq 0,$$

ce dont on déduit

$$\left\langle X^j, Y\right\rangle - (X^T X\hat{\theta})_j \leq \frac{\lambda}{2}.$$

Pour h négatif, le même raisonnement donne

$$\left\langle X^j, Y\right\rangle - (X^T X\hat{\theta})_j \geq -\frac{\lambda}{2}.$$

Les deux équations combinées donnent

$$\left|\left\langle X^j, Y\right\rangle - (X^T X\hat{\theta})_j\right| \leq \frac{\lambda}{2}. \tag{5.5}$$

On peut prouver (en regardant les conditions de Karusch-Kuhn-Tucker) que les conditions nécessaires (5.4) et (5.5) sont suffisantes, et qu'elles se résument à

$$0 \in \partial_{\hat{\theta}} f,$$

où $\partial_x f$ représente le sous-gradient de f en x, c'est à dire l'ensemble des directions h telles que pour tout y dans un voisinage de x,

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

Dans le cas orthonormé où $X^T X = I_k$, pour un $\hat{\theta}_j$ non nul, la condition (5.5) devient

$$\hat{\theta}_j + \frac{\lambda}{2}\mathrm{sg}(\hat{\theta}_j) = \left\langle X^j, Y\right\rangle .$$

On en déduit que $\mathrm{sg}(\hat{\theta}_j) = \mathrm{sg}(\langle X^j, Y\rangle)$, et

$$\hat{\theta}_j = \mathrm{sg}(\left\langle X^j, Y\right\rangle)(\left|\left\langle X^j, Y\right\rangle\right| - \frac{\lambda}{2}).$$

Les coordonées où $\hat{\theta}_j = 0$ vérifient

$$\left|\left\langle X^j, Y\right\rangle\right| \leq \frac{\lambda}{2}.$$

Ces conditions étant suffisantes, on déduit alors que (toujours dans le cas orthonormé),

$$\hat{\theta}_j = \mathrm{sg}(\left\langle X^j, Y\right\rangle)\left(\left|\left\langle X^j, Y\right\rangle\right| - \frac{\lambda}{2}\right)_+ = \mathrm{sg}(\hat{\theta}_{LS,j})\left(\left|\hat{\theta}_{LS,j}\right| - \frac{\lambda}{2}\right)_+,$$

où $\hat{\theta}_{LS}$ désigne l'estimateur par moindre carrés standard. L'estimateur Lasso dans ce cas effectue un seuillage doux : les coefficients trop petits de $\hat{\theta}_{LS}$ seront mis à 0 (comme en pénalisation ℓ_0), ceux suffisamment grands seront décalés vers 0 d'un facteur $\lambda/2$.

On peut aussi se représenter la parcimonie induite par la régularisation ℓ_1 en regardant le problème d'optimisation sous contrainte

$$\min_{\|u\|_1 \leq c_\lambda} \|Y - Xu\|^2,$$

équivalent au problème régularisé, et en faisant un dessin (la boule en norme 1 possède des "coins").

FAIRE LES 3 DESSINS : min dans la boule l1, lignes de niveaux ellipses qui tapent un coin, lignes de niveau l1 qui tapent un milieur de segment.

Performances en prédiction du Lasso

L'idée à retenir est que les performances en prédiction du Lasso sont conditionnées au fait que les colonnes de X sont à peu près orthogonales. On va donc supposer pour simplifier que les X^j sont normés (en pratique on peut toujours le faire), mais surtout que

$$\lambda_{min}(X^T X) \geq (1-\delta),$$

ce qui en particulier implique que

$$\|Xv\|^2 \geq (1-\delta)\|v\|^2,$$

une condition de quasi-isométrie. Sous ces conditions, on a le résultat suivant (dans le modèle Gaussien toujours).

Théorème 5.5

Si $\lambda \geq 2\sigma\sqrt{2(x+\log(k))}$, *on a, avec probabilité plus grande que* $1-2e^{-x}$,

$$\|X(\hat{\theta}_{LASSO}-\theta)\|^2 \leq \inf_u \left[\|X(\theta-u)\|^2 + \frac{\lambda^2\|u\|_0}{1-\delta}\right].$$

Démonstration. On aura besoin de deux lemmes.

Lemme 5.6

Si $\lambda \geq 2\sigma\sqrt{2(x+\log(k))}$, *alors*

$$\mathbb{P}\left(\|X^T\varepsilon\|_\infty \geq \frac{\lambda}{2}\right) \leq 2e^{-x}.$$

Démonstration. On se sert encore de la concentration Gaussienne : comme, pour $j=1,\dots,k$,

$$Z_j = (X^j)^T\varepsilon \sim \mathcal{N}(0,\sigma^2\|X^j\|^2) = \mathcal{N}(0,\sigma^2),$$

on a $\mathbb{P}(|Z_j| \geq \sigma\sqrt{2x}) \leq 2e^{-x}$. Une borne d'union donne le résultat. □

Lemme 5.7

Soit f *est une fonction convexe sur* $\mathbb{R}^k$. *Alors, pour tous* $x, y \in \mathbb{R}^k$ *et* $g_x \in \partial_x f$, $g_y \in \partial_y f$, *on a*

$$\langle g_y - g_x, y-x\rangle \geq 0.$$

Démonstration. Par définition, $f(y)-f(x) \geq \langle g_x, (y-x)\rangle$, et $f(x)-f(y) \geq \langle g_y, x-y\rangle$. Il suffit d'additionner. □

On passe à la preuve du Théorème. Dans le cadre d'une fonction de perte générale il faudrait passer par des majorations uniformes de processus (comme pour le Théorème 5.3), ce cas est notamment traité dans les ouvrages de S. Van de Geer. Pour les moindres carrés, on peut s'en sortir avec une astuce algébrique.

Soit $\hat{\theta}$ l'estimateur Lasso. Les conditions d'optimalité s'écrivent alors

$$2X^T(Y - X\hat{\theta}) = \lambda \hat{g},$$

où $\hat{g} \in \partial_{\hat{\theta}} = \|.\|_1$. On peut alors écrire, pour un concurrent potentiel $u \in \mathbb{R}^k$,

$$2\left\langle X^T Y, \hat{\theta} - u \right\rangle - 2\left\langle X^T X\hat{\theta}, \hat{\theta} - u \right\rangle = \lambda \left\langle \hat{g}, \hat{\theta} - u \right\rangle,$$

soit

$$2\left\langle Y, X(\hat{\theta} - u \right\rangle - 2\left\langle X\hat{\theta}, X(\hat{\theta} - u) \right\rangle = \lambda \left\langle \hat{g}, \hat{\theta} - u \right\rangle.$$

Comme $Y = X\theta + \varepsilon$, on obtient

$$2\left\langle X(\hat{\theta} - \theta), X(\hat{\theta} - u) \right\rangle = 2\left\langle \varepsilon, X(\hat{\theta} - u) \right\rangle - \lambda \left\langle \hat{g}, \hat{\theta} - u \right\rangle.$$

Le terme de gauche peut s'écrire $\|X(\hat{\theta} - \theta)\|^2 + \|X(\hat{\theta} - u)\|^2 - \|X(u - \theta)\|^2$. Par ailleurs, si $g \in \partial_u \|.\|_1$, on a $\left\langle \hat{g}, \hat{\theta} - u \right\rangle \geq \left\langle g, \hat{\theta} - u \right\rangle$, d'après le Lemme 5.7. On en déduit

$$\|X(\hat{\theta} - \theta)\|^2 + \|X(\hat{\theta} - u)\|^2 \leq \|X(u - \theta)\|^2 - \lambda \left\langle g, \hat{\theta} - u \right\rangle + 2\left\langle \varepsilon, X(\hat{\theta} - u) \right\rangle.$$

On explicite maintenant un choix de $g \in \partial_u \|.\|_1$. En notant J le support de u, n'importe quel $g = \mathrm{sg}(u) + h$, avec $h_J = 0$ et $\|h\|_\infty \leq 1$ convient. On prend $h = \mathrm{sg}(\hat{\theta} - u)_{J^c}$, et alors

$$\left\langle g, \hat{\theta} - u \right\rangle \geq -\|(\hat{\theta} - u)_J\|_1 + \|(\hat{\theta} - u)_{J^c}\|_1.$$

Cela donne

$$\|X(\hat{\theta} - \theta)\|^2 + \|X(\hat{\theta} - u)\|^2 \leq \|X(u - \theta)\|^2 + 2\left\langle \varepsilon, X(\hat{\theta} - u) \right\rangle + \lambda\|(\hat{\theta} - u)_J\|_1 - \lambda\|(\hat{\theta} - u)_{J^c}\|_1$$

On travaille maintenant le terme en ε. On a

$$2\left\langle \varepsilon, X(\hat{\theta} - u) \right\rangle \leq 2\|X^T\varepsilon\|_\infty \|(\hat{\theta} - u)\|_1 \leq \lambda\left(\|(\hat{\theta} - u)_J\|_1 + \|(\hat{\theta} - u)_{J^c}\|_1\right),$$

d'après le Lemme 5.6. On en déduit

$$\|X(\hat{\theta} - \theta)\|^2 + \|X(\hat{\theta} - u)\|^2 \leq \|X(u - \theta)\|^2 + 2\lambda\|(\hat{\theta} - u)_J\|_1.$$

Enfin, on remarque que

$$\|(\hat{\theta} - u)_J\|_1 \leq \sqrt{|J|}\|\hat{\theta} - u\|,$$

et donc

$$\begin{aligned} 2\lambda\|(\hat{\theta} - u)_J\|_1 &\leq 2\lambda\sqrt{|J|}\|\hat{\theta} - u\| \\ &\leq \frac{\lambda^2|J|}{1-\delta} + (1-\delta)\|\hat{\theta} - u\|^2 \\ &\leq \frac{\lambda^2|J|}{1-\delta} + \|X(\hat{\theta} - u)\|^2, \end{aligned}$$

par hypothèse de quasi-isométrie. On en déduit

$$\|X(\hat{\theta} - \theta)\|^2 \leq \|X(u - \theta)\|^2 + \frac{\lambda^2\|u\|_0}{1-\delta}.$$

□

On remarque que si $\|\theta_0\| = s$, le choix optimal $\lambda = 2\sigma\sqrt{2(x + \log(k))}$ donne

$$\|X(\hat{\theta}_{LASSO} - \theta)\|^2 \leq C\sigma^2 \frac{s(\log(k) + x)}{1 - \delta},$$

et on retrouve l'ordre de grandeur des vitesses minimax sur cette classe. Donc, si X est proche d'être orthogonale, la pénalisation ℓ_1 donne des performances comparables à la pénalisation ℓ_0, tout en étant réalisable. Le point crucial est cette condition d'orthogonalité. Elle peut être largement assouplie, mais ne reste pas gratuite (on réfère au cours de C. Giraud, *Introduction to High-Dimensional Statistics* pour le lecteur intéressé).

Prix à payer pour la relaxation convexe

Le premier défaut structurel du Lasso se rencontre dans les situations à forte colinéarité entre colonnes. Prenons le cas extrême où $\theta = (\theta_1, 0, \ldots, 0)^T$ et $X^1 = X^2$. Pour le critère idéal $\|X(\theta - u)\|^2$ pénalisé en norme 1, n'importe quel u tel que $u_1 + u_2 = \theta_1$ est convenable, là où la pénalisation ℓ_0 privilégiera $u_1 = 0$ ou $u_2 = 0$. En pratique, le Lasso sélectionne non-seulement les variables supports mais aussi les variables fortement corrélées avec, et fournit donc plutôt une "borne sup" sur le support (les conditions de corrélation entre variable garantissant de trouver le bon support peuvent être trouvées dans *On model selection consistency of the lasso*, P. Zhao et B. Yu). Néanmoins, on peut se servir de cette première approximation de support pour refaire tourner une procédure plus coûteuse (comme du ℓ_0) sur les variables sélectionnées pour pallier ce défaut).

Le deuxième défaut tient au rétrécissement des coefficients non-nuls vers 0. Dans le cas orthogonal, les coefficients non nuls sont $\mathrm{sg}(\hat{\theta}_{LS,j})\left(\left|\hat{\theta}_{LS,j}\right| - \frac{\lambda}{2}\right)$, dont les prédictions en performance sont souvent moins bonnes que les moindres carrés standards sur le support sélectionné. Une pratique courante consiste alors à réajuster un moindre carrés standard (ou ridge) sur les variables sélectionnées par le Lasso.

Terminons enfin sur le choix du paramètre λ. En pratique, la "trajectoire" entière pour λ dans un intervalle peut se calculer par l'algorithme LARS, on peut alors chercher des "sauts" dans les ensembles sélectionnés et tester plusieurs candidats par cross-validation par exemple.

5.2 Problèmes liés à la taille d'échantillon

Lorsque la taille d'échantillon n devient grande, plusieurs problèmes se posent en pratique.

— Celui de la *complexité algorithmique* des estimateurs proposés : pour n de l'ordre de 10000 (situation courante), les complexité d'ordre linéaire en n sont à privilégier.

— Celui de la mémoire à utiliser : si l'estimateur proposé nécessite le stockage en mémoire vive de l'ensemble des données (pouvant atteindre plusieurs Go assez facilement), vous risquez de ne pas pouvoir le calculer.

D'un point de vue pratique, les estimateurs à faible complexité, et pouvant être parallélisable sont privilégiés. C'est de nos jours une caractéristique cruciale à rechercher lorsque l'on conçoit une méthode.

D'un point de vue théorique, sur le principe on peut contraindre le risque minimax sur une classe à un risque minimax pris sur les estimateurs calculables en temps polynomial par exemple. Peu de bornes inférieurs ont été prouvées à ce jour, à l'exception notable de *Lower bounds on the performance of polynomial-time algorithms for sparse linear regression*, Y. Zhang, M. Wainwright, M. Jordan, qui montre grosso modo que les bornes supérieures pour le LASSO sont les bornes inférieures prises sur l'ensemble des estimateurs calculables en temps polynomial. On rappelle que ce n'est pas le risque minimax optimal pris sur l'ensemble des estimateurs.

5.2.1 Exemple : régression linéaire, design aléatoire

On se place encore dans le modèle

$$Y_i = \langle \theta, X_i \rangle + \varepsilon_i,$$

mais cette fois-ci on suppose que les (X_i, Y_i) sont i.i.d., et les ε_i le sont aussi et indépendants de (X_i, Y_i) (et de moyenne nulle et variance σ^2. Pour simplifier un peu les choses on suppose que $E(X_1) = 0$ et

$$\Sigma = \mathbb{E}(X_1 X_1^T) = I_d,$$

de sorte que le risque en prédiction s'écrive, pour $u \in \mathbb{R}^d$,

$$\begin{aligned} R(u) &= \mathbb{E}((Y_1 - \langle X_1, u \rangle)^2) = \mathbb{E}(\langle \theta - u, X_1 \rangle + \varepsilon_1)^2 \\ &= \mathbb{E}\left(\mathbb{E}\left(\langle \theta - u, X_1 \rangle + \varepsilon_1)^2 \mid X\right)\right) = \sigma^2 + \mathbb{E}\left(\langle \theta - u, X_1 \rangle\right)^2 \\ &= \sigma^2 + \|\theta - u\|^2. \end{aligned}$$

Le risque empirique correspondant est alors

$$R_n(u) = \frac{1}{n} \sum_{i=1}^n (Y_i - \langle u, X_i \rangle)^2 = \frac{1}{n} \|Y - X\theta\|^2,$$

dont un minimiseur est donné par

$$\hat{\theta}_{LS} = (XX^T)^{-1} X^T Y,$$

c'est à dire l'estimateur par moindre carrés ordinaire. On peut vérifier que

$$R(\hat{\theta}_{LS}) - R(\theta) = \sigma^2 \mathbb{E}\left(\mathrm{Tr}((X^T X)^{-1})\right),$$

et que c'est le risque minimax sur ce modèle (où la distribution de X est fixée toutefois). On peut remarquer que

$$\hat{\Sigma} = \frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^n X_i X_i^T \underset{p.s.}{\longrightarrow} \Sigma = I_d,$$

par la loi des grands nombres (avec la seule hypothèse $\mathbb{E}(\|X\|^2) < +\infty$. Intuitivement on s'attend donc à avoir $\mathbb{E}(\mathrm{Tr}((X^T X)^{-1})) \sim \frac{d}{n}$. C'est à peu près vrai si on rajoute des hypothèses sur la loi de X qui garantissent la bonne concentration de $\hat{\Sigma}$ vers Σ, par exemple en demandant que X_1 soit bornée ou a une loi spécifique.

Par exemple, si on suppose que $X_1 \sim \mathcal{N}(0, I_d)$, alors $(X^T X)^{-1}$ suit une *loi de Wishart inverse* de paramètres (n, I_d). On peut alors calculer son espérance

$$\mathbb{E}(\mathrm{Tr}((X^T X)^{-1})) = \frac{d}{n - d - 1}.$$

Bref, tout cela pour dire que pour le problème de régression linéaire à design aléatoire (connu) pas trop dégénéré, l'excès de risque en prédiction optimal est de l'ordre de $\sigma^2 \frac{d}{n}$ pour le régime des grands n (par rapport à d).

Descente de gradient sur R_n

L'estimateur $\hat{\theta}_{LS}$ nécessite l'inversion d'une matrice $d \times d$, opération dont la complexité est bornée inférieurement par d^2 (complexité équivalente à celle du produit matriciel). Lorsque d est grand (de l'ordre de 10000 disons), on peut s'inquiéter de la faisabilité en temps raisonnable de l'estimation par moindre carrés. On peut/doit alors se tourner vers des estimateurs "algorithmiques". Une manière de faire est d'essayer de minimiser R_n via une descente de gradient, c'est à dire, poser

$$u_{t+1} = u_t - \gamma_t \nabla_{u_t} R_n,$$

à partir d'un u_0 donné (disons 0), et en s'arrêtant au bout d'un certain moment (à déterminer).

PROPOSITION 5.8

Soient $\hat{\lambda}_{\min}$ et $\hat{\lambda}_{\max}$ les respectivement plus grande et plus petite valeur propre de $\hat{\Sigma}$. Si on choisit le pas de gradient $\gamma_t = \gamma < \frac{1}{4\hat{\lambda}_{\max}}$ constant, on a, pour tout $t \geq 0$,

$$\|u_{t+1} - \hat{\theta}_{LS}\|^2 \leq (1 - 2\gamma\hat{\lambda}_{\min})\|u_t - \hat{\theta}_{LS}\|^2.$$

Par conséquent, avec $u_0 = 0$, on a

$$\|u_T - \hat{\theta}_{LS}\|^2 \leq (1 - 2\gamma\hat{\lambda}_{\min})^T \|\hat{\theta}_{LS}\|^2.$$

Démonstration. On aura besoin d'une inégalité liée à la forte convexité de R_n. Il est immédiat que R_n est convexe. Par ailleurs, on a, pour $u \in \mathbb{R}^d$,

$$\nabla_u^2 R_n = 2\hat{\Sigma} \succcurlyeq 2\hat{\lambda}_{\min} I_d.$$

On en déduit que R_n est $2\hat{\lambda}_{\min}$-fortement convexe, c'est à dire que $u \mapsto R_n(u) - \hat{\lambda}_{\min}\|u\|^2$ est convexe. On en déduit alors que pour tous $u, v \in \mathbb{R}^d$,

$$R_n(u) \geq R_n(v) + \langle \nabla_v R_n, u - v \rangle + \hat{\lambda}_{\min}\|u - v\|^2,$$

menant à l'inéquation suivante

$$\left\langle \nabla_u R_n, u - \hat{\theta}_{LS} \right\rangle \geq R_n(u) - R_n(\hat{\theta}_{LS}) + \hat{\lambda}_{\min}\|\hat{\theta}_{LS} - u\|^2 \tag{5.6}$$

Un autre ingrédient de ce type de preuve est la Lipschitzianité des gradients. Dans notre cas, comme $\nabla_u R_n = \frac{2}{n}\sum_{i=1}^n (\langle u, X_i \rangle - Y_i) X_i$, on a immédiatement

$$\|\nabla_u R_n - \nabla_v R_n\| \leq 2\hat{\lambda}_{\max}\|u - v\|. \tag{5.7}$$

On en déduit alors que

$$R_n(u) \leq R_n(v) + \langle \nabla_v R_n, u - v \rangle + 2\hat{\lambda}_{\max}\|u - v\|^2,$$

menant à une inégalité spécifique pour la descente de gradient à un pas :

$$R_n(u_{t+1}) \leq R_n(u_t) - \gamma(1 - 2\hat{\lambda}_{\max}\gamma)\|\nabla_{u_t} R_n\|^2,$$

soit encore

$$\|\nabla_{u_t} R_n\|^2 \leq \frac{R_n(u_t) - R_n(u_{t+1})}{\gamma(1 - 2\hat{\lambda}_{\max}\gamma)} \leq \frac{R_n(u_t) - R_n(\hat{\theta}_{LS})}{\gamma(1 - 2\hat{\lambda}_{\max}\gamma)}, \tag{5.8}$$

dès lors que $\gamma < (2\hat{\lambda}_{\max})^{-1}$.

On peut alors commencer la preuve de l'inégalité de récursion. Pour $t \geq 0$, on peut écrire

$$\begin{aligned}\|u_{t+1} - \hat{\theta}_{LS}\|^2 &= \|u_t - \gamma\nabla_{u_t} R_n - \hat{\theta}_{LS}\|^2 \\ &= \|u_t - \hat{\theta}_{LS}\|^2 - 2\gamma \left\langle u_t - \hat{\theta}_{LS}, \nabla_{u_t} R_n \right\rangle + \gamma^2 \|\nabla_{u_t} R_n\|^2.\end{aligned}$$

Le deuxième terme à gauche se majore au moyen des inégalités de forte convexité (5.6), le troisième terme avec les inégalités de type Lipschitz (5.8). Cela donne

$$\|u_{t+1} - \hat{\theta}_{LS}\|^2 \leq \|u_t - \hat{\theta}_{LS}\|^2(1 - 2\gamma\hat{\lambda}_{\min}) + (R_n(u_t) - R_n(\hat{\theta}_{LS})) \left(\frac{\gamma}{(1 - 2\hat{\lambda}_{\max}\gamma)} - 2\gamma \right).$$

On conclut en remarquant que le dernier terme est négatif lorsque $\gamma < (4\hat{\lambda}_{\max})^{-1}$. □

On peut en déduire alors le corollaire suivant.

Corollaire 5.9

Pour M, λ_- et λ_+ des quantités positives, notons

$$A_{M,\lambda_-,\lambda_+} = \{w \mid 0 < \lambda_- \leq \hat{\lambda}_{\min} \leq \hat{\lambda}_{\max} \leq \lambda_+ \text{ et } \|\hat{\theta}_{LS}\|^2 \leq M\}.$$

Pour le choix $T = \frac{\log\left(\frac{nM}{\sigma^2 d}\right)}{-\log\left(1 - \frac{\lambda_-}{4\lambda_+}\right)}$, on a, sur A,

$$\|u_T - \hat{\theta}_{LS}\|^2 \leq \frac{\sigma^2 d}{n}.$$

L'idée derrière ce choix de temps d'arrêt est la suivante : vu que la meilleure erreur possible du point de vue statistique est en $\sigma^2 d/n$, il ne sert à rien d'optimiser au-delà. Dès lors, pour ce choix de T, la sortie de l'algorithme de descente de gradient sera quasiment optimale (à un facteur 2 près), en termes d'ordres de grandeurs en d/n. D'un point de vue computationnel le gain est certain : on se retrouve à calculer $\log(n)$ itérations, chacune requérant le calcul d'un gradient (en nd opérations), donc au total un nombre d'opérations en $O(dn\log(n))$, ce qui est plus raisonnable que le calcul direct de $\hat{\theta}_{LS}$ dans le régime d grand et $n >> d$.

Remarque 1 : On voit que la calibration du pas de gradient ainsi que le temps d'arrêt nécessitent la connaissance a priori de bornes sur $\hat{\lambda}_{\min}$, $\hat{\lambda}_{\max}$ et $\|\hat{\theta}_{LS}\|^2$. D'un point de vue théorique, on peut s'en sortir avec de la concentration entre $\hat{\Sigma}$ et Σ, ainsi que de la concentration sur $\|\hat{\theta}_{LS}\|$ (ou de la connaissance a priori de type $\theta \in \mathrm{B}(0, M)$, et cela donne in fine des inégalités oracles sur u_T (en minorant la proba de A). D'un point de vue pratique, en normalisant les colonnes de X on contrôle le $\hat{\lambda}_{\max}$, mais estimer le λ_- via le calcul de $\hat{\lambda}_{\min}$ peut s'avérer coûteux. Dans notre modèle où $\lambda_{\min} = 1$, on peut prendre les bornes inférieures en déviation (de manière générale, on ne pourra garantir l'optimalité que sur un modèle avec un $\lambda_{\min} \geq c > 0$).

Remarque 2 : On peut aussi directement travailler sur un évènement où $\nabla_u R_n$ et $\nabla_u R$ sont proches (uniformément en u), et dérouler la méthode de preuve en regardant cette fois $\|\theta - u_t\|^2$ et en travaillant sur la convexité de R. Cette fois-ci une connaissance de $\lambda_{\min}$ est requise. C'est cette approche qui est plutôt privilégiée en statistiques théoriques, on en donnera un exemple dans la partie suivante (gradient stochastique).

Descente de gradient stochastique

Dans le cas où n est vraiment très grand (plusieurs dizaines de millions par exemple), les algorithmes "batch" (nécessitant d'effectuer des calculs sur l'échantillon entier, donc d'avoir les n données en mémoire vive) peuvent s'avérer infaisables, même s'ils sont de faible complexité (comme la descente de gradient en $n\log(n)$.

Dans ces situations, plusieurs ruses opératoires peuvent mener à des estimateurs : par exemple en découpant l'échantillon en plusieurs morceaux traitables, donnant chaque morceau à différentes unités, et en recombinant à la fin (c'est notamment le cas des algorithmes parallélisables). On peut aussi viser des estimateurs "online", c'est à dire prenant les données une par une à la volée et actualisant la sortie (l'estimateur). On va donner un exemple de cette dernière stratégie via la descente de gradient stochastique, qui, massivement utilisée en machine learning, reste néanmoins analysable d'un point de vue statistique. Pour un panorama complet des méthodes appropriées dans un contexte de taille de données énorme, vous pouvez vous référer à un cours de "big data" (c'en est un des aspects).

Dans notre cadre de régression linéaire à design aléatoire, le principe de la descente de gradient stochastique peut s'énoncer comme suit : on voit les données $(X_i, Y_i)_{i=1,\dots,n}$ une par une, on part de $u_0 = 0$, et on actualise via

$$u_{t+1} = u_t - \gamma_t \hat{g}_t,$$

où cette fois-ci $\hat{g}_t$ va être une estimation du gradient $g_t = \nabla_{u_t} R$ basé sur la $(t+1)$-ème observation :

$$\hat{g}_t = 2\left(\langle X_{t+1}, u_t\rangle - Y_{t+1}\right) X_{t+1}.$$

Le point-clé est que les données utilisées pour évaluer le gradient doivent être indépendantes de celles utilisées pour arriver jusqu'à l'état t. On pourrait très bien utiliser des "mini-batchs" pour calculer ces estimations de gradient, c'est d'ailleurs une solution préférable en pratique (pour des questions de variance). On peut traduire cette propriété d'indépendance par

$$E_t\left(\hat{g}_t\right) = \nabla_{u_t} R = g_t,$$

où $E_t = \mathbb{E}\left(. \mid X_{1:t}\right)$ désigne l'espérance conditionnelle par rapport aux données ayant mené jusqu'à l'état t. Reste à calibrer les pas. Dans cette partie on regardera le modèle un peu plus général où

$$0 < \lambda_{\min} I_d \preccurlyeq \Sigma = \mathbb{E}(XX^T) \preccurlyeq \lambda_{\max} I_d.$$

On remarque que dans ce cadre l'excès de risque en prédiction s'écrit

$$\begin{aligned} R(u) - R(\theta) &= \mathbb{E}\left[(Y - \langle X, u\rangle)^2 - (Y - \langle X, \theta\rangle)^2\right] \\ &= (u-\theta)^T \Sigma (u - \theta). \end{aligned}$$

On admettra que le risque minimax reste de l'ordre de $\sigma^2 d/n$. Avec un peu plus d'hypothèses sur X et θ, on peut prouver facilement le résultat suivant sur une variante de SGD.

THÉORÈME 5.10

Supposons que $\mathbb{E}(\|X\|^4) = \kappa_4 < +\infty$, *et* $\theta \in \mathrm{B}(0, M)$, *pour un* M *connu. On pose alors*

$$\begin{aligned} u_0 &= 0, \\ u_{t+1} &= \pi_{\mathrm{B}(0,M)}\left(u_t - \gamma_t \hat{g}_t\right), \end{aligned}$$

en prenant $\gamma_t = \frac{1}{\lambda_{\min}(t+1)}$. *On a alors, pour tout* $t \geq 1$,

$$\mathbb{E}\left\|u_t - \theta\right\|^2 \leq \frac{32M^2\kappa_4 + 8\sigma^2 \mathrm{Tr}(\Sigma)}{\lambda_{\min}^2 t}.$$

Quelques remarques avant de passer à la preuve :

1. D'un point de vue complexité, on a besoin de calculer n gradients en dimension d, on fait donc au total $O(nd)$ opérations (à peu près comme la descente de gradient classique). En revanche, on n'aura jamais qu'à stocker un état courant (d valeurs) et un gradient (d autres valeurs). Le gain en mémoire est alors appréciable.
2. Du point de vue de la vitesse de convergence, le terme en M^2 est sous-optimal, et peut être enlevé au prix d'un terme traduisant l'attache à la condition initiale (en $\|\theta - u_0\|^2$) qui s'écrase beaucoup plus vite que le terme en σ^2. La preuve est beaucoup plus technique, le lecteur intéressé la trouvera dans Bach et Moulines 2011. En admettant que la bonne vitesse pour $\mathbb{E}\|\hat{\theta}_{SGD} - \theta\|^2$ est de l'ordre de $\sigma^2 \mathrm{Tr}(\Sigma)/(\lambda_{\min}^2 n)$, on peut borner l'erreur en prédiction en $\sigma^2(\lambda_{\max}/\lambda_{\min})^2 d/n$, ce qui est le bon ordre de grandeur (en d et n). En ce sens, les méthodes de gradient stochastique peuvent être considérées comme optimales.
3. La calibration du pas en $1/(\lambda_{\min}(t+1))$ est critique (quoique on puisse aller jusqu'à $1/(2\lambda_{\min}(t+1))$. Pour être optimale théoriquement, cette méthode nécessite la connaissance a priori de $\lambda_{\min}$. D'autres méthodes, utilisant des pas en $t^{-\alpha}$ où $\alpha \in [0, 1[$ peuvent être utilisées, avec moyennisation finale des étapes, qui ne nécessitent pas cette connaissance (voir l'article de Bach et Moulines 2011 et les travaux récents de F. Bach). Dans ces approches, on perd le côté "online" (il faut connaître la taille d'échantillon avant de lancer l'algorithme), mais on n'a toujours pas besoin de stocker le jeu de données en entier (ni toutes les étapes).
4. Le résultat donné est en espérance, c'est facilité par la relation $E_t \hat{g}_t = g_t$. Pour des résultats en déviation, des conditions supplémentaires sur le bruit doivent être demandées. De manière générale, la descente de gradient stochastique "simple" décrite au-dessus est de variance élevée, on lui préfère souvent des approches moyennisantes comme dans la remarque précédente ou par mini-batches, plus stables en pratique (et avec des bornes en déviation plus sympathiques).

Démonstration. Tout est basé sur à peu près la même récurrence que dans la preuve pour la descente de gradient sur R_n. Ici on tire profit de la forte convexité de R, qui donne

$$\langle \nabla_u R, u - \theta \rangle \geq \lambda_{\min} \|\theta - u\|^2. \tag{5.9}$$

On aura aussi besoin d'une borne sur les gradients, en espérance, si $u_t \in \mathrm{B}(0, M)$,

$$\begin{aligned} E_t\|\hat{g}_t\|^2 &= 4E_t \left\| \langle X_{t+1}, u_t - \theta \rangle X_{t+1} - \varepsilon_t X_t \right\|^2 \\ &\leq 4(u_t - \theta)^T E_t \left(X_{t+1} X_{t+1}^T X_{t+1} X_{t+1}^T \right) (u_t - \theta) + 4\sigma^2 \mathrm{Tr}(\Sigma) \\ &\leq 4\kappa_4 \|u_t - \theta\|^2 + 4\sigma^2 \mathrm{Tr}(\Sigma) \\ &\leq 16 M^2 \kappa_4 + 4\sigma^2 \mathrm{Tr}(\Sigma) = G^2, \end{aligned} \tag{5.10}$$

avec $G^2 = 16M^2\kappa_4 + 4\sigma^2\mathrm{Tr}(\Sigma)$. Par construction on a, pour tout t, $u_t \in \mathrm{B}(0, M)$. Ensuite, on peut écrire

$$\begin{aligned} E_t\|u_{t+1} - \theta\|^2 &= E_t \|\pi_{\mathrm{B}(0,M)}(u_t - \gamma_t \hat{g}_t) - \theta\|^2 \\ &\leq E_t \|u_t - \gamma_t \hat{g}_t - \theta\|^2 \\ &= \|u_t - \theta\|^2 - 2\gamma_t E_t \langle u_t - \theta, \hat{g}_t \rangle + \gamma_t^2 E_t \|\hat{g}_t\|^2 \\ &= \|u_t - \theta\|^2 - 2\gamma_t \langle u_t - \theta, g_t \rangle + \gamma_t^2 E_t \|\hat{g}_t\|^2 \\ &\leq (1 - 2\gamma_t \lambda_{\min}) \|u_t - \theta\|^2 + \gamma_t^2 G^2. \end{aligned}$$

On en déduit alors l'équation de récurrence suivante

$$\mathbb{E}\|u_{t+1} - \theta\|^2 \leq \left(1 - \frac{2}{t+1}\right) \mathbb{E}\|u_t - \theta\|^2 + \frac{G^2}{\lambda_{min}^2 (t+1)^2}. \tag{5.11}$$

Regardons ce qui se passe pour $t = 1$. On a d'une part

$$\|g_1\|^2 = \|E_1 \hat{g}_1\|^2 \leq E_1 \|\hat{g}_1\|^2 \leq G^2.$$

D'autre part, l'inégalité de convexité (5.9) se réécrit dans ce cas

$$\langle g_1, u_1 - \theta \rangle \geq \lambda_{\min} \|\theta - u_1\|^2.$$

Une application de Cauchy-Schwarz donne alors

$$\|\theta - u_1\|^2 \leq \frac{\|g_1\|^2}{\lambda_{\min}^2} \leq \frac{G^2}{\lambda_{\min}^2} \leq \frac{2G^2}{\lambda_{\min}^2 \times 1},$$

et la propriété est prouvée pour $t = 1$. Notons maintenant Δ_t la quantité $\mathbb{E}(\|u_t - \theta\|^2)$, et supposons $\Delta_t \leq \frac{2G^2}{\lambda_{\min}^2 t}$. L'équation de récurrence donne alors

$$\begin{aligned} \Delta_{t+1} - \frac{2G^2}{\lambda_{\min}^2 (t+1)} &\leq \left(1 - \frac{2}{t+1}\right) \Delta_t + \frac{G^2}{(t+1)^2} - \frac{2G^2}{\lambda_{\min}^2 (t+1)} \\ &\leq \frac{G^2}{\lambda_{\min}^2 t(t+1)} (2 - 4 + 1) \leq 0, \end{aligned}$$

ce dont on déduit le résultat. □

Chapitre 6

Intro à la stat non paramétrique

Dans certaines situations, on ne peut pas modéliser les observations par une famille paramétrée par $\mathbb{R}^d$, c'est notamment le cas en régression linéaire lorsque l'on ne connaît pas la loi de l'erreur. Si le paramètre d'intérêt de la loi générant les observations se résume à un vecteur de dimension finie (par exemple la moyenne), on peut s'en sortir à coup d'hypothèses supplémentaires sur P ne nécessitant pas forcémenent l'appartenance à une famille de lois de dimension finie (par exemple des conditions de moments, cf le chapitre Modèle linéaire).

En revanche, si le paramètre d'intérêt dans la loi générant les observations est par essence non-paramétrique (son éventuelle densité, son support, l'appartenance à une famille de lois, etc.), on est obligés d'employer des méthodes spécifiques. On présentera dans ce chapitre trois exemples classiques d'estimation non paramétrique et les méthodes qui vont avec : le test d'adéquation à une loi (Kolmogorov-Smirnov), l'estimation de densité, et l'estimation de support.

6.1 Adéquation à (une famille de) loi(s) : test de Kolmogorov-Smirnov

Le modèle sous-jacent au test de Kolmogorov-Smirnov est le suivant : on se donne $X_1, \dots, X_n$ variables aléatoires sur $\mathbb{R}$, i.i.d. de loi P, où P **a une densité** f. Le modèle complet s'écrit alors

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (P_f^{\otimes n})_{f \text{ densité}}\right).$$

C'est donc un modèle fortement non paramétrique (paramétré par les densités possibles sur $\mathbb{R}$). Le test de Kolmogorov-Smirnov vise à répondre à la question suivante : "étant donné une densité de référence f_0, les observations sont-elles générées suivant f_0 ?". Plus précisément, les hypothèses sont

$$\begin{array}{rcl} H_0 & : & P \sim f_0 d\lambda \\ H_1 & : & P \nsim f_0 d\lambda. \end{array}$$

Remarque : L'hypothèse d'avoir une densité n'est nécessaire que pour l'hypothèse nulle. On verra par la suite que le test de KS est consistant même si la loi dans l'alternative n'a pas de densité.

On va devoir construire un test qui satisfait les trois propriétés suivantes :

1. La statistique de test doit être facile à calculer.

2. La loi de la statistique de test sous H_0 de doit pas dépendre de f_0 (pour une calibration simple).
3. Le test doit être asymptotiquement consistant.

6.1.1 Fonction de répartition empirique et statistique de test

Une bonne statistique de test doit approcher ce que l'on cherche à caractériser. Ici, pour une loi P de densité f, on doit approcher n'importe quelle quantité caractérisant la loi. On a plusieurs choix possibles : essayer d'estimer la densité directement, estimer la fonction caractéristique, ou la fonction de répartition. C'est cette dernière stratégie qui est mise en oeuvre par le test de KS. Un estimateur de la fonction de répartition F est donné naturellement par la fonction de répartition emprique F_n, définie par

$$\forall t \in \mathbb{R} \quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(X_i).$$

(FAIRE DESSIN). C'est la fonction de répartition associée à la *mesure empirique* $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.
Remarque : La fonction de répartition empirique conserve toute l'information de l'échantillon, excepté l'ordre. On peut montrer que, pour le modèle à densité, la statistique d'ordre $(X_{(1)}, \dots, X_{(n)})$ est exhaustive. En ce sens, on ne perd rien (en terme d'information sur la densité sous-jacente) à considérer F_n.

La statistique du test de KS mesure juste l'écart entre la fonction de répartition empirique et la fonction de répartition cible, au sens de la norme infinie.

Definition 6.1 : Statistique de KS

La statistique de test de Kolmogorov-Smirnov, notée D_n est définie par

$$D_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_0(x) - F_n(x)| = \sqrt{n}\|F_n - F_0\|_\infty.$$

La normalisation en $\sqrt{n}$ se comprend bien : si on fixe x, on a $F_n(x) \sim \mathcal{B}(n, F_0(x))$ sous H_0, et le Théorème central limite donne alors $\sqrt{n}(F_n(x) - F_0(x)) \rightsquigarrow \mathcal{N}(0, F_0(x)(1 - F_0(x))$. Tout le travail théorique consiste à passer d'un théorème central limite ponctuel à un théorème central limite "uniforme". C'est possible avec la notion générale de "Classe de Donsker", on se contentera ici d'une version plus faible. Remarquons enfin qu'a priori la loi de D_n dépend de F_0. La suite montrera qu'il n'en est rien (si F_0 est continue).

6.1.2 Calculabilité et liberté de la statistique de KS

On va regarder les deux premiers points du cahier des charges, à savoir le calcul effectif de la statistique de KS et sa non-dépendance en F_0 (liberté), les deux étant liés.

On aura besoin de quelques outils concernant la fonction de répartition d'une loi à densité (et la fonction quantile correspondante).

Lemme 6.2

Soit P une loi sur $\mathbb{R}$ ayant pour fonction de répartition F. On définit la fonction quantile q par $q(t) = \inf\{x \mid F(x) \geq t\}$, pour $t \in]0,1[$. On a alors

1. *Si $U \sim \mathcal{U}(]0,1[$, $q(U) \sim P$.*

Par ailleurs, si P admet une densité sur $\mathbb{R}$, on a de plus

2. *F est continue sur $\mathbb{R}$.*
3. *$F \circ q = Id_{]0,1[}$.*
4. *Si $X \sim P$, $F(X) \sim \mathcal{U}(]0,1[)$.*

Démonstration. On commence par vérifier que pour $t \in]0,1[$, $\{F(x) \geq t\} = \{x \geq q(t)\}$. En effet, si $F(x) \geq t$, par définition de q on a bien $q(t) \leq x$. Réciproquement, soit $x \geq q(t)$. Comme F est continue à droite, on a $F(q(t)) \geq t$, et par croissance de F on a bien $F(x) \geq t$. Maintenant, si $U \sim \mathcal{U}(]0,1[)$ et $x \in \mathbb{R}$,

$$\mathbb{P}(q(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

Donc $Q(U) \sim P$.

F étant cadlàg, il faut vérifier la continuité à gauche. Comme $F(t) - F_-(t) = P(\{t\})$, et que $P << \lambda$, on en déduit que F est continue à gauche (et donc continue tout court).

Soit $t \in]0,1[$, et $x_n \to q(t)^+$. On a d'une part $F(x_n) \geq t$, et par continuité à droite $F(q(t)) \geq t$. Si $F(q(t)) > t$, la continuité à gauche de F donne $y < q(t)$ tel que $F(y) > t$, ce qui contredit la définition de $q(t)$. Donc $F(q(t)) = t$.

Si $X \sim P$, alors $F(X) \sim F(q(U))$, où $U \sim \mathcal{U}(]0,1[)$ d'après le premier point. Le troisième point permet de conclure. □

Pour analyser la statistique de KS, on commence par le petit résultat suivant (on rappelle que l'on se place dans le cas où F_0 est continue).

Lemme 6.3

Si $(X_{(1)}, \ldots, X_{(n)})$ représente l'échantillon trié par ordre croissant, on a

$$D_n = \sqrt{n} \sup_{i=0,\ldots,n-1} \left| \frac{i}{n} - F_0(X_{(i)}) \right| \vee \left| \frac{i}{n} - F_0(X_{(i+1)}) \right|,$$

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

On remarque alors que le calcul de D_n nécessite $2n-1$ calculs (à partir de l'échantillon trié). Sous cette forme, on se rend aussi compte que D_n est une statistique (la mesurabilité de la norme infinie n'étant pas garantie sur un espace de base non dénombrable).

Démonstration. Soit $t \in \mathbb{R}$. Si $t < X_{(1)}$, $F_n(t) = 0$, et $F_0(X_{(0)}) = 0$. Par croissance de F_0, on en déduit

$$0 - F_0(X_{(1)}) = F_n(t) - F_0(X_{(1)}) \leq F_n(t) - F_0(t) \leq 0 - F_0(X_{(0)}),$$

donc $\sup_{t\leq X_{(1)}}|F_n(t)-F_0(t)|\leq |F_0(X_{(0)})|\vee|F_0(X_{(1)})|=F_0(X_{(1)})$, le sup étant atteint en prenant $t\to X_{(1)}$ (F_0 est continue car P est à densité).

Maintenant, pour $t\in[X_{(i)},X_{(i+1)}[$, on a $F_n(t)=\frac{i}{n}$, et la croissance de F_0 induit

$$\frac{i}{n}-F_0(X_{(i+1)})\leq F_n(t)-F_0(t)\leq \frac{i}{n}-F_0(X_{(i)}),$$

et donc

$$\sup_{t\in[X_{(i)},X_{(i+1)}[}|F_n(t)-F_0(t)|=\left|\frac{i}{n}-F_0(X_{(i)})\right|\vee\left|\frac{i}{n}-F_0(X_{(i+1)})\right|,$$

la borne étant atteinte en regardant $t=X_{(i)}$ ou $t\to X_{(i+1)}$ (par continuité de F_0 toujours). D'où le résultat. □

On remarque que si l'on n'a pas continuité de F_0, on peut toujours remplacer $F_0(X_{(i+1)})$ par $F_0(X_{(i+1)}^-)$ dans le résultat précédent. L'avantage de supposer F_0 continue est que cela garantit que la loi de D_n ne dépend pas de F_0.

Théorème 6.4 : Liberté de la statistique de KS

Si F_0 est continue, alors

$$D_n\sim\sqrt{n}\sup_{i=0,\dots,n-1}\left|\frac{i}{n}-U_{(i)}\right|\vee\left|\frac{i}{n}-U_{(i+1)}\right|,$$

où $U_1,\dots,U_n$ sont i.i.d. $\mathcal{U}(]0,1[)$. En particulier la loi de D_n ne dépend pas de F_0.

Démonstration. En notant q_0 la fonction quantile de F_0, on a $(X_1,\dots,X_n)=(q_0(U_1),\dots,q_0(U_n))$, où les U_i sont i.i.d. $\mathcal{U}(]0,1[)$. Comme q_0 est croissante on a $(X_{(1)},\dots,X_{(n)})=(q_0(U_{(1)}),\dots,q_0(U_{(n)}))$, et donc

$$D_n\sim\sqrt{n}\sup_{i=0,\dots,n-1}\left|\frac{i}{n}-F_0(q_0(U_{(i)}))\right|\vee\left|\frac{i}{n}-F_0(q_0(U_{(i+1)}))\right|=\sqrt{n}\sup_{i=0,\dots,n-1}\left|\frac{i}{n}-U_{(i)}\right|\vee\left|\frac{i}{n}-U_{(i+1)}\right|,$$

en utilisant que $F_0\circ q_0=Id_{]0,1[}$ (continuité de F_0). □

Ce phénomène de liberté va servir à calibrer le seuil $t_{n,\alpha}$ dans le test de KS. En effet, si on connaît la loi de D_n, il suffit de prendre pour t_α le $1-\alpha$ quantile de D_n, et le test de KS s'écrit $\mathbb{1}_{D_n\geq t_{n,\alpha}}$. Malheureusement on ne connaît pas cette loi exactement. Plusieurs options s'offrent alors à nous :

1. Comme la loi de D_n est libre, on peut la simuler et approcher ses quantiles.
2. On peut utiliser de la concentration sur D_n.
3. On peut montrer que $D_n\rightsquigarrow D$, pour un D connu (maximum de pont Brownien), et utiliser les quantiles de D pour construire un test asymptotique.

Le 3- sortant du cadre d'un cours d'introduction aux statistiques, on montrera un peu de 2-.

6.1.3 Comportement asymptotique et déviations de la statistique de KS

Il s'agit ici d'étudier le comportement de D_n, ou $\|F_n - F_0\|_\infty$, sous H_0 et l'alternative. Une première étape est de montrer que $\|F_n - F_0\|_\infty \to 0$ lorsque $n \to +\infty$, sous H_0. Lorsque x est fixé, sous H_0, $F_n(x) \to F_0(x)$ par la loi des grands nombres. Il faut alors passer d'une loi des grands nombres ponctuelle à une loi des grands nombres uniforme. C'est l'objet du théorème de Glivenko-Cantelli.

Théorème 6.5 : Théorème de Glivenko-Cantelli

Si P a pour fonction de répartition F_0 continue, alors

$$\|F_0 - F_n\|_\infty \underset{n\to+\infty}{\longrightarrow} 0,$$

en probabilité.

Cela montre en particulier que D_n ne croît pas plus vite que $\sqrt{n}$ (de fait, elle converge en loi).

Démonstration. D'après le théorème sur la liberté de D_n, il suffit de montrer $\|F_0 - F_n\|_\infty \longrightarrow 0$ dans le cas où $P \sim \mathcal{U}(]0,1[)$ (c'est une commodité technique, on peut le montrer directement pour une fonction de répartition F_0 générale). On se donne N un autre entier, et on subdivise $]0,1[$ en $0 = x_{0,N} < x_{1,N} = \frac{1}{N} < x_{j,N} = \frac{j}{N} < x_{N,N} = 1$.

Pour $x \notin]0,1[$, on a $|F_n(x) - F_0(x)| = 0$. Maintenant, si $x_{j,N} \leq x \leq x_{j+1,N}$, pour $j \in [\![0, N-1]\!]$, on a

$$F_n(x) - F_0(x) \leq F_n(x_{j+1,N}) - F_0(x_{j,N}) = F_n(x_{j+1,N}) - F_0(x_{j+1,N}) + \frac{1}{N},$$
$$F_n(x) - F_0(x) \geq F_n(x_{j,N}) - F_0(x_{j+1,N}) = F_n(x_{j,N}) - F_0(x_{j,N}) - \frac{1}{N}.$$

On en déduit

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| \leq \max_{j=0,\dots,N} |F_n(j/N) - (j/N)| + \frac{1}{N}.$$

Comme $F_n(j/N) \sim \mathcal{B}(n, (j/N))$,¨ en utilisant Bienaymé-Cebicev et une borne d'union, on obtient

$$\mathbb{P}\left(\max_{j=0,\dots,N} |F_n(j/N) - (j/N)| \geq \frac{t}{2}\right) \leq \frac{N}{nt^2}.$$

En prenant $N = \lceil \frac{2}{t} \rceil$, on en déduit

$$\mathbb{P}\left(\|F_n - F_0\|_\infty \geq t\right) \leq \frac{\lceil \frac{2}{t} \rceil}{nt^2} \underset{n\to+\infty}{\longrightarrow} 0.$$

□

Remarque : L'hypothèse de continuité de F_0 n'est pas nécessaire (on peut adapter la preuve). En revanche elle est nécessaire pour avoir la "liberté" de D_n.

Corollaire 6.6 : Consistance du test de KS

Si $X_1, \dots, X_n$ sont i.i.d. de loi P, P ayant une fonction de répartition F continue différente de F_0, alors

$$D_n \to +\infty,$$

en P probabilité.

Remarque : Là encore l'hypothèse P continue peut être relâchée (même si elle ne l'est pas, on aura quand même $\|F - F_n\|_\infty \to 0$ en P proba).

Démonstration. Si $F_0 \neq F$, il existe alors $x \in \mathbb{R}$ tel que $|F_0(x) - F(x)| = \delta > 0$. On a alors

$$D_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)| \geq \sqrt{n}|F_n(x) - F_0(x)| \geq \sqrt{n}\delta - \sqrt{n}\|F_n - F\|_\infty.$$

D'après le théorème 6.5, on a $\|F - F_n\|_\infty \to_{n\to+\infty} 0$ en P probabilité. On a donc

$$P\left(\|F_n - F\|_\infty \leq \frac{\delta}{2}\right) \to_{n\to+\infty} 1,$$

ce qui entraîne

$$P\left(D_n \geq \frac{\sqrt{n}\delta}{2}\right) \to_{n\to+\infty} 1,$$

et donc $D_n \to_{n\to+\infty} +\infty$ en P probabilité. □

Revenons maintenant à la calibration du seuil sous H_0. Comme déjà dit, on peut calculer un seuil approché via simulations (la statistique de test étant libre). On peut aussi baser cette calibration sur une inégalité de déviation.

Théorème 6.7 : Déviations de la statistiques de KS

Pour tout $n > 0$ et $\varepsilon > 0$, on a

$$\mathbb{P}\left(D_n \geq \varepsilon\right) \leq 4 \exp\left(-\frac{\varepsilon^2}{8}\right).$$

On peut alors prendre comme seuil $t_\alpha = \sqrt{8 \log\left(\frac{4}{\alpha}\right)}$. Cette inégalité peut être améliorée : le $4e^{-\varepsilon^2/8}$ peut être ramené à un $2e^{-2\varepsilon^2}$ (inégalité DKW, résultat difficile), et là on ne peut pas faire mieux. Dans tous les cas de figure, on remarque que, d'un point de vue minimax,

$$\sup_P \mathbb{E}\|F_n - F\|_\infty \leq \frac{C}{\sqrt{n}},$$

c'est à dire que la vitesse d'estimation d'une fonction de répartition est une vitesse paramétrique. Cela s'explique par le fait que les fonctions de répartitions sont une classe fonctionnelle bien particulière : ce sont des fonctions croissantes. En utilisant

ce point et un argument comme dans la preuve de Glivenko-Cantelli, estimer une fonction de répartition en norme infinie revient à contrôler $\sup_{t\in\mathbb{R}} |(P-P_n)f_t|$, où $f_t = \mathbb{1}_{]-\infty,t]}$, c'est à dire des déviations sur un ensemble de fonctions unidimensionnel. Retrouver une vitesse paramétrique est alors naturel. On verra ensuite qu'estimer la densité (vue comme la dérivée de la fonction de répartition) est un problème **vraiment** non-paramétrique, pour lequel ce phénomène ne peut advenir.

Démonstration. On aura besoin du Lemme de symétrisation, outil classique en théorie des processus empiriques. On rappelle qu'une variable de Rademacher ε est définie par $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$.

Lemme 6.8 : Symétrisation

Soit $X_{1,t}, \dots, X_{n,t}$ suite de vecteurs aléatoires indépendants à t fixé, où $t \in \mathcal{T}$, $\varepsilon_1, \dots \varepsilon_n$ une suite de variables de Rademacher indépendantes (entre elles et des $X_{i,t}$), et $\Psi : \mathbb{R} \to \mathbb{R}$ convexe croissante. On a alors

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

Cette inégalité de symétrisation permet de se ramener à une borne sur

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

qui est souvent plus simple à étudier (dans le cas où $X_{i,t} \in \mathbb{R}$, c'est un sup de sommes de termes indépendants (les ε_i)). Cela permet aussi de majorer les espérances des supremum de processus empirique en utilisant des notions de "dimension" de l'espace parcouru par les $(X_{1,t}, \dots, X_{n,t})_{t\in\mathcal{T}}$ (le cas extrême étant $\mathcal{T} < +\infty$).

Preuve du lemme de symétrisation. On commence par se donner $(X'_{1,t}, \dots, X'_{n,t})$, copie indépendante de $(X_{1,t}, \dots, X_{n,t})$. On peut alors remarquer que $(X_{i,t} - X'_{i,t})$ est symétrique, et est donc de même loi que $\varepsilon_i(X_{i,t} - X'_{i,t})$. On écrit alors

$$\begin{aligned}
\mathbb{E}\left[\psi\left(\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n (X_{i,t} - \mathbb{E}(X_{i,t}))\right\|\right)\right] &= \mathbb{E}\left[\psi\left(\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n (X_{i,t} - \mathbb{E}(X'_{i,t}))\right\|\right)\right] \\
&\leq \mathbb{E}\left[\psi\left(\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n (X_{i,t} - X'_{i,t})\right\|\right)\right] \quad \text{Jensen} \\
&\leq \mathbb{E}\left[\psi\left(\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n \varepsilon_i(X_{i,t} - X'_{i,t})\right\|\right)\right] \\
&\leq \mathbb{E}\left[\psi\left(\frac{1}{2}\left(2\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n \varepsilon_i X_{i,t}\right\| + 2\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n \varepsilon_i X'_{i,t})\right\|\right)\right)\right] \\
&\qquad \psi \text{ croissante et } \varepsilon_i \sim -\varepsilon_i \\
&\leq \mathbb{E}\left[\psi\left(2\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n \varepsilon_i X_{i,t}\right\|\right)\right] \quad \psi \text{ convexe et } X_{i,t} \sim X'_{i,t}.
\end{aligned}$$

□

On retourne à la preuve du Théorème. On a $X_1, \dots, X_n$ i.i.d. de loi P, et on note $Y_{i,t} = \mathbb{1}_{X_i \leq t}$, pour $t \in \mathbb{R}$, de sorte que

$$\sqrt{n} D_n = \sup_{t \in \mathbb{R}} \left| \sum_{i=1}^n Y_{i,t} - \mathbb{E}(Y_{i,t}) \right| := Z.$$

Pour obtenir une inégalité de déviation sur Z, il est courant de regarder $\mathbb{E}(\exp(\lambda Z))$, pour $\lambda > 0$, et d'utiliser une inégalité de Markov ensuite (c'était la même recette pour l'inégalité de Hoeffding). On peut alors utiliser le Lemme de symétrisation, avec $\psi(x) = \exp(\lambda x)$, ce qui donne

$$\begin{aligned} \mathbb{E}\left(\exp(\lambda Z)\right) &\leq \mathbb{E}\left(\exp(2\lambda \sup_{t \in \mathbb{R}} \left| \sum_{i=1}^n \varepsilon_i Y_{i,t} \right|) \right) \\ &= \mathbb{E}\left(\sup_{t \in \mathbb{R}} \exp(2\lambda \left| \sum_{i=1}^n \varepsilon_i Y_{i,t} \right|) \right). \end{aligned}$$

Maintenant, on peut regarder plus précisément le terme intégré. De fait, si on réordonne $X_1, \dots, X_n$ en $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, on peut écrire

$$\begin{aligned} \sup_{t \in \mathbb{R}} \exp(2\lambda \left| \sum_{i=1}^n \varepsilon_i Y_{i,t} \right|) &= \sup_{t \in \mathbb{R}} \exp(2\lambda \left| \sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i \leq t} \right| \\ &= \sup_{t \in \mathbb{R}} \exp(2\lambda \left| \sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_{(i)} \leq t} \right| \\ &= \max_{k=1,\dots,n} \exp(2\lambda \left| \sum_{i=1}^k \varepsilon_{(i)} \right|), \end{aligned}$$

où les $\varepsilon_{(i)}$ sont réordonnés suivant l'ordre des X_i. Comme les ε_i sont échangeables, on a, pour toute permutation σ de $[\![1, \dots, n]\!]$,

$$(\varepsilon_{\sigma(1)}, \dots, \varepsilon_{\sigma(n)}) \sim (\varepsilon_1, \dots, \varepsilon_n).$$

On en déduit alors, les ε_i étant indépendants des X_i, que

$$\max_{k=1,\dots,n} \exp(2\lambda \left| \sum_{i=1}^k \varepsilon_{(i)} \right|) \sim \max_{k=1,\dots,n} \exp(2\lambda \left| \sum_{i=1}^k \varepsilon_i \right|),$$

et donc

$$\mathbb{E}\left[\sup_{t \in \mathbb{R}} \exp(2\lambda \left| \sum_{i=1}^n \varepsilon_i Y_{i,t} \right|) \Big| X_1, \dots, X_n \right] = \mathbb{E}\left[\max_{k=1,\dots,n} \exp(2\lambda \left| \sum_{i=1}^k \varepsilon_i \right|) \right].$$

Il s'agit maintenant de regarder les moments exponentiels de la marche aléatoire symétrique réfléchie. On peut commencer par utiliser l'égalité

$$\mathbb{E}\left[\max_{k=1,\dots,n} \exp(2\lambda \left| \sum_{i=1}^k \varepsilon_i \right|) \right] = \int_0^{+\infty} \mathbb{P}\left(\exp(2\lambda \left| \sum_{i=1}^k \varepsilon_i \right|) \geq u \right) du,$$

valable pour toute variable aléatoire positive, et regarder la probabilité que cette marche aléatoire dépasse un certain seuil $u > 0$. En notant A_k l'évènement

$$A_k = \left\{ \left| \sum_{i=1}^k \varepsilon_i \right| \geq u \right\} \cap \left(\bigcap_{j<k} \left\{ \left| \sum_{i=1}^j \varepsilon_i \right| < u \right\} \right),$$

on a

$$\left\{\max_{k=1,\dots,n}\exp(2\lambda\left|\sum_{i=1}^{k}\varepsilon_i\right|)\geq u\right\}=\bigcup_{k=1}^{n}A_k,$$

avec une union disjointe. Ensuite, à k fixé, on remarque que

$$\mathbb{P}\left\{\sum_{j=k+1}^{n}\varepsilon_j\geq 0\right\}=\mathbb{P}\left\{\sum_{j=k+1}^{n}\varepsilon_j\leq 0\right\}\geq\frac{1}{2},$$

par symétrie des ε_i. On en déduit alors

$$\mathbb{P}\left(\exp(2\lambda\left|\sum_{i=1}^{n}\varepsilon_i\right|)\geq u\Big|A_k\right)\geq\frac{1}{2}.$$

Maintenant, on peut écrire

$$\begin{aligned}\mathbb{P}\left(\left\{\max_{k=1,\dots,n}\exp(2\lambda\left|\sum_{i=1}^{k}\varepsilon_i\right|)\geq u\right\}\right)&=\sum_{k=1}^{n}\mathbb{P}(A_k)\\&\leq\sum_{k=1}^{n}\mathbb{P}(A_k)\left[2\mathbb{P}\left(\exp(2\lambda\left|\sum_{i=1}^{n}\varepsilon_i\right|)\geq u\Big|A_k\right)\right]\\&\leq 2\mathbb{P}\left(\exp(2\lambda\left|\sum_{i=1}^{n}\varepsilon_i\right|)\geq u\right).\end{aligned}$$

On en déduit

$$\begin{aligned}\mathbb{E}\left[\max_{k=1,\dots,n}\exp(2\lambda\left|\sum_{i=1}^{k}\varepsilon_i\right|)\right]&\leq 2\mathbb{E}\left[\exp(2\lambda\left|\sum_{i=1}^{n}\varepsilon_i\right|)\right]\\&\leq 2\mathbb{E}\left[\exp(2\lambda\sum_{i=1}^{n}\varepsilon_i)+\exp(-2\lambda\sum_{i=1}^{n}\varepsilon_i)\right]\\&\leq 4\mathbb{E}\left[\exp(2\lambda\sum_{i=1}^{n}\varepsilon_i)\right]\qquad\text{(symétrie)}\\&\leq 4\exp(2n\lambda^2)\qquad\text{(Hoeffding).}\end{aligned}$$

On conclut en prenant $\lambda=\varepsilon/(4\sqrt{n})$ et en utilisant l'inégalité de Markov. □

Extension ? : famille de lois (paramétrée), test asymptotique ou astuces

6.2 Estimation de densité

Les deux exemples standards en estimation non paramétriques sont

— la *régression* : on suppose qu'on observe $(X_1,Y_1),\dots,(X_n,Y_n)$ correspondant à n réalisations i.i.d. du modèle

$$Y=f(X)+\varepsilon,$$

où ε est un bruit indépendant de X et d'espérance nulle, et f est une fonction dans une classe $\mathcal{F}$ quelconque. On a vu dans la Section 2.3 que si $\mathcal{F}=\{\langle\theta,.\rangle\mid\theta\in\mathbb{R}^d\}$, alors ce problème est un problème de régression linéaire (paramétrique). Si $\mathcal{F}$ est une classe plus générale (par exemple ensemble des fonctions continues), cela devient un problème non-paramétrique (l'ensemble des paramètres n'est plus de dimension finie).

— *L'estimation de densité* : on suppose que l'on observe $X_1, \dots, X_n$ i.i.d. de densité inconnue f que l'on suppose appartenir à une classe de régularité.

Les méthodes et vitesses de convergences en régression et estimation de densité sont similaires, aussi on ne traitera ici que le cas de l'estimation de densité. Pour simplifier un peu, on prendra $X_1, \dots, X_n$ des variables aléatoires sur $]0,1[$, (on peut généraliser dans $\mathbb{R}^d$), et on regardera, pour un estimateur $\hat{f}$ de f, le risque quadratique intégré (MISE), c'est à dire la fonction de perte

$$\ell(\hat{f}, f) = \int_0^1 (\hat{f} - f)^2(x)dx.$$

Pour que cette perte soit bien définie, on suppose que la densité sous-jacente $f \in L^2(]0,1[)$.

6.2.1 Histogrammes, noyaux et consistance

D'après le paragraphe précédent on sait que F_n tend vers la bonne fonction de répartition, pour la norme infinie en probabilité. On peut alors être tenté de dériver F_n, ce qui donnerait $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ (soit la distribution empirique). Bien que cette dérivation ait un sens en termes de distributions, on ne peut pas espérer approcher f dans L^2 par ce biais.

Estimateur par histogrammes

Une première approche pour essayer de "lisser" $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ peut être de découper $]0,1[$ en N cases $(A_j)_{j=1,\dots,N}$ de longueur $1/N$, et de prendre pour estimateur de f l'estimateur par histogrammes correspondant

$$\begin{aligned}\hat{f}_{n,N}(x) &= \sum_{j=1}^N \frac{P_n(A_j)}{\mathrm{Vol}(A_j)} \mathbb{1}_{A_j}(x) = \sum_{j=1}^N \frac{\sum_{i=1}^n \mathbb{1}_{A_j}(X_i)}{n\mathrm{Vol}(A_j)} \mathbb{1}_{A_j}(x) \\ &= \sum_{j=1}^N \frac{(F_n(j/N) - F_n((j-1)/N))}{\mathrm{Vol}(A_j)} \mathbb{1}_{A_j}(x).\end{aligned}$$

Cela donne un estimateur de densité constant par morceaux. En notant $\mathcal{F}_A$ l'ensemble des fonctions L^2 constantes sur les $(A_j)_{j=1,\dots,n}$, on peut remarquer que à partition fixe, on aura

$$\hat{f}_{n,N}(x) \longrightarrow_{n\to+\infty} \bar{f}_N(x) = \pi_{\mathcal{F}_A}(f),$$

presque sûrement, où $\bar{f}_N$ est la projection de f au sens L^2 sur $\mathcal{F}_A$ s'écrivant

$$\bar{f}_N(x) = \sum_{j=1}^N \frac{P(A_j)}{\mathrm{Vol}(A_j)} \mathbb{1}_{A_j}(x) = \sum_{j=1}^N \left(\frac{1}{\mathrm{Vol}(A_j)} \int_{A_j} f(u)du \right) \mathbb{1}_{A_j}(x).$$

Dès lors, pour N grand $\bar{f}_N$ devrait être proche de f, et, lorsque n grandit, $\hat{f}_{n,N}$ devrait converger vers $\bar{f}_N$. Il reste à déterminer les régimes de croissance respectifs pour prouver la consistance de l'estimateur.

Proposition 6.9

Consistance des estimateurs par histogramme Supposons que $f \in L^2(]0,1[)$, et notons $h_n = 1/(N_n)$ (N_n représentant un choix de nombre de cases dépendant de n). Si $h_n \to 0$ et $nh_n \to +\infty$, alors

$$\|\hat{f}_n - f\|_{L^2} \to 0,$$

en probabilité.

Avant de passer à la preuve, quelques commentaires sur les hypothèses de consistance. L'hypothèse $h_n \to 0$ est là pour que le *biais* $\|\bar{f}_{N_n} - f\|$ tende vers 0. L'hypothèse $nh_n \to +\infty$ y est pour que le terme de variance $\|\bar{f}_{N_n} - \hat{f}_n\|$ tende vers 0 : elle prescrit que le nombre moyen de points dans une case (nh_n) doit tendre vers $+\infty$ pour pouvoir estimer efficacement le coefficient de chaque case. Cette heuristique se retrouvera pour quasiment tous les estimateurs non paramétriques, notamment ceux à noyaux. En dimension d, le nombre moyen de points par case devient nh_n^d, et l'hypothèse correspondante devient alors naturellement $nh_n^d \to +\infty$.

Démonstration. Commençons par regarder le terme de biais.

Analyse du terme de biais.

On aura besoin d'un résultat standard en analyse fonctionnelle.

Lemme 6.10 : Densité des C_c^∞ dans L^p

Si Ω est un ouvert de $\mathbb{R}^d$, alors les fonctions C^∞ à support compact inclus dans Ω sont denses dans $L^p(\Omega)$, pour tout $p \geq 1$.

Maintenant, si f est C^∞ à support compact, notons $L = \|f'\|_\infty$ sa constante de Lipschitz. On a alors, pour tout $x \in A_j$,

$$\begin{aligned} |\bar{f}_{N_n}(x) - f(x)| &= \frac{1}{\mathrm{Vol}(A_j)} \left| \int_{A_j} (f(u) - f(x)) du \right| \\ &\leq \frac{1}{\mathrm{Vol}(A_j)} \int_{A_j} L|u - x| du \leq L h_n. \end{aligned}$$

On en déduit alors

$$\|\bar{f}_{N_n} - f\|_{L^2} \leq L h_n.$$

Maintenant, si $f \in L^2$ et $g \in C_c^\infty$ sont telles que $\|f - g\|_{L^2} \leq \varepsilon$, on a

$$\begin{aligned} \|f - \bar{f}_{N_n}\|_{L^2} &\leq \|g - \bar{g}_{N_n}\|_{L^2} + \|(f - g) - \pi_{\mathcal{F}_A}(f - g)\|_{L^2} \\ &\leq 2\|f - g\|_{L^2} + \|g - \bar{g}_{N_n}\|_{L^2} \\ &\leq 2\varepsilon + \|g - \bar{g}_{N_n}\|_{L^2}. \end{aligned}$$

Comme $h_n \to 0$, on en déduit que pour tout $\varepsilon > 0$, $\limsup_n \|f - \bar{f}_{N_n}\|_{L^2} \leq 2\varepsilon$, soit $\|f - \bar{f}_{N_n}\|_{L^2} \to 0$.

Analyse du terme de variance

On remarque que si $x \in A_j$,

$$\hat{f}_n(x) \sim \frac{1}{n\text{Vol}(A_j)}\mathcal{B}(n, p_j),$$

où $p_j = P(A_j)$. On peut alors en déduire que

$$\mathbb{E}((\hat{f}_n(x) - \bar{f}_{N_n}(x))^2) = \frac{1}{n^2\text{Vol}^2(A_j)}\text{Var}(\mathcal{B}(n, p_j)) \leq \frac{p_j}{n\text{Vol}^2(A_j)}.$$

En intégrant on obtient

$$\int_0^1 \mathbb{E}((\hat{f}_n(x) - \bar{f}_{N_n}(x))^2)dx \leq \sum_{j=1}^{N_n} \frac{p_j}{n\text{Vol}(A_j)} \leq \frac{\sum_{j=1}^{N_n} p_j}{nh_n} = \frac{1}{nh_n}.$$

Par Fubini on en déduit que $\mathbb{E}(\|\hat{f}_n - \bar{f}_{N_n}\|_{L^2}^2) \to 0$ dès lors que $nh_n \to +\infty$, et donc la convergence en proba vers 0. □

Concluons cette partie sur la remarque suivante : à choix de cellules fixées A, estimer une densité par histogramme est en fait un problème d'estimation paramétrique. En effet, cela revient à vouloir estimer $\pi_{\mathcal{F}_A}(f)$, qui peut être décrit par un paramètre $\theta \in \mathbb{R}^{N_n}$ (avec $\theta_j = p_j/\text{Vol}(A_j)$). De fait, l'estimateur par histogramme est un M-estimateur : pour $\theta \in \mathbb{R}^{N^n}$ notons $f_\theta = \sum_{j=1}^{N_n} \theta_j \mathbb{1}_{A_j}$. Le critère théorique est

$$\begin{aligned} M(\theta) &= \int_0^1 (f - f_\theta)^2(u)du \\ &= \int_0^1 f^2(u)du - 2\int_0^2 f(u)f_\theta(u)du + \int_0^1 f_\theta^2(u)du, \end{aligned}$$

et $\bar{f}_{N_n}$ correspond bien à $\arg\min_\theta M(\theta)$. Comme le premier terme ne dépend pas de θ on peut choisir M de manière équivalente via

$$M(\theta) = -2\int_0^1 f(u)f_\theta(u)du + \int_0^1 f_\theta^2(u)du.$$

Le critère empirique associé est alors

$$\begin{aligned} M_n(\theta) &= -\frac{2}{n}\sum_{i=1}^n f_\theta(X_i) + \sum_{j=1}^{N_n} \text{Vol}(A_j)\theta_j^2 \\ &= -2\left(\sum_{j=1}^{N_n} \theta_j \frac{\left(\sum_{i=1}^n \mathbb{1}_{X_i \in A_j}\right)}{n}\right) + \sum_{j=1}^{N_n} \text{Vol}(A_j)\theta_j^2 \\ &= -2\left\langle \theta, (P_n(A_j))_{j=1,\dots,N_n} \right\rangle + \sum_{j=1}^{N_n} \text{Vol}(A_j)\theta_j^2. \end{aligned}$$

On peut alors vérifier que le M-estimateur associé est bien $\hat{f}_n$. Cette approche peut être utilisée pour donner des vitesses de convergence.

Estimateur par fenêtres glissantes

L'estimateur par fenêtres glissantes (ou estimateur de Rosenblatt) se base encore sur l'idée que l'on essaye plutôt d'estimer, pour un $x \in]0, 1[$, la valeur moyenne de $f(x)$ sur une petite case autour de x, mais cette fois-ci la case n'est pas fixe, elle

bouge continûment avec x. Plus précisément, pour une largeur de bande h, on définit l'estimateur par fenêtre glissante via

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n \mathbb{1}_{X_i \in [x-h, x+h]}}{2nh} = \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

Cet estimateur est un *estimateur à noyau* : en notant $K = \frac{1}{2}\mathbb{1}_{[-1,1]}$ (appelé noyau) et $K_h : u \mapsto \frac{1}{h}K\left(\frac{u}{h}\right)$ le noyau rescalé, l'estimateur par fenêtre glissante s'écrit

$$\begin{aligned}
\hat{f}_n(x) = (P_n * K_h)(x) &= \int_0^1 K_h(x-u) P_n(du) \\
&= \sum_{i=1}^n \frac{1}{n} K_h(x - X_i) \\
&= \frac{1}{n} \sum_{i=1}^n \frac{1}{2h} \mathbb{1}_{X_i \in [x-h, x+h]} \\
&= \frac{1}{2h}(F_n(x+h) - F_n(x-h)).
\end{aligned}$$

FAIRE DESSIN. Comme pour les histogrammes, le principe est de "lisser" la distribution empirique P_n en la convolant avec un noyau de manière à obtenir une densité. De ce fait on biaise le problème : à h fixé la limite de $\hat{f}_n$ sera $f_h = (P * K_h)$ et non pas f. Là encore, comme

$$f_h(x) = \frac{F(x+h) - F(x-h)}{2h} = \frac{1}{2h} \int_{x-h}^{x+h} f(u) du,$$

pour h assez petit tel que $[x-h, x+h] \subset]0,1[$, on aura bien $f_h \to f$ (sur l'ensemble des points de Lebesgue, donc presque partout) lorsque $h \to 0$. Sous les mêmes conditions que pour les histogrammes (taille de fenêtre qui tend vers 0, assez de points en moyenne dans chaque fenêtre), on a les mêmes garanties de convergence.

PROPOSITION 6.11 : CONSISTANCE DE L'ESTIMATION PAR NOYAUX

Si $f \in L^2(]0,1[)$ et h_n est une suite de fenêtres vérifiant $h_n \to 0$ et $nh_n \to +\infty$, alors l'estimateur à noyau $\hat{f}_n$ associé vérifie

1. *$\hat{f}_n$ est une densité,*
2. *$\mathbb{E}(\|\hat{f}_n - f\|_{L^2}^2) \longrightarrow_{n \to +\infty} 0$.*

Démonstration. Commençons par le premier point. $\hat{f}_n$ est bien mesurable et positive, il suffit donc de vérifier que $\int_0^1 \hat{f}_n(x)dx = 1$. Pour ce faire, on écrit

$$\begin{aligned}
\int_0^1 \hat{f}_n(x)dx &= \int_0^1 dx \int_0^1 K_h(x-u) P_n(du) \\
&= \int_0^1 P_n(du) \int_0^1 K_h(x-h)dx \qquad \text{(Fubini)} \\
&= \int_0^1 P_n(du) 1 = 1,
\end{aligned}$$

où on a utilisé $\int_0^1 K_h(x-u)dx = 1$ (c'était le but du rescaling). On passe maintenant à la consistance, en se basant sur la décomposition biais/ variance suivante :

$$\mathbb{E}(\|\hat{f}_n - f\|_{L^2}^2) = \|f - f_{h_n}\|_{L^2}^2 + \mathbb{E}(\|\hat{f}_n - f_{h_n}\|_{L^2}^2).$$

Anayse du terme de biais : Comme dans le cas par histogrammes on commence par supposer que f est C^∞ à support compact inclus dans $]0,1[$, et on note L sa constante de Lipschitz. L'inégalité fondamentale est la suivante : pour $x \in]0,1[$, on a

$$\begin{aligned}
|f_{h_n}(x) - f(x)| &= \left| \int_0^1 K_{h_n}(x-u) f(u) du - f(x) \right| \\
&= \left| \int_{x-h_n}^{x+h_n} K_{h_n}(x-u) f(u) du - f(x) \right| \qquad (n \text{ assez grand, support de } K_{h_n}) \\
&= \left| \int_{-h_n}^{h_n} K_{h_n}(v)(f(x-v) - f(x)) dv \right| \qquad (\text{chgt var et } \textstyle\int K_{h_n} = 1) \\
&\leq \int_{-h_n}^{h_n} K_{h_n}(v) |f(x-v) - f(x)| dv \\
&\leq L h_n \int_{-h_n}^{h_n} K_{h_n}(v) dv = L h_n \to 0 \qquad (f \text{ Lipschitz}).
\end{aligned}$$

On en déduit $\|f_{h_n} - f\|_{L^2} \leq L h_n \to 0$. Pour passer au cas général, on utilise l'inégalité d'Young, se traduisant ici par

$$\|(f-g) \star K_{h_n}\|_{L^2} \leq \|f-g\|_{L^2} \|K_{h_n}\|_1 = \|f-g\|_{L^2}.$$

Soit maintenant $f \in L^2$ et $\varepsilon > 0$. Il existe alors g fonction C^∞ à support compact inclus dans $]0,1[$ telle que $\|f-g\|_{L^2} \leq \varepsilon$. On peut alors écrire

$$\begin{aligned}
\|f - f_{h_n}\|_{L^2} &\leq \|g - g_{h_n}\|_{L^2} + \|f-g\|_{L^2} + \|f_{h_n} - g_{h_n}\|_{L^2} \\
&\leq \|g - g_{h_n}\|_{L^2} + \varepsilon + \|(f-g) * K_{h_n}\|_{L^2}\| \\
&\leq \|g - g_{h_n}\|_{L^2} + 2\varepsilon,
\end{aligned}$$

d'où on déduit que $\limsup_n \|f - f_{h_n}\|_{L^2} \leq 2\varepsilon$, et enfin $\|f - f_{h_n}\|_{L^2} \to 0$.

Analyse du terme de variance : On commence par remarquer que, pour $x \in]0,1[$,

$$\hat{f}_n(x) - f_{h_n}(x) = \frac{1}{n} \sum_{i=1}^n (K_{h_n}(x - X_i) - \mathbb{E}(K_{h_n}(x - X_i))),$$

de sorte que

$$\mathbb{E}(\hat{f}_n(x) - f_{h_n}(x))^2 = \frac{\mathrm{Var}(K_{h_n}(x - X_1)}{n},$$

ce qui mène, via une application de Fubini, à

$$\mathbb{E}(\int_0^1 \hat{f}_n(x) - f_{h_n}(x))^2) dx \leq \int_0^1 \frac{\mathrm{Var}(K_{h_n}(x - X_1)}{n} dx.$$

Par ailleurs, on peut majorer brutalement $\int_0^1 \mathrm{Var}(K_{h_n}(x-X_1)dx$ via

$$\begin{aligned}\int_0^1 \mathrm{Var}(K_{h_n}(x-X_1)dx &\leq \int_0^1 \mathbb{E}(K_{h_n}^2(x-X_1))dx \\ &= \int_0^1 dx \int_0^1 du \frac{K^2\left(\frac{x-u}{h_n}\right)}{h_n^2} f(u) \\ &= \frac{1}{h_n}\int_0^1 f(u)du \int_0^1 \frac{K^2\left(\frac{x-u}{h_n}\right)}{h_n} dx \\ &= \frac{1}{h_n}\int_0^1 f(u)du \int_{\mathbb{R}} K^2(y)dy \\ &= \frac{\|K\|_{L^2(\mathbb{R})}^2}{h_n},\end{aligned}$$

d'où on déduit

$$\mathbb{E}\left(\|\hat{f}_n - f_{h_n}\|^2\right) \leq \frac{\|K\|_{L^2}^2}{nh_n} \to 0.$$

□

De cette preuve on peut tirer quelques enseignements. Premièrement on peut avoir une borne pour une taille de n fixé pour $\mathbb{E}(\|\hat{f}_n - f\|^2)$ si f est Lipschitz. On détaillera la vitesse de convergence minimax associée dans la section qui suit. Dans un deuxième temps, on se rend compte que cette proposition reste valide si on prend pour noyau K une fonction qui vérifie

1. $\int K = 1$,
2. $\int K^2 < +\infty$,
3. K est à support compact.

On peut légèrement relâcher la dernière contrainte : il suffit que K décroisse suffisamment vite. En pratique, beaucoup d'autres noyaux peuvent être utilisés. Citons les exemples les plus célèbres. DESSINS

— Rectangulaire, $\frac{1}{2}\mathbb{1}_{|x|\leq 1}$.
— Triangulaire, $(1-|x|)\mathbb{1}_{|x|\leq 1}$.
— Parabolique (Epanechnikov), $\frac{3}{4}(1-x^2)\mathbb{1}_{|x|\leq 1}$.
— Biweight, $\frac{15}{16}(1-x^2)^2\mathbb{1}_{|x|\leq 1}$.
— Gaussien, $\frac{1}{\sigma\sqrt{2\pi}}\exp(-x^2/(2\sigma^2))$.

En pratique, le choix du noyau se fait souvent au doigt mouillé après quelques essais. D'un point de vue théorique on montrera plus loins que des noyaux particuliers sont préférables lorsque l'on sait que la densité cible est suffisamment régulière.

6.2.2 Vitesses de convergence sur des classes de régularité

Commençons par un résultat pessimiste : bien que les estimateurs à noyaux soient consistant sur l'ensemble des densités dans $L^2(]0,1[)$, on ne peut pas espérer que cette consistance soit uniforme.

Théorème 6.12 : Indécidabilité minimax du problème d'estimation de densité

Pour tout $n \geq 1$, on a

$$\inf_{\hat{f}_n} \sup_{f \in L^2(]0,1[)} \mathbb{E}(\|\hat{f}_n - f\|^2) \geq 1,$$

où l'infimum est pris sur l'ensemble des estimateurs possibles basés sur n réalisations i.i.d. de f.

Avant de passer à la preuve, quelques commentaires. Comparé au cadre paramétrique où on peut espérer des vitesses en d/n, le cadre non-paramétrique correpond au cas $d = +\infty$, en ce sens cette borne est cohérente avec l'intuition que l'espace des densité L^2 est trop gros pour espérer avoir une vitesse de convergence uniforme tendant vers 0. On verra par la suite que ce n'est pas forcément la dimension "linéaire" de l'espace considéré qui est en jeu : des vitesses de convergences peuvent être obtenues sur des espaces de dimension infinies sous certaines conditions. Techniquement parlant, cette borne inférieure est assez proche du "No-free Lunch Theorem" en classification : il s'agira de construire un gros sous-ensemble paramétrique de L^2 pour lequel aucun classifieur ne pourra être uniformément bon. La construction exacte pourra être réutilisée par la suite dans le cadre des densités régulières.

Démonstration. On commence par découper $]0,1[$ en m intervalles $A_j = [\frac{j}{m}, \frac{j+1}{m}[$, $j = 0, \dots, m-1$. Maintenant, on définit deux motifs de base sur $]0,1[$ par

$$f_0(u) = 2\mathbb{1}_{u \leq \frac{1}{2}},$$
$$f_1(u) = 2 - f_0(u),$$

de telle sorte que $\int f_0 = \int f_1 = 1$, et $\int (f_1 - f_0)^2 = 4$. Pour $\varepsilon \in \{0,1\}^m$, on définit la densité f_ε par

$$f_\varepsilon(u) = \sum_{j=0}^{m} f_{\varepsilon_j}\left(m(u - (j/m)\right) \mathbb{1}_{u \in A_j}.$$

FAIRE DESSIN. On se donne maintenant une loi a priori uniforme sur $\{0,1\}^m$ via $\tilde{\varepsilon} \sim \bigotimes_{j=1}^m \mathcal{B}(1/2)$, et $\tilde{X}_{1:n} = (\tilde{X}_1, \dots, \tilde{X}_n)$ tel que $\tilde{X}_{1:n} \mid \tilde{\varepsilon} \sim (f_{\tilde{\varepsilon}}(u)du)^{\otimes n}$. Comme en Bayésien classique, pour un estimateur $\hat{f}$, on aura toujours

$$\begin{aligned}\sup_{f \in L^2(]0,1[)} \mathbb{E}\|\hat{f} - f\|_{L^2}^2 &\geq \sup_{\varepsilon \in \{0,1\}^m} \mathbb{E}\|\hat{f} - f_\varepsilon\|_{L^2}^2 \\ &\geq \mathbb{E}(\|\hat{f}(\tilde{X}_{1:n}) - f_{\tilde{\varepsilon}}\|_{L^2}^2).\end{aligned}$$

Il s'agit maintenant de minorer un risque Bayésien. Regardons un peu la loi a posteriori. Le modèle étant dominé par $\mathcal{L}_n$, on peut écrire, pour $x = x_{1:n} \in]0,1[^n$ et $varepsilon \in \{0,1\}^m$,

$$\begin{aligned}Q_x(\{\varepsilon\}) &\propto \prod_{i=1}^{n} \sum_{j=0}^{m} f_{\varepsilon_j}\left(m(x_i - (j/m)\right) \mathbb{1}_{x_i \in A_j} \\ &\propto \prod_{j \in S(x)} \left(\prod_{x_i \in A_j} f_{\varepsilon_j}\left(m(x_i - (j/m)\right) \right),\end{aligned}$$

où $S(x) = \{j \in [\![0, m-1]\!] \mid x \cap A_j \neq \emptyset\}$ (l'ensemble des cases visitées). On en déduit alors que les $\varepsilon_j \mid \tilde{X}$ sont indépendantes, et que, si $j \notin S(\tilde{X})$, $\varepsilon_j \mid \tilde{X} \sim \mathcal{B}(0, 1/2)$. Si $\hat{f}$ est un estimateur de f, on peut alors minorer son risque a posteriori par

$$\begin{aligned}
\mathbb{E}\left(\|\hat{f} - f_{\tilde{\varepsilon}}\|_{L^2}^2 \mid \tilde{X}\right) &\geq \sum_{j \notin S(\tilde{X})} \int_{A_j} \mathbb{E}[(\hat{f}(u) - f_{\tilde{\varepsilon}_j}(m(u - (j/m)))^2 \mid \tilde{X}]du \\
&\geq \sum_{j \notin S(\tilde{X})} \int_{A_j} \mathrm{Var}(2\mathcal{B}(1/2))du \\
&\geq \left(1 - \frac{|S(\tilde{X})|}{m}\right) \\
&\geq \left(1 - \frac{n}{m}\right).
\end{aligned}$$

On en déduit alors que, pour tout $m \geq 1$ et estimateur $\hat{f}$,

$$\sup_{f \in L^2(]0,1[)} \mathbb{E}\|\hat{f} - f\|_{L^2}^2 \geq \left(1 - \frac{n}{m}\right) \longrightarrow_{m \to +\infty} 1,$$

d'où le résultat. □

Il est donc illusoire d'espérer des performances uniformes sur L^2 tout entier. En revanche, si l'on suppose que f est suffisamment régulière, on peut espérer des bornes uniformes.

Proposition 6.13 : Vitesse sur les densités Lipschitz

Soit Lip_L *l'ensemble des fonctions L-Lipschitz de* $]0, 1[$. *Si on note* $\hat{f}_n$ *l'estimateur à noyau (rectangulaire) pour une fenêtre* h_n, *on a*

$$\sup_{f \in \mathrm{Lip}_L} \mathbb{E}\|\hat{f}_n - f\|^2 \leq L^2 h_n^2 + \frac{1}{nh_n}.$$

Par conséquent, en choisissant $h_n = (2nL^2)^{-\frac{1}{3}}$, *on a*

$$\sup_{f \in \mathrm{Lip}_L} \mathbb{E}\|\hat{f}_n - f\|^2 \leq 3L^{2/3} n^{-\frac{2}{3}}.$$

Cette vitesse bizarre est typique du non-paramétrique : elle est plus lente que le "paramétrique" $n^{-1/2}$ et fait jouer pleinement la régularité de f. On verra par la suite que c'est un cas particulier d'une vitesse générale (dépendant de la régularité et de la dimension de l'espace de départ). La preuve est simple.

Démonstration. On reprend la preuve de la consistance. On commence par remarquer que

$$\|f_{h_n} - f\|_{L^2} \leq L h_n,$$

puis que $\mathbb{E}(\|\hat{f}_n - f_{h_n}\|^2) \leq \frac{\|K\|_{L^2(\mathbb{R})}^2}{nh_n} = \frac{1}{nh_n}$. L'inégalité biais/variance permet de conclure. En choisissant $h_n = (2nL^2)^{-1/3}$, on arrive à

$$\mathbb{E}\|\hat{f}_n - f\|^2 \leq L^2 h_n^2 + \frac{1}{nh_n} = L^{2/3} n^{-\frac{2}{3}} (4^{-1/3} + 2^{1/3}) \leq 3L^{2/3} n^{-\frac{2}{3}}.$$

□

On verra par la suite que cette vitesse est optimale. L'estimateur à noyau rectangulaire n'est pas le seul à atteindre cette borne. On peut par exemple définir un estimateur par projection comme suit. On se donne $\phi_0, \dots, \phi_n, \dots$ la base de Fourier $e^{2ik\pi x}$, et pour $f \in \mathrm{Lip}_L$, $\alpha_j(f) = \langle f, \phi_j \rangle$ ses coefficients de Fourier. On peut alors être tenté d'estimer ces coefficients via

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i),$$

et de considérer l'estimateur $\hat{f}_m(x) = \sum_{j=0}^{m} \hat{\alpha}_j \phi_j(x)$, pour un m bien choisi. La question est : comment choisir ce m ? Là encore c'est une histoire de décomposition biais/variance. On peut écrire, à m fixé,

$$\mathbb{E}\|\hat{f}_m - f\|_{L^2}^2 = \sum_{j=1}^{m} \mathbb{E}(|\hat{\alpha}_j - \alpha_j|^2) + \sum_{j \geq m+1} \alpha_j^2.$$

Pour un j fixé, on a

$$\begin{aligned} \mathbb{E}(|\hat{\alpha}_j - \alpha_j|^2) &= \mathbb{E}\left| \frac{1}{n} \sum_{i=1}^{n} (\phi_j(X_i) - \mathbb{E}(\phi_j(X_i))) \right|^2 \\ &\leq \frac{\mathbb{E}|\phi_j|^2(X)}{n} \leq \frac{1}{n}, \end{aligned}$$

et donc le terme de variance est majoré par m/n. Pour le terme de biais, comme f est L-Lipschitz, ses coefficients vérifient

$$\sum_{j \geq 0} j^2 |\alpha_j|^2 \leq L^2,$$

ce dont on peut déduire

$$\sum_{j \geq m+1} |\alpha_j|^2 \leq \frac{L^2}{m^2}.$$

En choisissant $m \sim n^{\frac{1}{3}}$, on retrouve une vitesse en $n^{-2/3}$, qui est la vitesse optimale sur cette classe. Cet exemple montre que la frontière entre paramétrique et non-paramétrique ne doit pas être considérée comme "stricte". Les estimateurs par projection fournissent une des passerelles entre ces deux mondes.

On peut aller plus loin dans les vitesses pour des fonctions plus régulières. Pour $\beta, L > 0$, on note $\Sigma(\beta, L)$ l'ensemble des fonctions f telles que

— f est $\ell = \lceil \beta \rceil$-différentiable sur $]0,1[$,

— pour tout $x, y \in]0,1[$, $|f^\ell(x) - f^\ell(y)| \leq L|x-y|^{\beta-\ell}$.

On appelle cette classe une classe β-Hölder. Associée à une classe β-Hölder, on peut introduire une classe de noyaux d'odre ℓ. Un noyau K est d'odre ℓ si, en plus d'être un noyau, il vérifie

— $\int_{\mathbb{R}} u^j K(u) du = 0$ (et est bien définie), pour tout $j \leq \ell$.

On peut prouver que de tels noyaux existent, pour tout ordre. En revanche ils ne seront pas forcément positifs (au delà de l'ordre 1 c'est même impossible). On peut alors prouver le résultat suivant.

Théorème 6.14 : Vitesses de convergence pour des densités régulières

Supposons $f \in \Sigma(\beta, L)$, et soit K un noyau d'ordre $\ell = \lceil \beta \rceil$, vérifiant par ailleurs $\int_{\mathbb{R}} |u|^\beta |K(u)| du < +\infty$. Soit $\hat{f}_n$ l'estimateur de noyau K, avec pour fenêtre $h_n \sim n^{-\frac{1}{2\beta+1}}$. On a alors

$$\sup_{f \in \Sigma(\beta,L)} \mathbb{E}(\|\hat{f}_n - f\|_{L^2}^2) \leq C(L,\beta) n^{-\frac{2\beta}{2\beta+1}},$$

où $C(L,\beta)$ est une constante ne dépendant que de β et de L. Par ailleurs, la vitesse $n^{-\frac{2\beta}{2\beta+1}}$ est la vitesse minimax sur cette classe.

Démonstration. La borne inférieure est trop technique pour être enseignée à ce niveau (bien que cela reste du Bayésien, en un peu plus évolué). Le lecteur intéressé est renvoyé au Tsybakov. La borne supérieure en revanche est plutôt facile. On décompose encore notre risque en un biais et une variance :

$$\mathbb{E}(\|\hat{f}_n - f\|_{L^2}^2) = \|f - \bar{f}\|_{L^2}^2 + \mathbb{E}(\|\bar{f} - \hat{f}_n\|^2),$$

avec $\bar{f} = K * f$. Le terme de variance se traite de la même manière que précédemment :

$$\mathbb{E}(\|\bar{f} - \hat{f}_n\|^2) \leq \int_0^1 \frac{\mathrm{Var}(K_{h_n}(x - X_1)}{n} dx \leq \frac{\|K\|_{L^2(\mathbb{R})}^2}{n h_n}.$$

Le terme de biais se base sur la décomposition suivante : pour un x fixé dans $]0,1[$ et un y tel que $|x - y| \leq h_n$, on a

$$f(y) = f(x) + f'(x)(y-x) + \sum_{j=2}^{\ell-1} f^{(j)}(x) \frac{(y-x)^j}{j!} + \frac{(y-x)^\ell}{\ell!} f^{(\ell)}(\tau(y-x)),$$

pour un $\tau \in]-1,1[$. On peut alors écrire

$$\begin{aligned}
\bar{f}(x) - f(x) &= \int_R K_h(x-z)(f(z) - f(x)) dz \\
&= \int_{x-h}^{x+h} \frac{1}{h} K\left(\frac{x-z}{h}\right)(f(z) - f(x)) dz \\
&= \int_{-1}^{1} K(u)(f(x+hu) - f(x)) du \\
&= \int_{-1}^{1} K(u)(\sum_{j=1}^{\ell-1} \frac{h^j u^j}{j!} f^{(j)}(x) + \frac{u^\ell h^\ell}{\ell!}(f^{(\ell)}(x+\tau hu) - f^{(\ell)}(x))) du \\
&= \frac{h^\ell}{\ell!} \int_{-1}^{1} K(u) u^\ell (f^{(\ell)}(x+\tau hu) - f^{(\ell)}(x)) du,
\end{aligned}$$

où on a retranché artificiellement $f^{(\ell)}(x))$ à l'avant dernière ligne. On en déduit alors, comme $f \in \Sigma(\beta, L)$,

$$\begin{aligned}
|\bar{f}(x) - f(x)| &\leq \frac{h^\ell}{\ell!} L \int_{\mathbb{R}} |K(u)||u|^\beta h^{\beta-\ell} \\
&\leq C(\beta) h^\beta L.
\end{aligned}$$

On conclut en remarquant que

$$\|\bar{f} - f\|_{L^2}^2 \leq C(\beta)^2 L^2 h^{2\beta},$$

et en prenant $h_n = n^{-\frac{1}{2\beta+1}}$. □

On remarque alors que, le cas Lipschitz correspondant au cas $\beta = 1$, la borne donnée précédemment est cohérente. Pour conclure, mentionnons que ce résultat peut s'étendre en dimension quelconque (pour des densités régulières définies sur un ouvert inclus dans $\mathbb{R}^d$). La vitesse non-paramétrique standard pour l'estimation d'une densité en norme p étant dans ce cas le fameux

$$n^{-\frac{\beta}{2\beta+d}},$$

qui est la vitesse minimax sur les classes de régularité β. On a donné les bornes pour la norme au carré précédemment, cela revient à remplacer β par 2β au numérateur. On peut remarquer que lorsque la régularité tend vers $+\infty$, à la limite on retrouve une vitesse paramétrique en $1/\sqrt{n}$. D'un point de vue estimation par projection, plus la régularité augmente moins on a besoin d'estimer de coefficient, la limite étant un nombre constant de coefficient, ce qui nous remet dans un cadre d'estimation paramétrique. La boucle est bouclée !

6.3 Estimation de support

Chapitre 7

Classif

7.1 Problème de classif

Def, fction de régression, classifieur de Bayes.

7.2 Apprentissage

Ppe, consistance, UC, no-free lunch, vitesses sur une classe. Biais Variance. Vitesses paramétriques et non-paramétriques

7.3 Ex paramétriques : classes de Vapnik

7.4 Ex non paramétrique

consistance kppv, vitesse histogrammes.