Statistiques

Gayral Léo
basées sur le cours de Clément Marteau

Table des matières

Chapitre 1

Estimation d'un paramètre réel

17/01

1.1. Vocabulaire

Définition 1.1 (Échantillon aléatoire)
Les variables aléatoires réelles $(X_i)_{1\leq i\leq n} \overset{\text{iid}}{\sim} \mathbb{P}$ forment un échantillon aléatoire de taille n.

Définition 1.2 (Estimation paramétrique)
Soit $\Theta \subset \mathbb{R}^p$ l'espace des paramètres envisagés. Cet espace est muni de $\theta \in \Theta \mapsto \mathbb{P}_\theta$ qui à un paramètre associe une loi de probabilités sur $\mathbb{R}$ (ou de façon équivalente une variable aléatoire sur l'espace probabilisé Ω non spécifié).

On suppose $\exists \theta^* \in \Theta,\ \mathbb{P}_{\theta^*} = \mathbb{P}$ avec $\mathbb{P}$ la loi qu'on cherche à estimer.

Définition 1.3 (Expérience statistique identifiable)
La paramétrisation $\theta \mapsto \mathbb{P}_\theta$ est injective : $\forall \theta, \theta' \in \Theta,\ \mathbb{P}_\theta \overset{\text{loi}}{\sim} \mathbb{P}_{\theta'} \ \Rightarrow\ \theta = \theta'$.

Définition 1.4 (Estimateur)
Soit $\phi : \mathbb{R}^n \to \Theta \subset \mathbb{R}^p$ mesurable. On définit $\hat{\theta}_n := \phi(X_1 ... X_n)$ l'estimateur associé à ϕ.

Autrement dit, un estimateur au rang n est une variable aléatoire $\hat{\theta}_n \in \Theta$ mesurable dans $\sigma(X_1 ... X_n)$.

Remarque 1.5
Soient $\theta \in \Theta$ et $\phi : \mathbb{R}^n \to \mathbb{R}$ mesurable : $\mathbb{E}_\theta[\phi] := \int\limits_{\mathbb{R}^n} \phi(x_1 ... x_n) d\mathbb{P}_\theta(x_1) ... d\mathbb{P}_\theta(x_n)$.

Cela revient à considérer une variable aléatoire mesurable selon $\sigma(X_1...X_n)$, en supposant $(X_i)_{1\leq i\leq n} \overset{\text{iid}}{\sim} \mathbb{P}_\theta$. Pour l'estimateur $\hat{\theta}_n$ associé à ϕ, on pose alors $\mathbb{E}_\theta[\hat{\theta}_n] := \mathbb{E}_\theta[\phi]$. On adopte la même convention pour $\text{Var}_\theta(\phi)$ et $\text{Var}_\theta(\hat{\theta}_n)$.

1.2. Propriétés d'un estimateur

1.2.1. Convergence, consistance.

Définition 1.6 (Estimateurs consistants)
La suite d'estimateurs $(\hat{\theta}_n)_{n\in\mathbb{N}}$ vérifie $\hat{\theta}_n \overset{\text{proba}}{\longrightarrow} \theta^*$.
Autrement dit, $\forall\epsilon > 0,\ \mathbb{P}(|\hat{\theta}_n - \theta^*| > \epsilon) \underset{n\to\infty}{\longrightarrow} 0$.

1.2.2. Biais, risque quadratique.

Définition 1.7 (Biais)
Un estimateur $\hat{\theta}_n$ est sans biais lorsque $\mathbb{E}_{\theta^*}[\hat{\theta}_n] = \theta^*$.

Définition 1.8 (Risque quadratique)
$R(\hat{\theta}_n, \theta^*) := \mathbb{E}_{\theta^*}[(\hat{\theta}_n - \theta^*)^2] = \text{Var}_{\theta^*}(\hat{\theta}_n) + (\mathbb{E}_{\theta^*}[\hat{\theta}_n] - \theta^*)^2$.
On dit que $(\mathbb{E}_{\theta^*}[\hat{\theta}_n] - \theta^*)^2$ est le terme de biais, nul ssi $\hat{\theta}_n$ est sans biais.

1.2.3. Robustesse.

Définition 1.9 (Robustesse)
Informellement, la robustesse d'un estimateur $\hat{\theta}_n$ associé à ϕ désigne le fait de préserver des bonnes propriétés lors de petites modifications dans les données et paramètres du modèle choisi.

Ainsi, il arrive que lors d'un sondage, une valeur extrême et rare apparaisse. On cherche à ce que ce genre de valeur ne change que de manière très faible la valeur de l'estimateur. On dit alors que l'estimateur est robuste.

Exemple 1.10
Si $X_n \sim \mathcal{N}(\theta^*, 1)$, alors les valeurs moyennes $\hat{\theta}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ donnent des estimateurs consistants, par loi des grands nombres.
Si on considère la petite modification $X_n \sim (1-\epsilon)\mathcal{N}(\theta^*, 1) + \epsilon \times \delta_{10^4}$, alors $(\overline{X}_n)$ ne sera plus consistante. Il faudrait ici considérer $\hat{\theta}_n = \text{mediane}(X_1...X_n)$.

1.3. Estimation par insertion

Définition 1.11 (Mesure empirique)
Soient (X_n) des variables aléatoires quelconques, à valeurs dans le même espace mesurable $(E, \mathcal{F})$. On a $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ la mesure empirique sur E : pour $\omega \in \Omega$ fixé, $\mathbb{P}_n(\omega)$ est une somme finie de diracs sur E.

Si on considère $E = \mathbb{R}$ et $f \in \mathcal{C}_b^0$, alors $\int f(x)d\mathbb{P}(x) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ une variable aléatoire bornée.

Théorème 1.12 (Varadarajan)
Presque-sûrement, on a la convergence $\mathbb{P}_n(\omega) \overset{\text{étroitement}}{\longrightarrow} \mathbb{P}_{\theta^*}$ des mesures empiriques : $\forall f \in \mathcal{C}_b^0(\mathbb{R},\mathbb{R})$, $\int_{\mathbb{R}} f(x)d\mathbb{P}_n(x) \xrightarrow[n\to\infty]{p.s.} \mathbb{E}_{\theta^*}[f]$.

1.4. Estimation par maximum de vraisemblance

Définition 1.13 (Vraisemblance)
Soit $X = (X_1...X_n) \in \mathbb{R}^n$. Lorsque la loi $\mathbb{P}_\theta$ est à densité f_θ, on pose $L_\theta : \mathbb{R} \to \mathbb{R}$ la densité de la variable X avec les $(X_i)_{1\leq i\leq n} \overset{\text{iid}}{\sim} \mathbb{P}_\theta$.
Lorsque la loi $\mathbb{P}_\theta$ est discrète, on pose $f_\theta(x) = \mathbb{P}_\theta(X_1 = x) = \mathbb{E}_\theta[\delta_{X_1=x}]$ et L_θ la loi jointe de X.

Autrement écrit, $L_\theta(x_1...x_n) = \prod_{i=1}^{n} f_\theta(x_i)$. On définit la vraisemblance de X par $\theta \mapsto L_\theta(X)$ une application de Θ vers les probabilités sur $\mathbb{R}$.

Définition 1.14 (Estimateur du maximum de vraisemblance)
$\widehat{\theta_{MV}} = \underset{\theta\in\Theta}{\operatorname{argmax}} L_\theta(X)$ une variable aléatoire dans Θ.

Remarque 1.15 (log-vraisemblance)
Pour simplifier les calculs, avec la convention $\ln(0) = -\infty$, on est souvent amené à considérer la log-vraisemblance $l_\theta(x) = \ln(L_\theta(x))$. En effet, par stricte croissance du log, $\widehat{\theta_{MV}} = \underset{\theta\in\Theta}{\operatorname{argmax}} l_\theta(X)$.

Remarque 1.16
$l_\theta(x) = \sum_{i=1}^{n} \ln(f_\theta(x))$. Par stricte croissance de $x \mapsto \frac{x-c}{n}$, on peut également écrire $\widehat{\theta_{MV}} = \underset{\theta\in\Theta}{\operatorname{argmax}} M_n(\theta, X)$ avec $M_n(\theta,x) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{f_\theta(x_i)}{f_{\theta^*}(x_i)}\right)$.
On a, indépendamment de n, $\mathbb{E}_{\theta^*}[M_n(\theta,x)] = M(\theta) = \int_{\mathbb{R}} \ln\left(\frac{f_\theta(x)}{f_{\theta^*}(x)}\right) f_{\theta^*}(x)dx$.

Définition 1.17 (Divergence de Kullback)
Soient $\mathbb{P} = p \times \mu$ et $\mathbb{Q} = q \times \mu$ deux mesures absolument continues pour μ.
On pose $K(\mathbb{P},\mathbb{Q}) = \int \ln\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P} = \int p \ln\left(\frac{p}{q}\right) d\mu$ lorsque $\mathbb{P} << \mathbb{Q}$ et $K(\mathbb{P},\mathbb{Q}) = \infty$ sinon.
Cette définition s'adapte bien au cas où $\mathbb{P}$ et $\mathbb{Q}$ discrètes.
Dans la remarque précédente, on a $M(\theta) = -K(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$.

Lemme 1.18
Si $\forall \theta \in \Theta,\ \mathbb{P}_\theta \overset{\text{loi}}{=} \mathbb{P}_\theta^* \ \Rightarrow\ \theta = \theta^*$, alors $\theta^* \in \Theta$ est l'unique paramètre où le maximum $M(\theta^*) = \max_{\theta \in \Theta} M(\theta)$ est atteint. Ceci est en particulier vrai si l'expérience est identifiable.

24/01

Démonstration.
En premier lieu, on a $M(\theta^*) = 0$.
D'autre part, comme $\ln(x) \leq 2(\sqrt{x} - 1)$, alors lorsque $\theta \neq \theta^*$ on a :
$M(\theta) \leq -\int_{\mathbb{R}} \left(\sqrt{f(x,\theta^*)} - \sqrt{f(x,\theta)}\right)^2 dx < 0.$ □

Définition 1.19 (Distance de Hellinger)
Soient $\mathbb{P}, \mathbb{Q}$ deux probabilités sur $(\Omega, \mathcal{F})$. On définit la distance H via :
$H^2(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int_\Omega \left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}}\right)^2 = 1 - \int_\Omega \sqrt{d\mathbb{P}} \times \sqrt{d\mathbb{Q}}.$
On pose $A(\mathbb{P}, \mathbb{Q}) = \int_\Omega \sqrt{d\mathbb{P}} \times \sqrt{d\mathbb{Q}}$ l'affinité de Hellinger.

1.5. Information de Fisher

Définition 1.20 (Modèle régulier)
On suppose que notre modèle vérifie les points suivants :

(1) $\Theta \subset \mathbb{R}$ un ouvert.
(2) $\forall \theta \in \Theta,\ f_\theta \in \mathcal{C}^0(\mathbb{R}, \mathbb{R}^{+*})$.
(3) $\forall x \in \mathbb{R}^n,\ \theta \mapsto L(x,\theta)$ est 2-dérivable (donc $\mathcal{C}^1$).
(4) $\left(x \mapsto \sup_{\theta \in \Theta} |\partial_\theta L(x,\theta)|\right) \in L^1(\mathbb{R}^n)$ pour la mesure de Lebesgue.
(5) $\left(x \mapsto \sup_{\theta \in \Theta} \left|\partial_\theta^2 L(x,\theta)\right|\right) \in L^1(\mathbb{R}^n)$ pour la mesure de Lebesgue.

Définition 1.21 (Fonction score)
Fonction score en $\theta \in \Theta$: $S_\theta : x \mapsto \partial_\theta l_\theta(x)$ la dérivée de la log-vraisemblance.

Proposition 1.22
Si X un modèle régulier, alors le score $S_\theta(X)$ est centré pour la loi $\mathbb{E}_\theta$.

Démonstration.
$\mathbb{E}_\theta[S_\theta(X)] = \sum_{i=1}^n \int_{\mathbb{R}} \partial_\theta(\ln(f)) \times f(x,\theta)dx = \sum_{i=1}^n \int_{\mathbb{R}} \partial_\theta f(x,\theta)dx$
$\mathbb{E}_\theta[S_\theta(X)] = \partial_\theta \left(\theta \mapsto \sum_{i=1}^n \int_{\mathbb{R}} f(x,\theta)dx\right) = \partial_\theta(\theta \mapsto n) = 0$ □

Définition 1.23 (Information de Fisher)
$I_n(\theta) := \mathrm{Var}_\theta(S_\theta(X)) = \mathbb{E}_\theta[(\partial_\theta l_\theta)^2(X)]$

Proposition 1.24
$I_n(\theta) = n \times I_1(\theta)$

Démonstration.
$I_n(\theta) = \mathrm{Var}_\theta\left(\sum_{i=1}^{n} S(X_i, \theta)\right) = \sum_{i=1}^{n} \mathrm{Var}_\theta(S(X_i, \theta)) = n\mathrm{Var}_\theta(S(X_1, \theta)).$ □

Proposition 1.25
$I_n(\theta) = -\mathbb{E}_\theta[\partial_\theta^2 l_\theta(X)]$

Démonstration.
cf. notes manuscrites. □

1.6. Inégalité de Cramér-Rao

Théorème 1.26
Soient X un modèle régulier tel que $I_1(\theta) < \infty$, g une fonction connue telle que $g(\theta)$ soit la quantité à approximer et $T_n(X)$ un estimateur sans biais de cette quantité.

Supposons que :

(1) $\partial_\theta\left(\int_{\mathbb{R}^n} T_n(x)L(x,\theta)dx\right) = \int_{\mathbb{R}^n} T_n(x)\partial_\theta L(x,\theta)dx$

(2) $\int_{\mathbb{R}^n} |T_n(x)\partial\theta L(\theta, x)|dx < \infty$

Alors $\mathrm{Var}_\theta(T_n(X)) \geq \frac{g'(\theta)^2}{I_n(\theta)}$.

En particulier, pour $g(\theta) = \theta$, on a $\mathrm{Var}_\theta(T_n(X)) = \mathbb{E}_\theta[(T_n(X)-\theta)^2] \geq \frac{1}{I_n(\theta)}$.

Démonstration.
cf. notes manuscrites. □

Définition 1.27 (Estimateur efficace)
On pose $B_n(\theta) = \frac{1}{I_n(\theta)}$ la borne de Cramér-Rao.

$\hat{\theta}_n$ un estimateur sans biais est efficace lorsque $\mathrm{Var}_\theta(\hat{\theta}_n) = B_n(\theta^*)$.

31/01

1.6.1. Efficacité Asymptotique.

Définition 1.28 (Biais)
Soit T_n un estimateur. On pose $b_n(\theta) = \mathbb{E}_\theta[T_n(X_1...X_n)] - \theta$ son biais.
Le risque quadratique peut s'écrire $R(T_n, \theta) = \mathrm{Var}_\theta(T_n) + b_n(\theta)^2$.

Remarque 1.29
Pour un estimateur T_n sans biais, l'inégalité de Cramér-Rao nous donne la minoration $\mathrm{Var}_\theta(T_n) \geq \frac{1}{I_n(\theta)}$. Avec un biais, on peut plus généralement obtenir $\mathrm{Var}_\theta(T_n) \geq \frac{(1+b_n'(\theta)^2}{I_n(\theta)}$.

Sous l'hypothèse (forte) que $\forall n,\ \theta \mapsto b_n(\theta)$ est dérivable et $b_n(\theta^*) \underset{n\to\infty}{\longrightarrow} 0$, on a alors $\underline{\lim}_{n\to\infty}\ (n \times \mathrm{Var}_{\theta^*}(T_n)) \geq \frac{1}{I_1(\theta^*)}$.

Remarque 1.30
Informellement, lorsque $b_n(\theta^*) >> \frac{1}{n}$, on cherchera à minimiser le biais si possible pour que le terme dominant dans le risque soit $\mathrm{Var}^*_\theta(T_n)$.

Lorsque $b_n(\theta^*) << \frac{1}{n}$, avec la remarque précédente, on a alors la minoration $\underline{\lim}_{n\to\infty}\ (n \times R(T_n, \theta^*)) \geq \frac{1}{I_1(\theta^*)}$.

Définition 1.31 (Efficacité asymptotique)
T_n un estimateur sur un modèle régulier, qui vérifie $\sqrt{n}(T_n - \theta^*) \xrightarrow{\text{loi}} \mathcal{N}\left(0, \frac{1}{I_1(\theta^*)}\right)$.

Dans ce cas, on a en particulier l'égalité optimale :

$$\underline{\lim}(nR(T_n, \theta^*)) = \underline{\lim}\, \mathbb{E}_{\theta^*}[(\sqrt{n}(T_n - \theta^*))^2] = \mathrm{Var}_{\theta^*}\left(\mathcal{N}(0, \tfrac{1}{I_1(\theta^*)})\right) = \tfrac{1}{I_1(\theta^*)}.$$

Théorème 1.32
On se place sous les hypothèses suivantes :

(1) $\forall x \in \mathbb{R},\ \theta \mapsto f(x,\theta)$ et $\theta \mapsto \ln(f(x,\theta))$ sont 2-dérivables.

(2) $D_2 : \theta \mapsto \partial^2_\theta[\ln(f)](x,\theta)$ est telle que $\sup_{x\in\mathbb{R}} |D_2(x,\theta) - D_2(x,\theta^*)| \underset{\theta\to\theta^*}{\longrightarrow} 0$.

(3) $\forall \theta \in \Theta,\ \int_{\mathbb{R}} |\partial_\theta f(x,\theta)|dx < \infty$ et $\int_{\mathbb{R}} |\partial^2_\theta f(x,\theta)|dx < \infty$.

(4) $\widehat{\theta_{MV}} \xrightarrow{\text{p.s.}} \theta^*$

Alors $\widehat{\theta_{MV}}$ est un estimateur asymptotiquement efficace.

Démonstration.
cf. notes manuscrites. □

Lemme 1.33 (Lemme de Slutsky)
On utilise dans la preuve précédente un résultat plus général :

Soient $X_n \xrightarrow{\text{loi}} X$ et $Y_n \xrightarrow{\text{proba}} a \in \mathbb{R}\backslash\{0\}$ des variables réelles sur un $(\Omega, \mathcal{F}, \mathbb{P})$.

Alors $\frac{X_n}{Y_n} \xrightarrow{\text{loi}} \frac{X}{a}$.

1.6.2. Estimateur de Hodge, Super-efficacité.

Définition 1.34 (Estimateur de Hodge)
Soit $\widehat{\theta_H} = \bar{X_n} \times \mathbb{1}_{|X_n| > \sqrt[4]{\frac{1}{n}}}$ un estimateur de $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$.

Remarque 1.35
Ponctuellement, à $\theta \in \mathbb{R}$ fixé, l'estimateur de Hodge est aussi sinon plus efficace que $\widehat{\theta_{MV}} = \bar{X}_n$.

Si $\theta > 0$, $\mathbb{P}(\widehat{\theta_H} = 0) \leq \mathbb{P}(\bar{X}_n < \sqrt[4]{\frac{1}{n}}) = \mathbb{P}\left(\mathcal{N}(0,1) < \sqrt{n}(\sqrt[4]{\frac{1}{n}} - \theta)\right) \to 0$ nous ramène au cas usuel. De même lorsque $\theta < 0$.

Si $\theta = 0$, $\mathbb{P}(\widehat{\theta_H} = 0) = \mathbb{P}(\mathcal{N}(0,1) \leq \sqrt[4]{\frac{1}{n}}) \to 1$, et on peut montrer que $\forall \alpha > 0,\ n^\alpha \widehat{\theta_H} \xrightarrow{\text{p.s.}} 0$.

Tout ceci montre que ponctuellement, on peut dépasser la vitesse $\frac{1}{\sqrt{n}}$ imposée par Cramér-Rao et l'efficacité asymptotique.

Remarque 1.36
Globalement, l'estimateur de Hodge est moins efficace que $\widehat{\theta_{MV}}$.

En effet, le risque de $\widehat{\theta_{MV}}$ est constant et donne $\sup_{\theta \in \mathbb{R}} R(\widehat{\theta_{MV}}, \theta) = \frac{1}{I_n(\theta)} = \frac{1}{n}$.

En revanche, $\sup_{\theta \in \mathbb{R}} R(\widehat{\theta_H}, \theta) \geq R(\widehat{\theta_H}, \theta_n)$ avec $\theta_n = \frac{1}{2\sqrt[4]{n}}$. Comme on ne peut pas atteindre de nombre dans $\left]0, \frac{1}{\sqrt[4]{n}}\right[$ avec $\widehat{\theta_H}$ et que θ_n est au milieu, $R(\widehat{\theta_H}, \theta_n) \geq \mathbb{E}_{\theta_n}\left[(\frac{1}{2\sqrt[4]{n}})^2\right] = \frac{1}{4\sqrt{n}}$. Dans cette zone, l'estimateur de Hodge est assez assez instable et met du temps à "décider" si θ vaut 0 ou non.

Un gain de vitesse ponctuel ne peut donc se faire qu'au détriment de la vitesse globale.

Chapitre 2

Intervalles de confiance

07/02

2.1. Introduction

Définition 2.1 (Intervalle de confiance de niveau $1-\alpha$)
Dans un cadre général, étant donné un estimateur $\hat{\theta}_n$ (ou bien un échantillon $X_1...X_n$), on cherche un intervalle aléatoire du type $IC = [g(\hat{\theta}_n), d(\hat{\theta}_n)]$, dit intervalle de confiance, tel que $\mathbb{P}(\theta \in IC) = 1-\alpha$.

Remarque 2.2
On peut élargir cette notion à des cas où $\mathbb{P} \geq 1-\alpha$, ou bien à des cas où l'(in)égalité est vraie asymptotiquement.

Selon le contexte, on peut par exemple chercher un intervalle symétrique de la forme $IC = [\hat{\theta}_n \pm v_\alpha]$, ou bien du type $IC = [m \times \hat{\theta}, M \times \hat{\theta}]$ avec $m < 1 < M$.

Généralement, lorsque $\hat{\theta}$ est sans biais, on aura $\mathbb{P}(\hat{\theta} \in IC) = 1$.

2.2. Cas d'une loi gaussienne (σ^2 connu)

Proposition 2.3 (Intervalle de confiance pour $\widehat{\theta_{MV}}$)
$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$:
$\widehat{\theta_{MV}} = \overline{X_n} \sim \mathcal{N}(\theta, \frac{\sigma^2}{n})$ donc $\frac{\sqrt{n}}{\sigma}(\widehat{\theta_{MV}} - \theta) \sim \mathcal{N}(0,1)$.
On cherche IC sous la forme $[\widehat{\theta_{MV}} \pm v_\alpha]$.
$\mathbb{P}(\theta \in IC) = \mathbb{P}(|\mathcal{N}(0,1)| \geq \frac{v_\alpha \sqrt{n}}{\sigma})$.

On pose $q_{1-\frac{\alpha}{2}}$ le $\left(1-\frac{\alpha}{2}\right)$-quartile de la gaussienne centrée réduite. Autrement dit, q est tel que $\int_{-\infty}^{q} e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} = 1-\frac{\alpha}{2}$.

Par symétrie, on a $q_{\frac{\alpha}{2}} = -q_{1-\frac{\alpha}{2}}$ et $\mathbb{P}\left(\mathcal{N}(0,1) \in [-q_{1-\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]\right) = 1-\alpha$.

Autrement dit, $v_\alpha = q_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$.

Remarque 2.4
Pour $\alpha = 5\%$, on a $q_{1-\frac{\alpha}{2}} \cong 1.96$.

Lorsque n augmente, v_α diminue. Lorsque $\alpha \in]0,1[$ diminue, v_α augmente.

2.3. Cas d'une loi gaussienne (σ^2 inconnu)

Remarque 2.5
Ici, σ est également inconnu, on a donc un espace de paramètres (θ, σ^2) à deux dimensions.

On cherche ici à évaluer un intervalle IC pour θ, en se basant sur l'approximation (biaisée) de la variance par $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$.

Définition 2.6 (Loi χ_d^2)
On pose χ_d^2 la loi de $\sum_{i=1}^{d} X_i^2$, où X_i sont des gaussiennes centrées-réduites iid.

Théorème 2.7 (Théorème de Cochran)

Soient E_i des sous-espaces orthogonaux de $\mathbb{R}^n$ de sorte que $\mathbb{R}^n = E_1 \overset{\perp}{\bigoplus} \ldots \overset{\perp}{\bigoplus} E_p$. On note $n_i = \dim(E_i)$.

Soit X un vecteur gaussien, de loi $\mathcal{N}(0, I_n)$ (où $I_n = (\mathrm{Cov}(X_i, X_j))$ la matrice de covariance). On pose $X_i = P_{E_i}(X)$ la projection orthogonale de X sur le sous-espace E_i. Alors :

(1) Les variables (X_i) sont indépendantes.

(2) Pour i quelconque, $||X_i||^2 \sim \chi_{n_i}^2$.

Démonstration.
On montre ici le cas $p = 2$. Le cas général se fait par une récurrence immédiate sur p.

Ainsi, $\mathbb{R}^n = E_1 \overset{\perp}{\bigoplus} E_2$. Soit (e_i) une base orthonormale de $\mathbb{R}^n$ adaptée à cette décomposition. On pose U la matrice (orthogonale) de passage de la base canonique à (e_i).

Alors $P_{E_1} = U^{-1} J_{n_1} U$ dans la base canonique, avec $J_{n_1} = (\mathbb{1}_{i=j\leq n_1})_{i,j\in[\![1,n]\!]}$.

(1) On remarque que, par anisotropie, UX est un vecteur gaussien de même loi que X. En particulier, $Y_1 = J_{n_1}UX$ (la projection sur les n_1 premières coordonnées) est indépendant de $Y_2 = (I_n - J_{n_1})UX$.

En revenant dans la base canonique, on en déduit $X_1 = U^{-1}Y_1$ et $X_2 = U^{-1}Y_2$ indépendants.

(2) $||X_1||^2 = ||U^{-1}Y_1||^2 = ||Y_1||^2$ est clairement la somme de n_1 gaussiennes centrées-réduites au carré, donc de loi $\chi^2_{n_1}$.

□

Corollaire 2.8
On s'intéresse à nouveau au cas σ^2 inconnu, avec les variables $\overline{X_n}$ et S_n^2 définies plus tôt.

(1) $\overline{X_n}$ et S_n^2 sont indépendantes

(2) $S_n^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$

Démonstration.
On considère l'échantillon X comme un vecteur gaussien de loi $\mathcal{N}(0, \sigma^2 \times I_n)$.
On pose $E_1 = \text{Vect}(e_1)$ où $e_1 = (1, ..., 1)$, et $E_2 = E_1^{\perp}$.
Avec les notations du théorème, on a $X_1 = \overline{X_n} \times e_1$ et $X_2 = (X_i - \overline{X_n})_{1 \leq i \leq n}$.
Il en découle $\overline{X_n} = \pi_1(X_1)$ et $S_n^2 = \frac{1}{n-1}||X_2||^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$ indépendantes.
□

Définition 2.9 (Loi de Students à d degrés de liberté)
Soient $U \sim \mathcal{N}(0,1)$ et $V \sim \chi^2_d$ indépendantes.
On pose T_d la loi de la variable $U \times \sqrt{\frac{d}{V}}$.

Remarque 2.10
On pose à nouveau $IC = [\widehat{\theta_{MV}} \pm v_\alpha]$. Il découle des résultats précédents :
$\mathbb{P}(\theta \in IC) = \mathbb{P}(\frac{\sqrt{n}}{S_n}|\widehat{\theta_{MV}} - \theta| \leq \frac{v_\alpha \sqrt{n}}{S_n}) = \mathbb{P}(|T_{n-1}| \leq \frac{v_\alpha \sqrt{n}}{S_n})$.
On pose $z_{1-\frac{\alpha}{2}}$ le quartiale de la loi de Students (symétrique) T_{n-1}. On a alors l'intervalle de confiance $IC = [\overline{X_n} \pm \frac{S_n}{\sqrt{n}} z_{1-\frac{\alpha}{2}}]$ au seuil $1 - \alpha$.
Contrairement au cas σ connu (où $|IC|$ est fixé et sa position est variable), ici, $|IC|$ est également aléatoire.

2.4. Cas des sondages

Remarque 2.11
Dans cette section, $X_i \overset{\text{iid}}{\sim} \mathcal{B}(\theta)$, $\Theta =]0, 1[$.
On a vu $\widehat{\theta_{MV}} = \overline{X_n}$. On cherche à nouveau $IC = [\overline{X_n} \pm v_\alpha]$.

En limite, avec le TCL, $\mathbb{P}_\theta\left(\left|\widehat{\theta_{MV}} - \theta\right| \leq v_\alpha\right) \to \mathbb{P}\left(|\mathcal{N}(0,1)| \leq v_\alpha \times \frac{\sqrt{n}}{\sqrt{\theta(1-\theta)}}\right)$.

En mettant de côté l'abus lorsqu'un considère $\mathbb{P}(|Z| \leq C\sqrt{n})$ pour un n arbitrairement grand/infini, on a donc $v_\alpha = q_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\theta(1-\theta)}{n}}$.

Avec ce v_α, on remarque que IC dépend du θ choisi. On peut contourner ce souci de 2 manières.

Proposition 2.12 (Première méthode : approximation de v_α)
Soit $\hat{\sigma}_n^2 = \overline{X_n}(1 - \overline{X_n}) \xrightarrow{\mathbb{P}_\theta} \theta(1-\theta)$. Par le lemme de Slutsky, on a donc $\frac{\sqrt{n}}{\hat{\sigma}_n}(\overline{X_n} - \theta) \xrightarrow{\text{loi}} \mathcal{N}(0,1)$. Il en découle $\mathbb{P}_\theta\left(\theta \in \left[\overline{X_n} \pm q_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_n}{\sqrt{n}}\right]\right) \to 1 - \alpha$.

Proposition 2.13 (Deuxième méthode : majoration de v_α)
Dans l'absolu, $\forall \theta \in]0,1[,\ \theta(1-\theta) \leq \frac{1}{4}$.

Ainsi, $\forall \theta \in]0,1[, \exists c \geq 1-\alpha,\ \mathbb{P}_\theta(\theta \in \left[\overline{X_n} \pm \frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right]) \to c$ nous donne tout de même un intervalle de confiance relativement satisfaisant.

Remarque 2.14
Lorsque $\alpha = 5\%$, on rappelle que $q_{1-\frac{\alpha}{2}} \cong 1.96$ et donc $v_\alpha \cong \frac{1}{\sqrt{n}}$.

En pratique, on dit que le TCL (et donc nos intervalles de confiance) est satisfaisant lorsque n est au moins de l'ordre de 30 ou 40.

Ainsi, avec une certitude de 95%, pour $n = 100$, on commet au plus 10% de marge d'erreur sur θ avec l'estimateur $\overline{X_n}$.

Pour $n = 10^3$ (resp. $n = 10^4$), on commet au plus 3% (resp. 1%) d'erreur.

Remarque 2.15
Dans un cadre plus général, on s'intéresse à une "statistique pivotale", une fonction $\psi(\hat{\theta}_n, \theta)$ dont la loi ne dépend pas de θ (ou éventuellement dont la loi limite ne dépend pas de θ).

Chapitre 3

Théorie des tests paramétriques

14/02

3.1. Mise en œuvre

Définition 3.1 (Hypothèse)
$H \subset \Theta$ (mesurable) une hypothèse.

Définition 3.2 (Test)
Soit $\psi : \mathbb{R}^n \to [\![0,1]\!]$ mesurable. On dit que $\psi(X_1...X_n)$ est un test.

Dans le cadre d'un test, ψ est associée à deux hypothèses H_0 et H_1. On dit que le test rejette H_0 lorque $\psi = 1$ et qu'il ne rejette pas H_0 lorsque $\psi = 0$.

Normalement, $H_0 \cap H_1 = \emptyset$ mais a priori, les ensembles ne sont pas complémentaires dans Θ.

Définition 3.3 (Erreur de première ou seconde espèce)
Pour un test donné, on définit $\mathbb{P}_{H_0}(\psi = 1) := \sup_{\theta \in H_0} \mathbb{P}_\theta(\psi = 1)$ l'erreur de première espèce, qui traduit la probabilité de rejeter à tort l'hypothèse H_0.

On définit de même l'erreur de seconde espèce $\mathbb{P}_{H_1}(\psi = 0)$, la probabilité de ne pas rejeter H_0 alors qu'on aurait du.

Remarque 3.4 (Cas bayésien)
On adopte dans ce cours un point de vue fréquentiste. Dans le cas bayésien, où on dispose d'une croyance (d'une probabilité) π sur l'espace de paramètres Θ, on peut considérer une autre définition de l'erreur via la définition :

$\mathbb{P}_H(\psi = 1) = \frac{1}{\pi(H)} \int\limits_{\theta \in H} \mathbb{P}_\theta(\psi = 1) d\pi(\theta)$.

Définition 3.5 (Test de niveau α)
Soit $\alpha \in]0, 1[$. ψ doit vérifier $\mathbb{P}_{H_0}(\psi = 1) \leq \alpha$. On le note alors ψ_α.

Remarque 3.6 (Recherche d'optimalité)
Dans un cadre général, l'objectif va être de minimiser l'erreur de seconde espèce parmi les test de niveau α.

3.2. Tests sur l'espérance d'un échantillon gaussien

3.2.1. Variance connue. On s'intéresse à $\mathcal{N}(\theta, \sigma^2)$, $\theta \in \mathbb{R}$ et σ fixé, connu. On pose $H_0 = \{\theta_0\}$ et $H_1 = \{\theta_1\}$ avec $\theta_0 < \theta_1$. On cherche un test sous la forme $\psi_\alpha = \mathbb{1}_{\overline{X_n} > \theta_0 + t_\alpha}$ où t_α est un réel fixé.

On veut $\mathbb{P}_{H_0}(\psi_\alpha = 1) = \mathbb{P}\left(\mathcal{N}(0,1) > \frac{\sqrt{n} \times t_\alpha}{\sigma}\right) = \alpha$ donc $t_\alpha = \frac{\sigma}{\sqrt{n}} q_{1-\alpha}$.

Dans ce cas, l'erreur de seconde espèce vérifie :

$\mathbb{P}_{\theta_1}(\psi_\alpha = 0) = \mathbb{P}\left(\mathcal{N}(0,1) \leq q_{1-\alpha} + \frac{\sqrt{n}}{\sigma}(\theta_0 - \theta_1)\right)$.

On remarque que, du moment que $\theta_1 - \theta_0 >> \sqrt{n}$, $\mathbb{P}_{\theta_1}(\psi_\alpha = 0) \underset{n \to \infty}{\longrightarrow} 0$.

Si on a simplement $\theta_1 - \theta_0 \geq \frac{\sigma(q_{1-\alpha} - q_\beta)}{\sqrt{n}}$, on conserve un contrôle au seuil β sur l'erreur de seconde espèce.

Avec le raisonnement inverse, avec des θ donnés, un contrôle au seuil β requiert un échantillon de taille $n \geq \left(\frac{\sigma(q_{1-\alpha} - q_\beta)}{\theta_1 - \theta_0}\right)^2$

3.2.2. Variance inconnue. Comme dans le chapitre précédent, on se place maintenant dans le cas où σ^2 est fixé mais inconnu. On considère cette fois-ci $H_0 =]-\infty, \theta_0]$ et H_1 son complémentaire, mais on conserve un test de type $\psi_\alpha = \mathbb{1}_{\overline{X_n} > \theta_0 + t_\alpha}$ où t_α. Ici, on a l'erreur de première espèce :

$\mathbb{P}_{H_0}(\psi_\alpha = 1) = \mathbb{P}_{\theta_0}(\psi_\alpha = 1) = \mathbb{P}\left(T_{n-1} > \frac{\sqrt{n}}{S_n} t_\alpha\right) = \alpha$

Avec $S_n^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \overline{X_n})^2$ l'estimateur de variance.

On a donc ici $t_\alpha = z_{1-\alpha} \times \frac{S_n}{\sqrt{n}}$ également aléatoire.

Cependant, dans ce cas, à n fixé, l'erreur de seconde espèce tends vers $1 - \alpha$ au fur et à mesure qu'on se rapproche à droite de θ_0 dans H_1. On est donc amenés à distinguer l'erreur commise pour un paramètre donné.

Définition 3.7 (Fonction de puissance)
Pour un test ψ, on définit $\pi : \theta \in H_1 \mapsto \mathbb{P}_\theta(\psi = 1)$ sa fonction de puissance.
L'erreur de seconde espèce est alors $\sup\limits_{H_1}(1 - \pi)$.

De façon générale, on cherchera alors à maximiser la fonction de puissance ponctuellement, ce qui revient à minimiser l'erreur de seconde espèce dans la mesure du possible.

Remarque 3.8
On remarque ici que malgré une certaine symétrie entre les deux hypothèses considérées, en changer l'ordre modifiera radicalement les résultats : il est important de choisir en priorité H_0 comme l'hypothèse qu'on aimerait tester et ne pas écarter (mais éventuellement rejeter avec une forte confiance si cela arrive).

28/02

3.3. Notion de p-valeur (TP)

Définition 3.9 (p-valeur)
On fixe des hypothèses H_0 et H_1 a priori.

Soient $T : \mathbb{R}^n \to \mathbb{R}$ mesurable et $(R_\alpha \subset \mathbb{R})_{\alpha\in]0,1[}$ des ensembles qui définissent des tests $\psi_\alpha = \mathbb{1}_{T(X_1...X_n)\in R_\alpha}$, de sorte que $\mathbb{P}_{H_0}(\psi_\alpha = 1) = \alpha$, ψ_α un test au seuil α.

Dans ce cas, la p-valeur est l'application $p_{val}(x) = \inf\{\alpha \in]0,1[,\ T(x) \in R_\alpha\}$ (ou de façon équivalente la variable $p_{val}(X_1...X_n)$).

Dans le cas où $\alpha \mapsto R_\alpha$ croissante pour l'inclusion, avec des lois suffisamment régulières, on peut montrer $p_{val}(x) = \mathbb{P}_{H_0,Y}(T(Y) \geq T(x))$ (où Y est de même loi que $X_1...X_n$, de paramètre dans H_0).

Remarque 3.10
La p-valeur est subordonnée à une famille de tests donnée. Elle permet en particulier de donner la plus forte confiance $1 - p_{val}(X)$ qu'on peut avoir pour ne pas rejeter H_0 lorsqu'il est vrai.

On peut donc définir un nouveau test au seuil α via $\Gamma_\alpha = \mathbb{1}_{p_{val}(X)\leq\alpha}$.

Ce test a l'avantage de ne plus "dépendre" du seuil alpha, on se limite à comparer p_{val} au seuil voulu. Plus informellement, p_{val} traduit donc la probabilité de rejeter H_0 pour un échantillon statistique donné.

14/03

3.4. Notions d'optimalité en théorie des tests

Dans un cadre général, optimiser un test revient à chercher un test dont la puissance sera maximale (éventuellement dans une certaine zone) parmi les tests (idéalement de niveau α) envisagés.

3.4.1. Test de Neyman-Pearson.

Dans cette sous-section, on considère $H_0 = \{\theta_0\} \neq H_1 = \{\theta_1\} \subset \Theta = \mathbb{R}$.

Définition 3.11 (Test uniformément plus puissant (UPP) au niveau α)
ψ un test de niveau α qui vérifie :

(1) ψ niveau exactement α : $\mathbb{P}_{\theta_0}(\psi = 1) = \alpha$

(2) ψ de puissance maximale α : $\mathbb{P}_{\theta_1}(\psi = 1) = \max_{\psi_\alpha} \mathbb{P}_{\theta_1}(\psi_\alpha = 1)$

Théorème 3.12
On pose $R^* = \{x \in \mathbb{R}^n,\ v(x, \theta_0, \theta_1) > t_\alpha\} \subset \mathbb{R}^n$ où $v(x, \theta_0, \theta_1) = \frac{L(x,\theta_1)}{L(x,\theta_0)}$ est le rapport de vraisemblance. On pose également $\psi = \mathbb{1}_{X \in R^*}$ son test associé. Supposons $t_\alpha > 0$ tel que $\mathbb{P}_{\theta_0}(\psi = 1) = \alpha$. Alors c'est un test UPP.

Démonstration.
cf. notes manuscrites. □

Définition 3.13 (Statistique exhaustive)
Soit $T : \mathbb{R}^n \to$ mesurable, appelée une statistique. On dit que T est exhaustive lorsque $\forall t \in \mathbb{R}$, la loi conditionnelle de X sachant $T(X) = t$ est indépendante du $\theta \in \Theta$ considéré.

Théorème 3.14 (Théorème de factorisation)
T est une statistique exhaustive $\Leftrightarrow \exists h : \mathbb{R}^n \to \mathbb{R}^+,\ \exists g : \mathbb{R}^2 \to \mathbb{R}^+$ mesurables telles que $L(x, \theta) = h(x) \times g(T(x), \theta)$.

Démonstration.
cf. notes manuscrites. □

Remarque 3.15
Soit T exhaustive : on peut mettre le rapport de vraisemblance sous la forme $v(x, \theta_0, \theta_1) = \frac{g(T(x),\theta_1)}{g(T(x),\theta_0)}$. Dans ce cas, v ne dépend plus vraiment de $x \in \mathbb{R}^n$ mais seulement de $T(x) \in \mathbb{R}$ qui contient l'information utile de l'échantillon statistique.

Dans le cas T exhaustif, on pourra donc voir v comme une fonction des trois paramètres réels $T(x)$, θ_0 et θ_1.

Exemple 3.16
On se place dans le cas $X_i \sim \mathcal{N}(\theta, \sigma^2)$ où σ est connu :

$$L(x, \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \times e^{-\frac{n}{2\sigma^2}\left(\frac{1}{n}||x||_2^2 - \overline{x_n}^2\right)} \times e^{-\frac{n}{2\sigma^2}(\overline{x_n}-\theta)^2}$$

Il en découle $v(x, \theta_0, \theta_1) = \exp\left(\frac{n(\theta_1-\theta_0)}{\sigma^2}\left[\overline{x_n} - \frac{\theta_1+\theta_0}{2}\right]\right)$.

Et alors $v > t_\alpha \ \Leftrightarrow\ \overline{x_n} > \frac{\theta_1+\theta_0}{2} + \frac{(\sigma/\sqrt{n})^2}{\theta_1-\theta_0} \times \ln(t_\alpha)$ lorsque $\theta_1 > \theta_0$.

3.4.2. Test du rapport de vraisemblance monotone. Dans cette sous-section, on considère $H_0 =]-\infty, \theta_0]$ et $H_1 =]\theta_0, +\infty[$, avec $\theta_0 \in \mathbb{R}$ fixé.

Définition 3.17 (Modèle à rapport de vraisemblance monotone (RVM))
Le modèle $(\mathbb{P}_\theta)_{\theta\in\Theta}$ est à RVM ssi :

$\exists T$ exhaustive, $\forall \theta < \theta' \in \Theta,\ T(x) \mapsto v = \frac{g(T(x),\theta')}{g(T(x),\theta)}$ est croissante sur $\mathbb{R}$.

Théorème 3.18
Soit $(\mathbb{P}_\theta)_{\theta\in\Theta}$ un modèle RVM et T exhaustive associée.

On pose $R^* = \{x \in \mathbb{R}^n,\ T(x) > t_\alpha\} \subset \mathbb{R}^n$. On considère $\psi = \mathbb{1}_{X\in R^*}$ le test associé. Si $\mathbb{P}_{\theta_0}(\psi = 1) = \alpha$, alors ψ est un test UPP de niveau α.

3.5. Retour sur le risque minimal

Remarque 3.19 (Rappels sur le risque d'une gaussienne)
Soient $X_i \sim \mathcal{N}(\theta, 1)$.

En ce qui concerne l'estimateur de Hodge, malgré une précision ponctuelle parfois meilleure que le maximum de vraisemblance, on a le pire risque $\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta[(\widehat{\theta_H} - \theta)^2]$ qui est de l'ordre de $\frac{1}{\sqrt{n}}$. A contrario, pour l'estimateur $\widehat{\theta_{MV}}$, le risque vaut uniformément $\frac{1}{n} = \frac{1}{I_n(\theta)}$.

Cette vitesse de convergence est-elle optimale ?

Proposition 3.20
Soient $s > 0$ et $\phi_n > 0$ des constantes fixées. On a la minoration :

$\inf_{\widehat{\theta_n}} \sup_{\theta\in\mathbb{R}} \phi_n^{-2}\mathbb{E}\left[\left|\widehat{\theta_n} - \theta\right|^2\right] \geq s^2\left(1 - \frac{1}{2}\sqrt{e^{12s^2n\phi_n^2} - 1}\right)$ avec $\widehat{\theta_n}$ qui parcourt l'ensemble des estimateurs sur un échantillon de taille n.

Avec $\phi_n = \frac{1}{\sqrt{n}}$ et $s > 0$ adapté, $\exists C > 0,\ \forall n \in \mathbb{N}^*,\ \inf_{\widehat{\theta_n}} \sup_{\theta\in\mathbb{R}} \mathbb{E}\left[\left|\widehat{\theta_n} - \theta\right|^2\right] \geq \frac{C}{n}$.

Autrement dit, à une constante multiplicative près, notre estimateur $\widehat{\theta_{MV}}$ est ici optimal en termes de vitesse de convergence du pire risque quadratique.

Démonstration.
cf. notes manuscrites. □

Chapitre 4

Modèle Linéaire

21/03

Dans ce contexte, on observe des couples (Y_i, Z_i) iid, dont la loi est donnée par $Y_i = f(Z_i) + \epsilon_i$, avec ϵ_i est un terme d'erreur aléatoire. Le but est ici de déterminer la fonction f inconnue.

On dit que Z_i est la variable explicative. Elle peut-être discrète (qualitative) ou continue (quantitative). On dit que Y_i est la variable à expliquer. On suppose par la suite que Y est à densité continue.

4.1. Cas linéaires simples

4.1.1. Cas des variables quantitatives.

Remarque 4.1
On présente ici le cas où f est sous la forme $az + b$, avec $\theta = (a, b) \in \mathbb{R}^2$ à estimer. Les raisonnements se transposent sans soucis dans le cas général.

Définition 4.2 (Méthode des moindres carrés)
L'approche générale la plus simple est la méthode des moindres carrés :
$\widehat{(a,b)}_{MC} = \text{argmin} \left(\sum_{i=1}^{n} (y_i - \alpha z_i - \beta)^2 \right)$.

Cette méthode donne naturellement $\hat{a} = \frac{\frac{<y,z>}{n} - \overline{y} \times \overline{z}}{\frac{<z,z>}{n} - \overline{z} \times \overline{z}} \cong \frac{\text{cov}(y,z)}{\text{var}(z)}$ et $\hat{b} = \overline{y} - \hat{a}\overline{z}$.

Remarque 4.3 (Estimation par maximum de vraisemblance)
On peut également estimer f par maximum de vraisemblance :
$\widehat{(a,b)}_{MV} = \underset{(\alpha,\beta) \in \mathbb{R}^2}{\text{argmax}}\, d\mathbb{P}(\forall 1 \leq i \leq n,\ \epsilon_i = y_i - \alpha z_i - \beta)$.

Dans le cas particulier où $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ indépendamment des Z_i (σ^2 connu), on a $\widehat{\theta_{MC}} = \widehat{\theta_{MV}}$.

28/03

Remarque 4.4 (Autres modèles linéaires)
Plus généralement, on peut chercher une fonction f sous la forme :
- $y = \sum a_i z^i$ avec $y, z \in \mathbb{R}$,
- $y = a \times e^z + b$ avec $y, z \in \mathbb{R}$,
- $y = \sum a_i z_i$ avec $y \in \mathbb{R}$ et $z \in \mathbb{R}^n$,
- ...

L'essentiel, comme on va le voir, est d'avoir une linéarité selon les paramètres à estimer pour déterminer f.

4.1.2. Cas des variables qualitatives : Analyse de variance (aov).

Définition 4.5
Dans le cas discret, on va avoir un modèle de la forme : $Y_{i,j} = \mu_i + \epsilon_{i,j}$.

On a ici $i \in I$ les modalités du système, les différentes valeurs qualitatives possible. A i fixé, on a $1 \leq j \leq n_i$ les observations faites sous la modalité i, et μ_i la valeur moyenne sous cette modalité.

Les remarques précédentes sur l'estimation s'appliquent également à ce cas.

Remarque 4.6
On peut également généraliser ce modèle à plusieurs modalités distinctes. On aura alors $Y_{i,j,k} = \alpha_i + \beta_j + \gamma_{i,j} + \epsilon_{i,j,k}$, avec $i \in I,\ j \in J,\ k \leq n_{i,j}$. Dans ce cas, $\gamma_{i,j}$ va traduire un effet croisé entre les modalités de I et de J, une certaine corrélation entre les deux modalités qui se renforcent ou se compensent.

On peut également combiner l'aov avec des variables informatives Z_i quantitatives sans soucis.

4.2. Modèle linéaire

Définition 4.7 (Modèle linéaire)
On remarque que dans tous les modèles précédents, on peut :
- Considérer $Y = (Y_i), \epsilon = (\epsilon_i) \in \mathbb{R}^n$ des vecteurs colonnes.
- Rassembler les paramètres (a_i) qui déterminent f dans un vecteur $\theta \in \mathbb{R}^p$.
- Construire une matrice $X \in M_{n,p}(\mathbb{R})$ à partir des variables informatives Z_i.

Ce faisant, on ramène le problème à un unique système matriciel : $Y = X\theta + \epsilon$.

C'est dans ce cas plus général qu'on se place par la suite.

Remarque 4.8 (Hypothèses sur X)
Par la suite on suppose qu'on a :
- Suffisamment d'observations : $p < n$
- Modèle régulier : $\mathrm{rg}(X) = p$

Remarque 4.9 (Hypothèses sur ϵ)
Par la suite on suppose qu'on a :
- Erreur centrée : $\mathbb{E}[\epsilon] = 0$
- Des coordonnées ϵ_i iid.
- Une variance finie : $\mathrm{Var}(\epsilon_i) < \infty$

En pratique, on supposera souvent (mais pas toujours) $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Définition 4.10 (Méthode des moindres carrés)
Dans ce contexte, notre estimateur se généralise en :
$\widehat{\theta_{MC}} = \underset{\theta \in \mathbb{R}^p}{\arg\min} ||Y - X\theta||^2$.

Théorème 4.11
$\widehat{\theta_{MC}} = \left({}^tXX\right)^{-1} {}^tXY$

Démonstration.
cf. notes manuscrites. □

Proposition 4.12
Soient θ^* la vrai valeur du paramètre et $\sigma^2 = \mathrm{Var}(\epsilon_i)$ (non nécessairement gaussienne) :
- $\mathbb{E}[\widehat{\theta_{MC}}] = \theta^*$
- $\mathrm{Var}(\widehat{\theta_{MC}}) := \left(\mathrm{Cov}(\widehat{\theta_{MC}}_i, \widehat{\theta_{MC}}_j)\right)_{1 \leq i,j \leq p} = \sigma^2({}^tXX)^{-1}$
- $\mathbb{E}\left[||\widehat{\theta_{MC}}^*_\theta||^2\right] = \sigma^2 \times \mathrm{tr}\left(({}^tXX)^{-1}\right)$.

Théorème 4.13
Supposons $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Posons $\widehat{\sigma^2} := \frac{1}{n-p}||Y - X\widehat{\theta_{MC}}||^2$. On a alors :
- $\widehat{\theta_{MC}}$ est un vecteur gaussien
- $\widehat{\sigma^2} \sim \sigma^2 \times \frac{\chi^2_{n-p}}{n-p}$

Démonstration.
cf. notes manuscrites. □

Remarque 4.14
Les résidus $Y_i - (X\widehat{\theta_{MC}})_i$ sont des approximations des ϵ_i.

4.2.1. Test de présence d'un sous-modèle.

Remarque 4.15
Dans le modèle théorique, on a $Y = X\theta^* + \epsilon = S^* + \epsilon$.

Si on décompose $X = (C_1...C_p)$, a priori, $S^* \in [X] := \text{Vect}(C_1...C_p)$. Chercher un sous-modèle revient à identifier les corrélations entre θ et Y.

Éliminer certains paramètres de θ revient alors à considérer une sous-matrice $X_0 = (C_{i_1}...C_{i_r})$, $r < p$. On compare alors l'hypothèse $H_0 : S^* \in [X_0]$ contre l'évènement plus large $H_1 : S^* \in [X]$.

Dans un cadre plus général, on peut également fusionner des paramètres a priori distincts θ_1 et θ_2 via une matrice du type $X_0 = ((C_1 + C_2), C_3...C_p)$.

Définition 4.16
Soient $Y = X\theta^* + \epsilon$ un modèle avec $X \in M_{n,p}(\mathbb{R})$ et $p_0 < p$. On considère $X_0 \in M_{n,p_0}$ telle que $[X_0] \subset [X]$ un sous-espace vectoriel : X_0 induit un modèle $Y = X_0\theta_0^* + \epsilon$.

Posons $\widehat{\theta}$ et $\widehat{\theta_0}$ les estimateurs des moindres carrés associés à ces modèles. On définit alors les variables $SCR = ||Y - X\widehat{\theta}||^2$ et $SCR_0 = ||Y - X_0\widehat{\theta_0}||^2$.

On remarque que $SCR_0 \geq SCR$.

Définition 4.17 (Loi de Fisher)
Soient $l, m \in \mathbb{N}^*$ des entiers non nuls. On définit la loi $F(l,m) \sim \frac{m}{l} \times \frac{\chi_l^2}{\chi_m^2}$, où les deux lois χ sont indépendantes l'une de l'autre.

Théorème 4.18
On suppose ϵ gaussien. Soit $\widehat{F} = \frac{n-p}{p-p_0} \times \frac{SCR_0 - SCR}{SCR}$.

Conditionnellement à H_0, on a alors $\widehat{F} \sim F(p - p_0, n - p)$.

Démonstration.
cf. notes manuscrites. □

Définition 4.19 (Test de Fisher)
Soit $q_{1-\alpha}$ le quantile d'une loi de Fisher $F(p-p_0, n-p)$. Soit le test de Fisher $\phi_\alpha = \mathbb{1}_{\widehat{F} > q_{1-\alpha}}$. C'est un test de niveau α pour l'hypothèse $H_0 : S^* \in [X_0]$.

Chapitre 5

Sélection de modèles

04/04

5.1. Cadre général

On se place dans le contexte du chapitre précédent, avec un modèle linéaire $Y = X\theta + \epsilon,\ X = \left(Z^{(1)}...Z^{(p)}\right) \in M_{n,p}(\mathbb{R}),\ p < n$.

On suppose ici $p >> 1$ et on cherche à identifier quelles variables $Z^{(i)}$ sont les plus pertinentes pour expliquer Y.

Définition 5.1 (Sous-modèle)
On appelle sous-modèle de $Y = X\theta + \epsilon$ une partie $m := \{i_1...i_r\} \subset [\![1, p]\!]$.

Soit alors $X_{(m)} := \left(Z^{(i_1)}...Z^{(i_r)}\right) \in M_{n,r}$. L'étude du sous-modèle m revient à estimer $\theta_{(m)}$ sous l'hypothèse $Y = X_{(m)}\theta_{(m)} + \epsilon'$.

Remarque 5.2
Soit $M \subset \mathcal{P}([\![1, p]\!])$. On suppose que le vrai modèle m^* est dans M.

Par exemple, $M = \{[\![1, m]\!],\ m \leq p\}$ peut traduire une volonté d'étudier les harmoniques d'un signal jusqu'à un certain rang m.

5.2. Approches naïves

Définition 5.3 (Coefficient d'ajustement)
$SCT = ||Y - \overline{Y}||^2 = \sum_{i=1}^{n} (y_i - \overline{y_n})^2 \sim \mathrm{Var}(Y)$ la variance empirique de Y.

$SCR_{(m)} = ||Y - \widehat{Y_{(m)}}||^2$ où $\widehat{Y_{(m)}} := X_{(m)}\widehat{\theta_{(m)}}_{MC}$.

On définit le coefficient $R^2_{(m)} = \frac{SCT - SCR_{(m)}}{SCT} \in]0, 1[$.

Remarque 5.4
Si $m' \subset m$, alors $SCR_{(m')} \geq SCR_{(m)}$ et donc $R^2_{(m')} \leq R^2_{(m)}$.

Ce critère peut s'avérer utile lorsqu'on compare m et m' de même cardinal $|m| = |m'|$ mais n'est pas pertinent dans un cadre général.

Définition 5.5 (Stratégie de régressions descendantes)
Cet algorithme a pour principe d'itérer des tests de Fisher. On considère X et Y fixés et connus de l'algorithme, ainsi qu'un seuil α (généralement 5%).

Soit $m \in M$ un modèle. On définit l'algorithme $SRD(m)$ comme suit :

(1) Entrée : $m = \{i_1 ... i_r\} \subset [\![1, p]\!]$.

(2) Pour tout $j \leq r$, on définit p_j comme la *p-valeur* du test de Fisher qui compare le modèle m à son sous-modèle $m \backslash \{i_j\}$ (cf. chap. précédent).

(3) Si $\forall j,\ p_j \leq \alpha$, alors on renvoie m.

(4) Sinon, soit $j = \operatorname{argmax} p_i$: on renvoie $SRD(m \backslash \{i_j\})$.

Remarque 5.6 (Stratégie de régressions montantes)
On peut globalement suivre le même principe en ajoutant pas à pas les colonnes au lieu de les retirer.

Dans les deux cas, l'algorithme ne repose sur aucune démonstration mathématique, bien que parfois pertinent en pratique et cohérent d'un point de vue informel.

5.3. Risque quadratique

Définition 5.7 (Risques en prédiction et en estimation)
Soit θ^* le vrai paramètre, dans le modèle $m^* \in M$.

— Risque en prédiction : $R(m) = \mathbb{E}[||X_{(m^*)}\theta^* - X_{(m)}\widehat{\theta_{(m)}}||^2]$

— Risque en estimation : $\bar{R}(m) = \mathbb{E}[||\theta^* - \widehat{\theta_{(m)}}||^2]$ où on prolonge $\widehat{\theta}$ par 0 sur les bonnes coordonnées dans $\mathbb{R}^p$.

Remarque 5.8
On va par la suite s'intéresser au risque en prédiction. On introduit pour cela $\mu^* = X_{(m^*)}\theta^*$ et $\mu_{(m)} = X_{(m)}\theta_{(m)} = P_{[X_{(m)}]}(\mu^*)$ la projection orthogonale de μ^* dans le sous-espace associé à $X_{(m)}$. Ces quantités sont ici fixées mais inconnues.

Proposition 5.9
Supposons l'erreur ϵ gaussienne de variance σ^2 connue pour chaque coordonnée. Alors $R(m) = \sigma^2(|m| + 1) + ||\mu_{(m)} - \mu^*||^2$.

Démonstration.
cf. notes manuscrites. □

Remarque 5.10
D'une part, pour $|m|$ croissant, assez naturellement, la variance $\sigma^2(|m|+1)$ augmente. D'autre part, pour m croissant pour l'inclusion, on a le biais $||\mu_{(m)} - \mu^*||^2$ qui diminue, sans pour autant qu'on en connaisse la valeur.
Soit $m_0 = \arg\min_{m \in M} R(m)$. Déterminer m_0 revient à chercher le meilleur compromis biais-variance. En pratique, cela dit, on n'a pas accès à μ^* : on va par la suite essayer d'estimer m_0.

5.4. Critère C_p de Mallows-Akaike

Définition 5.11
Soient $C_p(m) = ||Y - X_{(m)}\widehat{\theta_{(m)}}||^2 + 2\sigma^2(|m|+1)$ et $\tilde{C}_p(m) = C_p(m) - n\sigma^2$.

Proposition 5.12
$\mathbb{E}[||Y - \mu_{(m)}||^2] = (n - |m| + 1)\sigma^2 + ||\mu^* - \mu_{(m)}||^2$.
$\tilde{C}_p(m)$ est donc un estimateur sans biais de $R(m)$.

Définition 5.13
Soit $\widehat{m} := \arg\min_{m \in M} \tilde{C}_p(m) = \arg\min_{m \in M} C_p(m)$.
On peut montrer que le risque de $\widehat{m}$ va correctement approcher $R(m_0)$.

5.5. Critère AIC

On peut également vouloir minimiser la divergence de Kullback entre les lois de m et m^*.
On a $K(m, m^*) := K(\mathbb{P}_m, \mathbb{P}_{m^*}) = \frac{n}{2}\left(\ln\left(\frac{\sigma^2(m)}{\sigma^2}\right) + \frac{\sigma^2}{\sigma^2(m)} - 1\right) + \frac{||\mu^* - \mu_{(m)}||^2}{2\sigma^2(m)}$,
avec $\sigma^2(m) = \frac{1}{n-|m|}\mathbb{E}[||Y - X_{(m)}\widehat{\theta_{(m)}}||^2]$.
On a $AIC(m) = \ln\left(\frac{||Y - X_{(m)}\widehat{\theta_{(m)}}||^2}{n}\right) + \frac{n+|m|+1}{n-|m|-3}$ un estimateur sans biais de $K(m, m^*)$.

Chapitre 6

Statistiques en grande dimension

11/04

6.1. Introduction

On va ici continuer d'étudier le modèle linéaire. Cependant, contrairement au chapitre précédent, on travaille ici sous l'hypothèse $n < p$ (voire $n << p$) : on dispose de beaucoup plus de paramètres à déterminer ($\theta \in \mathbb{R}^p$) que d'observations ($Y \in \mathbb{R}^n$). On note également θ^* le vrai paramètre à estimer.

C'est par exemple le cas lors de l'étude de maladies rares, où on ne dispose que de quelques patients pour déterminer l'influence de beaucoup de facteurs. En particulier, ${}^tXX \in M_p(\mathbb{R})$ mais $\text{rg}({}^tXX) \leq n$ donc on ne peut pas calculer $\widehat{\theta_{MC}}$ aussi simplement qu'avant.

Définition 6.1 (Indice de parcimonie, sparsité)
Le support d'un vecteur $\theta \in \mathbb{R}^p$ est $J(\theta) = \{1 \leq j \leq p,\ \theta_j \neq 0\}$. L'indice de parcimonie du vecteur est $S(\theta) = \#J(\theta)$.

Définition 6.2
Dans le cas où $J(\theta^*)$ est connu, on définit l'estimateur :

$$\widehat{\theta^0} = \underset{\nu \in \mathbb{R}^p,\ J(\nu)=J(\theta^*)}{\arg\min} ||Y - X\nu||^2$$

Proposition 6.3
Soit $R(\theta, \theta') = \frac{1}{n}\mathbb{E}[||X(\theta - \theta')||^2]$ le risque.

Pour une erreur ϵ gaussienne, on a $R(\widehat{\theta^0}, \theta^*) = \frac{S(\theta^*)}{n}\sigma^2$.

Démonstration.
cf. notes manuscrites. □

6.2. Méthode LASSO

Définition 6.4
On suppose maintenant $J(\theta^*)$ inconnu mais θ^* parcimonieux ($S(\theta^*) << n$).
On définit l'estimateur $\tilde{\theta} = \underset{S(\nu)\leq S(\theta^*)}{\arg\min} ||Y - X\nu||^2$

Avec le paramètre $R(n)$ à fixer ultérieurement et si possible $S(\theta^*) \leq R(n)$.

On peut montrer que de façon équivalente on peut considérer :
$\tilde{\theta} = \underset{\nu\in\mathbb{R}^p}{\arg\min} \left(||Y - X\nu||^2 + \lambda(n)S(\nu)\right)$
Cette application n'est pas convexe, ce qui pose de gros problèmes en pratique.

Définition 6.5 (Alternative Least Absolut Shrinkage and Selection Operator (LASSO))
$\widehat{\theta_L} = \underset{\nu\in\mathbb{R}^p}{\arg\min} \left(\frac{1}{n}||Y - X\nu||_2^2 + \lambda||\nu||_1\right)$.
Cet estimateur a l'avantage d'être convexe et donc calculable.

Remarque 6.6 (Estimateur de Tychonov)
On peut également appliquer une nomre l^2 sur ν : $\widehat{\theta_R} = \underset{\nu\in\mathbb{R}^p}{\arg\min} \left(||Y - X\nu||_2^2 + \rho||\nu||_2\right)$.

Pour un choix judicieux de ρ, on a alors $\widehat{\theta_R} = ({}^tXX + \rho Id)^{-1}{}^tXY$.

Le choix d'une norme l^2 pour l'estimateur a un effet régularisant, analytique, tandis que le choix de la norme l^1 pour l'estimateur LASSO a pour effet de favoriser sa parcimonie.

6.3. Performances théoriques de l'opérateur LASSO

On suppose pour simplifier les calculs par la suite que $\forall j \leq p,\ \frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2 = 1$.

Théorème 6.7
On suppose ϵ gaussien. Soit $\delta > 0$ fixé. Supposons $\lambda(n) \geq \frac{4\sigma}{\sqrt{n}}\sqrt{2\ln(\frac{p}{\delta})}$. Avec probabilité supérieure à $1 - \delta$, on a :
$\frac{1}{n}||X(\widehat{\theta_L} - \theta^*)||_2^2 + \lambda||\widehat{\theta_L}||_1 \leq 3 \times \underset{\nu\in\mathbb{R}^p}{\inf} \left(\frac{1}{n}||X(\nu - \theta^*)||_2^2 + \lambda||\nu||_1\right)$

Démonstration.
cf. notes manuscrites. □

Corollaire 6.8
Pour $\lambda(n) = \frac{4\sigma}{\sqrt{n}}\sqrt{2\ln(\frac{p}{\delta})}$, il existe une constante C telle que, avec probabilité supérieure à $1-\delta$, on a $\frac{1}{n}||X(\widehat{\theta_L} - \theta^*)||^2 \leq C \times S(\theta^*)\lambda$.
Ceci garantit une prédiction consistante lorsque $S(\theta^*)\lambda(n) \underset{n\to\infty}{\longrightarrow} 0$.

Démonstration.
cf. notes manuscrites. □

6.4. Vitesse rapide avec contrainte sur X

Définition 6.9 (Hypothèse de compatibilité)
Pour un vecteur $\theta \in \mathbb{R}^p$ et $J \subset [\![1, p]\!]$ on définit $\theta_J = (\theta_j \times \chi_{j\in J})_{1\leq j\leq p}$.
Soient $\widehat{\Sigma} = \frac{1}{n}{}^tXX$ et $J_0 = J(\theta^*)$. On suppose que $\exists \phi_0 > 0,\ \forall \gamma \in \mathbb{R}^p$, si $||\gamma_{J_0^C}||_1 \leq 3||\gamma_{J_0}||_1$, alors on a $||\gamma_{J_0}||_1^2 \leq \frac{S(\theta^*)}{\phi_0^2}({}^t\gamma\widehat{\Sigma}\gamma) = \frac{S(\theta^*)}{n\phi_0^2}||X\gamma||^2$.

Lemme 6.10 (Condition suffisante sur l'hypothèse)
Si $\forall i,\ \widehat{\Sigma}_{i,i} = 1$ et $\forall i \neq j,\ |\Sigma_{i,j}| < \frac{1}{7\alpha S(\theta^*)}$, alors $\phi_0 = \sqrt{1 - \frac{1}{\alpha}}$ convient.

Démonstration.
cf. notes manuscrites. □

Théorème 6.11
On suppose X normalisé comme précédemment, ϵ gaussien et l'hypothèse de compatibilité satisfaite.
Soit $\delta > 0$ fixé. Supposons $\lambda(n) \geq \frac{4\sigma}{\sqrt{n}}\sqrt{2\ln(\frac{p}{\delta})}$. Il existe une constante C telle que, avec probabilité supérieure à $1-\delta$, on a $\frac{1}{n}||X(\widehat{\theta_L} - \theta^*)||^2 \leq C\frac{\lambda^2}{\phi_0^2}$.

Démonstration.
cf. notes manuscrites. □