

# ESTIMATION PARAMÉTRIQUE

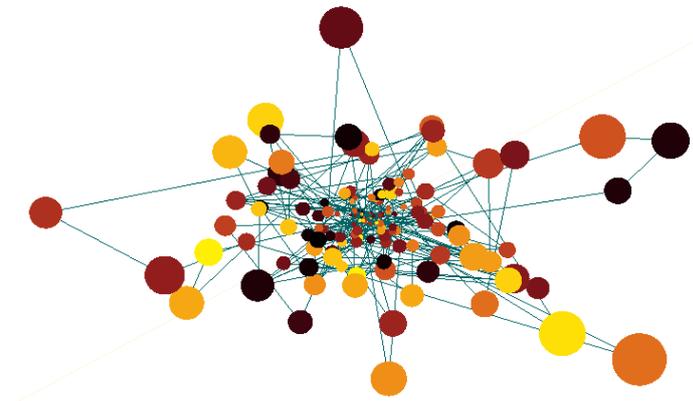
---

COURS DE MASTER 2

---

Bernard Delyon

21 septembre 2020





# Table des matières

<b>I</b>	<b>Introduction. Définitions</b>	<b>5</b>
I.1	Statistiques et paramètres	5
I.1.1	Statistiques et probabilités	5
I.1.2	Estimation paramétrique. Modèles et vraisemblance	5
I.1.3	Exercices	7
I.2	Quelques modèles statistiques et leur vraisemblance	7
I.2.1	Modèles statiques	7
I.2.2	Modèles de séries temporelles	9
I.2.3	Modèles spatiaux	13
I.2.4	Exercices et compléments	16
I.3	Principes d'estimation	18
I.3.1	Premiers concepts	18
I.3.2	Trois estimateurs classiques	19
I.3.3	Un résultat de convergence	23
I.3.4	Exercices et compléments	24
<b>II</b>	<b>Estimation semi-paramétrique</b>	<b>29</b>
II.1	Loi des grands nombres uniforme et autres outils	29
II.2	Fonctions d'estimation (Z-estimateurs)	32
II.2.1	Exemple : la méthode de la variable instrumentale	32
II.2.2	Convergence presque sûre	33
II.2.3	Normalité asymptotique	34
II.2.4	Fusion optimale de Z-estimateurs	36
II.2.5	Exercices et compléments	37
II.3	Estimateurs à minimum de contraste (M-estimateurs)	39
II.3.1	Exemple : régression non-linéaire	39
II.3.2	Propriétés asymptotiques	39
II.3.3	Le cas indépendant	41
II.3.4	Surajustement. Débiaisement du contraste	43
II.3.5	Exercices et compléments	44
II.4	Méthode des moments généralisée	46
II.4.1	Un exemple : le minimum de $\chi^2$	46
II.4.2	Définition et convergence	46
II.4.3	Exercices et compléments	48
II.5	Estimation sous contraintes	48
II.5.1	Exemple : estimation de matrice sous contrainte de rang	49
II.5.2	Convergence	49
II.5.3	Exemple : estimation de matrice sous contrainte de rang. Suite	51
II.5.4	Théorème de Wilks	52
II.5.5	Exercices	53
II.6	Bornes de grandes déviations et estimées des moments	53

<b>III Estimation paramétrique</b>	<b>57</b>
III.1 Comparaison des estimateurs	57
III.1.1 Risque, estimateur admissible et approche minimax.	57
III.1.2 L'approche de Rao-Blackwell-Lehmann-Scheffé	58
III.1.3 Exercices et compléments	60
III.2 L'estimateur au maximum de vraisemblance	63
III.2.1 Convergence presque sûre	63
III.2.2 Normalité asymptotique	64
III.2.3 Bornes exponentielles. Convergence des moments	69
III.2.4 Quelques exemples	70
III.2.5 Exercices et compléments	70
III.3 Borne de Cramér-Rao. Efficacité	72
III.3.1 Borne de Cramér-Rao	72
III.3.2 Le cas biaisé : Inégalité de van Trees	73
III.3.3 Efficacité.	75
III.3.4 Exercices et compléments	76
III.4 La méthode bayésienne	78
III.4.1 Estimateurs bayésiens	78
III.4.2 Le maximum a posteriori (MAP).	83
III.4.3 Exercices et compléments.	83
<b>IV Tests</b>	<b>91</b>
IV.1 Théorie non-asymptotique	91
IV.1.1 Introduction	91
IV.1.2 Approche UPP	92
IV.1.3 Dualité tests/régions de confiance.	96
IV.1.4 Approche bayésienne	97
IV.2 Seuils asymptotiques	97
IV.2.1 Cas paramétrique	97
IV.2.2 Cas semi-paramétrique	98
IV.3 Puissance asymptotique	99
IV.4 Exercices	102
<b>V L'approche martingale</b>	<b>103</b>
V.1 Le maximum de vraisemblance en situation générale	103
V.1.1 Théorie	103
V.1.2 Exemple : Chaînes de Markov stationnaires	105
V.1.3 Exercices et compléments	106
V.2 Processus autorégressif instable	107
V.3 Régression linéaire	108
<b>A Compléments sur les expériences régulières</b>	<b>111</b>
A.1 Liens entre les jeux d'hypothèses	111
A.2 Démonstration des théorèmes 29, 42. Régularité des scores	117
<b>B Démonstration du lemme 50</b>	<b>121</b>
<b>C Une borne sur les processus empiriques</b>	<b>123</b>
<b>D Théorèmes-limite pour les martingales</b>	<b>125</b>
<b>E Les critères MML, MDL, et BIC</b>	<b>127</b>

---

# INTRODUCTION. DÉFINITIONS

---



---

## I.1 Statistiques et paramètres

### I.1.1 Statistiques et probabilités

Le point de départ de la statistique est l'échantillon (les données, les observations. . .) ; c'est typiquement un tableau de chiffres ou de symboles (p. ex. âge, revenus et santé de 1000 personnes). Le but essentiel la statistique est de mesurer les dépendances entre variables ainsi que leur variabilité : elle cherche à démêler ce qui est *tendances systématiques* (p. ex. : le revenu augmente avec l'âge) des *variations aléatoires* (imprévisibles). La statistique va donc dans le sens contraire des probabilités :

**Probabilités** : inférer des propriétés des variables aléatoires à partir de la connaissance de leur distribution. Un résultat de probabilités est souvent un théorème-limite (lois des grands nombres, théorème-limite central, etc.).

**Statistiques** : inférer de l'information sur une distribution à partir de l'observation de variables aléatoires.

Un modèle statistique est une famille de lois de probabilités et une méthode statistique aura pour but d'en extraire une sous-famille grâce à un ensemble de données observé. Le cœur de la statistique est donc cette paire (modèles, données).

Les applications des statistiques sont pour bonne part la *prédiction/simulation* (voir les exemples du § I.2.1 : modèles de cinétique chimique, ou détection d'individus à risques par la régression logistique), la *détection de changement* (par observation d'un changement significatif du modèle estimé ; usure d'une machine...), le *codage* (si des données suivent un modèle paramétrique, il est bien plus économique de coder les paramètres et les *erreurs de prédictions* que les données elles-mêmes), et plus généralement l'*analyse*.

### I.1.2 Estimation paramétrique. Modèles et vraisemblance

**Cadre paramétrique.** Le modèle, **une expérience statistique**, consiste en une famille de lois de probabilités  $(P_\theta)_{\theta \in \Theta}$ , où  $\Theta$  est une partie de l'espace euclidien. On suppose que ces lois ont une densité par rapport à une mesure positive  $\sigma$ -finie<sup>1</sup> commune  $\mu(dy)$ , généralement la mesure de Lebesgue ou une mesure de comptage :

$$P_\theta(dy) = p_\theta(y)\mu(dy).$$

On dispose d'une observation  $Y$  tirée selon  $P_{\theta_*}$  et l'on cherche à estimer  $\theta_*$ . La vraisemblance de  $Y$  est  $p_\theta(Y)$  (cette fonction de  $\theta$  dépend du choix de  $\mu$ ), et son logarithme sera noté  $\mathcal{L}(\theta, Y) = \log(p_\theta(Y))$ <sup>2</sup>.

---

1. Rappelons que le théorème de Fubini n'est plus valide pour les mesures non  $\sigma$ -finies. Ceci interviendra dans un exemple du § III.2.4. Cette hypothèse sur  $\mu$  ne sera pas rappelée et sera implicite dans la suite.

2. Nous utilisons ici  $\mathcal{L}$  pour la log-vraisemblance. On trouve aussi la notation  $\mathcal{L}(\theta|Y) = p_\theta(Y)$  qui souligne que la vraisemblance est plutôt considérée comme une fonction de  $\theta$  ; cependant, dans la littérature, on parle autant de la vraisemblance de  $Y$  que de celle de  $\theta$ .

L'estimateur le plus populaire, celui du maximum de vraisemblance, est

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta, Y).$$

Si l'on dispose d'une suite iid d'observations  $Y = (Y_1, \dots, Y_n)$  toutes tirées selon  $P_{\theta_*}$ , l'expérience est associée à la famille des  $P_{\theta}^n$  et l'estimateur du maximum de vraisemblance, est

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \mathcal{L}(\theta, Y_i).$$

Il se justifie d'une part par le théorème de Neymann Pearson (théorème 36 p. 93) qui exprime que si l'on compare deux valeurs du paramètre, celle de plus grande vraisemblance mathématique est bien la plus vraisemblable au sens commun, et d'autre part par des résultats asymptotiques à grands échantillons (convergence presque sûre...). Il souffre cependant de trois objections potentielles :

1. Il peut être peu performant sur de petits échantillons.
2. Il est parfois trop difficile à calculer, typiquement dans le cas d'observations incomplètes.
3. En pratique, le caractère exagérément simplificateur du modèle  $(P_{\theta})_{\theta \in \Theta}$  fait que l'objectif final — p. ex. la prédiction en apprentissage (machine learning) — peut être mieux atteint par d'autres estimateurs le ciblant plus spécifiquement : le  $\theta$  qui conduit à la meilleure prédiction n'est pas forcément le plus vraisemblable pour le modèle considéré<sup>3</sup> ; il ne faut pas perdre de vue que cette approximation par un modèle est d'autant plus voyante que l'échantillon est grand, ce qui fait dire à Le Cam à la fin de sa contribution dans [32] : « The asymptotics fail precisely when one would feel they are applicable » (i.e. en présence d'un échantillon nombreux).

**Paramétrique et non-paramétrique.** Voici trois problèmes statistiques :

1. On observe une suite  $Y = (Y_1, \dots, Y_n)$  de variables iid gaussiennes. Estimer leur moyenne.
2. On observe une suite  $Y = (Y_1, \dots, Y_n)$  de variables iid uniformes sur  $[0, a]$ ,  $a$  est inconnu. Estimer leur moyenne.
3. On observe une suite  $Y = (Y_1, \dots, Y_n)$  de variables iid. Estimer leur densité.

Les modèles 1 et 2 sont **paramétriques** car le paramètre est de dimension finie et caractérise la distribution des variables. Le modèle 3 est dit **non-paramétrique** car la distribution n'est pas caractérisée par un paramètre de dimension finie, le paramètre est ici la densité commune aux variables.

**Paramétrique et semi-paramétrique.** Soit les problèmes statistiques :

1. On observe une suite  $Y = (Y_1, \dots, Y_n)$  de variables iid. Estimer l'espérance  $\theta$ .
2. Comme le précédent mais on suppose que la densité est symétrique.
3. Apprentissage : On observe une suite iid  $((Y_1, X_1), \dots, (Y_n, X_n))$ ,  $Y_i \in \{0, 1\}$ ,  $X_i \in \mathbb{R}^p$  ;  $X_i$  est la variable explicative qui servira à prédire la réponse  $Y_i$ . On dispose d'une suite paramétrée de fonctions à valeurs dans  $\{0, 1\}$ ,  $f_{\theta}$ ,  $\theta \in \Theta \subset \mathbb{R}^d$ . Trouver  $\theta$  qui minimise  $P(Y_1 \neq f_{\theta}(X_1))$ .
4. On observe une suite  $((Y_1, X_1), \dots, (Y_n, X_n))$  satisfaisant le modèle  $Y_i = f(\langle X_i, \theta \rangle) + e_i$  où  $\theta \in \mathbb{R}^d$  est le paramètre à estimer,  $f$  une fonction inconnue de  $\mathbb{R}$  dans  $\mathbb{R}$ ,  $X_i \in \mathbb{R}^d$  est une suite iid (variables explicatives), et  $e_i$  un bruit iid. Estimer  $\theta$ .

Pour ces problèmes, la loi est caractérisée par une paire  $(\theta, \eta)$  où  $\theta$  est le paramètre d'intérêt, de dimension finie, et  $\eta$  est le paramètre dit de nuisance, de dimension infinie ; dans l'exemple 4,  $\eta$  serait la paire  $(f, p)$  où  $p$  est la densité du bruit. On dit qu'il s'agit de modèles semi-paramétriques ; le terme de *méthode semi-paramétrique* fait généralement référence à une famille de méthodes particulières passant par une étape d'estimation (non-paramétrique) de  $\eta$ , dans l'exemple 2 ce serait l'estimation de la densité<sup>4</sup>. Nous traiteront dans la suite de problèmes semi-paramétriques mais pas de méthodes semi-paramétriques.

3. C'est la distinction entre *discriminative training* et *generative training* [145].

4. Ce problème est traité dans [135]. Il existe effectivement un estimateur strictement meilleur que  $\hat{\theta} = n^{-1} \sum Y_i$ .

### I.1.3 Exercices

**Exercice 1 (Données censurées).** Soit  $(X_i)_{1 \leq i \leq n}$  une suite de variables iid de loi  $\mathcal{E}(\theta^{-1})$  (exponentielles d'espérance  $\theta$ ). On observe  $Y_i = \min(X_i, c)$ ,  $c$  est connu. Donner le modèle statistique pour la suite  $Y_i$  et calculer l'estimateur au maximum de vraisemblance. *Indication* : Pour trouver  $\mu$  et  $p_\theta$ , on pourra les faire apparaître en exprimant, pour toute fonction  $f$  bornée,  $E[f(X_1)]$  sous la forme  $\int f(x)p_\theta(x)\mu(dx)$ .

**Exercice 2 (Chaîne de Markov).** Soit  $(Y_n)$  une chaîne de Markov telle que conditionnellement à  $Y_n$ ,  $Y_{n+1}$  suit une loi de Poisson de paramètre  $\theta_* Y_n + \theta_*$ .  $Y_0$  est déterministe connu et  $\theta_* \in ]0, 1[$ .

Exprimer la log-vraisemblance de  $Y_1, \dots, Y_n$  et donner l'estimateur au maximum de vraisemblance  $\hat{\theta}_n$  (on utilisera et l'on justifiera que  $P(Y_1, \dots, Y_n) = \prod_{i=1}^n P(Y_i | Y_{i-1})$ ). Préciser la mesure  $\mu(dy)$ .

Calculer la limite de  $E[Y_n]$ . On suppose que  $n^{-1}(Y_1 + \dots + Y_n)$  converge avec probabilité 1 vers cette limite, montrer que  $\hat{\theta}_n$  converge vers  $\theta_*$ .

**Exercice 3 (Un cas un peu spécial).** Soit la chaîne de Markov  $(Y_i)$ , où  $Y_{i+1}$  vaut  $Y_i$  avec probabilité  $\theta$  et  $Y_i/2$  avec probabilité  $1 - \theta$ .  $Y_0 = 1$ . Exprimer la log-vraisemblance de  $Y_1, \dots, Y_n$  (en fonction des variables  $Z_i = Y_{i-1}/Y_i - 1$ ). Préciser la mesure  $\mu(dy)$ .

**Exercice 4 (Maximum de vraisemblance mis en défaut).** On va observer ici que lorsque le nombre de paramètres approche le nombre d'observations, le maximum de vraisemblance peut donner de très mauvais résultats, même lorsque l'un des paramètres est facile à estimer.

Soit  $X_1, \dots, X_n, Y_1, \dots, Y_p$  des variables indépendantes avec  $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$ , et  $Y_i \sim \mathcal{E}(\sigma)$  (exponentielle d'espérance  $\sigma^{-1}$ ). On s'intéressera particulièrement au cas  $n = p$ . Le paramètre est ici  $\theta = (\sigma, \mu_1, \dots, \mu_n)$ .

Calculer  $\hat{\sigma}$ , l'estimateur au maximum de vraisemblance de  $\sigma$ . Discuter de sa valeur en fonction de  $n$  et  $p$  (distinguer  $n < p$ ,  $n = p$  et  $n > p$ ). En proposer un meilleur.

D'autres estimateurs seront proposés aux exercices 4 p. 24 et 9 p. 85. L'exercice 12 p. 86 en suggère un dernier bien plus sophistiqué (les  $\mu_k$  d'ici sont les  $\alpha_k$  là-bas).

**Exercice 5 (Loi uniforme).** On observe  $Y_i$ ,  $i = 1, \dots, n$ , iid, uniformes sur  $[0, \theta_*]$ .

1. Montrer que l'estimateur du maximum de vraisemblance est  $\hat{\theta}_n = \max Y_i$ .
2. Calculer  $P(\hat{\theta}_n > t)$  puis  $P(n\theta_*^{-1}(\theta_* - \hat{\theta}_n) > t)$  pour  $t > 0$  et montrer que

$$E[n^2(\theta_* - \hat{\theta}_n)^2] \longrightarrow 2\theta_*^2, \quad E[\hat{\theta}_n] = \frac{n}{n+1}\theta_* \quad (\text{I.1})$$

$$n(\theta_* - \hat{\theta}_n) \longrightarrow \theta_* \mathcal{E} \quad \text{en loi} \quad (\text{I.2})$$

où  $\mathcal{E}$  est une variable exponentielle de moyenne 1.

3. Dédire de (I.1) que l'estimateur  $\hat{\theta}'_n = (n+1)\hat{\theta}_n/n$  satisfait :  $E[n^2(\theta_* - \hat{\theta}'_n)^2] \longrightarrow \theta_*^2$ .

$\hat{\theta}'_n$  est donc, en terme d'erreur quadratique moyenne, strictement meilleur que l'estimateur au maximum de vraisemblance. On s'intéressera à l'exercice 8 p. 85 à l'estimation bayésienne de  $\theta_*$ .

## I.2 Quelques modèles statistiques et leur vraisemblance

Donnons quelques modèles qui seront utilisés par la suite pour illustrer le propos. Ces modèles peuvent être vus comme des éléments de base pour construire ceux qui sont réellement utilisés dans la pratique.

### I.2.1 Modèles statiques

**Familles exponentielles.** La famille exponentielle canonique standard complète associée à la mesure  $\mu$  sur  $\mathbb{R}^d$ , mesure non portée par un hyperplan, est définie par

$$P_\theta(dy) = e^{\theta^T y - Z(\theta)} \mu(dy) \quad (\text{I.3})$$

où  $Z$  se déduit du reste par la condition d'intégrale 1. L'ensemble  $\Theta$  où varie  $\theta$  est :

$$\Theta = \left\{ \theta : \int e^{\theta^T y} \mu(dy) < \infty \right\}. \quad (\text{I.4})$$

Les exemples typiques sont les lois binomiales, Poisson, gaussiennes, inverse gaussiennes, gamma, etc.

La famille est dite **régulière** si  $\Theta$  est ouvert. La condition d'hyperplan sur  $\mu$  implique que pour deux  $\theta$  différents, les mesures  $P_\theta$  sont différentes. On vérifie facilement les propriétés (du moins si  $\theta$  est intérieur à  $\Theta$ ) :

$$\begin{aligned}\nabla Z(\theta) &= E_\theta[Y] \\ \nabla^2 Z(\theta) &= E_\theta[YY^T] - E_\theta[Y]E_\theta[Y]^T = \text{Cov}_\theta(Y).\end{aligned}$$

La condition d'hyperplan sur  $\mu$  implique que  $Z$  est strictement convexe, et donc que  $m = \nabla Z(\theta)$  est en bijection avec  $\theta$ ; par conséquent, on peut également paramétrer la famille avec  $m$ .

Le  $\hat{\theta}$  du maximum de vraisemblance satisfait l'équation (exercice 2 p. 17) :

$$\frac{1}{n}(Y_1 + Y_2 + \dots + Y_n) = \nabla Z(\hat{\theta}). \quad (\text{I.5})$$

Noter que  $\hat{\theta}$  est également un estimateur par méthode de moment puisqu'il fait coïncider la moyenne empirique avec la moyenne théorique sous  $P_{\hat{\theta}}$ . Cet estimateur peut être difficile à obtenir si l'on ne possède pas d'expression simple pour la fonction  $Z$  (exercice 4 p. 17).

Le modèle général a la forme  $P_\theta(dy) = e^{\varphi(\theta)^T T(y) - Z(\theta)} \mu(dy)$ . Il est dit **canonique** si  $\varphi(\theta) = \theta$ , **standard** si  $T(y) = y$ , et **complet** si  $\Theta$  est bien tout l'ensemble défini par (I.4). La famille gaussienne  $\mathcal{N}(m, \sigma^2)$  avec  $\theta = (m, \sigma^{-2})$  n'est ni canonique ni standard.

Les modèles présentés dans la suite seront souvent issus de familles exponentielles (p. ex. le modèle autorégressif). On verra que les familles exponentielles régulières canoniques complètes non-nécessairement standard jouissent de propriétés exceptionnelles en ce qui concerne l'estimation du paramètre  $\nabla Z(\theta) = E[T(Y)]$  par  $\nabla Z(\hat{\theta})$  puisque le théorème de Lehmann-Schéffé s'y applique (théorème 20, exercice 13 p. 61) et que la borne de Cramér-Rao est atteinte (exercice 3 p. 77).

**Mélange de populations. Modèles à variables latentes.** Par exemple, dans le cas du mélange de gaussiennes, c'est une loi de probabilité de la forme

$$p(y) = \sum_{j=1}^P q_j g(y, \mu_j, \sigma_j)$$

où  $g(y, \mu_j, \sigma_j)$  est la densité de  $\mathcal{N}(\mu_j, \sigma_j^2)$  et les  $q_j$  sont des réels positifs ou nuls,  $\sum q_j = 1$ . C'est la distribution de la variable aléatoire obtenue en tirant un indice  $J$  au hasard selon les probabilités  $q_j$  puis en tirant  $Y$  selon la loi  $\mathcal{N}(\mu_J, \sigma_J^2)$ . On peut ainsi considérer  $J$  comme une variable non-observée, appelée dans la littérature *variable latente*. Un exemple en biologie :  $P = 2$ ,  $J$  vaut 1 si la cellule est saine et 2 sinon, et  $y$  est son diamètre;  $q_2$  sera la proportion de cellules infectées chez un individu concerné par l'expérience<sup>5</sup>.

Si le nombre  $P$  de populations est connu, les paramètres sont les  $(q_j, \mu_j, \sigma_j)$ ; sinon, on sort du cadre paramétrique.

L'étude du cas simple du  $n$ -échantillon d'un mélange de deux gaussiennes de moyenne  $\theta_1$  et  $\theta_2$ , de variance 1, avec  $q_1 = q_2 = 1/2$ , illustre bien la complexité de la vraisemblance.

Remarquons que dans le cas d'un mélange de lois issues d'une famille exponentielle,  $\sum_{j=1}^P q_j p_{\theta_j}(y)$ , la paire  $(e_J, Y)$  (où  $e_i$  est le  $i$ -ième vecteur de la base canonique) est issue d'une famille exponentielle dont les paramètres sont les  $(\log(q_j), \theta_j)$ . Ceci n'est pas d'un grand intérêt direct puisque  $J$  n'est pas observé, mais explique que des algorithmes d'estimation basés sur le concept de données manquantes (ici  $J$ ) peuvent donner de bons résultats (cf. exercice 13 p. 27).

**Modèles de régression. Apprentissage.** On observe des paires  $(X_i, Y_i)$ ,  $1 \leq i \leq n$ , dont la loi satisfait :

$$Y_i = n_\theta(X_i) + u_i, \quad E[u_i] = 0. \quad (\text{I.6})$$

5. On trouvera de nombreux exemples dans [110].

Les  $X_i$  sont considérés comme déterministes et la fonction  $n_\theta(x)$  est connue (comme fonction de  $\theta$  et  $x$ ) ;  $\theta$  est le paramètre. Par exemple  $Y_i$  est la concentration de produit actif dans un médicament conservé à température  $T_i$  pendant une durée  $D_i$  après fabrication,  $X_i = (D_i, T_i)$  et  $n_\theta(d, t) = \theta_1 e^{-\theta_2 d - \theta_3 t}$ . La suite  $Y_i$  est non-stationnaire. Le modèle est semi-paramétrique. Il devient paramétrique si l'on postule par exemple  $u_i \sim \mathcal{N}(0, \sigma^2)$ . Le cas le plus simple est le *modèle linéaire gaussien*

$$Y_i = X_i \theta + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2) \quad (\text{I.7})$$

où  $Y_i$  est scalaire,  $X_i$  est traditionnellement un vecteur ligne,  $\theta$  une colonne, et les  $u_i$  sont iid. Par exemple en imagerie par scanner médical (tomodensitométrie),  $\theta$  est la densité de l'objet (image tridimensionnelle vectorisée en colonne) et chaque  $Y_i$  un point de sa transformée de Radon : intégrale de  $\theta$  sur une droite  $D_i$ , mesurée par l'énergie reçue du rayon traversant. On a  $X_{ij} = 1_{j \in D_i}$ ,  $X_i$  est le vecteur de sélection des points de la droite  $D_i$ .

Un autre exemple extrêmement populaire est le modèle de *régression logistique* :  $Y_i \in \{0, 1\}$  et

$$P(Y_i = 1) = \frac{1}{1 + e^{-X_i \theta}} \quad (\text{I.8})$$

$X_i$  et  $\theta$  sont des vecteurs (ligne et colonne) de même dimension. Par exemple  $Y$  est l'apparition d'un cancer et  $X$  une mesure du tabagisme, des antécédents, ... Ou  $Y$  est le fait qu'une femme travaille et  $X$  son nombre d'enfants, le salaire de son mari, ... La vraisemblance est ici immédiate à écrire ; son expression est compliquée mais son logarithme est une fonction concave de  $\theta$  (exercice 11 p. 26), ce qui facilite sa maximisation.

L'estimateur le plus classique de  $\theta$  pour le modèle (I.6) est l'estimateur des *moindres carrés* (appelé *moindres carrés conditionnels* quand  $X_i$  est considéré aléatoire), qui correspond au maximum de vraisemblance si  $u_i$  est supposé gaussien :

$$\hat{\theta}_n = \operatorname{argmin}_\theta \sum (Y_i - n_\theta(X_i))^2. \quad (\text{I.9})$$

Dans le cas linéaire, éq. (I.7), on trouve l'estimateur OLS (ordinary least squares)

$$\hat{\theta} = (X^T X)^{-1} X^T Y \quad (\text{I.10})$$

où  $X$  est la matrice dont la  $i$ -ième ligne est  $X_i$  (design matrix) et  $Y$  le vecteur de  $i$ -ième composante  $Y_i$ .

Pour étudier l'asymptotique quand  $n \rightarrow \infty$  une approche commode est de supposer que la suite  $(X_i)$  est aléatoire stationnaire, considérant que  $n_\theta$  paramètre l'espérance de  $Y$  sachant  $X$ ,  $E[Y_i | X_i] = n_\theta(X_i)$ , ce qui revient à reformuler le modèle ainsi :

$$Y_i = n_\theta(X_i) + u_i, \quad E[u_i | X_i] = 0.$$

On verra que quand le nombre de données tend vers l'infini,  $\hat{\theta}_n$  tendra, sous certaines hypothèses, vers

$$\operatorname{argmin}_\theta E[(Y - n_\theta(X))^2], \quad (\text{I.11})$$

c.-à-d. le  $\theta$  qui minimise le carré moyen de l'erreur de prédiction quand  $(X, Y)$  est tiré selon sa loi.

Les exemples présentés plus haut montrent que l'objectif peut être soit  $\theta$  lui-même (l'objet tomographié), soit d'obtenir un estimateur qui conduise à un bon prédicteur  $\hat{Y}_i = n_{\hat{\theta}}(X_i)$  sur de nouvelles données. Les performances de l'estimateur seront mesurées donc différemment, ce qui peut changer beaucoup, particulièrement si la dimension de  $\theta$  est grande.

## I.2.2 Modèles de séries temporelles

Ces modèles sont généralement markoviens, et  $\theta$  paramétrera la probabilité de transition ; le calcul de la vraisemblance repose sur la formule de Bayes :  $p_\theta(y_1, \dots, y_n) = \prod p_\theta(y_i | y_{i-1}, y_{i-2}, \dots)$ .

**Le modèle autorégressif à moyenne mobile (ARMA( $p, q$ ))** est donné par la formule suivante :

$$Y_n = \sum_{k=1}^p a_k Y_{n-k} + \varepsilon_n + \sum_{k=1}^q b_k \varepsilon_{n-k}, \quad (\text{I.12})$$

où les  $\varepsilon_n$  sont des  $\mathcal{N}(0, \sigma^2)$  indépendantes. Les paramètres sont les  $a_k, b_k$  et  $\sigma$ . Si les  $b_i$  sont nuls, le modèle est simplement autorégressif.

Partant de conditions initiales, pour l'instant déterministes, pour  $Y$  et  $\varepsilon$ , spécifiquement  $(Y_1, \dots, Y_p)$  et  $(\varepsilon_p, \dots, \varepsilon_{p+1-q})$ , on peut obtenir  $(Y_n)_{n>p}$  par récurrence à l'aide du passé de  $(\varepsilon_n)_{n>p}$ ; on peut de même obtenir  $\varepsilon_n$  par récurrence à l'aide du passé de  $Y_n$  par la formule :

$$\varepsilon_n = Y_n - \sum_{k=1}^p a_k Y_{n-k} - \sum_{k=1}^q b_k \varepsilon_{n-k}. \quad (\text{I.13})$$

La tribu du passé de  $Y$  avant  $n$ ,  $\mathcal{F}_n = \sigma(Y_n, Y_{n-1}, \dots, Y_{p+1})$ , coïncide donc avec celle du passé de  $\varepsilon$ , et par conséquent  $\varepsilon_n$  est l'erreur de prédiction :  $\varepsilon_n = Y_n - E[Y_n | \mathcal{F}_{n-1}]$ . Comme la loi de  $Y_n$  sachant le passé est une gaussienne dont la moyenne est le membre de droite de (I.12) moins  $\varepsilon_n$ , on peut calculer facilement la vraisemblance en utilisant la formule de Bayes

$$P(Y_1, \dots, Y_n) = \prod_i P(Y_i | Y_1, \dots, Y_{i-1}).$$

Si l'on connaît les conditions initiales  $(Y_1, \dots, Y_p)$  et  $(\varepsilon_p, \dots, \varepsilon_{p+1-q})$ , alors tous les  $\varepsilon_k$  suivants peuvent être calculés à l'aide de la formule (I.13) et la vraisemblance de  $\theta = (a_1, \dots, a_p, b_1, \dots, b_q)$  s'écrit

$$\begin{aligned} \mathcal{L}(\theta, Y_{p+1}, \dots, Y_N) &= -\frac{1}{2\sigma^2} \sum_{n=p+1}^N \left( Y_n - \sum_{k=1}^p a_k Y_{n-k} - \sum_{k=1}^q b_k \widehat{\varepsilon}_{n-k} \right)^2 - N \log \sigma \\ &= -\frac{1}{2\sigma^2} \sum_{n=p+1}^N \widehat{\varepsilon}_n^2 - N \log \sigma \end{aligned}$$

où les  $\widehat{\varepsilon}_n$  sont les valeurs calculées (qui coïncident avec les vraies si le paramètre est le bon). On peut également voir cette expression comme la vraisemblance conditionnelle aux valeurs initiales mentionnées. Dans cette expression la dépendance en  $\theta$  est rendue très compliquée au travers du calcul des  $\widehat{\varepsilon}_i$ , sans compter le problème des conditions initiales inconnues en pratique pour les  $\varepsilon_k$ ; dans le cas autorégressif cette complication disparaît et l'estimation est relativement facile.

Si l'on note

$$Z_n = \begin{pmatrix} Y_n \\ \vdots \\ Y_{n-p+1} \end{pmatrix}, \quad \eta_n = \begin{pmatrix} \varepsilon_n \\ \vdots \\ \varepsilon_{n-q} \end{pmatrix}, \quad A = \begin{pmatrix} a_1 & a_2 & \dots & a_p \\ 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & b_1 & \dots & b_q \\ 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0 \end{pmatrix}$$

on a

$$Z_n = AZ_{n-1} + B\eta_n$$

ce qui permet de faire certains calculs de façon analogue au cas  $p = 1, q = 0$ , en particulier de représenter la loi stationnaire par

$$Z_n = B\eta_n + AB\eta_{n-1} + A^2B\eta_{n-2} + \dots, \quad (\text{I.14})$$

et également de voir que  $\widetilde{Z}_n = (Z_n, \varepsilon_n, \dots, \varepsilon_{n-q+1})$  admet la représentation markovienne  $\widetilde{Z}_n = \widetilde{A}\widetilde{Z}_{n-1} + \widetilde{B}\varepsilon_n$  pour un bon choix de  $\widetilde{A}$  et  $\widetilde{B}$ .

On peut également définir dans le même esprit des processus à valeurs entières positives (p. ex. nombre de cas de rougeole déclarés chaque jour<sup>6</sup>) par les équations de récurrence suivantes :

$$\Lambda_n = a_0 + \sum_{k=1}^p a_k Y_{n-k} + \sum_{k=1}^q b_k \Lambda_{n-k}$$

$$\text{Loi}(Y_n | Y_1, \dots, Y_{n-1}, \Lambda_1, \dots, \Lambda_n) = \mathcal{P}(\Lambda_n).$$

On pensera d'abord au cas simple  $p = 1, q = 0$ . Les coefficients doivent être positifs pour garantir la positivité de  $\Lambda_n$ , ce qui a l'inconvénient majeur d'imposer la positivité des corrélations du processus.

**Les processus de Poisson** sont très utilisés en théorie des files d'attente où il s'agit par exemple de représenter les instants d'arrivée de nouveaux clients (à un péage...). Si l'on appelle  $T_1, T_2, T_3, \dots$  les instants d'évènement, le processus est caractérisé par le fait que les durées entre deux évènements  $T_1, T_2 - T_1, T_3 - T_2, \dots$  sont des variables exponentielles iid de paramètre  $\lambda$ , appelé intensité. Le processus est le nombre  $N_t$  d'évènements avant  $t$ .

On montre que  $N_t$  est à accroissements indépendants, que le nombre d'évènements  $N_b - N_a$  suit une loi de Poisson  $\mathcal{P}(\tau)$ ,  $\tau = \lambda(b - a)$  ( $p_n = \tau^n e^{-\tau} / n!$ ), et que conditionnellement à  $N_b - N_a = n$  les instants d'évènement sont uniformément distribués sur l'intervalle (mesure de Lebesgue normalisée sur  $a \leq t_1 \leq \dots \leq t_n \leq b$ ).

On peut donc réaliser le processus sur  $[0, A]$  en générant  $N_A \sim \mathcal{P}(\lambda A)$  puis en tirant les  $N_A$  instants de sauts indépendamment et uniformément sur  $[0, A]$ ; ces  $T_i$  forment donc un processus de Poisson spatial (cf. § I.2.3). On peut le réaliser sur  $[0, +\infty[$  en superposant les réalisations obtenues de la même manière sur  $[0, A], [A, B], \dots$ .

**VRAISEMBLANCE.** Soit  $\tau > 0$  fixé,  $E_n = \{(t_1, \dots, t_n) : 0 \leq t_1 \leq \dots \leq t_n \leq \tau\}$  et  $\mu_n(dt)$  la mesure uniforme de masse 1 sur  $E_n$ . Soit  $Y = (T_1, \dots, T_N)$  l'observation du processus sur  $[0, \tau]$  alors pour toute fonction  $f$  on a, en vertu de ce qui vient d'être dit

$$E[f(Y)] = \sum_n P(N = n) E[f(Y) | N = n] = \sum_n e^{-\tau \lambda} \frac{\tau^n \lambda^n}{n!} \int_{E_n} f(t_1, \dots, t_n) \mu_n(dt_1, \dots, dt_n). \quad (\text{I.15})$$

Le choix  $\mu = \sum_n 1_{E_n} \mu_n$  sur  $\cup_n E_n$  conduit donc à la vraisemblance  $e^{-\tau \lambda} \lambda^N$  (au terme constant  $\frac{\tau^n}{n!}$  près). Cette vraisemblance ne dépend donc que du nombre d'évènements observés et pas de leur localisation.

**INTENSITÉ VARIABLE.** Lorsque l'intensité varie au cours du temps, on peut toujours définir un processus unique  $N_t$  à accroissements indépendants tel que  $N_b - N_a \sim \mathcal{P}(\Lambda_a^b)$ ,  $\Lambda_a^b = \int_a^b \lambda_s ds$ . Il suffit de prendre  $N_t = N_{\Lambda_t^0}$  où  $N_t^0$  est un processus de Poisson d'intensité 1. Cette fois-ci les instants de saut sur  $[a, b]$  ont, conditionnellement à  $N_b - N_a = n$  la loi de  $n$  v.a.iid de densité  $\lambda_t 1_{[a,b]} / \Lambda_a^b$ .

Si la fonction  $\lambda_s = \lambda_s(\theta)$  est paramétrée, p. ex.  $\lambda_s(\theta) = \theta_1 + \theta_2 e^{-\theta_3 s}$ , la vraisemblance du processus sur  $[a, b]$  est (un calcul plus formel devrait se faire comme en (I.15))

$$P_\theta(T_1, \dots, T_N) = P_\theta(T_1, \dots, T_N | N) P_\theta(N) \propto \left( \prod_{i=1}^N \lambda_{T_i}(\theta) \right) e^{-\int_a^b \lambda_t(\theta) dt} \quad (\text{I.16})$$

(on a omis un terme  $1/N!$  car il ne dépend pas de  $\theta$ )<sup>7</sup>. Notons que l'on a la formule classique suivante qui montre que  $\lambda_t$  s'interprète comme une mesure de la probabilité d'observer un évènement dans un avenir très proche :

$$P(\text{un saut sur } [t, t+h] | \mathcal{F}_t) = P(\mathcal{P}(\Lambda_t^{t+h}) \geq 1) = \lambda_t h + O(h^2) \quad (\text{I.17})$$

6. Pour un modèle plus complet, voir les équations (1) à (4) de [38]. Données disponibles sur la page www du premier auteur.

7. Mentionnons que l'on peut adjoindre à la mesure  $\lambda_t dt$  une somme de masses de Dirac,  $\sum_i p_i \delta_{\tau_i}$ ,  $p_i \leq 1$ . Interprétation : la simulation doit alors être complétée en ajoutant indépendamment un évènement à chaque instant  $\tau_i$  avec probabilité  $p_i$ . Un terme  $\prod_i p_i^{\varepsilon_i} (1 - p_i)^{1 - \varepsilon_i}$ , où  $\varepsilon_i$  vaut 1 si un évènement est observé en  $\tau_i$  et 0 sinon, doit être ajouté à la vraisemblance. Voir [42].

où  $\mathcal{F}_t$  est la tribu engendrée par les variables  $(N_s)_{s \leq t}$  (ou encore les  $T_i 1_{T_i \leq t}$ ).

Noter que si l'on n'observe pas toute la trajectoire, ce qui est courant, *la vraisemblance de l'observation devient très difficile à calculer*. Ce sera le cas dans beaucoup d'exemples qui vont suivre, et ceci motivera l'utilisation d'autres estimateurs que le maximum de vraisemblance.

**Processus à intensité stochastique.** Il s'agit d'une extension du cas précédent où  $\lambda_t$  peut dépendre du passé de la trajectoire (i.e. être une variable  $\mathcal{F}_t$ -mesurable) avec toujours l'interprétation donnée par (I.17). On peut définir ce processus comme celui dont la restriction à tout intervalle  $[0, b]$  a pour densité par rapport au processus de Poisson unité [41]

$$\left( \prod_{T_i \leq b} \lambda_{T_i} \right) e^{-\int_0^b (1-\lambda_t) dt}.$$

Il s'agit en fait de la loi conditionnelle au passé avant 0 puisque les  $\lambda_t$  risquent d'en dépendre. Le membre de droite de (I.16) est donc encore la vraisemblance.

Un exemple typique est un  $\lambda$  de la forme  $\lambda_t = \mu(t) + \sum_{T_i \leq t} \gamma(t - T_i)$ . Ce sont des processus de Hawkes [88]<sup>8</sup>. Des modèles de ce type sont utilisés pour les tremblements de terre (secousse initiale puis répliques [117, 60]), avec par exemple  $\gamma(x) = a(x+b)^{-p}$ ,  $\mu = cste$ . Ou encore  $\lambda_t = \psi(\theta t - N_t)$  où  $\psi$  est une fonction croissante  $> 0$  [118]; on trouvera dans [123] une application concernant les parasites du pin, et dans [113] une autre concernant les gangs à Los Angeles.

La simulation peut se faire par une sorte de méthode de rejet si l'on sait que  $\lambda$  est borné par une constante  $M$  connue : Il suffit de simuler un processus de Poisson  $T_n$  d'intensité  $M$  puis successivement pour chaque  $n$ , simuler une variable  $U_n \sim \mathcal{U}([0, M])$  et ne garder  $T_n$  que si  $U_n$  est inférieur à l'intensité  $\lambda_{T_n}$  en  $T_n$  calculée sur la base du passé de la nouvelle trajectoire. On vérifie facilement, au moins informellement, que (I.17) est satisfait.

**Un processus à sauts Markoviens** est caractérisé par sa matrice de taux de transition (ou générateur infinitésimal)  $A$ , de diagonale négative, positive ailleurs, et de sommes de ligne nulles.

Pour simuler ce processus, on procède comme suit : soit  $X(t) = i$  l'état courant, pour chaque état  $j \neq i$  on simule une variable exponentielle de paramètre  $A_{ij}$  et l'on fait la transition correspondant au minimum  $m$  de ces variables. On a alors l'état à l'instant  $X(t+m)$ , et  $X(s) = X(t)$  pour  $s \in [t, t+m[$ .

On vérifie que partant de  $i$ , l'état suivant  $j$  est choisi indépendamment du passé avec probabilité  $-A_{ij}/A_{ii}$ . Si l'on fait abstraction du temps, l'évolution peut donc se faire par une chaîne de Markov. La durée de séjour dans chaque état peut être simulée dans un deuxième temps, sa loi est une variable exponentielle indépendante du reste sauf de l'état courant  $i$  et de paramètre  $-A_{ii}$ .

On a la propriété suivante : La matrice de transition  $P^t$ ,  $P_{ij}^t = P(X_{t+s} = j | X_s = i)$ , vaut  $e^{tA}$ . En particulier :

$$P(X_{s+h} = j | X_s = i) = A_{ij}h + O(h^2), \quad i \neq j. \quad (\text{I.18})$$

La log-vraisemblance d'une trajectoire  $(Y_t)_{0 \leq t \leq T}$  se calcule informellement comme suit : Soit  $e_1, \dots, e_n$  la suite des états visités et  $t_1, \dots, t_n$  la suite des durées de séjour, la probabilité de la trajectoire est

$$\frac{A_{e_1 e_2}}{-A_{e_1 e_1}} \dots \frac{A_{e_{n-1} e_n}}{-A_{e_{n-1} e_{n-1}}} \left( -A_{e_1 e_1} e^{A_{e_1 e_1} t_1} \right) \dots \left( -A_{e_{n-1} e_{n-1}} e^{A_{e_{n-1} e_{n-1}} t_{n-1}} \right) e^{A_{e_n e_n} t_n}$$

le dernier terme correspondant à la probabilité de rester *au moins* un temps  $t_n$  dans  $e_n$ . D'où finalement

$$\mathcal{L}(A, (Y_t)_{0 \leq t \leq T}) = \sum_{ij} N_{ij} \log(A_{ij}) + \sum_i T_i A_{ii} = \sum_{ij} N_{ij} \log(A_{ij}) + \int_0^T A(t) dt$$

8. Soit  $\mu(t)$  une fonction positive et  $\gamma(t)$  une autre fonction positive, nulle sur  $\mathbb{R}_-$  et d'intégrale  $< 1$ . Le processus de Hawkes associé est construit de la façon suivante : simuler la première génération qui est un processus d'intensité  $\mu$ . Partant de chaque événement  $T_i$  de première génération, simuler ses enfants comme un processus de Poisson d'intensité  $\gamma(t - T_i)$  (ce sont donc des événements postérieurs à  $T_i$ ). Simuler ensuite la troisième génération de manière analogue en partant de chaque événement de la seconde, et continuer jusqu'à extinction des familles (ce qui arrive car  $\int \gamma(t) dt < 1$  : Le nombre d'enfants d'une génération est d'espérance  $\nu = \int \gamma$  et l'on montre que le nombre de descendants d'un parent donné est d'espérance  $1 + \nu + \nu^2 \dots = 1/(1 - \nu)$ ). Il est remarquable que ce processus soit à intensité stochastique avec la formule annoncée.

où  $N_{ij}$  est le nombre de fois que la transition  $i \rightarrow j$  a été effectuée,  $T_i$  est le temps total passé dans l'état  $i$ , et  $A(t) = A_{X(t)X(t)}$ .

EXEMPLE : LE PROCESSUS DU COIFFEUR. Cet exemple a pour but d'illustrer la modélisation de files d'attente ; la simplicité du principe provient de ce que tout est construit à partir de variables exponentielles, qui comme on l'a vu ont des propriétés tout à fait exceptionnelles. Nous ne faisons pas ici de démonstration, et renvoyons aux livres sur les files d'attente.

Le temps que le coiffeur met à couper les cheveux d'un client est exponentiel d'espérance 20 min. Les clients arrivent selon un processus de Poisson de taux 2/heure, mais n'entrent pas si les deux fauteuils d'attente sont occupés. L'état est le nombre de clients dans le salon,  $S = \{0, 1, 2, 3\}$ . Alors l'état est un processus à sauts markoviens avec

$$A = \begin{pmatrix} -2 & 2 & 0 & 0 \\ 3 & -5 & 2 & 0 \\ 0 & 3 & -5 & 2 \\ 0 & 0 & 3 & -3 \end{pmatrix}.$$

EXEMPLE : ÉVOLUTION DU CANCER DU POUMON [120, 93]. Il y a trois états : *pas de rechute*, *rechute*, *mort*. La combinaison des trois traitements utilisée pour chaque patient (*chimiothérapie* et/ou *radiothérapie* et/ou *thérapie hormonale*) est introduite via le vecteur  $z = (1_c, 1_r, 1_h) \in \{0, 1\}^3$  dépendant du patient et une paramétrisation de la matrice :

$$A_{ij} = \exp(\langle z, \theta_{ij} \rangle), \quad \theta_{ij} \in \mathbb{R}^3, \quad i \neq j.$$

Comme  $A_{21} = A_{3i} = 0$ , il y a 9 paramètres :  $(\theta_{12}, \theta_{13}, \theta_{23})$ . Les auteurs disposent de l'histoire de 300 patients (300 trajectoires) ; la vraisemblance est le produit des vraisemblances des trajectoires individuelles (les patients étant indépendants). L'estimation du modèle permet de comparer l'effet des traitements.

EXEMPLE : DYNAMIQUE DES ÉPIDÉMIES<sup>9</sup>. Soit le modèle SIR (Susceptible, Infectious, Recovered/Removed), où  $S$  est le nombre d'individus susceptibles d'être atteints par contagion,  $I$  le nombre d'individus infectés, et  $R$  le nombre d'individus récemment guéris et donc (provisoirement) immunisés,  $S + I + R = n$  :

	transition	taux
Guérison	$(S, I, R) \rightarrow (S, I - 1, R + 1)$	$\lambda_0 I$
Infection	$(S, I, R) \rightarrow (S - 1, I + 1, R)$	$\lambda_1 SI/n$
Sensibilisation	$(S, I, R) \rightarrow (S + 1, I, R - 1)$	$\lambda_2 R$ .

La première transition représente une guérison ( $I \rightarrow R$ ), le taux est proportionnel à  $I$  car on suppose que chaque individu infecté guérit indépendamment du reste en un temps exponentiel de paramètre  $\lambda_0$ . Le risque d'infection pour un individu sain étant proportionnel à la proportion d'individus infectés, le taux de la deuxième transition ( $S \rightarrow I$ ) (infection) est choisi proportionnel à  $SI/n$ . Si sur des données  $(S(t), I(t), R(t))_{t \leq T}$  on observe  $N_0, N_1, N_2$  transitions de chaque type, la log-vraisemblance est, à un terme indépendant des paramètres près

$$N_0 \log(\lambda_0) + N_1 \log(\lambda_1) + N_2 \log(\lambda_2) - \int_0^T \left\{ \lambda_0 I(t) + \lambda_1 S(t)I(t)/n + \lambda_2 R(t) \right\} dt.$$

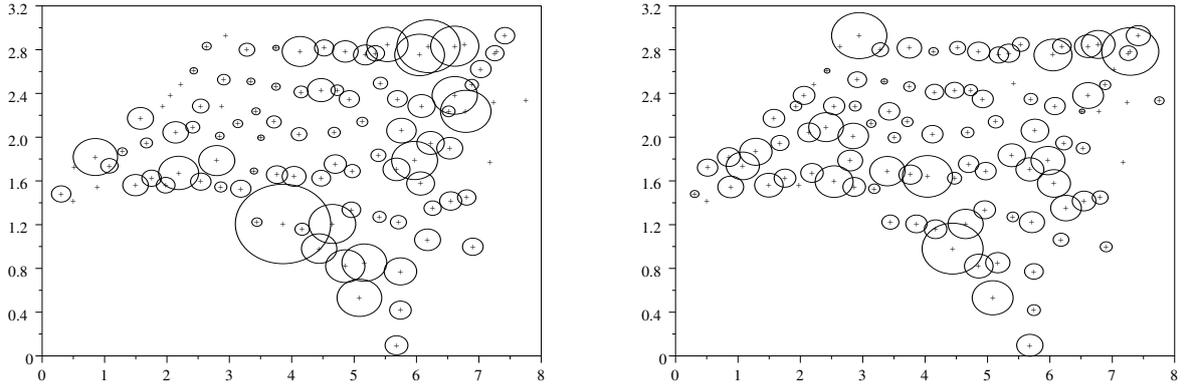
### 1.2.3 Modèles spatiaux

On ne parlera pas ici des champs gaussiens qui sont les modèles les plus utilisés : il s'agit de vecteurs gaussiens particuliers. S'ils sont classiquement utilisés pour modéliser les données spatiales, les réseaux bayésiens également en sont des cas particuliers.

On trouvera des compléments à cette section dans [78, 50] et dans un article célèbre de Julian Besag [33].

**Modèle conditionnellement binomial.** Considérons les données de mort subite de nouveau-né dans les 100 comtés de Caroline du Nord [58].

9. Pour un traité général sur le sujet, recommandons [62].



On dispose du nombre  $Y_a$  de morts subites de bébés de moins d'un an entre 1974 et 1978 dans le comté  $a$ , et du nombre  $n_a$  de naissances durant cette période,  $a$  variant de 1 à 100. On dispose du même tableau pour 1979-1984. Typiquement  $n_a$  est compris entre 500 et 30000 et  $Y_a$  entre 0 et 50. On a représenté sur la figure les données sur chaque période par une carte où l'on a placé pour chaque comté un cercle dont le diamètre est proportionnel à  $Y_a/n_a$  (la valeur moyenne de cette quantité est 0,002) ; l'unité spatiale est la centaine de kilomètres (la figure est verticalement dilatée). Une corrélation spatiale semble apparaître, elle serait due à l'existence d'une variable explicative spatialement corrélée (structure de la population).

On considère le modèle suivant (les  $n_a$  sont déterministes)

$$P(Y) = Z^{-1} \exp \left\{ \alpha \sum_a Y_a + \frac{1}{2} \sum_{a,b} c_{ab} (Y_a - qn_a)(Y_b - qn_b) \right\} \prod_a \frac{1}{Y_a!(n_a - Y_a)!} \quad (\text{I.19})$$

avec  $c_{aa} = 0$  et  $c_{ab} = c_{ba}$ . On vérifie facilement que la loi de  $Y_a$  sachant les autres valeurs est binomiale :

$$\text{Loi}(Y_a | Y_b, b \neq a) = \mathcal{B} \left( n_a, \frac{\lambda_a}{\lambda_a + 1} \right), \quad \lambda_a = e^{\alpha + \sum_b c_{ab}(Y_b - qn_b)} \quad (\text{I.20})$$

Ce modèle est visiblement surparamétré puisqu'il y a 4950 paramètres  $c_{ab}$  en jeu pour ici 200 observations. Une solution est de se donner une forme particulière :

$$c_{ab} = \gamma \mathbf{1}_{d(a,b) < \delta}$$

où  $d(a, b)$  est la distance entre les comtés  $a$  et  $b$ . Le modèle n'a plus pour paramètres que  $\alpha, \gamma, \delta$  et  $q$ . On peut également faire dépendre  $\alpha$  de la période,  $\alpha = \alpha_p$ , ce qui fait alors 5 paramètres.

En dehors de l'intérêt de l'estimation des paramètres, on peut également considérer ce modèle comme une base pour faire des tests ; par exemple, on pourrait tester les hypothèses  $\alpha_1 = \alpha_2$  (pas de différence significative entre les périodes) ou encore  $\gamma = 0$  (pas de corrélation significative entre comtés proches).

Il s'avère utile d'ajouter un terme  $\beta \sum n_a Y_a$  à l'exponentielle de (I.19) ce qui revient à remplacer  $\alpha$  par  $\alpha + \beta n_a$  dans (I.20).

La vraisemblance  $P(Y)$  (ou  $P(Y^1)P(Y^2)$  puisqu'on a deux périodes) est trop compliquée pour que l'on puisse envisager de calculer son maximum sur le paramètre  $\theta = (\alpha, \beta, \gamma, q)$ . On choisit plutôt de maximiser la pseudo-vraisemblance<sup>10</sup>, une des méthodes classiques que l'on verra au § I.3.2.

On obtient alors les estimées  $\hat{\delta} = 0,64$  (ce qui fait une moyenne de 7 termes non nuls par ligne de la matrice  $c$ ),  $\hat{\alpha} = -6.1$ ,  $\hat{\gamma} = 0.01165$ ,  $\hat{q} = 0.00188$  et  $\hat{\beta} = -1.64 \cdot 10^{-5}$ . Les paramètres estimés sont ici considérés comme significativement non nuls. Le fait que  $\beta \neq 0$  vient possiblement de ce que  $n_a$  est corrélé avec des variables explicatives manquantes. Noter que l'espérance de  $Y_a/n_a$  en absence de dépendance,  $e^{\hat{\alpha}}/(1 + e^{\hat{\alpha}}) = 0.0022$ , est proche de  $\hat{q}$  ce qui confirme l'interprétation du terme  $qn_a$  comme un recentrage.

10. On pourrait utiliser la méthode présentée à l'exercice 4 p. 17, mais cela reste très difficile à mettre en œuvre.

**Les processus ponctuels spatiaux.** Ils permettent de modéliser des distributions de points dans un espace  $E$ , par exemple la distribution de certaines espèces dans les forêts [144], de maladies dans des régions [63, 64]; voir aussi [4] pour des exemples d'études en biologie. Soit  $E$  (généralement une partie compacte de  $\mathbb{R}^d$ ) un espace muni d'une mesure finie  $\mu$  sans atome : une réalisation du **processus de Poisson spatial** de mesure d'intensité  $\mu$  est un ensemble  $X = \{X_1, \dots, X_N\}$  où  $N$  est tiré selon une loi de Poisson de paramètre  $\mu(E)$  et les  $X_i$  sont tirés indépendamment dans  $E$  selon  $\mu(\cdot)/\mu(E)$  (l'absence d'atome implique que les  $X_i$  sont différents). Donc pour toute fonction  $\varphi$  définie sur  $\cup_{n \geq 0} E^n$  symétrique de ses arguments (fonction d'ensemble), borélienne<sup>11</sup> et bornée

$$E[\varphi(X)] = \sum_{n=0}^{\infty} \frac{e^{-\mu(E)}}{n!} \int \dots \int \varphi(x_1, \dots, x_n) \mu(dx_1) \dots \mu(dx_n). \quad (\text{I.21})$$

Le nombre de points présents dans une partie mesurable  $D$  suit une loi de Poisson de paramètre  $\mu(D)$  [99]. Une propriété simple mais importante est le principe de superposition donnant la loi de la réunion de deux processus de Poisson indépendants :  $\mathcal{P}(\mu) \cup \mathcal{P}(\nu) \sim \mathcal{P}(\mu + \nu)$ .

On considère souvent une famille paramétrée  $\mathcal{P}(q_\theta(y)\mu(dy))$  ( $q_\theta$  est une fonction positive avec  $\mu(q_\theta) < \infty$ )

$$X \sim \mathcal{P}(q_\theta(y)\mu(dy))$$

et l'estimation utilisera que la vraisemblance du nouveau modèle est, à une constante près, pour une observation  $X = \{X_1, \dots, X_n\}$

$$\left( \prod_{k=1}^n q_\theta(X_k) \right) e^{-\int q_\theta(y)\mu(dy)}.$$

Par exemple  $X$  désigne des emplacements d'arbres,  $q_\theta(y) = \exp(\theta_0 + \theta_1 \xi_1(y) + \theta_2 \xi_2(y))$  où  $\xi_1(y)$  et  $\xi_2(y)$  sont deux caractéristiques de la nature du sol mesurées en  $y$  et le paramètre  $\theta$  quantifie leur influence sur la présence ou non d'un arbre en  $y = X_k$ .

Une autre approche consiste à définir la loi d'un processus ponctuel par sa densité  $p_\theta$  par rapport au processus de Poisson  $\mathcal{P}(\mu)$  lui-même. Par exemple le processus de **Strauss** défini par

$$X \sim p_\theta(x)\mathcal{P}(\mu), \quad p_\theta(x) = Z(\theta)^{-1} \beta^{n(x)} \gamma^{s(x)}, \quad (\text{I.22})$$

où  $n = n(x)$  est le nombre de points de  $x = \{x_1, \dots, x_n\}$ ,  $s(x)$  est le nombre de paires points à distance inférieure à un seuil  $r$  donné,  $\theta = (\log \beta, \log \gamma)$ , et  $Z(\theta)$  est la constante de normalisation. On suppose ici  $\mu(E) = 1$  car  $\beta$  règle l'intensité du Poisson. La maximisation de la vraisemblance est rendue difficile par le fait que la fonction  $Z(\theta)$  est inconnue (voir l'exercice 4 p. 17). L'estimation par méthode de moments sera étudiée à l'exercice 12 p. 26.

Une généralisation est le processus de **Gibbs** dont la densité, paramétrée ici par  $\theta = (\beta, \gamma)$ , est

$$Z(\theta)^{-1} \beta^{n(x)} \exp\left(-\sum_{i < j} \Phi(\gamma(x_i - x_j))\right)$$

où  $\Phi$  est une certaine fonction [109].

Une autre famille importante est formée des modèles de **Neyman-Scott**, appelés encore **cluster point process**, originellement utilisés pour modéliser la distribution des galaxies [116] : un processus de Poisson spatial  $X$  (typiquement non-observé) réalise les centroïdes (centres de galaxies) et autour de chaque point  $X_k$  les observations (planètes)  $Y_k$  sont distribuées (dans un second temps, c.-à-d. conditionnellement à  $X$ ) selon une certaine loi choisie typiquement comme une loi de Poisson dont l'intensité est de la forme  $\mu_0(x - X_k)dx$  où  $\mu_0$  est une fonction fixe (p. ex. une densité gaussienne multipliée par un facteur, les paramètres étant le facteur et la variance; c'est le modèle de Thomas<sup>12</sup>).

11. Au sens où  $\varphi(x)1_{|x|=n}$  coïncide avec une fonction borélienne sur  $\mathbb{R}^d$ , voir p.ex. [106].

12. Par exemple dans [107] :  $Y = (Y_1, \dots, Y_p)$  est l'ensemble des cas de leucémie dans l'état de New-York entre 1978 et 1982 (ensemble de points du plan). Il s'agit de voir si les cas de leucémie sont plus fréquents au voisinage de sites d'enfouissement de déchets. L'observation semble montrer que ces points ne sont pas uniformément répartis mais forment des groupes. Les auteurs présupposent l'existence d'un vecteur  $X = (X_1, \dots, X_n)$ , ensemble des centroïdes des groupes, et tentent d'étudier sa loi conditionnelle aux observations  $Y_i$ .

Le processus de **Cox** est un processus de Poisson dont le taux est aléatoire (il est donc en vérité Poisson conditionnellement au taux). Un cas classique est le processus de Cox log-gaussien : on se donne un processus gaussien sur  $\mathbb{R}^d$ ,  $G(x)$ , puis on pose  $\Lambda(x) = \exp(G(x))$ ; la mesure  $\Lambda(x)dx$  est le taux du processus. Le processus  $G(x)$  peut être modélisé soit en se donnant sa fonction de covariance<sup>13</sup>, par exemple  $c(x, y) = a \exp(-b\|x - y\|)$ , et sa fonction moyenne s'il n'est pas centré, soit comme  $G(x) = g_0(x) + \sum U_i g_i(x)$  où les  $g_i$  sont des fonctions fixes et les  $U_i$  sont des iid  $\mathcal{N}(0, 1)$ .

Un processus de Neymann-Scott dont les lois conditionnelles aux centroïdes sont poissonniennes est un processus de Cox en vertu du principe de superposition (une réunion de processus de Poisson indépendants est un processus de Poisson dont le taux est la somme des taux individuels); dans le cas particulier considéré plus haut, son intensité aléatoire est  $\sum_k \mu_0(x - X_k)dx$ .

Les paramètres apparaissant dans ces modélisations sont typiquement les paramètres de modélisation des covariances ( $a$  et  $b$  dans l'exemple), ou des paramètres intervenant dans  $\mu_0$  (p. ex. loi gaussienne centrée de variance inconnue), ou des paramètres de régression s'il y a des variables explicatives en chaque point (vitesse du vent, altitude...), p. ex.  $\Lambda(x) = \exp(\langle \theta, y(x) \rangle + G(x))$ . Pour tous ces modèles, les calculs de vraisemblance peuvent être très compliqués ce qui pousse à utiliser des méthodes d'estimation différentes du maximum de vraisemblance [4].

**Processus de Poisson spatio-temporel.** Il s'agit simplement de considérer un processus de Poisson sur l'espace produit  $E \times [0, T]$ .

Souvent la mesure d'intensité s'écrit  $\mu(dx)\lambda(dt)$  où  $\lambda$  est la mesure de Lebesgue sur  $[0, T]$ , ce qui revient à dire que les instants d'évènement forment un processus de Poisson standard d'intensité  $\mu(E)$ , et que pour un tel instant l'évènement est tiré sur  $E$  selon  $\mu(dx)/\mu(E)$ .

Ce modèle est par exemple utilisé pour représenter la distribution spatio-temporelle des orages [119]; cet article modélise également les averses, au travers de leur loi conditionnelle aux orages : chaque orage génère un nombre poissonnien d'averses, distribuées uniformément en son voisinage, de durée et d'intensité aléatoires; c'est un processus de Neyman-Scott.

## I.2.4 Exercices et compléments

**Exercice 1 ( $C_p$ -Mallows et estimateur SURE<sup>14</sup> de Stein).** Soit des observations  $Y = (Y_1, \dots, Y_n)$ , les paramètres  $\mu_i = E[Y_i]$  et un estimateur  $\hat{\mu}_i = \hat{\mu}_i(Y)$ . On peut penser au cas de la régression, linéaire ou non. On note le risque instantané et le risque moyen

$$R = \sum_i (\mu_i - \hat{\mu}_i)^2, \quad \bar{R} = E[R].$$

- Vérifier l'équation  $(\mu_i - \hat{\mu}_i)^2 = -(Y_i - \mu_i)^2 + (Y_i - \hat{\mu}_i)^2 + 2(\hat{\mu}_i - \mu_i)(Y_i - \mu_i)$ . En déduire que

$$R = \sum_i (Y_i - \hat{\mu}_i)^2 + 2(Y_i - \mu_i)\hat{\mu}_i + \xi \tag{I.23}$$

où  $\xi$  est une variable dont l'espérance est la somme des variances des  $Y_i$  changée de signe.

- Montrer que dans le cas linéaire, i.e.  $\hat{\mu}(y) = \mu_0 + Py$ ,  $P \in \mathbb{R}^{n \times n}$  (p.ex. estimateur OLS (I.10), ridge (III.40)), si  $Y$  a pour variance  $\sigma^2 Id$ , alors l'estimateur suivant de  $\bar{R}$  est non biaisé :

$$\hat{R} = \sum_i (Y_i - \hat{\mu}_i)^2 + 2\sigma^2 Tr(P) - n\sigma^2. \tag{I.24}$$

- Montrer que dans le cas de l'estimateur OLS (I.10) on a  $Tr(P) = p$ , où  $p$  est le nombre de variables. Le  $C_p$ -Mallows est l'estimateur de  $\sigma^{-2}\bar{R}$  qui s'en déduit :

$$C_p = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - X_i \hat{\beta})^2 + 2p - n.$$

13. Une fonction  $c(x, y)$  définie sur  $\mathbb{R}^d \times \mathbb{R}^d$  est une covariance valide si pour toute suite finie  $x_1, \dots, x_n$  la matrice  $c(x_i, x_j)$  est définie positive.

14. Stein's Unbiased Risk Estimate.

4. (SURE) Montrer que si  $y \mapsto \hat{\mu}(y)$  est régulière (on ne détaille pas les hypothèses), et si  $Y \sim \mathcal{N}(\mu, \sigma^2 Id)$ , alors  $Cov(\hat{\mu}_i, Y_i) = \sigma^2 E(\frac{\partial \hat{\mu}_i}{\partial y_i}(Y))$ . Ceci conduit à l'estimateur non-biaisé de  $\bar{R}$

$$\hat{R} = \sum_i (Y_i - \hat{\mu}_i)^2 + 2\sigma^2 \sum_i \frac{\partial \hat{\mu}_i}{\partial y_i}(Y) - n\sigma^2. \quad (\text{I.25})$$

*Indication* : On fera une intégration par parties dans l'intégrale gaussienne.

L'intention de Mallows était de mettre en place un critère permettant de comparer les performances de modèles linéaires ayant moins de variables explicatives ;  $\sigma$  est traditionnellement calculé sur la base du modèle le plus compliqué. Nous approfondirons p. 43. Une application de (I.25) à l'estimateur lasso, éq. (III.37), se trouve dans [149], dont un cas particulier est traité à l'exercice 11 p. 61.

On peut voir (I.25) comme résultant grossièrement de (I.24) après linéarisation de  $\hat{\mu}$ , car  $Cov(\hat{\mu}_i, Y_i) = E[(\hat{\mu}_i(Y) - \hat{\mu}_i(\mu))(Y_i - \mu_i)]$ . Ceci rend défendable son utilisation dans des cas non gaussiens proches de la linéarité.

Le terme  $\text{div}(\hat{\mu}(Y))$  de (I.25) est interprété comme une mesure du nombre de degrés de liberté [70].

**Exercice 2 (Familles exponentielles).** Démontrer la formule (I.5). Vérifier que le processus de Strauss du § I.2.3 est une famille exponentielle.

**Exercice 3 (Régression non-linéaire).** Montrer que si (I.6) est satisfait pour  $\theta = \theta_*$ , alors  $\theta_*$  est solution de (I.11).

**Exercice 4 (Vraisemblance non normalisée).** On dispose d'une observation  $Y^0$  d'une variable  $Y$  de densité (par rapport à une mesure fixe  $\mu(dy)$ ) de la forme  $e^{Q(y,\theta)}/Z(\theta)$  où la fonction  $Q$  est connue et simple à calculer, contrairement à  $Z(\theta)$  ; par exemple les variables de (I.19) avec  $\theta = (\alpha, \gamma, \delta, q)$  ou le processus de Strauss du § I.2.3.

1. Montrer que le gradient de la log-vraisemblance vaut

$$\nabla_{\theta} Q(Y^0, \theta) - E_{\theta}[\nabla_{\theta} Q(Y, \theta)].$$

2. Montrer que pour tout  $\psi \in \Theta$  fixe, la log-vraisemblance vaut, à une quantité indépendante de  $\theta$  près

$$Q(Y^0, \theta) - \log E_{\psi}[e^{Q(Y,\theta) - Q(Y,\psi)}]$$

où  $E_{\psi}$  est la loi associée à  $\psi$ . Noter que cette équation permet aussi d'obtenir le premier point.

La première équation permet d'estimer  $\theta$  au maximum de vraisemblance par méthode de gradient si l'on dispose d'algorithmes permettant de simuler des échantillons pour le calcul des espérances (p. ex. méthodes MCMC). La méthode de gradient fournit une suite  $\theta_n$  convergeant vers la limite cherchée et la deuxième équation permet de mesurer l'évolution de la vraisemblance lorsque l'on s'approche de la limite, en calculant le deuxième terme avec un  $\psi$  fixe, proche de la limite (i.e.  $\psi = \theta_n$  pour un  $n$  grand). Cette méthode est très utilisée [89, 114] mais généralement moins efficace que celle des vraisemblances partielles lorsque le problème se prête à cette dernière (présentée plus bas p. 20).

3. On suppose maintenant que  $Y = (Y_V, Y_H)$  où seul  $Y_V^0$  est observé. Montrer que le gradient de la log-vraisemblance de  $Y_V^0$  vaut

$$E_{\theta}[\nabla_{\theta} Q(Y, \theta) | Y_V = Y_V^0] - E_{\theta}[\nabla_{\theta} Q(Y, \theta)].$$

La première espérance s'obtiendra typiquement par simulation de réalisations de  $Y_H$  conditionnellement à  $Y_V = Y_V^0$ . C'est un algorithme proposé pour les modèles graphiques (équation (17.37) de [11] ; voir aussi § 17.4.4 du même livre).

**Exercice 5 (Files d'attente).** Des clients arrivent à un guichet selon un processus de Poisson de paramètre  $\lambda_a$  et sont servis à leur tour avec un temps de service exponentiel de paramètre  $\lambda_s$ . Montrer que la vraisemblance des observations sur  $[0, t]$  vaut

$$\lambda_a^{N_a} \lambda_s^{N_s} e^{-\lambda_a t - \lambda_s(t-t_0)},$$

où  $N_a$  est le nombre de clients arrivés,  $N_s$  est le nombre de clients servis, et  $t_0$  le temps pendant lequel aucun client n'est présent (observer que le nombre de client dans le système forme un processus de Poisson dont on déterminera les transitions en s'inspirant du processus du coiffeur).

## I.3 Principes d'estimation

### I.3.1 Premiers concepts

**Le cadre asymptotique.** On se donne une suite infinie d'observations  $Y_1, Y_2, \dots$  et une suite d'estimateurs  $\hat{\theta}_n = \hat{\theta}_n(Y_1, \dots, Y_n)$ .

La suite  $Y_i$  n'est pas forcément indépendante, par exemple s'il s'agit d'une série temporelle. Les premières questions qui se posent sont alors typiquement la convergence de  $\hat{\theta}_n$  vers une limite  $\theta_*$  désirée, puis la normalité asymptotique de  $\sqrt{n}(\hat{\theta}_n - \theta_*)$ . Il peut arriver que la vitesse de convergence soit plus rapide que  $n^{-1/2}$  comme c'est le cas de l'estimateur de l'exercice 5 p. 7 ( $P_\theta$  est la loi uniforme sur  $[0, \theta]$ ).

Le cadre abstrait général est celui d'une **suite d'expériences statistiques**  $(P_\theta^n)_{n \geq 1}$ . Souvent  $P_\theta^n$  représentera la loi de  $Y_1, \dots, Y_n$  sous le paramètre  $\theta$ , mais rien n'impose en fait que  $P_\theta^n$  soit la restriction de  $P_\theta^{n+1}$  à une certaine tribu.

**Consistance.** Une suite d'estimateurs  $\hat{\theta}_n = \hat{\theta}_n(Y_1, \dots, Y_n)$  est dite consistante (ou convergente) si  $\hat{\theta}_n \rightarrow \theta_*$  en probabilité, et fortement consistante si la convergence a lieu presque sûrement.

**Biais.** Le biais d'un estimateur  $\hat{\theta}$  de  $\theta_*$  est  $b(\hat{\theta}) = E[\hat{\theta}] - \theta_*$ , et s'il est nul  $\hat{\theta}$  est dit non-biaisé<sup>15</sup>. Les estimateurs non-biaisés sont rares en pratique. Ce concept n'est pas invariant par changement de variables, car si  $\hat{\theta}$  est non-biaisé,  $f(\hat{\theta})$  sera généralement un estimateur biaisé de  $f(\theta_*)$  (exercice 3 p. 24).

**Variance asymptotique.** Si l'on a une limite en loi du type  $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, V)$ , on dit que  $V$  est la variance asymptotique de l'estimateur. On est souvent capable de calculer  $V$ , contrairement à la variance de chaque  $\hat{\theta}_n$ . Ceci permet d'avoir une évaluation de l'erreur d'estimation, et surtout de comparer différents estimateurs (en comparant leur variance asymptotique<sup>16</sup>).

**Limite en loi et changement de variables.** Le théorème suivant exprime que si un estimateur est consistant, sa distribution asymptotique est essentiellement la même que celle de toute fonction  $g$  de ce dernier car  $g(\hat{\theta}_n) \simeq g(\theta_*) + \nabla g(\theta_*)(\hat{\theta}_n - \theta_*)$  :

1 - THÉORÈME (Delta-method)

Soit  $\hat{\theta}_n \in \mathbb{R}^d$  une suite d'estimateurs satisfaisant

$$r_n(\hat{\theta}_n - \theta_*) \rightarrow X \quad \text{en loi}$$

pour une certaine distribution  $X$  et une certaine suite  $r_n$  tendant vers l'infini. Soit également une fonction  $g$  différentiable en  $\theta_*$ . Alors

$$r_n(g(\hat{\theta}_n) - g(\theta_*)) \rightarrow \nabla g(\theta_*)X \quad \text{en loi.}$$

*Démonstration.* On a

$$g(\theta_* + h) - g(\theta_*) = \nabla g(\theta_*)h + \|h\|\varepsilon(h)$$

pour une certaine fonction  $\varepsilon$  qui tend vers 0 quand  $\|h\|$  tend vers 0. Par conséquent

$$g(\hat{\theta}_n) - g(\theta_*) = \nabla g(\theta_*)(\hat{\theta}_n - \theta_*) + \|\hat{\theta}_n - \theta_*\|\varepsilon(\hat{\theta}_n - \theta_*). \quad (\text{I.26})$$

Il ne suffit plus qu'à tout multiplier par  $r_n$  et à utiliser que le produit d'une suite convergeant en loi par une suite convergeant en probabilité vers 0 converge en probabilité vers 0 (lemme de Slutsky). ■

15. On verra au § III.1.1 une définition un peu différente dans le cas où l'on s'intéresse à un risque non quadratique ; pour l'instant celle-ci nous suffit.

16. Ceci a été mis en pratique pour la première fois par Laplace en 1818 pour estimer  $\theta_*$  dans le modèle de régression  $y_i = \theta_*x_i + e_i$ , les  $e_i$  étant iid centrés de densité symétrique  $\varphi$ . Il compare  $\hat{\theta}_{LS} = \sum_i y_i x_i / \sum_i x_i^2$  (moindres carrés) à  $\hat{\theta}_{MS} = \arg \min \sum_i |y_i - \theta x_i|$ . Il trouve que  $\hat{\theta}_{LS} - \theta_*$  est proche d'une gaussienne de variance  $\int x^2 \varphi(x) dx / \sum_i x_i^2$  tandis que  $\hat{\theta}_{MS} - \theta_*$  est proche d'une gaussienne de variance  $1/(4\varphi(0)^2 \sum_i x_i^2)$ . La méthode à choisir dépend donc du résultat de la comparaison entre  $1/(4\varphi(0)^2)$  et  $\int x^2 \varphi(x) dx$ . Ceci est bien raconté par Stephen Stigler [133].

On déduit facilement du théorème, ou directement de l'équation (I.26), est que si  $g$  est injective, si  $\nabla g(\theta_*)$  est inversible et si

$$r_n(g(\widehat{\theta}_n) - g(\theta_*)) \longrightarrow Y \quad \text{en loi,}$$

alors

$$r_n(\widehat{\theta}_n - \theta_*) \longrightarrow \nabla g(\theta_*)^{-1} Y \quad \text{en loi.}$$

Ceci s'applique à l'estimateur de modèle exponentiel (2) puisque le théorème-limite central s'applique à  $\nabla Z(\widehat{\theta})$ , et l'on trouve que  $\sqrt{n}(\widehat{\theta}_n - \theta_*)$  converge en loi vers  $\mathcal{N}(0, \text{Cov}_{\theta_*}(Y)^{-1})$ .

### I.3.2 Trois estimateurs classiques

On présente ici informellement l'esprit dans lequel sont construits les estimateurs qui seront considérés dans le chapitre suivant. Les trois formes proposées recouvrent l'essentiel des estimateurs connus. Rappelons que même dans le cas paramétrique, l'estimateur au maximum de vraisemblance n'est pas forcément préféré (cf. la discussion p. 6).

**Minimum de contraste.  $M$ -estimateurs.** Ce type d'estimateur généralise celui des moindres carrés (I.9) et celui du maximum de vraisemblance ; sa forme est :

$$\widehat{\theta}_n = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^n K(\theta, Y_i). \quad (\text{I.27})$$

On suppose que  $\theta_*$  minimise la fonction de contraste :

$$\theta_* = \operatorname{argmin}_{\theta} k(\theta), \quad k(\theta) = E[K(\theta, Y_i)].$$

Dans le cas de (I.9), on a  $k(\theta) = E[(Y_1 - n_{\theta}(X_1))^2]$  qui est bien la quantité à minimiser, et l'on minimise la version empirique de ce contraste.

En toute généralité, les estimateurs à minimum de contraste sont ceux pour lesquels

$$K_n(\widehat{\theta}_n, \omega) - \min_{\theta} K_n(\theta, \omega) \longrightarrow 0 \quad (\text{I.28})$$

où la fonction  $K_n$  doit avoir une limite quand l'échantillon grandit :

$$\lim_n K_n(\theta, \omega) = k(\theta), \quad p.s. \quad (\text{I.29})$$

avec toujours  $\theta_* = \operatorname{argmin}_{\theta} k(\theta)$ .

MOINDRES CARRÉS ET ESTIMATEUR ROBUSTE POUR UN PARAMÈTRE DE TRANSLATION.

On se donne le modèle  $Y_i = \theta_* + \varepsilon_i$  où les  $\varepsilon_i$  ont une distribution symétrique par rapport à 0. On a les deux estimateurs

$$\begin{aligned} \widehat{\theta} &= \operatorname{argmin}_{\theta} \sum_{i=1}^n (Y_i - \theta)^2 = \frac{1}{n} \sum_i Y_i \\ \widehat{\theta}_R &= \operatorname{argmin}_{\theta} \sum_{i=1}^n \psi(Y_i - \theta) \end{aligned} \quad (\text{I.30})$$

où  $\psi(x) = x^2 1_{|x| \leq \alpha} + (2\alpha|x| - \alpha^2) 1_{|x| > \alpha}$  pour un certain  $\alpha$ .  $\widehat{\theta}_R$  satisfait une équation implicite qui l'exprime comme la moyenne des données projetées :  $\widehat{\theta}_R = n^{-1} \sum_i \max(\min(Y_i, \widehat{\theta}_R + \alpha), \widehat{\theta}_R - \alpha)$ . On le retrouvera au § II.3.1.

PSEUDO-VRAISEMBLANCE. VRAISEMBLANCE CONDITIONNELLE. VRAISEMBLANCE PARTIELLE <sup>17</sup>.

La vraisemblance partielle consiste à remplacer la vraisemblance par un produit de marginales extraites des données ; les marginales ont rarement une expression simple mais un exemple est donné à

17. Voir par exemple [55]. Chaque terme est un cas particulier du précédent.

l'exercice 7 p. 25.

La vraisemblance conditionnelle consiste à remplacer la vraisemblance par un produit de termes dont chacun est la probabilité d'une partie des données conditionnellement à une autre (éventuellement vide), l'exemple le plus classique étant

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{a \in A} P_{\theta}(Y_a | Y_b, b \neq a) \quad (\text{I.31})$$

ce qui peut conduire, comme on va le voir, à des formules beaucoup plus simples à traiter. Elle est couramment utilisée en fiabilité ([74] p.10) et en processus spatiaux ([33] § 6.1). La vraisemblance partielle est donc un cas particulier de vraisemblance conditionnelle, où l'on ne conditionne par rien. Dans l'expression ci-dessus, la famille  $(Y_a)_{a \in A}$  représente a priori une seule réalisation, par exemple d'un processus spatial, le produit ne faisant qu'augmenter si l'on en a plusieurs indépendantes.

Reprenons le modèle conditionnellement binomial (I.19) :

$$P_{\theta}(Y) = Z^{-1} \exp \left\{ \alpha \sum Y_a + \beta \sum n_a Y_a + \frac{1}{2} \sum_{a \neq b} c_{ab} (Y_a - qn_a)(Y_b - qn_b) \right\} \prod_a \frac{1}{Y_a! (n_a - Y_a)!}$$

où  $c_{ab} = \gamma 1_{d(a,b) < \delta}$  et  $\theta = (\alpha, \beta, \gamma, \delta, q)$ . Comme  $Z$  dépend de  $\theta$ , la vraisemblance est difficile à calculer. Strauss et Ikeda [136]<sup>18</sup> proposent d'utiliser la pseudo-vraisemblance (I.31) qui est simple à calculer car la loi de  $Y_a$  sachant les autres est binomiale (cf. (I.20)) de paramètres  $n_a$  et  $\lambda_a / (1 + \lambda_a)$  avec

$$\lambda_a = \exp \left\{ \alpha + n_a \beta + \gamma \sum_b 1_{d(a,b) < \delta} (Y_b - qn_b) \right\}.$$

La pseudo-vraisemblance obtenue coïncide avec la vraisemblance d'un modèle de régression logistique habituel où chaque individu  $a$  a pour réponse  $Y_a$  et pour variable explicative

$$X_a = \left( n_a, \sum_{b \neq a} 1_{d(a,b) < \delta} Y_b, \sum_{b \neq a} 1_{d(a,b) < \delta} n_b \right)$$

et les paramètres sont  $(a, \beta, \gamma, -\gamma q)$  (tout du moins si  $\delta$  est fixé). On peut donc faire sa maximisation à l'aide d'un logiciel standard. Dans les résultats présentés au § I.2.3, le paramètre  $\delta$  a été estimé par essais successifs de sorte à maximiser la pseudo-vraisemblance obtenue à  $\delta$  fixé.

Pour la convergence, voir l'exercice 14 p. 28. D'autres exemples se trouvent aux exercices 7 p. 25, 10 p. 25, et également 5 p. 44 pour une pseudo-vraisemblance non conditionnelle.

**PSEUDO-VRAISEMBLANCE EN APPRENTISSAGE.** Considérons le modèle de mélange de populations de la page 8, avec pour but de fournir une méthode pour prédire  $J$  sachant  $Y$ . On dispose ici de données étiquetées  $(Y_i, J_i)$ . Supposons pour simplifier les  $q_i$  connus. Une alternative au maximum de vraisemblance est de maximiser  $\sum_i \log(p_{\theta}(J_i | Y_i))$ . Alors que la modélisation initiale, qui est celle de l'analyse discriminante, présente plutôt  $J$  comme une variable explicative et  $Y$  comme une réponse, on voit que la pseudo-vraisemblance utilisée renverse les rôles. On qualifie parfois la première approche (maximum de vraisemblance) de *générative*, et la seconde de *discriminative*. Noter que dans le cas d'un mélange de deux gaussiennes, la loi de  $J$  sachant  $Y$  correspond à un modèle de régression logistique (éq. I.8 où  $(X, Y)$  joue le rôle de  $(Y, J)$ ); dans le cas de plus de deux gaussiennes, c'est un modèle logistique multinomial<sup>19</sup>.

**Fonctions d'estimation. Z-estimateurs.**  $\hat{\theta}_n$  est cette fois solution de

$$n^{-1} \sum_{i=1}^n H(\hat{\theta}_n, Y_i) = 0 \quad (\text{I.32})$$

ou plus généralement

$$H_n(\hat{\theta}_n, \omega) \longrightarrow 0 \quad (\text{I.33})$$

18. Ils travaillent sur les données *monastère* présentées à l'exercice 10 p. 25.

19. Pour des compléments concernant les approches génératives et discriminatives, on peut consulter [145]. Pour la comparaison directe de l'analyse discriminante et de la régression logistique, voir [122].

pour une certaine fonction  $H_n$ . L'hypothèse étant que le vrai paramètre  $\theta_*$  satisfait

$$h(\theta_*) = 0, \quad h(\theta) = \lim_n H_n(\theta, \omega). \quad (\text{I.34})$$

On suppose que la limite existe p.s.

Ce type d'estimateur généralise l'estimateur à minimum de contraste en prenant  $H = \nabla_\theta K$ . Noter cependant que si  $K$  n'est pas dérivable, ou que plusieurs extremums existent, l'estimateur à minimum de contraste ne sera pas associé à une fonction d'estimation.

EXEMPLE : RÉGRESSION LINÉAIRE. On s'intéresse au salaire  $y_i$  d'un individu  $i$  en fonction de son nombre d'années d'études  $a_i$ . On postule le modèle de régression simple

$$y_i = \alpha_* + \beta_* a_i + e_i. \quad (\text{I.35})$$

où  $e_i$  est indépendant de  $a_i$ . La méthode des moindres carrés qui consiste à minimiser en  $\alpha$  et  $\beta$

$$\sum_i (y_i - \alpha - \beta a_i)^2$$

conduit à (I.32) avec

$$\begin{aligned} H_0(\theta, X_i) &= \begin{pmatrix} y_i - \alpha - \beta a_i \\ a_i(y_i - \alpha - \beta a_i) \end{pmatrix}, & X_i &= (y_i, a_i), & \theta &= (\alpha, \beta) \\ h_0(\theta) &= \begin{pmatrix} E[y_1 - \alpha - \beta a_1] \\ E[a_1 y_1 - \alpha a_1 - \beta a_1^2] \end{pmatrix}. \end{aligned} \quad (\text{I.36})$$

On a bien  $h_0(\theta_*) = 0$ .

EXEMPLE : MÉTHODE DE LA VARIABLE INSTRUMENTALE POUR LE CAS D'UNE VARIABLE EXPLICATIVE NON-MESURÉE. Dans l'exemple qui suit, le Z-estimateur ne peut se réduire à M-estimateur.

On veut maintenant prendre en compte les capacités intrinsèques  $c_i$  (variable non-mesurée) de l'individu. On postule le modèle de régression

$$y_i = \alpha_* + \beta_* a_i + \gamma_* c_i + e_i.$$

Les triplets  $(a_i, c_i, e_i)$  sont indépendants et  $e_i$  est centré indépendant de  $(a_i, c_i)$ . Le paramètre d'intérêt,  $\beta_*$ , représente l'influence du nombre d'années d'étude *indépendamment* des qualités de l'étudiant. On suppose que  $c_i$  est sur une échelle telle que  $E[c_i] = 0$ , si bien que  $E[y_i] = \alpha_* + \beta_* E[a_i]$ .

La corrélation entre  $a_i$  et  $c_i$  fait que  $E[y_i|a_i] \neq \alpha_* + \beta_* a_i$ ; on ne peut donc pas régler la question en rejetant le terme  $\gamma_* c_i$  dans le bruit, c.-à-d. considérer le modèle (I.35). L'estimation de  $\alpha_*$  et  $\beta_*$  peut néanmoins se faire par l'utilisation d'une variable explicative décorrélée de  $c_i$  et de  $e_i$ , par exemple le niveau de salaire des parents  $s_i$ . En effet on voit que la paire d'équations

$$\begin{aligned} E[y_i] &= \alpha_* + \beta_* E[a_i] \\ E[y_i s_i] &= \alpha_* E[s_i] + \beta_* E[s_i a_i] \end{aligned}$$

permet d'estimer  $\alpha_*$  et  $\beta_*$  à partir d'un nombre infini d'échantillons en remplaçant les espérances par leur estimée empirique. On obtient un Z-estimateur où la fonction d'estimation est une modification de (I.36) qui ne correspond plus à un minimum de contraste :

$$H(\theta, X_i) = \begin{pmatrix} y_i - \alpha - \beta a_i \\ y_i s_i - \alpha s_i - \beta a_i s_i \end{pmatrix}, \quad X_i = (y_i, a_i, s_i), \quad \theta = (\alpha, \beta)$$

qui est bien d'espérance nulle si  $\alpha = \alpha_*$  et  $\beta = \beta_*$ . On a

$$h(\theta) = E[H(\theta, X_i)] = \begin{pmatrix} 1 & E[a_1] \\ E[s_1] & E[a_1 s_1] \end{pmatrix} \begin{pmatrix} \alpha_* - \alpha \\ \beta_* - \beta \end{pmatrix}$$

Pour avoir unicité de la solution de  $h(\theta) = 0$ , il faut que la matrice soit inversible, c.-à-d. que  $s$  ne soit pas décorrélé de  $a$ .

Comme il a été mentionné, la corrélation entre les variables  $a$  et  $c$  fait que si l'on estime  $(\alpha^*, \beta^*)$  par des moindres carrés basés sur le modèle  $y_i = \alpha_* + \beta_* a_i + e'_i$ , le résultat sera essentiellement différent.

EXEMPLE : PROCESSUS DE STRAUSS. Pour ce processus, cf. p. 15, la vraisemblance est difficile à calculer. On montre (exercice 12 p. 26) que la fonction suivante est une fonction d'estimation pour la paire  $(\beta, \gamma)$  :

$$H(\beta, \gamma, X) = \left( \begin{array}{c} n(X) - \beta \int \gamma^{s(\{X, \xi\})} \mu(d\xi) \\ 2s(X) - \beta \int s(X, \xi) \gamma^{s(\{X, \xi\})} \mu(d\xi) \end{array} \right).$$

**Méthode des moments.** Elle consiste à exprimer  $\theta_*$  comme fonction de statistiques simples des données (moyenne, variance, corrélations...)

$$\theta_* = f(m, \sigma, r_{12}, \dots)$$

et d'obtenir un estimateur en remplaçant les statistiques exactes par les statistiques empiriques :

$$\hat{\theta} = f(\hat{m}, \hat{\sigma}, \hat{r}_{12}, \dots).$$

EXEMPLE. Soit le modèle de régression linéaire de (I.7)

$$Y_i = X_i \theta + u_i, \quad u \sim \mathcal{N}(0, \sigma^2 Id).$$

Il faut également estimer  $\sigma^2$ . Si  $Q$  est la matrice de la projection orthogonale sur l'orthogonal des colonnes de  $X$  (matrice ayant  $X_i$  en  $i$ -ième ligne), alors  $QY = Qu$  ce qui fait que

$$E[\|Qy\|^2] = q\sigma^2$$

où  $q$  est le rang de  $Q$ , d'où  $\hat{\sigma}^2 = \frac{1}{q} \|Qy\|^2$ . C'est l'estimateur non-biaisé habituel, différent de l'estimateur au maximum de vraisemblance; l'estimateur REML pour les modèles à effets aléatoires repose sur cette approche.

MÉTHODE DES MOMENTS GÉNÉRALISÉE. Dans le cas paramétrique iid avec observations scalaires et  $\Theta \subset \mathbb{R}^d$ , on peut par exemple chercher  $\hat{\theta}_n$  solution de

$$\int y^k p_\theta(y) \mu(dy) = \frac{1}{n} \sum_{i=1}^n Y_i^k, \quad k = 1 \dots d$$

relation qui exprime, implicitement,  $\theta$  comme fonction des premiers moments empiriques. Dans les cas favorables, il y a une unique solution car il y a autant d'équations que d'inconnues. Cet estimateur peut se réécrire avec une fonction d'estimation.

L'utilisation de moments supplémentaires peut donner un meilleur estimateur,

$$\hat{\theta}_n = \arg \min_{\theta} \sum_{k=1}^K p_k \left( E_\theta[Y^k] - \frac{1}{n} \sum_{i=1}^n Y_i^k \right)^2$$

où  $K \geq d$  les  $p_k$  sont à définir (de manière à avoir un bon estimateur). C'est la méthode des moments généralisée. Il s'agit d'un estimateur à minimum de contraste, mais qui n'est pas de la forme (I.27).

*Remarque.* La méthode des moments a un pendant non-paramétrique en estimation de densité qui consiste à choisir la distribution de densité  $\pi(y)$  par rapport à  $\mu^n$  pour  $Y = (Y_1, \dots, Y_n)$  qui soit d'entropie maximale sous la contrainte que les moments coïncident avec les moments empiriques et éventuellement sous d'autres contraintes (indépendance des  $Y_i$  sous  $\pi$ ...). Rappelons que l'entropie de  $\pi$  vaut

$$E_\pi[-\log \pi(Y)] = - \int \log(\pi(y_1, \dots, y_n)) \pi(y_1, \dots, y_n) \mu(dy_1) \dots \mu(dy_n).$$

Maximiser l'entropie s'interprète comme choisir la distribution qui ajoute le moins d'information possible, qui soit en quelque sorte la plus uniforme. Par exemple si l'on cherche la distribution d'entropie maximale, d'espérance marginale commune  $m$ , et de variance marginale commune  $\sigma^2$  (en pratique estimées empiriquement sur  $(Y_1, \dots, Y_n)$ ), on trouve la gaussienne  $\mathcal{N}(m1_n, \sigma^2 Id_n)$ ; si l'on impose de surcroît les valeurs des  $p$  premières corrélations, on trouve un processus autorégressif d'ordre  $p$ .

### I.3.3 Un résultat de convergence

Des hypothèses adéquates sur les fonctions  $K_n(\theta, \omega)$  et  $H_n(\theta, \omega)$  vont permettre d'obtenir la convergence de  $\hat{\theta}_n$  vers  $\theta_*$ . Elles se résument en deux points

- ▶ Convergence uniforme des fonctions
- ▶ Unicité de la solution du problème limite.

L'objectif sera alors de vérifier que ces conditions sont satisfaites avec probabilité 1. La vérification de la première fera intervenir la loi des grands nombres uniforme, qui est le premier point du chapitre suivant. Énonçons le théorème de nature déterministe que nous allons exploiter :

#### 2 - THÉORÈME

Soit  $\theta \mapsto H_n(\theta)$  (resp.  $K_n(\theta)$ ) une suite de fonctions non-nécessairement continues définies sur un compact  $\Theta$  à valeurs dans un espace vectoriel normé (resp. dans  $\mathbb{R}$ ). On suppose que cette suite converge **uniformément** vers une limite **continue**  $h(\theta)$  (resp.  $k(\theta)$ ). Soit  $\theta_n$  une suite de points de  $\Theta$  :

- ▶ Si  $H_n(\theta_n) \rightarrow 0$  alors  $h(\theta_n) \rightarrow 0$ .  
Si en outre l'équation  $h(\theta) = 0$  admet une solution unique  $\theta_*$ , alors  $\theta_n \rightarrow \theta_*$ .
- ▶ Si  $K_n(\theta_n) - \min_{\theta} K_n(\theta) \rightarrow 0$  alors  $k(\theta_n) \rightarrow \min(k)$ .  
Si de plus le minimum de  $k$  est atteint en un unique point  $\theta_*$ , alors  $\theta_n \rightarrow \theta_*$ .

On a donc en particulier convergence p.s. du Z-estimateur (resp. M-estimateur) si ces hypothèses sont vérifiées avec probabilité 1 pour les fonctions  $\theta \mapsto H_n(\theta, \omega)$  (resp.  $\theta \mapsto K_n(\theta, \omega)$ ).

*Démonstration.* La convergence de  $h(\theta_n)$  et  $k(\theta_n)$  est conséquence immédiate de la convergence uniforme des fonctions et la convergence de  $\theta_n$  vient de ce que dans le cas contraire on peut extraire de  $(\theta_n)$  une suite convergente ayant une limite différente de  $\theta_*$ , ce qui entraîne facilement une contradiction car  $h$  et  $k$  sont continues. ■

**Exemple.** On peut appliquer ce résultat à l'estimateur de la médiane dans le cas d'un espace métrique général, donné par le contraste :

$$K_n(\theta) = \frac{1}{n} \sum_{i=1}^n d(Y_i, \theta).$$

L'équicontinuité implique la convergence uniforme sur tout compact. Tout va donc bien si  $\Theta$  est compact et si  $E[d(Y_1, \theta)]$  a un minimum unique  $\theta_* \in \Theta$ . Si seulement les parties fermées bornées de  $\Theta$  sont compactes, on peut se ramener au cas compact en notant que la suite  $\theta_n$  est p.s. bornée car  $d(0, \theta_n) \leq K_n(\theta_n) + K_n(0) \leq 2K_n(0)$  (le compact dans lequel on se place dépend donc de  $\omega$ , mais peu importe).

On peut faire une discussion analogue avec la moyenne de Fréchet :  $K_n(\theta) = \frac{1}{n} \sum_{i=1}^n d(Y_i, \theta)^2$ .

**Comment obtenir la convergence uniforme ?** C'est le point délicat essentiel. L'utilisation de la loi des grands nombres permet en général d'obtenir assez naturellement la convergence simple (pour tout  $\theta$ ). Le théorème suivant donne deux résultats simples permettant de passer de la convergence ponctuelle à la convergence uniforme de fonctions<sup>20</sup> ; ils sont toutefois peu utilisés car les hypothèses sont fortes :

#### 3 - THÉORÈME

Soit  $f_n$  une suite de fonctions croissantes sur un intervalle compact  $I \subset \mathbb{R}$  convergeant en tout point d'une partie dense  $D \subset I$  vers la restriction à  $D$  d'une fonction  $f$  continue sur  $I$ , alors la convergence est uniforme.

Soit  $f_n$  une suite de fonctions convexes sur un ouvert convexe  $O \subset \mathbb{R}^d$  convergeant vers une certaine limite en tout point d'une partie dense de  $O$ , alors elle converge en tout point de  $O$  et la convergence est uniforme sur tout compact.

20. Le premier est classique. Pour le second, voir [127], Th.I.8.

Dans ce théorème, comme dans le suivant, il est intéressant de s'être ramené à une partie dense car la convergence ponctuelle de  $H_n(\theta)$  vers  $h(\theta)$  résultera de la loi des grands nombres, ce qui impose de ne pas avoir à manipuler un nombre plus que dénombrable d'ensembles de probabilité un (c.-à-d. que si  $A_\theta = \{\omega : H_n(\theta) \rightarrow h(\theta)\}$  et  $P(A_\theta) = 1$  pour tout  $\theta$ , et si  $D$  est dénombrable, alors  $P(\cap_{\theta \in D} A_\theta) = 1$ ).

Ce théorème s'applique bien dans le cas de l'estimateur (I.30) avec  $f_n(\theta) = \frac{1}{n} \sum_{i=1}^n \psi(Y_i - \theta)$ . On a également concavité de la log-vraisemblance dans le cas du modèle de régression logistique ordonnée (exercice 11 p. 26).

En raison du caractère trop restrictif de ce théorème, on présentera dans la suite le théorème 4 qui est spécifiquement adapté aux suites de fonctions aléatoires qui sont des moyennes empiriques.

### I.3.4 Exercices et compléments

**Exercice 1.** À quelle famille appartient l'estimateur (I.9)?

**Exercice 2.** On considère le modèle pour la paire  $(X, Y)$  :

$$Y = (X + \alpha)e + \mu, \quad X \sim \mathcal{E}(\lambda), \quad e \sim \mathcal{N}(0, 1).$$

Le paramètre est  $\theta = (\alpha, \lambda, \mu)$ . Proposer un estimateur simple pour  $\lambda$  et pour  $\mu$  basé sur un  $n$ -échantillon  $(X_i, Y_i)_{1 \leq i \leq n}$ . Calculer  $E[(Y - \mu)^2(X - \lambda^{-1})]$ . En déduire un estimateur de  $\alpha$ . A quel type d'estimateur a-t-on affaire? On pourra comparer, en termes de complexité, à l'estimation de  $\alpha$  au maximum de vraisemblance (à  $\mu$  et  $\lambda$  connus).

**Exercice 3.** Que peut-on dire de l'estimateur (scalaire)  $\hat{\theta}$  si à la fois  $\hat{\theta}$  et  $\hat{\theta}^2$  sont non-biaisés?

**Exercice 4 (Estimation de la variance. Maximum de vraisemblance du maximum de vraisemblance).** Soit  $Y_i$  une suite iid,  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ , et  $S = \sum_{i=1}^n (Y_i - \bar{Y})^2$ .

1. On suppose  $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ .

(a) Quelle est l'estimateur au maximum de vraisemblance de  $\theta = (\mu, \sigma^2)$ ?

(b) Soit  $X$  observation unique de loi  $\Gamma(\beta, p)$ , i.e. de densité  $(x/\beta)^p e^{-x/\beta} x^{-1} / \Gamma(p)$ . Quel est l'estimateur au maximum de vraisemblance de  $\beta$  lorsque  $p$  est connu?

(c) On sait que  $S$  est  $\sigma^2$  fois un  $\chi_{n-1}^2$ , c.-à-d. qu'il suit la loi  $\Gamma(\beta, p)$  avec  $\beta = 2\sigma^2$ ,  $p = (n-1)/2$ . En déduire un autre estimateur de  $\sigma^2$  basée sur l'observation de  $S$ . Quelle est la propriété bien connue de ce dernier?

2. S'inspirer du point précédent pour proposer une stratégie analogue concernant l'estimation de  $\sigma$  dans le modèle exposé à l'exercice 4 p. 7, dans le cas  $n < p$ . On rappelle qu'une somme de  $n$  variables iid de loi  $\mathcal{E}(\lambda)$  suit une  $\Gamma(1/\lambda, n)$ .

**Exercice 5.** Soit  $X_1, \dots, X_n$  une suite iid de loi  $\mathcal{N}(\theta_*, 1)$ . On estime  $\theta_*$  par

$$\hat{\theta}_n = \arg \min_{\theta} \sum_{i=1}^n \frac{1}{3} |\theta - X_i|^3 + \frac{1}{2} |\theta - X_i|^2.$$

1. Exprimer ce M-estimateur comme un Z-estimateur (i.e. donner la fonction  $H$ ). Vérifier sa validité (i.e.  $h(\theta_*) = 0$ ).

2. On suppose maintenant que  $X_i$  suit une loi de densité  $p(x - \theta_*)$  où  $p$  est une densité connue. Proposer une condition simple sur  $p$  pour que l'estimateur reste valide.

**Exercice 6 (Contrastes en estimation de densité).** Soit  $(Y_i)$  une suite iid et  $p_\theta(y)$ ,  $\theta \in \mathbb{R}^d$ , un modèle paramétrique pour la densité de la loi de  $Y_i$ . On suppose donc que  $Y_i \sim p_{\theta_*}(y) dy$ . Dans cet exercice, on ne cherchera pas à faire de démonstration rigoureuse de convergence, mais on calculera ce qui est demandé, en supposant que les hypothèses nécessaires à un comportement attendu sont satisfaites.

1. Soit le contraste

$$K_n(\theta) = \int p_\theta(y)^2 dy - \frac{2}{n} \sum_{i=1}^n p_\theta(Y_i).$$

- (a) Calculer la fonction  $k(\theta)$ . Cette fonction dépend bien entendu de la loi commune aux  $Y_i$ , c.-à-d. de  $\theta_*$ . Vérifier que  $\theta_*$  minimise bien  $k$ , ce qui justifie l'utilisation de ce contraste.
- (b) Quelle sera la limite de la suite  $\hat{\theta}_n$  si la loi de  $Y_i$  a une densité  $q$  qui n'appartient pas au modèle? On demande simplement une formulation géométrique simple.
- (c) Généraliser à  $K_n(\theta) = \int u(y)p_\theta(y)^2 dy - \frac{2}{n} \sum_{i=1}^n u(Y_i)p_\theta(Y_i)$  pour une fonction  $u \geq 0$ . Le meilleur  $u$  est calculé à l'exercice 4 p. 44.

2. Pour quelle valeur de  $c$  le contraste suivant

$$K_n(\theta, Y) = \int p_\theta(y)^4 dy - \frac{c}{n} \sum_{i=1}^n p_\theta(Y_i)^3$$

pourra-t-il raisonnablement conduire à une estimation asymptotiquement consistante de  $\theta_*$ ?

**Exercice 7 (Pseudo-vraisemblance [57]).** Soit  $X_i$ ,  $1 \leq i \leq n$ , une suite de variables gaussiennes centrées réduites telles que  $Cor(X_i X_j) = \theta$ ,  $i \neq j$ . On observe  $Y_i = 1_{X_i > 0}$  et l'on cherche à estimer  $\theta$ . On pourra admettre que pour  $i \neq j$ , si  $\theta_*$  désigne le vrai paramètre,

$$P(Y_i = 1, Y_j = 1) = q(\theta_*), \quad q(\theta) = \frac{1}{4} + \frac{\arcsin(\theta)}{2\pi}.$$

Expliquer ce qui justifie le choix de la pseudo-vraisemblance suivante

$$\mathcal{L}_p(\theta, Y) = \left( \sum_{i < j} 1_{Y_i = Y_j} \right) \log(2q(\theta)) + \left( \sum_{i < j} 1_{Y_i \neq Y_j} \right) \log(1 - 2q(\theta))$$

et donner l'équation dont  $\hat{\theta}$  est solution.  $\theta_*$  est-il le minimum unique de la fonction  $k(\theta)$ ?

**Exercice 8 (Vraisemblance conditionnelle).** Soit  $Y_1, \dots, Y_n$  un processus AR(1) (i.e. ARMA(1,0)). Ecrire la vraisemblance de  $Y_2, \dots, Y_n$  conditionnellement à  $Y_1$ . En déduire l'estimateur au maximum de vraisemblance (conditionnel à la valeur initiale  $Y_1$  comme au § I.2.2)  $\hat{\theta}_1$ . En déduire également la loi de  $Y_i$  sachant  $(Y_j)_{j \neq i}$ , puis l'estimateur au maximum de vraisemblance conditionnelle associé  $\hat{\theta}_2$  (équation (I.31) avec  $A = \{2, 3, \dots, n-1\}$ ). Montrer que si  $(Y_i)_{i \geq 1}$  est stationnaire ergodique (i.e. la loi des grands nombres s'applique) alors  $\hat{\theta}_1$  et  $\hat{\theta}_2$  convergent chacun vers une limite que l'on explicitera, limites généralement différentes si le processus n'est pas un AR(1).

**Exercice 9 (Un mauvais contraste. Validation croisée)<sup>21</sup>.** Soit  $(Y_i)$  une suite iid dont on cherche à estimer la densité. On a classiquement une famille d'estimateurs  $\hat{p}_h(x)$  indexée par un paramètre  $h$  qu'il va falloir choisir (largeur des intervalles de l'histogramme, largeur de fenêtre de l'estimateur à noyau, etc).

La famille  $\hat{p}_h$  est désormais donnée (i.e. on raisonne conditionnellement aux  $Y_i$ ). Une autre suite  $(Z_i)$  iid de même loi que les  $Y_i$  sera utilisée pour choisir le meilleur  $h$ . Comme ce meilleur  $h$  risque de dépendre du point  $x$ , on se propose de considérer le contraste local suivant inspiré du maximum de vraisemblance :

$$K(h, Z) = \sum_{i=1}^n 1_{a \leq Z_i \leq b} \log \hat{p}_h(Z_i)$$

où  $[a, b]$  est un petit intervalle contenant  $x$ . On note  $p_*$  la densité commune aux  $Y_i$ . Expliquer pourquoi ce contraste est valide si  $a = -\infty$  et  $b = +\infty$  (utiliser l'inégalité de Jensen, cf. la discussion autour de (III.3)), et observer que ceci n'a pas de raison de se produire si  $[a, b]$  ne contient pas le support de  $p_*$ . Proposer une modification normalisée de  $K$  qui corrige ce problème.

**Exercice 10 (Vraisemblance conditionnelle).** Strauss et Ikeda [136] travaillent sur les données *monastère* de Sampson : un tableau  $n \times n$ ,  $n = 18$ , de variables  $X_{ij}$  valant 1 si le moine  $i$  déclare bien s'entendre avec  $j$  et 0 sinon. Un modèle classique pour ce type de données (social network) est le suivant

$$P_\theta(X_{ij}, 1 \leq i, j \leq n) = Z(\theta)^{-1} \exp \left\{ \theta_1 \sum_{i,j} X_{ij} + \theta_2 \sum_{i,j} X_{ij} X_{ji} \right\}.$$

21. Inspiré de [84]. Une discussion générale sur le choix de  $h$  par validation croisée se trouve dans [3] p. 540 et suivantes.

Il s'agit d'un des modèles les plus simples parmi ceux considérés dans l'article. Si  $\theta_2 = 0$ , les  $X_{ij}$  sont indépendants et  $p = (1 + e^{-\theta_1})^{-1}$  est la probabilité qu'un lien se tisse entre deux individus ;  $\theta_2$  mesure la réciprocité des relations. Proposer une pseudo-vraisemblance basée sur un maximum de vraisemblance conditionnelle.

**Exercice 11 (Régression catégorielle ordonnée [121]).** La réponse  $Y$  est à valeurs dans  $\{1, \dots, m\}$ , p.ex. {guérison, élimination des symptômes, aucun effet, aggravation} ; on note  $X \in \mathbb{R}^p$  le vecteur (ligne) des variables explicatives, p.ex. (âge, dose, ...), ici déterministes. Le modèle postule l'existence de  $\mu_0 = -\infty < \mu_1 \dots < \mu_m < \mu_{m+1} = +\infty$ , de  $\beta \in \mathbb{R}^p$ , et d'une variable  $Z$ , dite latente car non-observée telle que  $Y = k$  si  $\mu_k \leq Z + X\beta < \mu_{k+1}$ . Dit plus crûment,  $X\beta$  est l'effet déterministe qui fait croître  $Y$  et  $Z$  est l'aléa. Si  $F$  (connue) est la fonction de répartition de  $Z$ , ceci revient à dire que

$$P(Y = k) = F(\mu_{k+1} - X\beta) - F(\mu_k - X\beta)$$

ce qui résume le modèle, paramétré par  $(\mu_1, \dots, \mu_m, \beta)$ . Les choix classiques pour  $F$  sont soit  $F(t) = 1/(1 + e^{-t})$  (modèle logistique), soit la loi gaussienne (modèle probit).

On rappelle qu'une fonction  $\varphi \geq 0$  est log-concave si l'ensemble  $\{\varphi > 0\}$  est convexe et que  $\log \varphi$  y est concave. On vérifiera que  $(x, y) \mapsto 1_{x < y}$  est log-concave sur  $\mathbb{R}^2$ , et que le produit de deux fonctions log-concaves l'est encore. Le théorème de Prékopa-Leindler affirme alors que si  $(x, y) \mapsto \varphi(x, y)$  est log-concave sur  $\mathbb{R}^{n \times p}$ , alors  $x \mapsto \int \varphi(x, y) dy$  l'est sur  $\mathbb{R}^n$ .

Démontrer que si la densité  $f(t) = F'(t)$  est log-concave, alors la log-vraisemblance d'un échantillon  $(X_1, \dots, X_n)$  est une fonction concave de  $(\mu_1, \dots, \mu_m, \beta)$ . Le théorème 3 s'applique donc.

**\*Exercice 12 (Un Z-estimateur pour les processus de Poisson ponctuels).** Soit  $(S, \mathcal{A}, \mu)$  un espace mesuré, la mesure  $\mu$  est supposée finie et sans atome, et soit  $\mathcal{P}_\mu$  la loi du processus de Poisson ponctuel d'intensité  $\mu$  sur  $S$ . Soit  $X \mapsto f_\theta(X)$  une famille paramétrée de fonctions définies sur  $\cup_n S^n$ , et

$$P_\theta = Z(\theta)^{-1} f_\theta(X) \mathcal{P}_\mu$$

une famille paramétrique de lois absolument continues par rapport à  $\mathcal{P}_\mu$ . On suppose que pour tout  $\theta$

$$f_\theta(X) > 0 \implies \forall x \in X, f_\theta(X \setminus \{x\}) > 0.$$

La constante de normalisation  $Z(\theta)$  est inconnue et difficile à calculer si bien que le maximum de vraisemblance pose de sérieux problèmes à réaliser.

1. Soit une fonction  $(X, \xi) \mapsto \varphi(X, \xi)$  définie sur  $(\cup_n S^n) \times S$ , mesurable bornée telle que pour tout  $\xi$  la fonction  $X \mapsto \varphi(X, \xi)$  soit symétrique en ses variables (fonction d'ensemble). Montrer que

$$E_\mu E_\theta \left[ \varphi(X, \xi) \frac{f_\theta(X \cup \{\xi\})}{f_\theta(X)} \right] = \frac{1}{\mu(S)} E_\theta \left[ \sum_{x \in X} \varphi(X \setminus \{x\}, x) \right] \quad (\text{I.37})$$

où  $E_\mu$  désigne l'espérance sous  $\mu(d\xi)/\mu(S)$  et  $E_\theta$  est l'espérance sous  $P_\theta$ .

*Indication* : Commencer par le cas  $f_\theta = 1$  en utilisant (I.21) ; terminer utilisant la définition de  $E_\theta$  et en jouant sur  $\varphi$ .

2. On considère le processus de Strauss défini par  $f_\theta(x) = \beta^{n(x)} \gamma^{s(x)} / Z(\theta)$ , où  $n = n(x)$  est le nombre de points de  $x = \{x_1, \dots, x_n\}$ ,  $s(x)$  est le nombre de paires points à distance inférieure à un seuil  $r$  donné,  $\theta = (\log \beta, \log \gamma)$ , et  $Z(\theta)$  est une constante de normalisation. On suppose ici  $\mu(S) = 1$  car  $\beta$  règle l'intensité du Poisson où  $r$  est connu et  $\theta = (\beta, \gamma)$ . Utiliser les deux fonctions

$$\varphi_1(X, \xi) = 1, \quad \varphi_2(X, \xi) = \sigma(X, \xi) = \sum_{x \in X} 1_{d(x, \xi) \leq r}$$

pour justifier la validité du Z-estimateur basé sur  $p$  réalisations du processus, donnant  $(\beta, \gamma)$  comme solution du système

$$\begin{aligned} \sum_{i=1}^p n(X_i) &= \beta \sum_{i=1}^p \int \gamma^{\sigma(X_i, \xi)} \mu(d\xi) \\ 2 \sum_{i=1}^p s(X_i) &= \beta \sum_{i=1}^p \int \sigma(X_i, \xi) \gamma^{\sigma(X_i, \xi)} \mu(d\xi). \end{aligned}$$

3. On se propose de montrer la validité de l'estimateur à minimum de contraste basé sur

$$K(\theta, X) = \int \frac{f_\theta(X \cup \{\xi\})}{f_\theta(X)} \mu(d\xi) - \sum_{x \in X} \log \frac{f_\theta(X)}{f_\theta(X \setminus \{x\})}.$$

On peut l'interpréter comme un estimateur de maximum de vraisemblance partielle (voir [27], ou [106] p.109).

- (a) (Cette question n'est pas utilisée dans la suite) En considérant  $\varphi(X, \xi) = \nabla_\theta \log \frac{f_\theta(X \cup \{\xi\})}{f_\theta(X)}$  montrer que  $E_{\theta_*}[\nabla_\theta K(\theta, X)]_{\theta=\theta_*} = 0$ . On ne détaillera pas les questions d'intégrabilité.
- (b) Montrer que pour tout  $\theta$  on a

$$E_{\theta_*}[K(\theta_*, X)] \leq E_{\theta_*}[K(\theta, X)].$$

On utilisera la formule (I.37) avec  $\varphi(X, \xi) = \log \frac{f_\theta(X \cup \{\xi\})}{f_\theta(X)}$ , on posera  $u_\theta(X, \xi) = \frac{f_\theta(X \cup \{\xi\})f_{\theta_*}(X)}{f_\theta(X)f_{\theta_*}(X \cup \{\xi\})}$ , et l'on utilisera l'inégalité  $\log u \leq u - 1$ .

Montrer que  $E_{\theta_*}[K(\theta_*, X)] = E_{\theta_*}[K(\theta, X)]$  si et seulement si  $(P_{\theta_*} \times \mu)(\{u_\theta(X, \xi) = 1\}) = 1$ .

**Exercice 13 (Algorithme EM pour les variables partiellement observées).** Soit un couple de variables aléatoires  $(X, Y)$  de densité  $p_{\theta_*}(x, y)$  par rapport à une mesure de référence produit notée par simplicité  $dxdy$ . On est dans le cadre paramétrique habituel. On observe seulement  $X$ ; l'estimateur du maximum de vraisemblance  $\hat{\theta}$  satisfait donc

$$\int p_\theta(X, y)dy \leq \int p_{\hat{\theta}}(X, y)dy \tag{I.38}$$

pour tout  $\theta$ . Malheureusement cette maximisation est difficile à réaliser en pratique.

- (i) Montrer en utilisant l'inégalité de Jensen que pour tout  $\theta$

$$\int \log p_\theta(X, y)p_{\hat{\theta}}(X, y)dy \leq \int \log p_{\hat{\theta}}(X, y)p_{\hat{\theta}}(X, y)dy. \tag{I.39}$$

On pourra également vérifier que, sous des hypothèses de régularité, la condition d'ordre 1 pour maximiser le membre de gauche de (I.38) et de (I.39) est la même.

- (ii) Vérifier que (I.39) se réécrit

$$E_{\hat{\theta}}[\log p_\theta(X, Y)|X] \leq E_{\hat{\theta}}[\log p_{\hat{\theta}}(X, Y)|X]. \tag{I.40}$$

Montrer que ceci revient à dire que  $\hat{\theta}$  satisfait l'équation de point fixe

$$\hat{\theta} = F(\hat{\theta}), \quad F(\theta) = \arg \max_{\theta'} E_\theta[\log p_{\theta'}(X, Y)|X]. \tag{I.41}$$

- (iii) L'algorithme EM consiste à calculer la suite

$$\theta_{n+1} = F(\theta_n). \tag{I.42}$$

Il est difficile de donner un théorème précis et général concernant la convergence, qui en particulier peut avoir lieu vers un minimum local. Démontrer néanmoins que le membre de gauche de (I.38) augmente d'une étape à la suivante.

*Indication* : On vérifiera que (I.42) implique que (I.39) est satisfait avec  $\theta = \theta_n$  et  $\hat{\theta} = \theta_{n+1}$ , puis on utilisera l'inégalité de Jensen.

- (iv) Soit  $X = (X_1, \dots, X_n)$ , chaque  $X_i$  étant un mélange de deux (pour simplifier) gaussiennes  $\mathcal{N}(\mu_i, 1)$  choisies avec probabilité  $q$  et  $1 - q$ , avec  $\theta = (q, \mu_1, \mu_2)$ , et  $Y = (Y_1, \dots, Y_n) \in \{1, 2\}^n$  la variable cachée correspondante (cf. p. 8). Vérifier que (I.42) prend une forme simple, tandis que (I.38) est problématique à résoudre directement. Vérifier que cela reste vrai si l'on considère des lois de variance inconnue  $\mathcal{N}(\mu_i, \sigma_i^2)$ ,  $\theta = (q, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ .

L'expression de  $F$  explique le terme EM : Expectation-Maximization.

En pratique, l'espérance dans (I.41) sera calculée soit explicitement comme au point (iv), soit par simulation de réalisations de  $Y$  conditionnellement à  $X$  sous  $P_\theta$  (algorithme MCEM).

La convergence peut être très lente, et sensible à l'initialisation. Il existe des méthodes d'accélération, basées sur des solutions numériques de problèmes de point fixe, cf. [111] chap. 4, [72].

**Exercice 14 (Pseudo-vraisemblance. Vraisemblance partielle).** Soit des variables  $(X_i, Y_i)_{i \in I}$ . On suppose que la loi de  $X_i$  sachant  $Y_i$  ne dépend pas de  $i$ , et admet une densité  $p_\theta(x|y)$  par rapport à la mesure  $\mu(dx)$ . La loi de  $Y_i$  sera notée  $Q_{\theta_*}^i$ ; on pourra commencer par le cas où cette loi est indépendante de  $i$ . Soit l'estimateur

$$\hat{\theta} = \arg \max_{\theta} \sum_{i \in I} \log p_\theta(X_i|Y_i).$$

Dans le cas de  $n$  échantillons indépendants, il faut ajouter une somme sur  $n$ . Calculer la fonction de contraste et montrer qu'elle admet un maximum unique au vrai paramètre  $\theta_*$  à moins que pour un autre  $\theta$ , pour tout  $i$  et  $Q_{\theta_*}^i$ -presque tout  $y$  on ait  $p_\theta(x|y) = p_{\theta_*}(x|y)$   $\mu(dx)$ -p.p. (on calculera  $k(\theta_*) - k(\theta)$  et l'on utilisera l'inégalité de Jensen; pour une fonction strictement convexe, il n'y a égalité que pour une variable aléatoire p.s. constante).

EXEMPLE. Voici une situation simple où  $p_\theta(x|y)$  est facile à calculer mais pas  $p_\theta(x)$ : prenons  $I = \{1, 2, \dots, n\}$ ,  $X_i \in \{-1, 0, 1\}$ ,  $p_\theta(X_1, \dots, X_n) = Z_\theta \exp\{-\theta \sum_{i=1}^{n-1} (X_i - X_{i+1})^2\}$ , et  $Y_i = (X_{i-1}, X_{i+1})$ . La loi conditionnelle  $p_\theta(x|y)$  sous  $P_\theta$  est facile à calculer, alors qu'il n'existe pas de formule simple pour  $p_\theta((X_i)_{i \in I})$  car la normalisation  $Z_\theta$  est incalculable.

**Exercice 15 (Stabilisation de variance).** Soit  $\hat{\theta}_n$  une suite d'estimateurs de  $\theta_* \in \mathbb{R}$  tels que  $\sqrt{n}(\hat{\theta}_n - \theta_*)$  converge en loi vers la gaussienne  $\mathcal{N}(0, \sigma^2(\theta_*))$ . On dit qu'une transformation  $h$  stabilise la variance si  $\sqrt{n}(h(\hat{\theta}_n) - h(\theta_*))$  converge en loi vers  $\mathcal{N}(0, 1)$ .

1. Trouver une relation entre  $\sigma$  et  $h$ . On supposera  $h$  continuellement dérivable en  $\theta_*$ .
2. Vérifier que si  $\hat{\theta}_n = n^{-1} \sum_{i=1}^n X_i$ , où les  $X_i$  sont des  $\mathcal{B}(1, \theta)$  iid, alors  $h(\theta) = 2 \arcsin \sqrt{\theta}$  convient. Que vaut  $h$  si les  $X_i$  sont des Poisson i.i.d?

**Exercice 16 (Jackknife).** Soit  $X_n$  une suite iid et  $\hat{\theta}_n(X_1, \dots, X_n)$  une suite d'estimateurs d'une certaine quantité. On considère

$$\check{\theta}_n = n\hat{\theta}_n - \frac{n-1}{n} \sum_{i=1}^n \hat{\theta}_n^{(i)} \quad \text{où} \quad \hat{\theta}_n^{(i)} = \hat{\theta}_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$$

On suppose que le biais de  $\hat{\theta}_n$  est de la forme  $\sum_{k \geq 1} b_k n^{-k}$ . Montrer que le biais de  $\check{\theta}_n$  a un développement analogue sans terme en  $n^{-1}$ .

La comparaison de la variance des deux estimateurs, terme généralement prépondérant dans l'erreur, est plus compliquée, mais très importante.

On vérifie facilement que si  $\hat{\theta}_n$  est l'estimateur de la variance  $n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$  alors  $\check{\theta}_n$  est l'estimateur non-biaisé de l'exercice 4 p. 24.

# II

---

## ESTIMATION SEMI-PARAMÉTRIQUE

---

---

### II.1 Loi des grands nombres uniforme et autres outils

Le but de cette partie est d'introduire les différents outils mathématiques dont la combinaison permettra d'obtenir les résultats de convergence des estimateurs. Au vu du théorème 2, il faudra montrer, si l'on prend le cas de l'estimateur des moments, que les fonctions  $H_n(\theta, \omega) = n^{-1} \sum_{i=1}^n H(\theta_n, Y_i)$  convergent uniformément vers  $h(\theta)$ . Comme le théorème 3 ne s'applique que rarement, ceci sera fait à l'aide de la loi des grands nombres uniforme dont nous donnons ici une version générale dans le cadre qui nous intéresse. Cette version généralise au cas dépendant un théorème classique consacré aux suites de variables indépendantes (p.ex. § 19.8 de [21] ou Th.3.1(vi) de [3]), avec une condition de régularité affaiblie. Mentionnons enfin qu'on verra au § II.6 que l'inégalité de Sobolev (II.58) fournit une approche assez abordable pour traiter cette question d'uniformité d'une façon plus quantifiée ; elle conduira en particulier au théorème 26.

Une autre approche basée sur l'entropie métrique et la dimension de Vapnik-Chervonenkis est plus sophistiquée mais très fructueuse ; on en trouvera une présentation simple et opérationnelle dans [17], qui peut introduire à des exposés plus pointus comme [8] chap.3, [21] chap.19, ou [22] ; elle est essentiellement restreinte aux suites indépendantes. Nous ne l'aborderons pas.

On considérera le jeu d'hypothèses suivant :

**(GNU)**  $(\Theta, d)$  est un espace métrique compact.  $E$  est un espace vectoriel normé séparable<sup>1</sup>.  $H(\theta, y) : \Theta \times \mathbb{R}^m \rightarrow E$  est une fonction telle que pour tout  $\theta, y \mapsto H(\theta, y)$  soit mesurable.  $(Y_i)_{i \geq 1}$  est une suite de variables aléatoires. Ils satisfont

(1) Il existe une variable aléatoire  $Y$  telle que pour tout  $\theta_0 \in \Theta$

$$\frac{1}{n} \sum_{i=1}^n H(\theta_0, Y_i) \longrightarrow E[H(\theta_0, Y)] \quad p.s. \quad (\text{II.1})$$

De plus

$$E \left[ \sup_{\theta \in \Theta} \|H(\theta, Y)\| \right] < \infty \quad (\text{II.2})$$

et pour toute boule  $B$  de centre  $\theta_0$ , la fonction  $\psi(y) = \sup_{\theta \in B} \|H(\theta, y) - H(\theta_0, y)\|$  satisfait

$$\frac{1}{n} \sum_{i=1}^n \psi(Y_i) \longrightarrow E[\psi(Y)] \quad p.s. \quad (\text{II.3})$$

La mesurabilité des sup fait partie des hypothèses.

---

1. L'espace  $E$  contient un ensemble dénombrable dense. Si  $E$  est séparable, toute fonction mesurable  $f$  à valeurs dans  $E$  est limite de fonctions étagées, et si sa norme est dans  $L_1$  elle a une espérance correctement définie dans  $E$ , avec pour toute forme linéaire  $\lambda$ ,  $\lambda(f) = \int \lambda(f) d\mu$ . Le cas non séparable est plus compliqué. Voir [68], ou encore [129].

(2) Pour tout  $\theta_0 \in \Theta$ , presque sûrement, la fonction  $\theta \mapsto H(\theta, Y)$  est continue au point  $\theta_0$  (cet ensemble de probabilité 1 peut dépendre de  $\theta_0$ ).

L'hypothèse (GNU1) concerne essentiellement l'applicabilité de la loi des grands nombres; elle est bien entendu satisfaite (sous (II.2)) par les suites iid, mais aussi par les chaînes de Markov Harris récurrentes [5]. L'hypothèse (GNU2) accepte certaines situations où la fonction  $H$  est discontinue en  $\theta$ , par exemple si  $\Theta = [a, b]$ ,  $Y$  a une loi diffuse et  $H(\theta, Y) = 1_{Y \leq \theta} f(\theta, Y)$  où  $f$  est continue en  $\theta$ ; Wald semble être le premier, en 1949, à avoir souligné ce point dans un cadre général [138].

#### 4 - THÉORÈME (LOI DES GRANDS NOMBRES UNIFORME)

On se place sous l'hypothèse (GNU). Alors la fonction

$$h(\theta) = E[H(\theta, Y)]$$

est continue et l'on a avec probabilité 1

$$\limsup_n \sup_{\theta \in \Theta} \left\| h(\theta) - \frac{1}{n} \sum_{i=1}^n H(\theta, Y_i) \right\| = 0. \quad (\text{II.4})$$

*Démonstration.* L'idée est de se ramener par compacité au cas trivial où  $\Theta$  est fini. Pour tout  $\theta_0 \in \Theta$ , considérons la fonction

$$f_{\theta_0}(\eta) = E \left[ \sup_{d(\theta, \theta_0) < \eta} \|H(\theta, Y) - H(\theta_0, Y)\| \right].$$

Alors  $f_{\theta_0}(\eta)$  tend vers 0 quand  $\eta$  tend vers 0, en raison du théorème de Lebesgue. Ceci implique en particulier la continuité de  $h(\theta)$  car bien entendu

$$\sup_{d(\theta, \theta_0) < \eta} \|h(\theta) - h(\theta_0)\| = \sup_{d(\theta, \theta_0) < \eta} \|E[H(\theta, Y) - H(\theta_0, Y)]\| \leq f_{\theta_0}(\eta).$$

Fixons  $\varepsilon > 0$ . Pour tout  $\theta_0$ , il existe  $\eta(\theta_0) > 0$  tel que  $f_{\theta_0}(\eta(\theta_0)) < \varepsilon$ . Les boules ouvertes de centre  $\theta$  et de rayon  $\eta(\theta)$  forment un recouvrement d'ouverts de  $\Theta$  dont on peut extraire par compacité un sous-recouvrement fini :

$$\Theta = \cup_{j=1}^J B_j, \quad B_j = \{\theta : d(\theta, \theta_j) < \eta(\theta_j)\}.$$

On écrit alors pour  $\theta \in \Theta$ , et  $j = j(\theta)$  le plus petit indice tel que  $\theta \in B_j$  :

$$\frac{1}{n} \sum_{i=1}^n H(\theta, Y_i) - h(\theta) = \frac{1}{n} \sum_{i=1}^n H(\theta, Y_i) - H(\theta_j, Y_i) + \frac{1}{n} \sum_{i=1}^n H(\theta_j, Y_i) - h(\theta_j) + (h(\theta_j) - h(\theta)).$$

Ces trois termes sont des fonctions de  $\theta$  dont il s'agit de majorer la norme uniforme, sans oublier que  $j$  dépend de  $\theta$ . Celle du troisième est inférieure à  $\varepsilon$ ; celle du second tend vers 0 quand  $n$  tend vers l'infini (car  $J$  est fini); seul le premier terme pose difficulté. Sa norme uniforme est inférieure à :

$$\varphi_n = \sup_j \frac{1}{n} \sum_{i=1}^n \sup_{\theta \in B_j} \|H(\theta, Y_i) - H(\theta_j, Y_i)\|$$

si bien qu'avec probabilité 1, en vertu de (II.3)

$$\lim_n \varphi_n = \sup_j f_{\theta_j}(\eta) \leq \varepsilon.$$

On a montré que le membre de gauche de (II.4) est asymptotiquement inférieur à  $2\varepsilon$ ; comme  $\varepsilon$  est arbitraire, il est nul. ■

La proposition suivante est une conséquence utile de la loi des grands nombres uniforme :

5 - PROPOSITION

Soit  $\theta_n$  une suite de variables aléatoires à valeurs dans  $\Theta$  (non-nécessairement compact) convergeant presque sûrement vers une limite déterministe  $\theta_*$ . Si (GNU) est satisfaite sur un voisinage compact  $\Theta_*$  de  $\theta_*$ , alors presque sûrement :

$$\lim_n \frac{1}{n} \sum_{i=1}^n H(\theta_n, Y_i) = h(\theta_*).$$

*Démonstration.* Soit  $\theta'_n$  qui vaut  $\theta_n$  si  $\theta_n \in \Theta_*$  et  $\theta_*$  sinon, alors

$$\frac{1}{n} \sum_{i=1}^n H(\theta_n, Y_i) = \frac{1}{n} \sum_{i=1}^n \left( H(\theta_n, Y_i) - H(\theta'_n, Y_i) \right) + \frac{1}{n} \sum_{i=1}^n \left( H(\theta'_n, Y_i) - h(\theta'_n) \right) + h(\theta'_n).$$

Le premier terme est nul pour  $n$  assez grand, le deuxième tend vers 0 en raison de la loi des grands nombres uniforme, et le dernier converge vers  $h(\theta_*)$ . ■

Les Z-estimateurs ont la forme  $M_n(\hat{\theta}_n) = 0$  avec  $M_n(\theta) = \frac{1}{n} \sum_{i=1}^n H(\theta, Y_i)$ . Noter que l'on a en général plus ou moins facilement la normalité asymptotique de  $\sqrt{n}M_n(\theta_*)$  comme conséquence d'une version du théorème-limite central. C'est alors l'équation  $M_n(\hat{\theta}_n) - M_n(\theta_*) \sim \nabla M_n(\theta_*)(\hat{\theta}_n - \theta_*)$  qui va permettre de déduire la normalité asymptotique de  $\hat{\theta}_n - \theta_*$ , car  $\nabla M_n(\theta_*)$  converge en vertu de la loi des grands nombres (et  $M_n(\hat{\theta}_n) = 0$ ). Le **score**  $M_n(\theta_*)$  fournit donc une mesure de l'erreur d'estimation, à une normalisation près. Le résultat suivant formalise cette discussion avec un jeu d'hypothèses adéquat :

6 - THÉORÈME (SCORE ET ERREUR D'ESTIMATION)

Soit  $M_n(\theta, \omega)$  des fonctions mesurables de  $\mathbb{R}^d \times \Omega$  dans  $\mathbb{R}^d$ . Soit  $\theta_* \in \mathbb{R}^d$  possédant un voisinage  $\{\theta : \|\theta - \theta_*\| < \varepsilon\}$  sur lequel pour tout  $n \geq 1$  et presque tout  $\omega$  l'application  $\theta \mapsto M_n(\theta, \omega)$  est de classe  $C^1$ . Soit  $\theta_n(\omega)$  une suite de variables aléatoires à valeurs dans  $\mathbb{R}^d$ . On suppose que

$$\begin{aligned} \theta_n &\xrightarrow{P} \theta_* \\ \sqrt{n} M_n(\theta_n) &\xrightarrow{P} 0 \\ \sqrt{n} M_n(\theta_*) &\xrightarrow{d} \mathcal{D}, \quad \text{pour une certaine distribution } \mathcal{D} \\ \sup_{\|\theta - \theta_*\| \leq \varepsilon} \|\nabla M_n(\theta) - J(\theta)\| &\xrightarrow{P} 0 \end{aligned}$$

pour une certaine fonction  $\theta \mapsto J(\theta)$  continue au point  $\theta_*$ . Alors, si  $J_* = J(\theta_*)$  est inversible,

$$\sqrt{n}(\theta_n - \theta_*) + \sqrt{n}J_*^{-1}M_n(\theta_*) \xrightarrow{P} 0. \quad (\text{II.5})$$

*Démonstration.* Le théorème des accroissements finis donne pour un  $\theta'_n$  situé sur le segment  $[\theta_n, \theta_*]$  liant  $\theta_n$  à  $\theta_*$  :

$$M_n(\theta_n) = M_n(\theta_*) + \nabla M_n(\theta'_n)(\theta_n - \theta_*) = M_n(\theta_*) + (J_* + r_n)(\theta_n - \theta_*) \quad (\text{II.6})$$

où  $r_n = \nabla M_n(\theta'_n) - J_*$  tend en probabilité vers 0; en réalité, comme  $M_n$  est vectoriel, on doit appliquer le théorème coordonnée par coordonnée et le vecteur  $\theta'_n$  est différent d'une ligne à l'autre de  $\nabla M_n(\theta'_n)$ . L'équation (II.6) équivaut à

$$\theta_n - \theta_* + J_*^{-1}M_n(\theta_*) = (J_* + r_n)^{-1}M_n(\theta_n) + (J_*^{-1} - (J_* + r_n)^{-1})M_n(\theta_*).$$

Le produit par  $\sqrt{n}$  des deux termes de gauche tend vers 0. ■

Notons que ce théorème conduit directement à la normalité asymptotique de l'estimateur au maximum de vraisemblance (dans le cas d'observations indépendantes), la partie difficile pouvant être de montrer la convergence de  $\theta_n$  et la convergence uniforme en probabilité de la dernière hypothèse ; nous y reviendrons.

## II.2 Fonctions d'estimation (Z-estimateurs)

On considère l'estimateur solution de :

$$\sum_{i=1}^n H(\hat{\theta}_n, Y_i) = 0. \quad (\text{II.7})$$

On va montrer que sous des hypothèses raisonnables,  $\hat{\theta}_n$  converge p.s vers  $\theta_*$ , avec normalité asymptotique de  $\sqrt{n}(\hat{\theta}_n - \theta_*)$ . On verra le rôle important joué par la matrice de bruit  $I = \text{Var}(\sum_{i=1}^n H(\theta_*, Y_i))/n$ , qui sera définie de façon un peu différente, et par la matrice de sensibilité  $J = E[\nabla H(\theta_*, Y)]$ .

Il existe de nombreux exemples de fonctions d'estimation (et d'estimateurs à minimum de contraste) dans des cas de chaînes de Markov ou de processus ponctuels dépendant d'un paramètre [28, 81], situations où le maximum de vraisemblance est souvent très difficile à mettre en œuvre (cf. § I.3.2 et les exercices 7 p. 25, 10 p. 25, 14 p. 28). Ces estimateurs sont également appelés Z-estimateurs.

### II.2.1 Exemple : la méthode de la variable instrumentale

Présentons une situation typique de fonction d'estimation qui n'est pas le gradient d'une fonction de contraste. Soit le processus

$$y_n = \theta_* y_{n-1} + e_n = \sum_{k \geq 0} \theta_*^k e_{n-k} \quad (\text{II.8})$$

où  $|\theta_*| < 1$  et le bruit  $e_n$  est supposé stationnaire centré, de variance finie et seulement  $q$ -dépendant, c.-à-d. que  $e_{n+q}$  est indépendant de  $(e_n, e_{n-1}, \dots)$ . C'est par exemple le cas, si  $e_n$  est un processus MA<sup>2</sup>. Dans ce cas l'estimateur des moindres carrés

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \theta y_{i-1})^2 = \frac{\hat{r}_1}{\hat{r}_0}$$

n'est généralement pas consistant ( $\hat{r}_i$  désigne la  $i$ -ième corrélation empirique), car (II.8) implique que  $r_1 = \theta_* r_0 + E[y_{n-1} e_n] \neq \theta_* r_0$ . En multipliant l'équation (II.8) par  $y_{n-q}$  et en prenant l'espérance, il vient

$$E[y_n y_{n-q}] = \theta_* E[y_{n-1} y_{n-q}]$$

en raison de la  $q$ -dépendance, ou même, plus simplement, de la  $q$ -décorrélation. On peut donc estimer  $\theta_*$  par  $\hat{r}_q / \hat{r}_{q-1}$ . C'est le Z-estimateur basé sur la fonction

$$H(\theta, Y_k) = (y_k - \theta y_{k-1}) y_{k-q}, \quad Y_k = (y_k, \dots, y_{k-q}). \quad (\text{II.9})$$

Comme  $E[H(\theta, Y_k)] = r_q - \theta r_{q-1}$ ,  $\theta_*$  est la seule valeur qui annulera l'espérance à moins que  $r_q$  et  $r_{q-1}$  soient nuls.

On généralise facilement au cas où  $y$  a une partie autorégressive plus compliquée,

$$y_n = \sum_{i=1}^p \theta_i y_{n-i} + e_n \quad (\text{II.10})$$

---

2. On a considéré dans ce paragraphe un exemple plus général que l'ARMA, car dans de nombreuses applications,  $y_k$  est la réponse d'un système physique à une excitation  $e_k$ , p.ex. les vibrations d'un pont sur lequel roulent des voitures. La partie MA est associée à la nature de l'excitation, tandis que la partie AR représente la façon dont le système répond ; c'est seulement cette dernière qui donne des informations sur son état réel. Il est alors naturel de remplacer l'hypothèse MA par l'hypothèse plus générale de  $q$ -dépendance,  $q$  étant dans notre exemple le temps que met un véhicule à traverser le pont multiplié par la fréquence d'échantillonnage.

en considérant la fonction vectorielle  $H(\theta, Y_k)$  dont les coordonnées sont les

$$H(\theta, Y_k)_j = \left( y_k - \sum_i \theta_i y_{k-i} \right) y_{k-q+1-j}, \quad j = 1, \dots, p \quad (\text{II.11})$$

$$= e_k(\theta) t_{k-q,j}. \quad (\text{II.12})$$

La dernière écriture suggère la philosophie générale de la méthode instrumentale : les  $e_k(\theta)$  sont les erreurs de prédiction sous l'hypothèse  $\theta$ , et les  $t_{kj}$  sont les instruments,  $p$  variables du passé avant  $k$ ; on a bien par  $q$ -dépendance que  $e_k(\theta_*) \perp t_{k-q,j}$  puisque  $t_{k-q,j}$  ne dépend que du passé avant  $k - q$ .

## II.2.2 Convergence presque sûre

### 7 - THÉORÈME

On se place sous (GNU) et l'on pose  $h(\theta) = E[H(\theta, Y)]$ . Toute suite d'estimateurs  $\hat{\theta}_n \in \Theta$  telle que

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(\hat{\theta}_n, Y_i) = 0, \quad p.s. \quad (\text{II.13})$$

satisfait p.s.

$$\lim_n h(\hat{\theta}_n) = 0.$$

Si de plus  $\theta_*$  est la seule solution de  $h(\theta) = 0$ , alors  $\hat{\theta}_n$  tend p.s. vers  $\theta_*$ .

*Démonstration.* Il s'agit d'une conséquence immédiate des théorèmes 4 et 2. ■

**Le cas non-compact.** On s'est placé sous l'hypothèse de compacité de  $\Theta$ . Cette hypothèse est restrictive mais a le mérite de la simplicité. Si  $\Theta$  n'est pas compact on a besoin d'une autre version du théorème où l'on se place sous l'hypothèse suivante, qui fait malheureusement intervenir une propriété des estimateurs, mais qui est plus générale que (GNU) :

**(GNU')** *Il existe une suite croissante de compacts  $\Theta_p$  dont la réunion des intérieurs fait  $\Theta$ , et sur chacun desquels est satisfait (GNU).*

*De plus la suite  $\hat{\theta}_n$  reste presque sûrement dans un compact (pouvant dépendre de  $\omega$ ).*

Cette hypothèse se vérifiera facilement si par exemple  $\|\hat{\theta}_n\|$  peut être borné par une expression faisant intervenir des moyennes empiriques (c'est particulièrement immédiat dans le cas de l'estimateur du paramètre d'une famille exponentielle, formule (I.5)).

### 8 - THÉORÈME

Le théorème 7 reste vrai si l'on remplace (GNU) par (GNU').

*Démonstration.* Notons d'abord que comme les intérieurs des  $\Theta_p$  recouvrent le compact où la suite  $\hat{\theta}_n$  reste confinée, il existe un  $p(\omega)$  tel que  $\Theta_{p(\omega)}$  contienne toute la suite. Considérons pour tout  $p$  l'estimateur  $\hat{\theta}_n^{(p)}$  qui vaut  $\hat{\theta}_n$  si  $\hat{\theta}_n \in \Theta_p$  et sinon  $\theta_*$ . Le théorème 7 implique la convergence presque sûre de  $\hat{\theta}_n^{(p)}$  vers  $\theta_*$ . On a donc convergence de  $\hat{\theta}_n$  vers  $\theta_*$  sur l'ensemble  $A_p = \{\omega : \forall n, \hat{\theta}_n \in \Theta_p\}$ . Comme  $P(\cup_p A_p) = 1$  la convergence a lieu avec probabilité 1. ■

### II.2.3 Normalité asymptotique

(FE)  $H(\theta, y)$  est une fonction  $\mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}^d$ . Il existe un voisinage  $\Theta_0$  de  $\theta_*$ , sur lequel, p.s. pour tout  $i$ , la fonction  $\theta \mapsto H(\theta, Y_i)$  est de classe  $C^1$ . Le triplet  $(\Theta_0, (Y_i)_{i \geq 0}, \nabla H(\theta, y))$  satisfait (GNU).

Par commodité on a supposé  $H$  définie pour  $\theta \in \mathbb{R}^d$ , alors que pour notre problème, la fonction  $H$  n'est définie que sur  $\Theta$ ; il suffit, pour appliquer le théorème qui suit, de lui donner la valeur 0 ailleurs, car en réalité, seul  $\Theta_0$  compte désormais.

#### 9 - THÉORÈME

On se place sous (FE). On suppose de plus que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n H(\theta_*, Y_i) \xrightarrow{d} \mathcal{N}(0, I) \quad (\text{II.14})$$

pour une certaine matrice  $I$ , et que la matrice de sensibilité  $J = E[\nabla H(\theta_*, Y)]$  est inversible. Alors toute suite  $\hat{\theta}_n$  convergeant en probabilité vers  $\theta_*$  et telle que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n H(\hat{\theta}_n, Y_i) \xrightarrow{P} 0$$

satisfait :

$$\sqrt{n} \left( \theta_n - \theta_* - \frac{1}{n} J^{-1} \sum_{i=1}^n H(\theta_*, Y_i) \right) \xrightarrow{P} 0 \quad (\text{II.15})$$

$$\sqrt{n} (\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, J^{-1} I J^{-T}). \quad (\text{II.16})$$

*Remarques.* 1/ Par analogie avec la théorie du maximum de vraisemblance, l'inverse de la matrice  $J^{-1} I J^{-T}$  est parfois appelé matrice d'information de Godambe (Godambe information matrix).

2/ La variance de  $\hat{\theta}_n$  sera donc faible si  $I$  est petit et  $J$  grand : peu de bruit et forte sensibilité.

3/ De manière générale, si  $\theta_n - \theta_* - \sum_{i=1}^n u(Y_i) = o_P(n^{-1/2})$  pour une certaine fonction  $u$ , on dit que l'estimateur  $\hat{\theta}_n$  est asymptotiquement linéaire.

*Démonstration.* C'est une application directe du théorème 6 avec  $M_n(\theta) = n^{-1} \sum_{i=1}^n H(\theta, Y_i)$ , où la quatrième hypothèse est une conséquence de la loi des grands nombres uniforme appliquée à  $n^{-1} \sum \nabla H(\theta, Y_i)$  pour  $\theta$  dans un voisinage compact de  $\theta_*$ . ■

**Régions de confiance.** Si l'on dispose d'estimateurs  $\hat{I}_n \xrightarrow{P} I$  (p. ex. formule (II.23) plus bas ; dans le cas dépendant, il peut être difficile de trouver un  $\hat{I}_n$  satisfaisant) alors l'hypothèse (II.14) implique que

$$\mathcal{R}_\alpha = \left\{ \theta : \left\| \hat{I}_n^{-1/2} \sum_{i=1}^n H(\theta, Y_i) \right\|^2 \leq n \chi_d^2(1 - \alpha) \right\} \quad (\text{II.17})$$

est une région de confiance pour  $\theta_*$  de risque asymptotique  $\alpha$  (i.e. son risque réel  $\alpha_n$  tend vers  $\alpha$ ). Rappelons que  $\chi_d^2(1 - \alpha) \simeq d - 2 \log \alpha$  [92]. Comme de plus sous les hypothèses du théorème  $\hat{J}_n = n^{-1} \sum \nabla H(\hat{\theta}_n, Y_i) \xrightarrow{P} J$  on a également la région

$$\mathcal{R}_\alpha = \left\{ \theta : n(\hat{\theta}_n - \theta)^T \hat{V}^{-1} (\hat{\theta}_n - \theta) \leq \chi_d^2(1 - \alpha) \right\}, \quad \hat{V} = \hat{J}_n^{-1} \hat{I}_n \hat{J}_n^{-T}. \quad (\text{II.18})$$

On a également que la région (intervalles de confiances simultanés)

$$\mathcal{R} = \left\{ \theta : \forall k, \sqrt{n} |(\hat{\theta}_n - \theta)_k| \leq \hat{\sigma}_k \sqrt{\chi_1^2(1 - \alpha_k)} \right\}, \quad \hat{\sigma}_k^2 = \hat{V}_{kk}$$

est de risque asymptotique inférieur à  $1 - \prod_k (1 - \alpha_k)$ . Pour démontrer ceci, notons que pour tout vecteur gaussien  $X \sim \mathcal{N}(0, V)$ , on a en vertu de l'inégalité de Khatri (1967)<sup>3</sup> :

$$P\left(\forall k, X_k^2 \leq \sigma_k^2 \chi_1^2(1 - \alpha_k)\right) \geq \prod_k P\left(X_k^2 \leq \sigma_k^2 \chi_1^2(1 - \alpha_k)\right) = \prod_k (1 - \alpha_k).$$

Le membre de gauche étant la limite de  $P(\theta_* \in \mathcal{R})$ , ceci prouve bien le niveau de confiance asymptotique. Si  $\varepsilon$  est le niveau attendu et que tous les  $\alpha_k$  sont égaux, on a  $\alpha_k \simeq \varepsilon/d$ , et on trouve  $\chi_1^2(1 - \alpha_k) \simeq \log(d/\varepsilon)$ .

**Estimation pratique de  $I$ .** L'estimation de  $J$  ne posera généralement pas de problème, l'estimateur empirique étant raisonnable. Il en est tout autrement pour  $I$ . Si la suite  $Y_i$  est i.i.d, on aura  $I = E[H(\theta_*, Y_1)H(\theta_*, Y_1)^T]$ . En revanche, si la suite  $Y_i$  n'est pas indépendante, la situation est plus compliquée. Commençons par l'exemple du § II.2.1 où les  $H(\theta_*, Y_1)$  sont  $q$ -dépendantes.

ESTIMATION DE  $I$  DANS LE CAS ARMA. Considérons le processus (II.10) du § II.2.1, qui est un peu plus général que l'ARMA( $p, q$ ), avec l'équation d'estimation (II.11). Pour simplifier les écritures on se restreint à (II.8, II.9), i.e.  $p = 1$ . Si l'on suppose la suite  $Y_k$  stationnaire, le calcul de  $J$  ne pose pas problème

$$J = -E[y_{k-1}y_{k-q}] = -E[y_1y_q]$$

qui peut s'estimer facilement sur les données. Concernant  $I$ , notons que par stationnarité

$$\begin{aligned} \frac{1}{n} E\left[\left(\sum_{i=1}^n H(\theta_*, Y_i)\right)^2\right] &= \frac{2}{n} E\left[\sum_{i < j} H(\theta_*, Y_i)H(\theta_*, Y_j)\right] + \frac{1}{n} E\left[\sum_{i=1}^n H(\theta_*, Y_i)^2\right] \\ &= \frac{2}{n} E\left[\sum_{i=1}^n (n-i)H(\theta_*, Y_0)H(\theta_*, Y_i)\right] + E[H(\theta_*, Y_0)^2]. \end{aligned} \quad (\text{II.19})$$

Comme  $H(\theta_*, Y_0)H(\theta_*, Y_i) = e_0 e_i y_{-q} y_{i-q}$ , son espérance est nulle pour  $i \geq q$  (noter que la  $q$ -décorrélation est ici insuffisante, puisqu'on considère des moments d'ordre 4), et sinon vaut  $E[e_q e_{q+i} y_0 y_i]$  (stationnarité), et en faisant tendre  $n$  vers l'infini on obtient, en supposant que le membre de gauche de (II.19) converge bien vers  $I$ ,

$$I = 2 \sum_{i=1}^q E[e_q e_{q+i} y_0 y_i] + E[e_q^2 y_0^2]. \quad (\text{II.20})$$

Tout comme  $J$ , cette quantité peut facilement s'estimer sur les données tant que  $q$  est connu et pas trop grand.

ESTIMATION DE  $I$  DANS LE CAS GÉNÉRAL. Supposons la suite  $Y_i$  stationnaire. En reprenant l'équation (II.19) on voit que, sous des hypothèses de mélange raisonnables :

$$I = R(0) + \sum_{i>0} R(i) + R(i)^T, \quad R(i) = \text{Cov}(H_1, H_{i+1}), \quad H_i = H(\theta_*, Y_i). \quad (\text{II.21})$$

L'estimation de  $I$  pose problème car il y a une infinité de  $R(i)$  à estimer. Si l'on met les estimées empiriques dans la formule (II.21) l'estimateur obtenu n'est pas consistant. Il faut prendre un estimateur fenêtré [73], ce qui n'est pas très facile car il y a un paramètre sensible à régler pour équilibrer biais et variance.

Une autre stratégie consiste à utiliser quand même un estimateur non-consistant, comme par exemple  $\tilde{I} = n^{-2} \sum S^2$ , où la somme contient les  $2n - 1$  sommes  $S$  de la forme  $\sum_{i=1}^k H_i$  ou bien  $\sum_{i=k}^n H_i$ , mais à démontrer que la statistique  $(n\check{n})^{-1/2} \sum_{i=1}^n H_i$  est néanmoins asymptotiquement pivotale, ce qui est parfois le cas [98, 26]. Sa loi peut alors s'estimer par simulation, ce qui permet d'utiliser cette statistique pour des tests.

3. Pour tout vecteur gaussien centré  $(Y_1, \dots, Y_n)$  et des  $c_k$  réels,  $P(|Y_k| \leq c_k, k = 1 \dots n) \geq \prod_k P(|Y_k| \leq c_k)$ . C'est une conséquence de l'inégalité de Royen (<http://arxiv.org/abs/1512.08776>) : Si  $E$  et  $F$  sont convexes symétriques,  $P(Y \in E \cap F) \geq P(Y \in E)P(Y \in F)$ .

## II.2.4 Fusion optimale de Z-estimateurs

Il se peut que naturellement on ait  $\dim(H) > \dim(\theta)$ , en particulier si l'on dispose de nombreuses variables instrumentales ; par exemple si l'on définit  $H$  comme dans (II.11) mais pour  $j$  allant de 1 à  $p'$ ,  $p' > p$ . Dans ce cas l'équation  $\sum_i H(\theta, Y_i) = 0$  sera génériquement sans solution.

Autre exemple : Soit  $N_1, N_2, \dots, N_m$  les effectifs des modalités d'une variable au cours de  $n$  expériences indépendantes. Supposons que les probabilités d'apparition des modalités soient paramétrées par un vecteur  $\theta \in \Theta \subset \mathbb{R}^d$ ,  $d \leq m - 1$  ; on a donc  $m$  fonctions  $p_1(\theta), \dots, p_m(\theta)$  dont la somme fait 1. Par exemple le modèle de Hardy-Weinberg pour la fréquence de génotypes dans la population, qui pour  $m = 3$  considère  $(p_1, p_2, p_3) = (\theta^2, 2\theta(1 - \theta), (1 - \theta)^2)$ .

On dispose naturellement des fonctions d'estimation  $H_k(\theta, y) = p_k(\theta) - 1_{y=k}$ ,  $k = 1 \dots m$ . Ce qui conduit à résoudre simultanément

$$p_k(\hat{\theta}_n) = N_k/n, \quad k = 1, \dots, m - 1,$$

soit  $m - 1$  équations pour  $d$  inconnues.

Une méthode pour se ramener à autant d'équations que d'inconnues est de se donner une matrice  $K \in \mathbb{R}^{\dim(H) \times \dim(\theta)}$  et de résoudre

$$\sum_i K^T H(\theta, Y_i) = 0.$$

On vérifie aussitôt que la variance asymptotique de  $\hat{\theta}$  vaut alors

$$V_K = (K^T J)^{-1} K^T I K (K^T J)^{-T}$$

avec  $I$  et  $J$  définies au théorème 9. Noter que si  $K$  dépend de  $\theta$ , tout en étant de classe  $C^2$  au voisinage de  $\theta_*$ , la formule reste valide, à condition d'y mettre sa valeur pour  $\theta = \theta_*$ . On montre que, si  $I$  est inversible,  $V_K$  est toujours supérieure à  $(J^T I^{-1} J)^{-1}$  qui est sa valeur si  $K = I^{-1} J$ . L'inversibilité de  $I$  n'est en fait pas réellement nécessaire et l'on a plus généralement le résultat suivant que nous énonçons de façon informelle (cf. l'exercice 5 p. 37) :

*Si  $J^T$  est de rang plein, alors toute solution  $K$  de l'équation  $J = IK$  fournit une fonction d'estimation projetée  $K^T H(\theta, y)$  asymptotiquement optimale au sens où  $V_K$  est minimale.*

Noter que la nouvelle fonction d'estimation  $\tilde{H}(\theta, y) = K^T H(\theta, y)$  satisfait  $I = J$ . Cette méthode permet d'obtenir de bons estimateurs dans des cadres assez généraux de recherche d'estimateurs efficaces ([115] p. 28).

Une autre méthode de fusion consiste à combiner linéairement des estimateurs obtenus avec des fonctions d'estimation différentes. On voit à l'exercice 3 que le résultat est asymptotiquement le même que de prendre l'estimateur résultant de la combinaison des fonctions d'estimation.

Dans certains cas paramétriques,  $I$  et  $J$  s'expriment directement en fonction de  $\theta_*$ , auquel cas une estimée préliminaire  $\hat{\theta}$  de  $\theta_*$  conduira à une matrice  $K(\hat{\theta})$  convenable. Sinon il faut les estimer, éventuellement à l'aide d'une première estimée  $\hat{\theta}$  de  $\theta_*$ , par exemple avec les deux premières formules de l'encadré (II.23) dont la convergence se justifie avec les hypothèses (GNU). Si  $\hat{\theta}$  converge à vitesse  $n^{-\frac{1}{2}}$ , la seconde estimée  $\tilde{\theta}$  pourra se calculer en annulant  $\sum \tilde{H}(\theta, Y_i)$  ou par une méthode de Newton au premier ordre avec la formule suivante (par souci de clarté on se place dans le cas  $I$  inversible) :

$$\begin{aligned} \tilde{\theta} &= \hat{\theta} - \left( \sum_i \nabla \tilde{H}(\hat{\theta}, Y_i) \right)^{-1} \left( \sum_i \tilde{H}(\hat{\theta}, Y_i) \right), \quad \tilde{H}(\theta, y) = \hat{J}^T \hat{I}^{-1} H(\theta, y) \\ &= \hat{\theta} - (\hat{J}^T \hat{I}^{-1} \hat{J})^{-1} \left( \frac{1}{n} \sum_i \tilde{H}(\hat{\theta}, Y_i) \right) \end{aligned} \quad (\text{II.22})$$

si  $\hat{J}$  correspond à (II.23). Soit au final

On suppose les  $Y_i$  indépendantes.

Calcul de la nouvelle fonction d'estimation :

$$\begin{aligned}\hat{J} &= \frac{1}{n} \sum_{i=1}^n \nabla H(\hat{\theta}, Y_i) \\ \hat{I} &= \frac{1}{n} \sum_i H(\hat{\theta}, Y_i) H(\hat{\theta}, Y_i)^T - \frac{1}{n^2} \left( \sum_i H(\hat{\theta}, Y_i) \right) \left( \sum_i H(\hat{\theta}, Y_i) \right)^T \\ \tilde{H}(\theta, y) &= \hat{J}^T \hat{I}^{-1} H(\theta, y).\end{aligned}\tag{II.23}$$

Calcul de l'estimateur :

$$\begin{aligned}\tilde{\theta} &= \hat{\theta} - \left( \sum_i \tilde{H}(\hat{\theta}, Y_i) \tilde{H}(\hat{\theta}, Y_i)^T \right)^{-1} \sum_i \tilde{H}(\hat{\theta}, Y_i) \\ \text{ou bien } \sum_i \tilde{H}(\tilde{\theta}, Y_i) &= 0.\end{aligned}$$

Ceci conduira à une vitesse optimale. Si les  $Y_i$  sont dépendantes, il faut revoir l'estimation de  $I$  (cf. la fin du § II.2.3).

## II.2.5 Exercices et compléments

**Exercice 1.** On considère l'exercice 2 p. 24 pour lequel l'estimateur est  $\hat{\lambda} = n / \sum X_i$ ,  $\hat{\mu} = \sum Y_i / n$  et  $\hat{\alpha} = -2/\hat{\lambda} + \frac{1}{2}\hat{\lambda}^2 \sum (Y_i - \hat{\mu})^2 (X_i - \hat{\lambda}^{-1}) / n$ . Montrer qu'il s'agit d'un Z-estimateur et que les matrices  $I$  et  $J$  peuvent se calculer explicitement (on ne fera qu'initier ce calcul un peu laborieux).

**Exercice 2.** On considère un processus autorégressif d'ordre 1, ce qui nous place dans la situation du § II.2.1 avec  $q = 1$ , et  $H$  donné par (II.9). On suppose le processus stationnaire ergodique. Exprimer  $J$  en fonction de la variance de  $y_1$ . Vérifier que les variables  $H(\theta_*, Y_k)$  sont orthogonales et en déduire l'expression de  $I$  en fonction des variances de  $y_1$  et de  $\epsilon_1$ . En déduire une expression de la variance asymptotique de  $\hat{\theta}$  en fonction de  $\theta_*$  seulement.

**Exercice 3 (Fusion).** Soient  $H_1(\theta, y)$  et  $H_2(\theta, y)$  deux fonctions d'estimation satisfaisant les hypothèses du théorème 9. Soit  $\hat{\theta}_1$  et  $\hat{\theta}_2$  les deux estimées correspondantes. Soit  $0 < p < 1$  et  $\hat{\theta}_3$  l'estimée obtenue avec  $H_3 = pH_1 + (1-p)H_2$ . On suppose que  $H_3$  satisfait également (II.14). Montrer, en utilisant le théorème 6 que  $\sqrt{n}(\hat{\theta}_3 - p\hat{\theta}_1 - (1-p)\hat{\theta}_2)$  converge en probabilité vers 0.

**Exercice 4.** Soit le processus (chaîne de Markov)

$$Y_i = e^{-\theta_* Y_{i-1}} + u_i$$

où les  $u_i$  sont iid uniformes sur  $[-1, 1]$ . On suppose que  $Y_i$  est stationnaire ergodique. On sait que  $\theta_* \in [1, 2]$ . On se propose d'estimer  $\theta_*$  à partir de  $n$  premières observations par  $\hat{\theta}_n$  solution de l'équation

$$\sum_{i=1}^n (Y_i - e^{-\theta Y_{i-1}}) Y_{i-1} = 0$$

1. Démontrer la convergence presque sûre de  $\hat{\theta}_n$  vers  $\theta_*$ .
2. On admet la normalité asymptotique. Calculer la variance asymptotique de  $\hat{\theta}_n$  en fonction de  $E[Y_i^2]$  et  $E[Y_i Y_{i-1}^2]$ . Proposer un estimateur de cette variance.

**Exercice 5.** Il s'agit de démontrer la minoration de  $V_K$  du § II.2.4. On rappelle qu'une matrice symétrique  $S \in \mathbb{R}^{d \times d}$  est  $\geq 0$  si pour tout  $x \in \mathbb{R}^d$ ,  $x^T S x \geq 0$ , et que  $S \geq T$  si  $S - T \geq 0$ . On rappelle que  $I \in \mathbb{R}^{m \times m}$  et  $J \in \mathbb{R}^{m \times d}$  où  $m$  est la dimension de  $H$  et  $d$  celle de  $\theta$ ; on a donc  $d < m$ .

Il faut donc montrer que si  $K_0$  est solution de  $J = IK_0$ , et si  $J$  est de rang plein, on a  $V_{K_0} \leq V_K$  pour tout  $K$ , c'est-à-dire que

$$(K_0^T IK_0)^{-1} \leq (K^T IK_0)^{-1} K^T IK (K^T IK_0)^{-T}.\tag{II.24}$$

1. Montrer que si  $J = IK_0$  est de rang plein,  $K_0$  aussi et  $K_0^T IK_0$  est bien inversible.  
*Indication* : Vu les dimensions de  $J$ , dire qu'elle est de rang plein revient à dire que  $Jx = 0 \Rightarrow x = 0$ . Pour montrer l'inversibilité on vérifiera que l'équation  $x^T K_0^T IK_0 x = 0$  est sans solution non-nulle; utiliser une racine carrée de  $I$ .

2. Montrer que si  $S$  et  $T$  sont symétriques  $\geq 0$ , et  $M$  une matrice rectangulaire, alors  $S \leq T$  implique  $M^T S M \leq M^T T M$ , si les dimensions sont compatibles. En déduire que (II.24) se réduit à montrer :

$$IK_0(K_0^T IK_0)^{-1} K_0^T I \leq I. \quad (\text{II.25})$$

pour toutes matrices  $I$  carrée symétrique et  $K_0$  rectangulaire, du moment que l'inverse existe.

3. Montrer que pour toute matrice rectangulaire  $J$  on a si l'inverse existe  $J(J^T J)^{-1} J^T \leq Id$ .  
*Indication*. Une matrice  $P$  telle que  $P^T = P = P^2$  est une matrice de projection orthogonale; par conséquent pour tout  $x$ ,  $x^T P x \leq \|x\|^2$ .

4. En déduire (II.25). *Indication* : On utilisera une racine carrée symétrique de  $I$ .

**Exercice 6 (Estimateur de la médiane)**. On suppose que la fonction de répartition de  $Y$  est continue. Étudier l'estimateur donné par l'équation ( $\text{signe}(0) = 0$ ) :

$$\sum_{i=1}^n \text{signe}(\hat{\theta}_n - Y_i) = 0.$$

On pourra le faire en utilisant soit le théorème 4 soit le théorème 3.

**Exercice 7 (Maximum de vraisemblance empirique (MVE) <sup>4</sup>)**. On est dans le cas où  $\dim(H) > \dim(\theta)$ . On considère le vecteur  $(\hat{\theta}, \hat{p}_1, \dots, \hat{p}_n)$  solution du problème

$$\max \sum_{i=1}^n \log p_i \quad \text{sous} \quad \sum_{i=1}^n p_i H(\theta, Y_i) = 0, \quad \sum_{i=1}^n p_i = 1. \quad (\text{II.26})$$

On cherche donc à réaliser exactement les contraintes au prix d'une légère modification de la mesure empirique (sous la contrainte  $\sum_{i=1}^n p_i = 1$ , le premier terme est maximum pour  $p_i = 1/n$ ).

La première question a pour but de chercher les conditions du premier ordre pour le problème (II.26). On admettra ensuite que des hypothèses adéquates sont mises sur  $H$  pour que, avec probabilité 1, ces conditions soient nécessaires et suffisantes et conduisent à une unique solution  $\hat{\theta}$ .

1. Vérifier que les conditions du premier ordre pour (II.26) sont ( $\gamma$  est le vecteur des multiplicateurs de Lagrange)

$$\sum_i p_i H(\theta, Y_i) = 0, \quad p_i = \frac{1}{n + n\gamma^T H(\theta, Y_i)}, \quad \sum_i p_i \gamma^T \nabla H(\theta, Y_i) = 0.$$

Vérifier que les deux premières équations impliquent  $\sum p_i = 1$ .

2. Montrer qu'on peut mettre le maximum de vraisemblance empirique dans le cadre des Z-estimateurs à condition de considérer le paramètre  $(\theta, \gamma)$ . Que vaut  $\gamma_*$ ?
3. Soit

$$\hat{J}(\theta) = \sum_i \hat{p}_i \nabla H(\theta, Y_i), \quad \hat{I}(\theta) = \sum_i \hat{p}_i H(\theta, Y_i) H(\theta, Y_i)^T, \quad \bar{H}(\theta) = \frac{1}{n} \sum_i H(\theta, Y_i).$$

Vérifier que  $\hat{\theta}$  est solution de  $\hat{J}(\hat{\theta})^T \hat{I}(\hat{\theta})^{-1} \bar{H}(\hat{\theta}) = 0$  (multiplier l'équation  $\hat{p}_i + \hat{p}_i \gamma^T H(\hat{\theta}, Y_i) = n^{-1}$  par  $H(\hat{\theta}, Y_i)^T$  et sommer...); ce qui est assez inattendu. Le MVE peut donc être vu comme une façon un peu spéciale d'estimer  $I$  et  $J$  dans la méthode du § II.2.4.

4. INVARIANCE. Soit  $M(\theta)$  une matrice inversible. Que se passe-t-il d'insatisfaisant dans la méthode du § II.2.4 (i.e. la résolution de  $J(\theta_*) I(\theta_*)^{-1} \bar{H}(\theta) = 0$ ) si l'on remplace la fonction  $H(\theta, y)$  par  $M(\theta) H(\theta, y)$ , qui se produit pas pour le MVE?
5. FAIBLESSE DE LA MÉTHODE. Il a été observé expérimentalement que dans des situations où il n'y a pas de valeur  $\theta_*$  telle que  $E[H(\theta_*, Y)] = 0$  (modélisation trop approximative), la solution de (II.26) peut donner des poids très inégaux aux observations et conduire à des solutions insatisfaisantes. Ceci rend la méthode peu robuste.

---

4. On trouvera des compléments dans le livre de Owen ou dans [100], ou encore [49].

## II.3 Estimateurs à minimum de contraste (M-estimateurs)

### II.3.1 Exemple : régression non-linéaire

Reprenons le modèle de la page 8 :

$$Y_i = n_\theta(X_i) + e_i, \quad E[e_i] = 0. \quad (\text{II.27})$$

L'estimateur des moindres carrés est

$$\hat{\theta}_n = \operatorname{argmin} \sum (Y_i - n_\theta(X_i))^2. \quad (\text{II.28})$$

L'estimateur robuste de Huber est

$$\hat{\theta}_n = \operatorname{argmin} \sum \psi(Y_i - n_\theta(X_i)) \quad (\text{II.29})$$

où  $\rho$  est une fonction qui commence en parabole, et se prolonge linéairement :  $\psi(x) = x^2 1_{|x| \leq \alpha} + (2\alpha|x| - \alpha^2) 1_{|x| > \alpha}$ . Cet estimateur est moins sensible aux données aberrantes ; plus précisément dans le cas du problème du paramètre de translation du § I.3.2 (c.-à-d.  $n_\theta(x) = \theta$ ), il minimise la variance asymptotique la pire calculée lorsque la loi des  $Y_i$  varie dans un certain voisinage de la distribution gaussienne, voisinage constitué d'un mélange de la loi gaussienne et d'une autre loi (contamination) produisant les données aberrantes ; la constante  $\alpha$  dépend du poids attribué à la contamination dans le mélange, voir [12] § 4 pour des arguments précis, et [143] pour les détails pratiques.

Attention, la consistance (i.e.  $\nabla k(\theta_*) = 0$ , ici  $E[\psi'(e_1)] = 0$ ) n'est pas automatique, elle nécessite une hypothèse supplémentaire, par exemple que  $e_1$  ait une loi symétrique.

### II.3.2 Propriétés asymptotiques

On considère l'estimateur au minimum de contraste :

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \sum_{i=1}^n K(\theta, Y_i). \quad (\text{II.30})$$

#### 10 - THÉORÈME

On se place sous (GNU) pour  $Y_i$  et la fonction  $K(\theta, y)$ . On pose

$$k(\theta) = E[K(\theta, Y)].$$

Alors la suite  $k(\hat{\theta}_n)$  converge p.s. vers le minimum de la fonction  $k$ .

Si de plus  $\theta_*$  est le minimum unique de  $k(\theta)$  (identifiabilité) alors  $\hat{\theta}_n$  tend p.s. vers  $\theta_*$ .

Les conclusions restent vraies si l'on remplace (GNU) par (GNU').

*Démonstration.* Sous (GNU) c'est une conséquence immédiate du théorème 4 et du théorème 2. L'extension à (GNU') se fait comme à la démonstration du théorème 8. ■

Ce théorème est malheureusement inopérant dans des situations où  $K(\theta, Y)$  peut prendre des valeurs très grandes, voire infinies. C'est le cas lorsque l'on réalise le maximum de vraisemblance pour la famille  $P_\theta = \mathcal{U}([0, \theta])$  ; en effet, on a dans ce cas

$$K(\theta, Y_i) = -\log p_\theta(Y_i) = -\log(\theta^{-1} 1_{Y_i \in [0, \theta]})$$

et dès que  $\theta < \theta_*$ , on est sûr de trouver un  $Y_i$  pour lequel  $K(\theta, Y_i) = +\infty$ , ce qui en revanche est bon pour l'estimation. Pour traiter ces situations, on a le résultat suivant :

11 - THÉORÈME

On considère les hypothèses suivantes :

- (a) Pour tout  $A \geq 0$ , la suite  $(Y_i)_{i \geq 1}$  et la fonction  $(\theta, y) \mapsto \min(A, K(\theta, y))$  satisfont (GNU).
- (b) La fonction  $k(\theta) = E[K(\theta, Y)]$  a un unique minimum en  $\theta_*$ . En particulier  $k(\theta_*) < +\infty$ .
- (c) On a p.s.  $\frac{1}{n} \sum_{i=1}^n K(\theta_*, Y_i) \rightarrow k(\theta_*)$ .

Sous (a,b,c),  $\hat{\theta}_n$  tend p.s. vers  $\theta_*$ .

On peut également remplacer dans cet énoncé (GNU) par (GNU').

*Remarque.* Comme en vertu de (a)  $E[K(\theta, Y)_-]$  est fini, la fonction  $k(\theta) = E[K(\theta, Y)_+] - E[K(\theta, Y)_-]$  est bien définie mais peut valoir  $+\infty$ .

*Démonstration.* On pose pour tout  $A \geq 0$  (valeur qui tendra vers  $+\infty$ )

$$K_n^A(\theta) = \frac{1}{n} \sum_{i=1}^n \min(A, K(\theta, Y_i))$$

$$k^A(\theta) = E[\min(A, K(\theta, Y))]$$

$$K_n(\theta) = \frac{1}{n} \sum_{i=1}^n K(\theta, Y_i).$$

Alors pour tout  $n$  et tout  $A$

$$K_n^A(\hat{\theta}_n) \leq K_n(\hat{\theta}_n) \leq K_n(\theta_*).$$

La convergence uniforme des fonctions  $K_n^A$  vers  $k^A$  implique alors que

$$\overline{\lim}_n k^A(\hat{\theta}_n) \leq \overline{\lim}_n \left( K_n^A(\hat{\theta}_n) + \|K_n^A - k^A\|_\infty \right) \leq \overline{\lim}_n K_n(\theta_*) = k(\theta_*)$$

la dernière identité provenant de (c). Ceci arrive avec probabilité 1 pour tout  $A$  entier. Soit  $\omega$  appartenant à cet ensemble de probabilité 1. Pour toute valeur d'adhérence  $\hat{\theta}_\infty$  de la suite  $\hat{\theta}_n(\omega)$ , on a donc par continuité de  $k^A$

$$k^A(\hat{\theta}_\infty) \leq k(\theta_*).$$

On peut faire tendre  $A$  vers l'infini et obtenir (théorème de convergence monotone)  $k(\hat{\theta}_\infty) \leq k(\theta_*)$ , ce qui prouve que  $\hat{\theta}_\infty = \theta_*$ , et donc que  $\hat{\theta}_n$  converge vers  $\theta_*$ .

L'extension à (GNU') se fait comme dans la démonstration du théorème 8. ■

Revenons à (II.28), mais supposons de manière plus générale que  $Y_i = f(X_i) + e_i$  pour une certaine fonction  $f$  (i.e. le vrai modèle n'appartient pas à la famille paramétrique). On suppose également que  $e_i$  est indépendant de  $X_i$

$$k(\theta) = E[(Y_1 - n_\theta(X_1))^2] = E[(f(X_1) + e_1 - n_\theta(X_1))^2] = E[(f(X_1) - n_\theta(X_1))^2] + E[e_1^2].$$

On aura donc convergence si par exemple la suite  $(X_i, e_i)$  est iid et :

$\forall x$ , la fonction  $\theta \mapsto n_\theta(x)$  est continue

$$E\left[\sup_{\theta \in \Theta} n_\theta(X_1)^2\right] < \infty$$

$$E[e_i^2] < \infty$$

$$E\left[(f(X_1) - n_\theta(X_1))^2\right] \text{ est minimum pour un unique } \theta = \theta_*.$$

Pour avoir la normalité asymptotique d'un estimateur par minimum de contraste, il suffit d'appliquer le théorème 9 à la fonction  $H(\theta, y) = \nabla_{\theta} K(\theta, y)$  car on a :

$$\sum_{i=1}^n \nabla_{\theta} K(\hat{\theta}_n, Y_i) = 0.$$

Noter que c'est la matrice de dérivée seconde de  $k$  qui va intervenir dans l'expression de la variance asymptotique. Dans les cas où  $K$  n'est pas assez régulière, l'indépendance des  $Y_i$  peut sauver la situation, c'est ce que nous présentons dans la section suivante.

### II.3.3 Le cas indépendant

Il se trouve que dans le cas indépendant, on peut affaiblir substantiellement les hypothèses pour la normalité asymptotique, ce qui permet, par exemple, de traiter le cas de la médiane où  $K(\theta, y) = |y - \theta|$ . Le jeu d'hypothèses est le suivant :

**(ME)**  $(\Theta, d)$  est un espace métrique compact.  $K(\theta, y) : \Theta \times \mathbb{R}^m \rightarrow E$  est une fonction telle que pour tout  $\theta, y \mapsto K(\theta, y)$  soit mesurable.  $(Y_i)_{i \geq 1}$  est une suite de variables iid. Il existe une fonction  $\dot{K}$  telle que, pour tous  $\theta$  et  $\theta'$  appartenant à un voisinage de  $\theta_*$ , avec probabilité 1

$$|K(\theta, Y_1) - K(\theta', Y_1)| \leq \dot{K}(Y_1) \|\theta - \theta'\|$$

et

$$E[\dot{K}(Y_1)^2] < \infty.$$

Presque sûrement  $\theta \mapsto K(\theta, Y_1)$  est différentiable en  $\theta_*$  et

$$E[\|\nabla K(\theta_*, Y_1)\|^2] < \infty.$$

De plus la fonction  $\theta \mapsto k(\theta) = E[K(\theta, Y_1)]$  admet un développement de Taylor à l'ordre deux en son minimum  $\theta_*$  de la forme

$$k(\theta_* + h) - k(\theta_*) = \frac{1}{2} h^T J h + o(|h|^2) \tag{II.31}$$

et la matrice de sensibilité  $J$  est inversible.

#### 12 - THÉORÈME

On se place sous (ME). Soit un M-estimateur  $\hat{\theta}_n$  convergeant en probabilité vers  $\theta_*$ , tel que

$$\sum_{i=1}^n K(\hat{\theta}_n, Y_i) - \min_{\theta} \sum_{i=1}^n K(\theta, Y_i) \xrightarrow{P} 0.$$

Alors

$$\sqrt{n} \left( \hat{\theta}_n - \theta_* - \frac{1}{n} J^{-1} \sum_{i=1}^n \nabla K(\theta_*, Y_i) \right) \xrightarrow{P} 0 \tag{II.32}$$

$$\sqrt{n} (\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, J^{-1} I J^{-T}), \quad I = E[\nabla K(\theta_*, Y_1) \nabla K(\theta_*, Y_1)^T]. \tag{II.33}$$

*Démonstration.* La trame de la démonstration suit [22] §3.2. Elle repose ici sur le théorème 65, un résultat délicat démontré à l'appendice C. Appliquons ce théorème avec

$$L_i(h) = K(\theta_* + h, Y_i) - K(\theta_*, Y_i) - \nabla K(\theta_*, Y_i) h - k(\theta_* + h) + k(\theta_*).$$

On a  $C(\dot{L}) \leq E[(2\dot{K}(Y_1) + \|\nabla K(\theta_*, Y_1)\|)^2]$ . Pour le calcul de  $C(L)$ , qui dépend de  $r$ , on a

$$\begin{aligned} C(L)^2 &= E \left[ \sup_{\|h\| \leq r} (K(\theta_* + h, Y_1) - K(\theta_*, Y_1) - \nabla K(\theta_*, Y_1)h - k(\theta_* + h) + k(\theta_*))^2 \right] \\ &\leq r^2 E \left[ \sup_{\|h\| \leq r} \|h\|^{-2} \left( K(\theta_* + h, Y_1) - K(\theta_*, Y_1) - \nabla K(\theta_*, Y_1)h - k(\theta_* + h) + k(\theta_*) \right)^2 \right]. \end{aligned}$$

Par application du théorème de Lebesgue, l'espérance tend vers 0 quand  $r \rightarrow 0$  et donc

$$C(L) \leq r\eta(r)$$

pour une certaine fonction  $r \mapsto \eta(r)$  qui tend vers 0 lorsque  $r \rightarrow 0$ . Donc pour  $r$  assez petit, la première inégalité de (C.7) devient

$$E \left[ \sup_{\|h\| < r} n^{-1} \left| \sum_{i=1}^n L_i(h) \right|^2 \right] \leq C \log \left( \frac{1}{\eta(r)} \right) r^2 \eta(r)^2 \leq r^2 \tilde{\eta}(r)^2 \quad (\text{II.34})$$

pour une autre fonction  $\tilde{\eta}$  qui tend également vers 0. Si l'on pose  $M_n(\theta) = \frac{1}{n} \sum_i K(\theta, Y_i)$ , il vient donc pour tout  $\theta$  tel que  $\|\theta - \theta_*\| \leq r$  :

$$|M_n(\theta) - M_n(\theta_*) - \nabla M_n(\theta_*)(\theta - \theta_*) - k(\theta) + k(\theta_*)| \leq n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r$$

avec  $E[X_r^2] \leq 1$ . La régularité de  $k$  en  $\theta_*$  implique alors

$$|M_n(\theta) - M_n(\theta_*) - \nabla M_n(\theta_*)(\theta - \theta_*) - \frac{1}{2}(\theta - \theta_*)^T J(\theta - \theta_*)| \leq n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r + o(\|\theta - \theta_*\|^2).$$

En appliquant cette inégalité avec  $\theta = \hat{\theta}_n$ , et avec  $\theta = \theta' = \theta_* - J^{-1} \nabla M_n(\theta_*)$ , il vient

$$\begin{aligned} |M_n(\hat{\theta}_n) - M_n(\theta_*) - \nabla M_n(\theta_*)(\hat{\theta}_n - \theta_*) - \frac{1}{2}(\hat{\theta}_n - \theta_*)^T J(\hat{\theta}_n - \theta_*)| &\leq n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r + o(\|\hat{\theta}_n - \theta_*\|^2) \\ |M_n(\theta') - M_n(\theta_*) + \frac{1}{2} \nabla M_n(\theta_*)^T J^{-1} \nabla M_n(\theta_*)| &\leq n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r + o(\|\nabla M_n(\theta_*)\|^2) \end{aligned}$$

qui est vérifié sur  $A_n \cap B_n$ ,  $A_n = \{\|\hat{\theta}_n - \theta_*\| \leq r\}$ ,  $B_n = \{\|J^{-1} \nabla M_n(\theta_*)\| \leq r\}$ . Puis par différence de ces deux équations

$$\begin{aligned} -M_n(\hat{\theta}_n) + M_n(\theta') + \frac{1}{2} \left( \hat{\theta}_n - \theta_* + J^{-1} \nabla M_n(\theta_*) \right)^T J \left( \hat{\theta}_n - \theta_* + J^{-1} \nabla M_n(\theta_*) \right) \\ \leq 2n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r + o(\|\hat{\theta}_n - \theta_*\|^2) + o(\|\nabla M_n(\theta_*)\|^2) \end{aligned}$$

et puisque  $\hat{\theta}_n$  minimise  $M_n$  à  $o(n^{-1})$  près

$$\begin{aligned} \frac{1}{2} \left( \hat{\theta}_n - \theta_* + J^{-1} \nabla M_n(\theta_*) \right)^T J \left( \hat{\theta}_n - \theta_* + J^{-1} \nabla M_n(\theta_*) \right) \\ \leq 2n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r + o(\|\hat{\theta}_n - \theta_*\|^2) + o(\|\nabla M_n(\theta_*)\|^2) + n^{-1} o_P(1) \\ = 2n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r + o(\|\hat{\theta}_n - \theta_*\|^2) + n^{-1} o_P(1) \end{aligned}$$

donc, on a finalement obtenu

$$1_{A_n \cap B_n} \|\hat{\theta}_n - \theta_* + J^{-1} \nabla M_n(\theta_*)\|^2 \leq 2n^{-\frac{1}{2}} r \tilde{\eta}(r) X_r + o(\|\hat{\theta}_n - \theta_*\|^2) + n^{-1} o_P(1). \quad (\text{II.35})$$

Noter que si l'on savait que  $\|\hat{\theta}_n - \theta_*\|^2 = O_P(1/n)$ <sup>5</sup>, la démonstration serait terminée en choisissant  $r = r_n = \alpha_n / \sqrt{n}$ , où  $\alpha_n \rightarrow \infty$  est choisi tel que  $\alpha_n \tilde{\eta}(\alpha_n / \sqrt{n})$  tende vers 0, car alors après multiplication par  $n$ , le membre de droite tend vers 0, tandis que  $P(A_n \cap B_n) \rightarrow 1$ .

Il ne reste donc plus qu'à démontrer que  $\|\hat{\theta}_n - \theta_*\|^2 = O_P(n)$ . Pour simplifier un peu les calculs qui suivent, notons que le même raisonnement que précédemment avec cette fois

$$L_i(h) = K(\theta_* + h, Y_i) - K(\theta_*, Y_i) - k(\theta_* + h) + k(\theta_*)$$

5. La notation  $X_n = O_P(1)$  signifie que  $P(|X_n| > A) \rightarrow 0$  quand  $A \rightarrow \infty$  uniformément en  $n$ .  $X_n = O_P(a_n)$  signifie que  $a_n^{-1} X_n = O_P(1)$ . Si  $X_n = O_P(1)$  et  $\varepsilon_n \rightarrow 0$  en probabilité (i.e.  $\varepsilon_n = o_P(1)$ ), alors  $\varepsilon_n X_n = o_P(1)$ .

conduit aux mêmes équations sauf que  $\eta$  et  $\tilde{\eta}$  sont désormais simplement bornés et les termes en  $\nabla M_n$  disparaissent, si bien que (II.35) devient une équation un peu plus simple ( $B_n$  a disparu) qui nous suffira :

$$1_{A_n} \|\hat{\theta}_n - \theta_*\|^2 \leq n^{-\frac{1}{2}} r X_r + o(\|\hat{\theta}_n - \theta_*\|^2) + n^{-1} \xi_n, \quad \xi_n = O_P(1).$$

Comme par hypothèse,  $\|\hat{\theta}_n - \theta_*\|$  converge en probabilité vers 0, on en déduit, en faisant passer à gauche le terme  $o(\|\hat{\theta}_n - \theta_*\|^2)$ , que pour une v.a.  $\rho_n$  qui converge vers 1 p.s.

$$1_{A_n} \rho_n \|\hat{\theta}_n - \theta_*\|^2 \leq n^{-\frac{1}{2}} r X_r + n^{-1} \xi_n, \quad \xi_n = O_P(1).$$

Pour tout  $R > 0$ , la variable  $U_n = 1_{|\xi_n| \leq R \rho_n} \|\hat{\theta}_n - \theta_*\|^2$  satisfait

$$E[1_{U_n \leq r} U_n] \leq cn^{-\frac{1}{2}} \sqrt{r} + n^{-1} R.$$

Par application du lemme 66 avec  $C_1 = cn^{-\frac{1}{2}}$ ,  $C_2 = n^{-1} R$  et  $x = R^2/n$ , on obtient que  $P(nU_n > R^2) \leq (4c + 2)/R$ , et donc

$$P(n\rho_n \|\theta_n - \theta_*\|^2 > R^2) \leq P(nU_n > R^2) + P(\xi_n > R) \leq \frac{4c + 2}{R} + P(\xi_n > R)$$

qui tend bien vers 0 quand  $R \rightarrow \infty$  uniformément en  $n$ , ce qui achève la démonstration puisque  $\rho_n$  converge vers 1 en probabilité. ■

### II.3.4 Surajustement. Débiaisement du contraste

Lorsque l'on veut, par exemple, estimer les coefficients d'un processus autorégressif d'ordre inconnu  $d_*$  (qui n'existe que si le processus en est effectivement un), l'idée naturelle est d'essayer tous les ordres, disons  $d = 1, \dots, 50$ ,  $\hat{\theta}_n^d \in \mathbb{R}^d$ , et de choisir celui qui donne la meilleure vraisemblance. Pour  $d \leq d_*$  la vraisemblance augmentera franchement, et ensuite elle continuera à croître doucement, car on maximise à chaque fois sur un ensemble plus grand ; l'ordre 50 sera toujours choisi avec, si  $n$  est modeste — i.e. juste un peu supérieur à 50 —, une erreur de prédiction minuscule sur les données ayant servi à l'estimation, mais bien plus grande sur un autre jeu de données de même loi, c'est le *surajustement* (*overfitting*).

Ce problème de l'évaluation du risque des estimateurs à des fins de comparaison, qui déjà a été évoqué à l'exercice 1 p. 16 ( $C_p$ -Mallows et estimateur SURE), est extrêmement classique en estimation, paramétrique ou non. Les statisticiens ont trouvé une approche spécifique aux estimateurs à minimum de contraste, qui fait l'objet de cette section.

On a en somme une famille croissante d'espaces de paramètres  $\Theta_d$ , et l'on voudrait choisir le meilleur des  $\hat{\theta}_n^d$ , au sens de la minimisation de  $k(\hat{\theta}_n^d)$ . L'interprétation proposée par Akaike (1973) est que pour  $d > d_*$ , la comparaison des  $K_n(\hat{\theta}_n^d)$  (au lieu de  $k(\hat{\theta}_n^d)$ ) est vaine à cause du biais introduit par l'estimation ; l'idée est alors de débiaiser l'estimée du contraste, idée qui remonte à Mallows en 1964 [108] (exercice 1 p. 16), afin de prendre en compte le surajustement. Le débiaisement introduit sera faible de sorte qu'il n'affectera pas le comportement pour  $d \leq d_*$ , mais corrigera la diminution du contraste pour  $d > d_*$ . On se place dans la suite dans le cas  $d > d_*$ , ce qui signifie que  $\theta_* \in \Theta_d$ .

Comme  $\nabla k(\theta_*) = 0$ , on a par un développement en  $\theta_*$  du contraste  $k$  :

$$k(\hat{\theta}_n) = k(\theta_*) + \frac{1}{2}(\hat{\theta}_n - \theta_*)^T J(\hat{\theta}_n - \theta_*) + o_P\left(\frac{1}{n}\right). \quad (\text{II.36})$$

Par ailleurs, par un développement limité de  $K_n$  en  $\hat{\theta}_n$ , et en appliquant la loi des grands nombres uniforme à la dérivée seconde de  $K_n$  (cf. la démonstration du théorème 9 avec  $H = \nabla K$ ), il vient

$$K_n(\hat{\theta}_n) = K_n(\theta_*) - \frac{1}{2}(\hat{\theta}_n - \theta_*)^T J(\hat{\theta}_n - \theta_*) + o_P\left(\frac{1}{n}\right) \quad (\text{surajustement}) \quad (\text{II.37})$$

si bien que par soustraction de ces deux équations

$$k(\hat{\theta}_n) = \tau + K_n(\hat{\theta}_n) + (\hat{\theta}_n - \theta_*)^T J(\hat{\theta}_n - \theta_*) + o_P\left(\frac{1}{n}\right), \quad (\text{II.38})$$

$$\tau = k(\theta_*) - K_n(\theta_*). \quad (\text{II.39})$$

La quantité  $K_n(\hat{\theta}_n)$  sous-estime donc le contraste  $k(\hat{\theta}_n)$  par un terme qui, au vu du théorème 9, est asymptotiquement d'espérance  $Tr(IJ^{-1})$ , auquel s'additionne  $\tau$ . Noter que  $\tau$  est indépendant de  $d \geq d_*$ . On a donc l'estimée asymptotiquement non-biaisée de  $k(\hat{\theta}_n) - \tau$  :  $K_n(\hat{\theta}_n) + \frac{1}{n}Tr(IJ^{-1})$ .

Si l'on considère le contraste normalisé ou non  $Q(\theta, \omega) = (\frac{1}{n}) \sum_i K(\theta, Y_i)$ , on peut réécrire le critère indépendamment de la normalisation, d'une façon un peu informelle :

$$\boxed{\begin{aligned} \hat{\theta}^d &= \arg \min_{\theta \in \Theta_d} Q(\theta) \\ TIC(d) &= 2Q(\hat{\theta}^d) + 2Tr(\hat{J}^{-1}\hat{I}), \quad \hat{J} \simeq E[\nabla^2 Q(\theta_*)], \quad \hat{I} \simeq Var(\nabla Q(\theta_*)) \end{aligned}}$$

où le  $\simeq$  signale une estimation (cf. la discussion du § II.2.3).  $TIC(d)$ , appelé *critère de Takeuchi* (dire Takéouchi), est une estimée corrigée du contraste en  $\hat{\theta}_n$ . Dans le cas du critère de vraisemblance, on verra que  $I = J$  ce qui fait que  $Tr(J^{-1}I) = d$ , c'est le critère d'Akaike (AIC). Dans le cas d'un modèle linéaire  $Y = X\beta_* + u$  où la variance de  $u$  est connue, on retrouve, pour l'estimateur OLS, le  $C_p$ -Mallows (Ex. 1 p. 16).

Par exemple si l'on estime un processus ARMA par moindres carrés des résidus et que ces derniers sont seulement supposés décorrés, la correction peut être sensiblement différente du  $2d$  de AIC obtenu sous l'hypothèse paramétrique gaussienne, car le calcul de  $I$  fait intervenir des moments croisés d'ordre 4 du bruit, possiblement très différents de ceux donnés par l'hypothèse gaussienne (cf. la formule (II.20))<sup>6</sup>. Deux autres exemples sont proposés dans les exercices 7 plus bas et 2 p. 53.

Nous poursuivrons cette discussion page 68.

### II.3.5 Exercices et compléments

**Exercice 1 (Régression non-linéaire).** Étudier la variance asymptotique de l'estimateur des moindres carrés (II.28). On supposera que  $Y_i - n_{\theta_*}(X_i)$  est indépendant de  $X_i$ .

**Exercice 2 (Estimation de la médiane).** Montrer que le théorème 12 s'applique à l'estimation de la médiane de  $Y$  par minimum de contraste avec  $K(\theta, y) = |\theta - y|$ , si la loi de  $Y$  a une densité par rapport à la mesure de Lebsgue sur un voisinage de  $\theta_*$ . Calculer la variance asymptotique.

**Exercice 3.** Soit le processus

$$X_k = f(\theta_*, X_{k-1}) + e_k$$

où les  $e_n$  sont iid  $\mathcal{N}(0, \sigma^2)$ . On supposera que  $X_k$  est stationnaire ergodique. On considère  $\hat{\theta}_n$  l'estimateur de  $\theta_*$  aux moindres carrés basé sur l'observation de  $(X_k)_{0 \leq k \leq n}$ . Proposer des conditions sur  $f$  pour avoir convergence presque sûre, normalité asymptotique (utiliser le théorème 68 et la remarque qui suit) et donner l'expression de la variance asymptotique.

Étudier le cas  $f(\theta, x) = \theta|x|$ , et  $|\theta_*| < 0,9$ .

**Exercice 4.** Calculer en fonction de  $u$  la variance asymptotique de l'estimateur du point (1.c) de l'exercice 6 p. 24. Montrer que  $u = 1/p_{\theta_*}$  est optimal (on fera un calcul informel, oubliant les hypothèses techniques nécessaires).

*Indication* : On exprimera les matrices  $I$  et  $J$  en fonction de  $p_{\theta_*}$ ,  $\partial_{\theta} p_{\theta_*}$  et  $v = u \partial_{\theta} p_{\theta_*} - \int u p_{\theta_*} \partial_{\theta} p_{\theta_*}$ .

**Exercice 5 (Pseudo-vraisemblance).** Le modèle autorégressif avec erreur LARCH est le suivant (on se restreint ici à des ordres petits pour simplifier) :

$$\begin{aligned} Y_i &= a_* Y_{i-1} + X_i, \\ X_i &= e_i (b_* + c_* X_{i-1}) \quad E[e_i] = 0, \quad Var(e_i) = 1. \end{aligned}$$

Seuls les  $Y_i$  sont observés. Les  $e_i$  sont iid  $\theta_* = (a_*, b_*, c_*)$ . On suppose  $|a_*|, |c_*| < 1$  de sorte que le modèle est stable (moments d'ordre deux bornés).

<sup>6</sup>. C'est simple à observer dans le cas de l'estimation d'un AR(1). Pour des exemples de processus faiblement mais non fortement ARMA (i.e. bruits décorrés mais non indépendants), voir [75].

Même si les  $e_i$  sont  $\mathcal{N}(0, 1)$ , la vraisemblance est incalculable. On préfère utiliser la pseudo-vraisemblance basée sur  $M_i = E_\theta[Y_i | \mathcal{F}_{i-1}]$  et  $V_i = \text{Var}_\theta(Y_i | \mathcal{F}_{i-1})$  :

$$\mathcal{L}(\theta, Y) = -\frac{1}{2} \sum_i \frac{(Y_i - M_i)^2}{V_i} - \log V_i, \quad \theta = (a, b, c)$$

$$M_i = aY_{i-1},$$

$$V_i = (b + c(Y_{i-1} - aY_{i-2}))^2$$

Malheureusement, l'estimation va très mal se passer car la fonction  $\mathcal{L}$  est trop chaotique : soit  $i$  tel que  $|Y_i| < |Y_{i-1}|$ , on peut poser  $a = Y_i/Y_{i-1}$ , puis placer  $V_i$  arbitrairement proche de 0 en choisissant  $b$  et  $c$ , ce qui conduit à une pseudo-vraisemblance infinie (à vérifier!). On va plutôt choisir le contraste [137]

$$K(\theta, Y) = \frac{1}{n} \sum_i \frac{(Y_i - aY_{i-1})^2 + h}{V_i + h} + \log(V_i + h)$$

où  $h > 0$  est fixé à l'avance. On suppose dans la suite que l'hypothèse (GNU) est satisfaite. On va montrer que, pour tout choix de  $h > 0$ ,  $k(\theta) = E[K(\theta, Y)]$  est minimum pour  $\theta = \theta_*$ . Un calcul direct de  $\nabla k(\theta)$  donne bizarrement un jeu d'équations sans issue. Nous proposons dans la suite une autre méthode.

Commencer par démontrer que  $k(\theta) \geq k_0(\theta)$

$$k_0(\theta) = E \left[ \frac{(b_* + c_* X_{i-1})^2 + h}{V_i + h} + \log(V_i + h) \right]$$

avec égalité seulement si  $a = a_*$ . Puis en utilisant que  $\frac{a}{x} + \log(x) \geq 1 + \log(a)$ , on montrera que  $k_0(\theta)$  atteint son minimum en  $\theta_*$ . En conclure que  $k$  est minimum en  $\theta_*$ . Montrer qu'un autre point  $\theta' = (a', b', c')$  est minimum seulement si  $a' = a_*$  et si presque sûrement  $(b_* + c_* X_{i-1})^2 = (b' + c' X_{i-1})^2$ . Ceci prouve que  $\theta_*$  est l'unique minimum; on a donc consistance de l'estimateur.

**Exercice 6 (Moindres carrés fonctionnels).** On se propose d'appliquer l'estimation au minimum de contraste à la régression non paramétrique. On observe :

$$y_k = f_*(x_k) + \xi_k, \quad k = 1, \dots, n$$

où  $f_*$  est une fonction inconnue, les  $x_k$  sont des variables iid uniformes sur  $[0, 1]$  et les  $\xi_k$  sont iid  $\mathcal{N}(0, 1)$ . On sait que  $f_* \in \Theta$  où

$$\Theta = \{f : f \text{ continue sur } [0, 1]; |f(0)| \leq 1; \forall x, y \in [0, 1], |f(x) - f(y)| \leq |x - y|\}$$

et l'on considère l'estimateur des moindres carrés :

$$\hat{f}_n = \underset{f \in \Theta}{\operatorname{argmin}} \sum (y_k - f(x_k))^2.$$

1. Montrer, en appliquant simplement un théorème classique, que  $\Theta$  est un ensemble compact pour la norme uniforme.
2. Montrer que presque sûrement  $\|\hat{f}_n - f_*\|_\infty \rightarrow 0$ .

**Exercice 7 (Une application du critère d'Akaike-Takeuchi).** Soit le modèle

$$y_{ij} = f(t_i) + e_{ij}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq J.$$

Les  $e_{ij}$  sont iid centrés. On cherche à approcher  $f$  (inconnue) par une fonction de la famille  $(g_\theta)_{\theta \in \mathbb{R}^d}$ . On commencera par traiter le cas  $d = 1$ . On propose le contraste suivant

$$k(\theta) = \frac{1}{2} \sum_i (f(t_i) - g_\theta(t_i))^2 - f(t_i)^2.$$

1. Proposer un estimateur à minimum de contraste. Pourquoi a-t-on retranché le terme  $f(t_i)^2$  dans le contraste?
2. Proposer un estimateur  $\hat{\sigma}^2$  de la variance commune aux  $e_{ij}$ , valide dès que  $J \geq 2$  (commencer par le cas  $J = 2$ ). Attention : on ne suppose pas que  $f$  appartienne à la famille  $g_\theta$ .
3. Plusieurs familles sont en fait disponibles avec des  $d$  différents; on se propose d'utiliser le critère de Takeuchi. Proposer un estimateur du deuxième terme de  $TIC(d)$  (p. 44) utilisant  $\hat{\sigma}^2$ .

## II.4 Méthode des moments généralisée

La méthode des moments, cf. § I.3.2, relève des fonctions d'estimation § II.2. L'idée de la méthode des moments généralisée est de choisir la valeur de  $\theta$  qui minimise une certaine distance entre des statistiques observées (moments) et des statistiques prédites par  $\theta$ . Cette méthode a été proposée par Lars Peter Hansen en 1982 pour traiter le cas  $\dim(H) > \dim(\theta)$ .

### II.4.1 Un exemple : le minimum de $\chi^2$

$N_1, N_2, \dots, N_m$  désignent les effectifs des modalités d'une variable au cours de  $n$  expériences indépendantes; notons

$$\hat{p}_j = N_j/n$$

les probabilités empiriques. Les probabilités d'apparition des modalités sont paramétrées par un vecteur  $\theta \in \Theta \subset \mathbb{R}^d$ ,  $d \leq m - 1$ ; une telle situation a été présentée au début du § II.2.4, le modèle de Hardy-Weinberg. On a donc  $m$  fonctions  $p_1(\theta), \dots, p_m(\theta)$  dont la somme fait 1. L'estimateur au minimum de  $\chi^2$  est :

$$\hat{\theta}_n = \operatorname{argmin}_{\theta} n \sum_{j=1}^m \frac{(\hat{p}_j - p_j(\theta))^2}{p_j(\theta)} = \operatorname{argmin}_{\theta} \sum_{j=1}^m \frac{\hat{p}_j^2}{p_j(\theta)}. \quad (\text{II.40})$$

Cet estimateur ne rentre pas dans les catégories considérées jusqu'ici puisqu'il est basé sur une combinaison de statistiques. Il tente d'annuler au mieux les fonctions d'estimation  $\hat{p}_j - p_j(\theta)$ , l'annulation simultanée étant désespérée si  $d < m - 1$  en raison du nombre de paramètres.

On va voir que sous des hypothèses générales et si les données sont tirées selon  $p(\theta_*)$ ,  $\hat{\theta}_n$  est asymptotiquement gaussien et que la statistique

$$T_n = n \sum_{j=1}^m p_j(\hat{\theta}_n)^{-1} (\hat{p}_j - p_j(\hat{\theta}_n))^2$$

suit asymptotiquement un  $\chi^2$  à  $m - \dim(\theta) - 1$  degrés de liberté; ceci permettra de tester s'il existe bien un  $\theta_*$  tel que les données soient tirées selon  $p(\theta_*)$  (car sous l'alternative  $T_n \rightarrow \infty$ ).

### II.4.2 Définition et convergence

Un estimateur des moments généralisé est un estimateur de la forme

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} H_n(\theta)^T S_n H_n(\theta), \quad H_n(\theta) = \frac{1}{n} \sum_{i=1}^n H(\theta, Y_i) \quad (\text{II.41})$$

où les matrices symétriques  $S_n$  peuvent dépendre des observations, et généralement  $\dim(H) > \dim(\theta)$ . On va voir qu'un bon choix de  $S_n$  est l'inverse d'une estimée de la variance de  $H_n(\theta_*)$ , par exemple :

$$S_n^{-1} = \frac{1}{n} \sum_{i=1}^n H(\theta_o, Y_i) H(\theta_o, Y_i)^T - H_n(\theta_o) H_n(\theta_o)^T$$

où  $\theta_o$  est une estimée initiale.

#### 13 - THÉORÈME

On suppose que (GNU) ou (GNU') est satisfaite. On suppose également que  $S_n$  converge presque sûrement vers une limite  $S(\omega) > 0$ . S'il existe  $\theta_*$  tel que  $h(\theta_*) = 0$ , alors  $h(\hat{\theta}_n)$  converge presque sûrement vers 0, et si  $\theta_*$  est unique,  $\hat{\theta}_n$  converge presque sûrement vers  $\theta_*$ .

*Démonstration.* Noter qu'il existe  $N(\omega)$  et  $\varepsilon(\omega) > 0$  tels que  $S_n > \varepsilon(\omega)Id$  si  $n > N(\omega)$ . Comme  $H_n(\theta_*)^T S_n H_n(\theta_*)$  tend vers zéro, alors  $H_n(\hat{\theta}_n)^T S_n H_n(\hat{\theta}_n)$  aussi, et donc  $H_n(\hat{\theta}_n)$  tend vers zéro. Il suffit maintenant d'appliquer le théorème 7 et le théorème 8. ■

14 - THÉORÈME

On suppose que  $H$  satisfait l'hypothèse (FE) (p. 34). On suppose également que  $S_n$  converge presque sûrement vers une limite  $S$  déterministe, et que la matrice de sensibilité  $J = \nabla_{\theta} h(\theta_*)$  satisfait  $J^T S J > 0$ . On suppose que  $\hat{\theta}_n$  converge en probabilité vers  $\theta_* \in \text{int}(\Theta)$ . Si

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n H(\theta_*, Y_i) \xrightarrow{d} \mathcal{N}(0, I)$$

alors

$$\sqrt{n} (\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, J_*^{-1} J^T S I S J J_*^{-1}), \quad J_* = J^T S J \quad (\text{II.42})$$

$$\sqrt{n} H_n(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}(0, Q I Q^T), \quad Q = Id - J J_*^{-1} J^T S. \quad (\text{II.43})$$

Si  $I$  est inversible, un choix de  $S$  qui minimise la variance dans (II.42) est  $I^{-1}$  (toujours si  $J^T S J > 0$ ). Si ce n'est pas le cas mais si l'espace image de  $I$  contient celui de  $J$ , c-à-d que l'équation  $IK = J$  admet une solution, toute matrice telle que  $ISI = I$  et  $J^T S J > 0$  minimise la variance, avec pour valeur  $(K^T I K)^{-1}$ .

Pour un tel choix de  $S$ , on a la convergence en loi

$$n H_n(\hat{\theta}_n)^T S_n H_n(\hat{\theta}_n) \xrightarrow{d} \chi_{\text{Rang}(I) - \dim(\theta)}^2.$$

*Démonstration.* Rappelons que (FE) implique que  $\nabla_{\theta} H_n(\hat{\theta}_n) \rightarrow J$  p.s. (proposition 5). Il suffit alors d'appliquer le théorème 6 avec

$$M_n(\theta) = \nabla_{\theta} H_n(\hat{\theta}_n)^T S_n H_n(\theta).$$

Les hypothèses sont immédiatement vérifiées avec  $\mathcal{D} = \mathcal{N}(0, J^T S I S J)$  et  $J_* = J^T S J$ , et (II.42) s'ensuit. Pour (II.43) nous exploitons (II.5) et massivement le lemme de Slutsky :

$$\begin{aligned} \sqrt{n} H_n(\hat{\theta}_n) &= \sqrt{n} H_n(\theta_*) + \sqrt{n} \nabla_{\theta} H_n(\hat{\theta}'_n) (\hat{\theta}_n - \theta_*) \\ &= \sqrt{n} H_n(\theta_*) + \sqrt{n} J (\hat{\theta}_n - \theta_*) + r_n, \quad r_n \xrightarrow{P} 0 \\ &= \sqrt{n} H_n(\theta_*) - \sqrt{n} J J_*^{-1} M_n(\theta_*) + r'_n \quad r'_n \xrightarrow{P} 0 \\ &= (Id - J J_*^{-1} J^T S) \sqrt{n} H_n(\theta_*) + r''_n \quad r''_n \xrightarrow{P} 0 \end{aligned}$$

ce qui prouve (II.43).

Pour l'optimalité de l'inverse de  $I$  nous renvoyons à l'exercice 1 p. 48. Consacrons-nous à la dernière affirmation. Soit  $R$  une racine carrée symétrique de  $S$ , la limite en loi de la statistique est la même que celle de  $n \|R H_n(\hat{\theta}_n)\|^2$ ; le vecteur  $\sqrt{n} R H_n(\hat{\theta}_n)$  converge en loi vers  $\mathcal{N}(0, P)$ ,  $P = R Q I Q^T R$ . On vérifie facilement que  $Q I Q^T = I - I K (K^T I K)^{-1} K^T I$ , puis que  $P$  est une matrice de projection orthogonale ( $P = P^T = P^2$ ). La loi de  $\mathcal{N}(0, P)$  est donc celle de  $PX$ ,  $X \sim \mathcal{N}(0, Id)$ .  $\|PX\|^2$  est un  $\chi^2$  à  $\text{Trace}(P) = \text{Trace}(IS) - \dim(\theta)$  degrés de libertés; comme  $IS$  est un projecteur sur l'espace image de  $I$  sa trace est bien le rang de  $I$ . ■

**Bilan.** Si l'espace image de  $I$  ne contient pas celui de  $J$ , il y a des directions asymptotiquement de bruit nul et de sensibilité non nulle, il faut donc prendre  $S$  très grande dans ces directions. Il s'agit d'une situation beaucoup plus compliquée (mais très favorable).

Le choix  $S = KK^T$  minimise également la variance asymptotique et choisir un tel  $S$  revient à calculer  $\widehat{\theta}_n$  comme solution de  $K^T H_n(\theta, Y) = 0$ , et pour cette fonction d'estimation le théorème 9 donne bien la variance attendue (cf. aussi le § II.2.4).

L'ajout à  $S$  d'une matrice positive  $R$  telle que  $RJ = 0$  ne change pas la variance asymptotique.

**Retour au minimum de  $\chi^2$ .** On trouve  $I = \text{Diag}(p(\theta_*)) - p(\theta_*)p(\theta_*)^T$  qui n'est pas inversible car  $I1_d = 0$ . La matrice  $S = \text{Diag}(p(\theta_*))^{-1}$  satisfait bien  $ISI = I$ . L'hypothèse  $J^T S J > 0$  doit être vérifiée au cas par cas mais sera très généralement satisfaite car  $J \in \mathbb{R}^{d \times (d-1)}$  et  $S$  est de rang  $d$ . Comme  $\theta_*$  est inconnu, on remplace  $\theta_*$  par  $\theta$  dans l'expression de  $S$ , ce qui conduit à (II.40). Le théorème permet de montrer que cet estimateur reste optimal, ceci fait l'objet de l'exercice 2.

### II.4.3 Exercices et compléments

**Exercice 1.** On se place dans le cadre du théorème 14. On suppose l'existence d'une solution  $K$  à l'équation  $IK = J$ . Utiliser l'équation (II.24) pour démontrer que

$$(K^T IK)^{-1} \leq (J^T S J)^{-1} J^T S I S J (J^T S J)^{-1}.$$

**Exercice 2 (Minimum de  $\chi^2$ ).** On reprend l'exemple du § II.4.1. On suppose que les variables suivent effectivement une loi  $p_1(\theta_*), \dots, p_m(\theta_*)$ .

1. Soit l'équation d'estimation  $H(\theta, Y_i) = (1_{Y_i=k} - p_k(\theta))_{1 \leq k \leq m}$ . Calculer  $I$  et montrer (informellement) que le choix  $S = \text{Diag}(p_k(\theta_*)^{-1}_{1 \leq k \leq m})$  dans (II.41) conduit à une variance optimale.  $\theta_*$  étant inconnu, on choisit de résoudre l'équation (II.40), ce qui fait l'objet du reste de l'exercice.
2. Quelle est la fonction  $H$  pour que (II.40) ait la forme (II.41) avec  $S_n = Id$  ?
3. Quelle est, pour tout  $\theta$ , la limite quand  $n$  tend vers l'infini de  $H_n(\theta, Y)^T S_n H_n(\theta, Y)$  ?
4. À quelle condition a-t-on identifiabilité (i.e. unicité de la solution de  $h(\theta_*) = 0$ ) ?
5. Donner des hypothèses raisonnables sur  $p(\theta)$  pour que l'estimateur au minimum de  $\chi^2$  converge vers la bonne solution.
6. Donner des hypothèses raisonnables pour qu'il y ait normalité asymptotique. Que valent  $I$  et  $J$  ?
7. Montrer que la variance asymptotique vaut  $(J^T J)^{-1}$  (on remarquera que  $I$  est de la forme  $Id - xx^T$  avec  $\|x\| = 1$ , et  $J^T x = 0$ ).
8. Quel est le rang de  $I$  ? Pourquoi ? Montrer que  $S = Id$  minimise bien la variance asymptotique.

## II.5 Estimation sous contraintes

Si l'on sait que  $\theta_*$  satisfait une équation du type  $g(\theta_*) = 0$ , on pourra utiliser le minimum de contraste sous contraintes :

$$\widehat{\theta}_n^c = \arg \min_{g(\theta)=0} K_n(\theta). \quad (\text{II.44})$$

On peut avoir

$$K_n(\theta) = \frac{1}{n} \sum K(\theta, Y_i), \quad (\text{II.45})$$

mais on peut aussi définir  $\widehat{\theta}_n^c$  par projection d'un premier estimateur  $\widehat{\theta}_n$

$$K_n(\theta) = (\widehat{\theta}_n - \theta)^T S (\widehat{\theta}_n - \theta) \quad (\text{II.46})$$

où  $S$  une matrice à choisir ; on verra que le meilleur choix de  $S$  est l'inverse de la variance asymptotique de  $\widehat{\theta}_n - \theta_*$ .

Afin de pouvoir traiter les deux formes (II.45) et (II.46), les deux théorèmes énoncés plus bas se placeront directement sous l'hypothèse de convergence uniforme des fonctions, hypothèse qui est satisfaite sous (GNU) dans le cas (II.45), et toujours satisfaite dans le cas (II.46) dès que  $\widehat{\theta}_n$  converge presque sûrement.

Nous traitons ici du minimum de contraste sous contrainte mais on peut également considérer un algorithme de fonction d'estimation sous contraintes ; ceci est fait à l'exercice 4 p. 53.

L'application des résultats de cette section aux tests de significativité relatifs à l'hypothèse nulle «  $g(\theta_*) = 0$  » est présentée au § IV.2.

### II.5.1 Exemple : estimation de matrice sous contrainte de rang

Supposons que  $\theta \in \mathbb{R}^{p \times q}$  soit une matrice, typiquement une matrice de variance-covariance, avec  $\theta = E[Y_i]$ , et que l'on sache que le rang de  $\theta$  est  $r < \min(p, q)$ . Il est logique de considérer l'estimateur

$$\hat{\theta}_n^c = \arg \min_{Rg(\theta)=r} \|\theta - \bar{Y}\|_F^2, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \|M\|_F^2 = \sum_{ij} M_{ij}^2 \quad (\text{norme de Frobenius}). \quad (\text{II.47})$$

La solution de (II.47) s'obtient classiquement au travers de la décomposition en valeurs singulières de  $\bar{Y}$ , c'est la base de la théorie de l'analyse en composante principales. Cette écriture ressemble à (II.46), avec  $\hat{\theta}_n = \bar{Y}$ , mais on a aussi, par une décomposition élémentaire,

$$\hat{\theta}_n^c = \arg \min_{Rg(\theta)=r} \frac{1}{n} \sum_{i=1}^n \|\theta - Y_i\|_F^2$$

qui correspond maintenant à (II.45).

### II.5.2 Convergence

#### 15 - THÉORÈME

On suppose que les fonctions  $K_n$  convergent uniformément sur  $\Theta$  vers une fonction  $k(\theta)$  ayant un minimum unique  $\theta_*$  sur  $\Theta$ , que  $g(\theta_*) = 0$  et que  $g$  est continue. Alors  $\hat{\theta}_n^c$  solution de (II.44) converge presque sûrement vers  $\theta_*$ .

*Remarque.* Pour le cas (II.45), la convergence uniforme sera garantie sous (GNU) ou (GNU').

*Démonstration.* Il suffit d'appliquer directement le théorème 2 avec  $\Theta \cap \{g = 0\}$  au lieu de  $\Theta$ . ■

Si  $g$  et  $K_n$  sont de classe  $C^1$  et si  $\hat{\theta}_n^c$  est intérieur à  $\Theta$ , on a la propriété classique

$$\begin{aligned} \nabla K_n(\hat{\theta}_n^c) + \lambda_n^T \nabla g(\hat{\theta}_n^c) &= 0 \\ g(\hat{\theta}_n^c) &= 0 \end{aligned}$$

où  $\lambda_n$  est le multiplicateur de Lagrange (vecteur colonne de la dimension de  $g$ ;  $\nabla K_n(\theta)$  est une ligne et chaque ligne de  $\nabla g(\theta)$  est le gradient d'une composante); ceci va nous ramener aux fonctions d'estimation, où maintenant le paramètre est  $(\theta, \lambda)$ . On utilisera les hypothèses suivantes

**(EC1)**  $g$  est à valeur dans  $\mathbb{R}^m$ .  $g$  est de classe  $C^1$  au voisinage de  $\theta_* \in \text{int}(\Theta)$ , et  $\nabla g$  y est de rang  $m$ .

**(EC2)** Il existe un voisinage compact  $\Theta_*$  de  $\theta_*$  tel que presque sûrement la fonction  $\theta \rightarrow K_n(\theta)$  est de classe  $C^2$  sur  $\Theta_*$ , et les fonctions  $\theta \rightarrow \nabla^2 K_n(\theta)$  y convergent uniformément vers  $\nabla^2 k$ <sup>7</sup>.

*Remarque.* (EC2) peut être vérifiée à l'aide du théorème 4. Ces hypothèses font que le théorème des fonctions implicites permet de faire, au voisinage de  $\theta_*$ , un changement de paramétrisation  $\theta \mapsto \theta'$  tel que  $(\theta'_i)_{i \leq m} = \nabla g(\theta)$ ; ceci ramène le problème au cas fondamental où les contraintes sont simplement

7. Si les suites  $\nabla^2 K_n$  et  $K_n$  convergent uniformément, alors  $\nabla K_n$  également. En effet si  $f$  est de classe  $C^2$  sur un intervalle  $[a, b]$  de longueur  $h$  alors  $\|f'\|_\infty \leq 2\|f\|_\infty/h + h\|f''\|_\infty$ ; ceci vient de ce que, si  $c \in (a, b)$  est tel que  $f'(c) = (f(b) - f(a))/h$ , alors  $|f'(x)| \leq |f'(c)| + |f'(x) - f'(c)| \leq 2\|f\|_\infty/h + h\|f''\|_\infty$ .

l'annulation des premières coordonnées,  $\widehat{\theta}_n^c$  étant alors un estimateur des coordonnées restantes. Nous n'exploitons pas cette remarque dans la suite, la méthode directe étant plus simple.

16 - THÉORÈME

On se place sous les hypothèses du théorème 15 et sous (EC1,EC2). Alors la suite  $(\widehat{\theta}_n^c, \lambda_n)$  converge p.s vers  $(\theta_*, 0)$ . Posons

$$J = \nabla^2 k(\theta_*), \quad G = \nabla g(\theta_*), \quad J' = \begin{pmatrix} J & G^T \\ G & 0 \end{pmatrix}.$$

Si  $\sqrt{n} \nabla K_n(\theta_*)$  converge en loi vers une certaine limite et si  $J'$  est inversible (en particulier  $J > 0$  suffit car  $G$  est de rang plein), alors, en notant  $\widehat{\theta}_n$  l'estimateur sans contrainte,

$$J' \begin{pmatrix} \widehat{\theta}_n^c - \theta_* \\ \lambda_n \end{pmatrix} = \begin{pmatrix} -\nabla K_n(\theta_*)^T \\ 0 \end{pmatrix} + \frac{r_n}{\sqrt{n}} = \begin{pmatrix} J(\widehat{\theta}_n - \theta_*) \\ 0 \end{pmatrix} + \frac{r'_n}{\sqrt{n}}$$

avec  $(r_n, r'_n) \xrightarrow{P} 0$ . On a en particulier

$$\widehat{\theta}_n^c - \theta_* = P(\widehat{\theta}_n - \theta_*) + \frac{r''_n}{\sqrt{n}} \tag{II.48}$$

où  $P$  est le projecteur (orthogonal pour la norme définie par  $J$ )

$$P = Id - J^{-1}G^T(GJ^{-1}G^T)^{-1}G \tag{II.49}$$

sur l'espace contraint linéarisé au voisinage de  $\theta_*$ . En particulier si  $\sqrt{n} \nabla K_n(\theta_*) \rightarrow \mathcal{N}(0, I)$  en loi, alors  $\sqrt{n}(\widehat{\theta}_n^c - \theta_*) \rightarrow \mathcal{N}(0, V_c)$  en loi avec :

$$V_c = PJ^{-1}IJ^{-1}P^T. \tag{II.50}$$

*Démonstration.* Comme  $\widehat{\theta}_n^c$  converge vers  $\theta_*$ , alors  $(\nabla K_n(\widehat{\theta}_n^c), \nabla g(\widehat{\theta}_n^c))$  converge vers  $(0, \nabla g(\theta_*))$  (cf. la proposition 5) et la condition de rang plein implique que

$$\lambda_n = -(\nabla g(\widehat{\theta}_n^c) \nabla g(\widehat{\theta}_n^c)^T)^{-1} \nabla g(\widehat{\theta}_n^c) \nabla K_n(\widehat{\theta}_n^c)^T$$

tend vers 0. Appliquons maintenant le théorème 6 avec

$$\theta' = \begin{pmatrix} \theta \\ \lambda \end{pmatrix}, \quad M_n(\theta') = \begin{pmatrix} \nabla K_n(\theta)^T + \nabla g(\widehat{\theta}_n^c)^T \lambda \\ g(\theta) \end{pmatrix}$$

Si  $\widehat{\theta}_n^c$  n'est pas encore dans le voisinage de  $\theta_*$  où  $g$  est de classe  $C^1$ , on remplace  $\nabla g(\widehat{\theta}_n^c)$  par 0; il aurait pu sembler plus naturel de mettre  $\theta$  au lieu de  $\widehat{\theta}_n^c$  dans  $\nabla g$ , mais il faut alors supposer  $g$  de classe  $C^2$  pour finir la démonstration. Les hypothèses sont bien vérifiées et la première égalité s'en déduit immédiatement. La seconde est l'application de la première en absence de contrainte (cas où tout ce qui précède reste valide).

L'inverse de  $J'$  s'écrit

$$\begin{pmatrix} J & G^T \\ G & 0 \end{pmatrix}^{-1} = \begin{pmatrix} J^{-1} - J^{-1}G^T S G J^{-1} & J^{-1}G^T S \\ S G J^{-1} & -S \end{pmatrix}, \quad S = (GJ^{-1}G^T)^{-1}$$

ce qui implique l'expression pour  $P$ . La dernière affirmation est immédiate. ■

On voit en particulier que  $V_c$  peut très bien être strictement plus grande que la variance asymptotique  $V = J^{-1}IJ^{-1}$  de  $\widehat{\theta}_n$ , car  $J$  dans (II.49) rend la projection  $P$  oblique, mais que  $\widehat{\theta}_n^c$  est toujours meilleur

que  $\widehat{\theta}_n$  si  $J = Id$  car  $P$  est alors une projection orthogonale. La variance asymptotique de  $\widehat{\theta}_n^c$  est nulle dans la direction des colonnes de  $G$  (i.e.  $V_c G = 0$ ), en raison des contraintes satisfaites. On a également  $V_c \leq V$  si  $I = J$  (on trouve  $V_c = PJ^{-1}$  et  $V = J^{-1}$ ). L'usage de l'estimateur contraint est donc risqué en dehors de ces situations (notons que pour le maximum de vraisemblance on a bien  $I = J$ ).

**Méthode en deux temps.** Si l'on veut améliorer un estimateur  $\widehat{\theta}_n$  existant en prenant en compte la contrainte  $g(\theta) = 0$ , une solution à la fois bonne et simple est de faire

$$\boxed{\widehat{\theta}_n^c = \arg \min_{g(\theta)=0} \frac{1}{2}(\theta - \widehat{\theta}_n)^T V^{-1}(\theta - \widehat{\theta}_n)} \quad (\text{II.51})$$

où  $V$  est la variance de  $\sqrt{n}(\widehat{\theta}_n - \theta_*)$  (ou une estimée... ). On voit facilement que  $I = J$ , et (II.50) et (II.48) deviennent

$$V_c = V - VG^T(GVG^T)^{-1}GV \quad (\text{II.52})$$

$$\widehat{\theta}_n^c - \theta_* = V_c V^{-1}(\widehat{\theta}_n - \theta_*) + \frac{r''}{\sqrt{n}}. \quad (\text{II.53})$$

L'exercice 3 p. 53 montre que le choix proposé pour  $V$  est optimal, car un autre choix, disons  $W$  pour la matrice de pondération de (II.51) conduit à une variance de

$$(Id - WG^T(GWG^T)^{-1}G)V(Id - G^T(GWG^T)^{-1}GW) \quad (\text{II.54})$$

qui est supérieure à  $V_c$  (appliquer l'exercice avec  $A = WG^T(GWG^T)^{-1}$ ).

Noter que  $V$  ne dépend pas de  $g$ , ce qui va être exploité à la fin du § II.5.3 pour le calcul effectif de  $\widehat{\theta}_n^c$ .

Noter également que l'on est ici dans le cas  $I = J$ , ce qui sera utile dans la suite.

### II.5.3 Exemple : estimation de matrice sous contrainte de rang. Suite

Vérifions que la condition  $\text{Rang}(\theta) = r$  se réécrit au voisinage de  $\theta_*$  sous la forme  $g(\theta) = 0$  où  $g$  est  $C^1$  et  $\nabla g$  de rang correspondant à son nombre de composantes. Pour ceci, notons que  $\theta_*$  a une sous matrice  $r \times r$  de rang  $r$ ,  $\theta_{*,I,J} = (\theta_{*ij})_{i \in I, j \in J}$ , et qu'au voisinage de  $\theta_*$ ,  $\theta_{I,J}$  est, par continuité du déterminant, toujours de rang plein. Notons le  $\mathcal{V}$ . Sur  $\mathcal{V}$ , les matrices de rang  $r$  vont satisfaire bien entendu

$$\det((\theta_{ij})_{i \in I \cup \{k\}, j \in J \cup \{l\}}) = 0, \quad k \notin I, l \notin J \quad (\text{II.55})$$

$$I \left( \begin{array}{ccc|c} & J & & l \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \hline & & & \\ \times & \times & \times & \times \end{array} \right)$$

Réciproquement, si une matrice  $\theta \in \mathcal{V}$  satisfait cet ensemble (II.55) de  $(p-r)(q-r)$  conditions, alors le mineur  $\theta_{I,J}$  est maximum (ne peut être augmenté sans annuler le déterminant) ce qui prouve que le rang est  $r$ . On a ici  $(p-r)(q-r)$  conditions de la forme  $g_{kl}(\theta) = 0$ ,  $k \notin I$ ,  $l \notin J$ , et pour tout  $k_0 \notin I$ ,  $l_0 \notin J$  la dérivée de  $g_{kl}$  dans la direction  $E_{k_0 l_0}$  (la matrice nulle partout sauf un 1 en  $(k_0, l_0)$ ) vaut zéro si  $(k_0, l_0) \neq (k, l)$  (évident) et sinon, comme la dérivée du déterminant par rapport au coefficient est le cofacteur associé, cette dérivée vaut  $\det((\theta_{ij})_{i \in I, j \in J})$  si  $(k_0, l_0) = (k, l)$ , nombre non nul par hypothèse. En résumé, pour chaque  $(k_0, l_0)$ , on a trouvé une direction pour laquelle seule la fonction  $g_{k_0, l_0}$  varie au premier ordre, et pas les autres. Le gradient de  $g$  est donc bien de rang plein.

**Bilan.** Au vu de ce qui précède, le choix de la norme de Frobenius dans (II.47) est bon si les coefficients des matrices sont décorrélés de même variance, sinon il faut adapter le critère en utilisant par exemple la méthode en deux temps; noter que l'optimisation sous contrainte de rang peut se faire numériquement sans expliciter  $g$  (cf. [53] § 3.3), ce qui fait que comme  $V$  est indépendant de  $g$ , cette dernière fonction n'apparaît pas dans l'algorithme. L'estimation de  $V$  ne pose pas de problème dans le cas de données indépendantes.

## II.5.4 Théorème de Wilks

Le théorème qui suit permet de construire une statistique pour tester l'hypothèse  $H_0 : \langle g(\theta_*) = 0 \rangle$ , le membre de droite de (II.56). Il faut pour cela se placer dans le cas  $I = J$ , ce qui sera le cas pour le maximum de vraisemblance. On a vu une méthode générique pour se ramener à  $I = J$ , équation (II.51); on va voir également un exemple en régression non-linéaire.

### 17 - THÉORÈME

On se place sous les hypothèses du théorème 16 avec toujours  $J = \nabla^2 k(\theta_*)$ . Si  $\sqrt{n} \nabla K_n(\theta_*)$  converge en loi vers  $\mathcal{N}(0, J)$ , alors

$$2n(K_n(\hat{\theta}_n^c) - K_n(\hat{\theta}_n)) \longrightarrow \chi_m^2 \quad (\text{II.56})$$

où  $m$  est la dimension de  $g$ , c.-à-d. la différence de dimension effective entre  $\theta$  et  $\theta^c$ .

Le cas  $g(\theta) = \theta - \theta_*$  conduit à la région de confiance :  $\mathcal{R}_\alpha = \left\{ \theta : 2 \sum_{i=1}^n K_n(\theta) - K_n(\hat{\theta}_n) \leq \chi_d^2(1 - \alpha) \right\}$ .

*Démonstration.* On peut appliquer le théorème 16. L'équation (II.48) donne

$$\begin{aligned} \hat{\theta}_n^c - \hat{\theta}_n &= -(Id - P)(\hat{\theta}_n - \theta_*) + \frac{r_n''}{\sqrt{n}} \\ &= -J^{-1}G^T(GJ^{-1}G^T)^{-1}G(\hat{\theta}_n - \theta_*) + \frac{r_n''}{\sqrt{n}}. \end{aligned}$$

Comme  $\hat{\theta}_n - \theta_* = n^{-1/2}J^{-1/2}X_n$  avec  $X_n \rightarrow \mathcal{N}(0, Id)$ , il vient,

$$\sqrt{n}(\hat{\theta}_n^c - \hat{\theta}_n) = -J^{-1}G^T(GJ^{-1}G^T)^{-1}GJ^{-1/2}X_n + r_n''$$

Par ailleurs, la formule de Taylor au voisinage de  $\hat{\theta}_n$  donne, pour un certain  $\theta'_n$  du segment  $[\hat{\theta}_n - \hat{\theta}_n^c]$

$$\begin{aligned} 2n(K_n(\hat{\theta}_n^c) - K(\hat{\theta}_n)) &= n(\hat{\theta}_n - \hat{\theta}_n^c)^T \nabla_{\theta}^2 K_n(\theta'_n)(\hat{\theta}_n - \hat{\theta}_n^c) \\ &= \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^c)^T (J + r_n) \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^c) \end{aligned}$$

par application de la proposition 5. En remplaçant, on trouve que la loi limite est la même que celle de

$$\begin{aligned} X_n^T J^{-1/2} G^T (GJ^{-1}G^T)^{-1} GJ^{-1} G^T (GJ^{-1}G^T)^{-1} GJ^{-1/2} X_n \\ = X_n^T J^{-1/2} G^T (GJ^{-1}G^T)^{-1} GJ^{-1/2} X_n \end{aligned}$$

qui est le carré de la norme de  $PX_n$ ,  $P = (GJ^{-1}G^T)^{-1/2}GJ^{-1/2}$ . Comme  $PP^T = Id_m$ , la loi de  $PX_\infty$  si  $X_\infty \sim \mathcal{N}(0, Id_d)$  est  $\mathcal{N}(0, Id_m)$ . Par conséquent  $PX_n$  converge en loi vers  $\mathcal{N}(0, Id_m)$ , ce qui prouve bien le résultat. ■

**Le cas non-compact.** Si  $\Theta$  n'est pas compact, en vertu des remarques faites au § II.2.2 les théorèmes restent vrais si leurs hypothèses sont satisfaites pour tout compact  $\Theta_0 \subset \Theta$  et que presque sûrement la suite  $\hat{\theta}_n$  reste confinée dans un compact de  $\Theta$  (le compact pouvant dépendre de l'évènement).

**Application.** Revenons au modèle de régression non-linéaire

$$\begin{aligned} Y_i &= n_{\theta_*}(X_i) + e_i \\ E[e_i | X_i] &= 0 \end{aligned}$$

avec l'estimateur (II.29) p. 39 :

$$\begin{aligned} K_n(\theta) &= \frac{1}{n} \sum_{i=1}^n \tau \psi(Y_i - n_{\theta}(X_i)) \\ \hat{\theta}_R &= \arg \min_{\theta} K_n(\theta) \\ \tau &= E[\psi''(e_1)] / E[\psi'(e_1)^2], \end{aligned} \quad (\text{II.57})$$

la constante  $\tau$  ayant été ajoutée pour pouvoir obtenir  $I = J$ . On suppose que les  $e_i$  et les  $X_i$  forment deux suites iid, indépendantes, que la loi de  $e_1$  est symétrique, et que les conditions d'intégrabilité nécessaires au calcul de  $\tau$  sont satisfaites ainsi que les hypothèses du théorème 16<sup>8</sup>. Alors dans ce cas on a bien  $I = J$  et le théorème de Wilks s'applique. La constante  $\tau$  peut être estimée à posteriori car sa valeur n'influe pas sur l'estimateur.

## II.5.5 Exercices

**Exercice 1.** Vérifier que l'on a bien  $I = J$  pour l'estimateur (II.57). Donner des conditions sur la fonction  $n_\theta(x)$  assurant que les hypothèses du théorème 16 seront satisfaites.

**Exercice 2 (Une application du critère d'Akaike-Takeuchi [128]).** On reprend l'exemple (II.57), sans constante  $\tau$  ajoutée en facteur du contraste. Montrer que le critère de Takeuchi (page 44) devient

$$TIC(d) = \sum_{i=1}^n \psi(Y_i - n_{\hat{\theta}_d}(X_i)) + d \frac{E[\psi'(e_1)^2]}{E[\psi''(e_1)]}.$$

**Exercice 3.** Soit  $A, G, S$  trois matrices,  $S$  étant définie positive. Montrer que

$$(Id - AG)S(Id - AG)^T \geq S - SG^T(GSG^T)^{-1}GS$$

si l'inverse existe (i.e.  $G$  est de rang ligne plein), avec égalité si  $A = SG^T(GSG^T)^{-1}$ .

*Indication :* Poser  $A = SG^T(GSG^T)^{-1} + R$ , et développer.

**Exercice 4.** On propose l'algorithme de fonction d'estimation avec contrainte suivant :

$$\begin{aligned} H_n(\hat{\theta}_n^c)^T + \lambda_n^T \nabla g(\hat{\theta}_n^c) &= 0 \\ g(\hat{\theta}_n^c) &= 0. \end{aligned}$$

On suppose que  $(\hat{\theta}_n^c, \lambda_n)$  converge vers  $(\theta_*, 0)$ , que  $\sqrt{n}H_n(\theta_*)$  converge en loi vers  $\mathcal{N}(0, I)$ , et que  $\nabla H_n(\theta_*)$  converge vers  $J$ .

1. Donner la variance asymptotique de  $\hat{\theta}_n^c$  (on pourra se contenter de reprendre informellement la démonstration du théorème en prenant garde que  $J$  n'est plus symétrique).
2. Soit  $M$  une matrice inversible. Montrer que si l'on remplace  $H_n(\hat{\theta}_n^c)^T$  par  $H_n(\hat{\theta}_n^c)^T M^T$  dans l'équation d'estimation, la meilleure variance asymptotique est

$$S - SG^T(GSG^T)^{-1}GS, \quad S = J^{-1}IJ^{-T}$$

obtenue avec  $M = J^T I^{-1}$  (ce résultat est à mettre en parallèle avec le § II.2.4). Utiliser l'exercice 3 p. 53.

## II.6 Bornes de grandes déviations et estimées des moments

La convergence en loi de  $\sqrt{n}(\hat{\theta}_n - \theta_*)$  vers une gaussienne ne dit rien des moments de ces variables. On propose ici une borne qui permet d'obtenir la convergence de ces moments. Elle sera appliquée à l'estimateur au maximum de vraisemblance.

Le plan d'action est le suivant : Si  $\hat{\theta}$  est un estimateur à minimum de contraste, l'idée va être de majorer  $E[f(\hat{\theta})]$  (typiquement  $f(\hat{\theta}) = \|\hat{\theta} - \theta_*\|^p$ ) par  $E[\sup_\theta f(\theta) \exp\{K(\theta, \omega) - K(\theta_*, \omega)\}]$ . Le sup est ensuite remplacé par une intégrale via l'inégalité de Sobolev (II.58), et le calcul se termine après permutation de l'intégrale et de l'espérance. Ceci court-circuite la méthode de chaînage employée dans [2].

L'expression  $f$  différentiable sur  $\Theta$  signifiera dans la suite  $f$  différentiable sur l'intérieur de  $\Theta$  et continue sur  $\Theta$ . On dira qu'un ouvert  $\Omega \subset \mathbb{R}^d$  possède la **propriété de cône** s'il existe un cône tronqué  $C$

8. A propos de  $E[\psi''(e_1)]$ . Si l'on utilise effectivement la fonction  $\psi$  donnée p. 39, sa dérivée seconde est  $2\delta_\alpha + 2\delta_{-\alpha}$  (masses de Dirac), il faut donc une densité continue  $p_e$  pour  $e_1$ , puis  $E[\psi''(e_1)] = 2(p_e(-\alpha) + p_e(\alpha))$ ; on obtient le même résultat en calculant  $\frac{d}{dt} E[\psi'(e_1 + t)]|_{t=0}$ .

d'intérieur non-vide et de sommet 0 tel pour tout point  $\theta \in \Omega$ , il existe une rotation  $R$  telle que  $\theta + RC \subset \Omega$ ; ceci signifie qu'il n'y a pas de pointe à la frontière de  $\Omega$ . Rappelons que pour toute fonction  $g$  différentiable et tout  $q > d$ , on a l'inégalité de Sobolev<sup>9</sup> :

$$\|g\|_\infty^q \leq C_q \int_\Omega \left( \|\nabla g(z)\|^q + |g(z)|^q \right) dz \quad (\text{II.58})$$

où  $C_q$  ne dépend que de  $q$  et de  $C$ . Pour  $\varepsilon \leq 1$  on peut appliquer la même inégalité à  $g(\varepsilon z)$  sur  $\Omega/\varepsilon$  (qui vérifie la condition de cône avec même  $C$ ) et l'on obtient

$$\begin{aligned} \|g\|_\infty^q &\leq C_q \int_{\Omega/\varepsilon} \left( \varepsilon^q \|\nabla g(\varepsilon z)\|^q + |g(\varepsilon z)|^q \right) dz \\ &= C_q \varepsilon^{-d} \int_\Omega \left( \varepsilon^q \|\nabla g(z)\|^q + |g(z)|^q \right) dz. \end{aligned}$$

Il sera plus pratique d'utiliser une version légèrement différente de cette formule : si  $h$  est une fonction strictement positive, alors en définissant  $g = h^{1/q}$ , il vient

$$\|h\|_\infty \leq C_q \varepsilon^{-d} \int_\Omega \left( \varepsilon^q \|\nabla \log h(z)\|^q + 1 \right) |h(z)| dz, \quad \varepsilon \leq 1. \quad (\text{II.59})$$

Afin de simplifier la démonstration du théorème qui va suivre, il est plus simple d'introduire un lemme où tout est normalisé, ce qui permet de s'affranchir des constantes  $n, \sigma$  et  $\eta$  qui apparaissent ensuite, et de mieux synthétiser l'information :

18 - LEMME

On suppose que  $T \subset \mathbb{R}^d$  est la fermeture de son intérieur, lequel possède la propriété de cône.  $T$  n'est pas nécessairement compact. Soit  $L(t, \omega)$  variable aléatoire mesurable sur  $T \times \Omega$ . On suppose que p.s.  $L$  est différentiable en  $t$ . On suppose que pour un  $q > d$ , un  $M > 0$ , et pour tout  $t \in T$

$$E \left[ \|\nabla_t L(t)\|^q e^{-L(t)} 1_{L(t) \geq 0} \right] \leq M(1 + \|t\|)^q e^{-\|t\|^2/2} \quad (\text{II.60})$$

$$E \left[ e^{-L(t)_+} \right] \leq M e^{-\|t\|^2/2} \quad (\text{II.61})$$

où l'on a noté  $L(t)$  pour  $L(t, \omega)$ . Soit une variable aléatoire  $\hat{t}(\omega)$  satisfaisant

$$L(\hat{t}(\omega), \omega) \leq 0 \quad p.s.$$

alors il existe une constante  $C_{q,T}$  ne dépendant que de  $q$  et de  $T$  telle que pour tout  $1 \leq i \leq d$  et tout  $\alpha \geq 0$

$$E \left[ \exp(\alpha \hat{t}_i(\omega)) \right] \leq C_{q,T} M(1 + \alpha^d) e^{\frac{\alpha^2}{2}}. \quad (\text{II.62})$$

La constante  $C_{q,T}$  satisfait la propriété suivante : pour tout  $t \in \mathbb{R}^d$  et  $a \geq 1$ , on a  $C_{q,aT+t} \leq C_{q,T}$ .

*Démonstration.* Posons pour tout  $\delta > 0$

$$\begin{aligned} r(t) &= r(t, \omega) = e^{-\rho(L(t, \omega))}, \quad \rho(x) = \frac{x^2 1_{x > 0}}{\delta + x} \\ f(t) &= \exp \{ \alpha (t_i - \alpha) \}. \end{aligned}$$

9. Corollaire 5.16 de [24]. On voit bien dans [43] comment on démontre d'abord cette inégalité pour  $\Omega = \mathbb{R}^d$  de manière relativement élémentaire (Théorème de Morrey, p.166) : on montre d'abord que pour tout cube  $Q$  parallèle aux axes de côté  $r$  et tout  $x \in Q$  on a  $|g(x) - \frac{1}{|Q|} \int_Q g(y) dy| \leq \frac{r^{1-d/q}}{1-d/q} \|\nabla g\|_{L^q(Q)}$  et l'on conclut en prenant  $r = 1$ . On traite ensuite le cas général en utilisant un opérateur de prolongement (Théorème IX.7) permettant d'étendre une fonction définie sur  $\Omega$  en une fonction définie sur  $\mathbb{R}^d$  sans trop augmenter le terme de droite de (II.58).

La fonction  $\rho(x)$  a été choisie de sorte à être  $\geq 0$ , croissante, dérivable, nulle sur  $\mathbb{R}_-$ , à dérivée  $\leq 1$  et telle que  $\|\rho(x) - x_+\|_\infty \leq \delta$ . En vertu de l'inégalité de Sobolev (II.59) et puisque  $r(\hat{t}) = 1$  on a

$$\begin{aligned} E[f(\hat{t})] &= E[f(\hat{t})r(\hat{t})] \\ &\leq E[\|rf\|_\infty] \\ &\leq C_q \varepsilon^{-d} E \int (\varepsilon^q \|\nabla \log(r(z)f(z))\|^q + 1) r(z)f(z) dz \\ &\leq 2^q C_q \varepsilon^{-d} E \int (\varepsilon^q \|\nabla L(z, \omega)\|^q \mathbf{1}_{L(z, \omega) \geq 0} + \alpha^q \varepsilon^q + 1) r(z)f(z) dz \end{aligned}$$

d'où en faisant tendre  $\delta$  vers 0 (et en notant  $e_i$  le  $i$ -ième vecteur de la base canonique) :

$$\begin{aligned} E[f(\hat{t})] &\leq 2^q C_q \varepsilon^{-d} E \int (\varepsilon^q \|\nabla L(z, \omega)\|^q \mathbf{1}_{L(z, \omega) \geq 0} + 1 + \alpha^q \varepsilon^q) e^{-L(z, \omega)} f(z) dz \\ &\leq 2^q C_q \varepsilon^{-d} \int \left\{ \varepsilon^q M(1 + \|z\|)^q + M + M\varepsilon^q \alpha^q \right\} e^{-\frac{\|z\|^2}{2} + \alpha(z_i - \alpha)} dz \\ &= 2^q C_q \varepsilon^{-d} e^{-\frac{\alpha^2}{2}} M \int \left\{ \varepsilon^q (1 + \|z + \alpha e_i\|^q) + \varepsilon^q \alpha^q \right\} e^{-\|z\|^2/2} dz \\ &\leq C'_q \varepsilon^{-d} e^{-\frac{\alpha^2}{2}} M (\varepsilon^q (1 + \alpha^q) + 1 + \varepsilon^q \alpha^q). \end{aligned}$$

Il ne reste plus qu'à prendre  $\varepsilon = 1/(1 + \alpha)$  (qui est bien  $\leq 1$ ).

La dernière affirmation du lemme vient de ce que  $aT + t_0$  satisfait une meilleure propriété de cône que  $T$ . ■

### 19 - THÉORÈME

On suppose que  $\Theta \subset \mathbb{R}^d$  est la fermeture de son intérieur, lequel possède la propriété de cône.  $\Theta$  n'est pas nécessairement compact.

Soit  $K_n(\theta, \omega)$  une suite de variable aléatoire mesurables sur  $\Theta \times \Omega$  (fonctions de contraste). On suppose que p.s.  $K_n$  est différentiable par rapport à  $\theta$ . On suppose que pour un  $q > d$ , certains  $M, \sigma, \eta > 0$ , un  $\theta_* \in \Theta$ , pour tout  $\theta \in \Theta$

$$\begin{aligned} E[\|\nabla_\theta K_n(\theta)\|^q e^{-\eta(K_n(\theta) - K_n(\theta_*))} \mathbf{1}_{K_n(\theta) \geq K_n(\theta_*)}] \\ \leq M(\sqrt{n} + n\|\theta - \theta_*\|)^q e^{-n\|\theta - \theta_*\|^2/2\sigma^2} \end{aligned} \quad (\text{II.63})$$

$$E[e^{-\eta(K_n(\theta) - K_n(\theta_*))}] \leq M e^{-n\|\theta - \theta_*\|^2/2\sigma^2} \quad (\text{II.64})$$

où l'on a noté  $K_n(\theta)$  pour  $K_n(\theta, \omega)$ . Soit  $\hat{\theta}_n$  qui satisfait

$$K_n(\hat{\theta}_n, \omega) \leq K_n(\theta_*, \omega) \quad p.s.$$

alors il existe une constante  $C$  dépendant des données  $(q, M, \sigma, \eta$  et  $\Theta)$ , mais pas de  $\theta_*$ , telle que pour tout  $1 \leq i \leq d$  et tout  $\alpha \geq 0$

$$P(\sqrt{n}|\hat{\theta}_{ni} - \theta_{*i}| \geq \alpha) \leq C(1 + \alpha^d) e^{-\frac{\alpha^2}{2\sigma^2}}. \quad (\text{II.65})$$

En particulier tout moment de  $\sqrt{n}(\hat{\theta}_n - \theta_*)$  est borné (quand  $n$  varie) et la suite  $\sqrt{n}\|\hat{\theta}_n - \theta_*\|/\sqrt{\log n}$  est bornée avec probabilité 1.

*Démonstration.* En appliquant le lemme 18 à la fonction

$$L(t, \omega) = \eta \left( K_n \left( \frac{\sigma}{\sqrt{n}} t, \omega \right) - K_n(\theta_*, \omega) \right)$$

sur  $T = \sqrt{n}(\Theta - \theta_*)/\sigma$ , et en posant  $\hat{t} = \sqrt{n}(\hat{\theta}_n - \theta_*)/\sigma$ , il vient

$$E[e^{\alpha\sqrt{n}(\hat{\theta}_n - \theta_*)_i/\sigma}] \leq C_1(1 + \alpha^d)e^{\frac{\alpha^2}{2}}.$$

où la constante  $C_1$  dépend maintenant également de  $(M, \sigma, \eta)$ . On en déduit

$$P(\sqrt{n}(\hat{\theta}_n - \theta_*)_i \geq \alpha) \leq E \left[ \exp \left\{ \frac{\alpha\sqrt{n}}{\sigma^2} (\hat{\theta}_i - \theta_{*i} - \alpha/\sqrt{n}) \right\} \right] \leq C_1(1 + \alpha^d \sigma^{-d})e^{\frac{\alpha^2}{2\sigma^2}}. \quad \blacksquare$$

EXEMPLE. Ce théorème sera appliqué à l'estimateur au maximum de vraisemblance, § III.2.3. Donnons ici un exemple simple en régression non-linéaire. On vérifie facilement que si les  $e_i$  de (II.27) sont  $\mathcal{N}(0, \sigma_e^2)$ , si  $\theta \mapsto n_\theta(x)$  est de classe  $C^1(\Theta)$ , et si

$$\forall x, \theta, \theta', \quad C_1 \|\theta - \theta'\| \leq |n_\theta(x) - n_{\theta'}(x)| \leq C_2 \|\theta - \theta'\|$$

pour certains  $0 < C_1 < C_2 < \infty$ , alors les hypothèses plus fortes mais plus simples suivantes sont satisfaites pour l'estimateur (II.28)

$$\begin{aligned} E[\|\nabla_\theta K_n(\theta)\|^q] &\leq M(\sqrt{n} + n\|\theta - \theta_*\|)^q, \quad \text{pour un } q > 2d \\ E[e^{-\eta(K_n(\theta) - K_n(\theta_*))}] &\leq M e^{-\lambda n \|\theta - \theta_*\|^2} \end{aligned}$$

((II.63) se déduit de l'inégalité de Cauchy-Schwartz). Les détails sont laissés en exercice (prendre  $\eta = 1/\sigma_e^2$ ).

# III

---

## ESTIMATION PARAMÉTRIQUE

---

---

On est désormais dans le cadre paramétrique général du § I.1, sauf que dans un premier temps il sera plus simple de ramasser les observations en un bloc  $Y = (Y_1, \dots, Y_n)$  qui sera considéré comme une seule observation. On travaillera ainsi sur la base d'une seule observation, quitte à particulariser dans le cas où elle prend la forme  $Y = (Y_1, \dots, Y_n)$  où les  $Y_i$  sont iid.

On cherche donc à estimer un paramètre inconnu  $\theta_*$  à partir d'une observation  $Y$  quand on se donne une famille de lois candidates  $P_\theta(y) = p_\theta(y)\mu(dy)$  pour  $Y$ ,  $\theta \in \Theta$ .

Nous avons vu dans l'introduction, p. 6, que les résultats que nous allons présenter ne doivent pas être pris excessivement au premier degré dans les applications pratiques, car le modèle est généralement approximatif, d'autant plus que l'échantillon est grand ; *d'un point de vue pratique, la vraisemblance doit être vue comme un contraste* (d'où par exemple l'estimateur sandwich (III.11)). Les modèles paramétriques sont néanmoins à la base du développement d'un grand nombre d'algorithmes (filtre de Kalman,...), et la théorie paramétrique permet d'expliquer le comportement d'estimateurs associés de manière détaillée et conceptuellement parlante ; elle sert également à la compréhension de situations plus compliquées (et plus réalistes) comme l'estimation de modèles non paramétriques, ou semi-paramétriques.

On trouvera des approfondissements concernant ce chapitre, par exemple dans [16, 21, 2, 13], références rangées par ordre de complexité croissante.

### III.1 Comparaison des estimateurs

#### III.1.1 Risque, estimateur admissible et approche minimax.

Soit une fonction de perte  $r(\theta, \theta') \geq 0$ , nulle pour  $\theta = \theta'$ , typiquement  $r(\theta, \theta') = |\theta - \theta'|^2$ . Le risque d'un estimateur  $\hat{\theta}$  est la fonction suivante, perte moyenne de l'estimateur lorsque le vrai paramètre est  $\theta$  :

$$R_\theta(\hat{\theta}) = E_\theta[r(\theta, \hat{\theta}(Y))].$$

C'est cette fonction qui caractérise la qualité de l'estimateur. La comparaison d'estimateurs est donc une comparaison de fonctions, c'est pourquoi il est difficile de mettre en avant un estimateur optimal, même théoriquement.

On dit que  $\hat{\theta}$  est **admissible** s'il n'existe pas d'autre estimateur  $\hat{\theta}'$  strictement meilleur au sens où :  $R_\theta(\hat{\theta}') \leq R_\theta(\hat{\theta})$  pour tout  $\theta$  et  $R_{\theta_0}(\hat{\theta}') < R_{\theta_0}(\hat{\theta})$  pour un certain  $\theta_0$ . On verra que les estimateurs bayésiens sont généralement admissibles, ce qui fournit une grande classe d'exemples.

Par ailleurs, un estimateur  $\hat{\theta}$  est dit **minimax sur**  $\Theta$  si et seulement si pour tout autre estimateur  $\hat{\theta}'$  on a

$$\sup_{\theta \in \Theta} R_\theta(\hat{\theta}) \leq \sup_{\theta \in \Theta} R_\theta(\hat{\theta}').$$

Pour éviter que les sup ne soient infinis, il faut généralement supposer  $\Theta$  compact.

On verra dans l'exercice 11 p. 86 que l'estimateur minimax pour l'observation de  $n$  variables de

Bernoulli  $\mathcal{B}(1, \theta)$  indépendantes a un risque constant égal à  $\frac{1}{4(1+\sqrt{n})^2}$ , à comparer avec  $\frac{\theta(1-\theta)}{n}$  pour le maximum de vraisemblance : l'abaissement du risque maximal au seuil minimax se paye par une uniformisation du risque à un niveau excessivement élevé, alors que l'estimateur au maximum de vraisemblance a un risque maximal quasiment équivalent,  $\frac{1}{4n}$  ; la morale est que si les estimateurs exactement minimax sont souvent asymptotiquement uniformément médiocres, les estimateurs asymptotiquement minimax peuvent éviter ce défaut. Nous reviendrons sur cette question délicate au § III.3.3.

Il est difficile de trouver des estimateurs minimax et si possible admissibles. De plus, il y a rarement unicité, ce qui fait que ces deux critères ne permettent pas de caractériser un estimateur privilégié.

Un estimateur sera dit **uniformément de risque minimal** (uniformly minimum risk, UMR) si sa fonction de risque est partout inférieure aux autres fonctions de risque

$$\forall \theta \in \Theta, \quad \forall \hat{\theta}', \quad R_\theta(\hat{\theta}) \leq R_\theta(\hat{\theta}')$$

où  $\hat{\theta}'$  varie parmi tous les estimateurs. C'est une propriété extrêmement forte ; un estimateur UMR est bien entendu optimal. Mais en général un tel estimateur n'existe pas car le membre de droite a pour chaque  $\theta$  un minimum nul : il suffit de prendre  $\hat{\theta}' = \theta$ . On va voir au paragraphe suivant un phénomène tout-à-fait remarquable : si l'on restreint raisonnablement la classe des estimateurs considérés, il peut exister un estimateur UMR (on imposera aux estimateurs d'être sans biais, ce qui interdit le choix  $\hat{\theta}' = \theta$ ).

On voit qu'il est assez difficile de trouver des critères qui vont permettre de mettre en évidence un estimateur meilleur que tous les autres. La méthode bayésienne qui consiste à mettre une distribution de probabilité  $\pi$  sur  $\theta$  et à chercher un estimateur qui minimise le risque moyen  $\int E_\theta[r(\theta, \hat{\theta})]\pi(d\theta)$ , donne une solution théoriquement simple au problème, voir § III.4.1 ; mais il reste à savoir quelle distribution poser sur  $\theta$ . Des liens classiques entre estimateurs bayésiens et minimax seront explorés.

Pour certains aspects de la théorie qui ne seront pas abordés ici (sauf pour les tests sans biais, plus bas), il est naturel de modifier la notion de biais en fonction du choix du risque : Un estimateur  $\hat{\theta}$  est dit **r-non-biaisé** si et seulement si

$$\forall \theta, \theta' \in \Theta, \quad E_\theta[r(\theta, \hat{\theta}(Y))] \leq E_\theta[r(\theta', \hat{\theta}(Y))]. \quad (\text{III.1})$$

Les estimateurs non-biaisés (au sens habituel) sont les quadratique-non-biaisés.

### III.1.2 L'approche de Rao-Blackwell-Lehmann-Scheffé

Cette théorie a pour but de mettre en évidence des estimateurs UMR dans un cadre non asymptotique. On a besoin d'introduire quelques concepts qui décrivent comment une statistique concentre l'information nécessaire à l'estimation de  $\theta_*$ .

Une statistique  $S(Y)$  est dite **exhaustive** (ang. : **sufficient**) si la loi conditionnelle  $P_\theta(Y|S(Y))$  ne dépend pas de  $\theta$ <sup>1</sup> ; une fois que  $S(Y)$  est connue, aucune autre information sur  $\theta$  ne peut être inférée de  $Y$ , point qui sera précisé mathématiquement plus bas. Le problème d'estimation se réduit à estimer  $\theta_*$  à partir d'une observation de  $S$ , ayant à sa disposition une famille lois (pour  $S$ ) indexées par  $\theta$  ; cette réduction diminue souvent considérablement le vecteur d'observation, car  $S$  aura typiquement la dimension de  $\theta$ , mais son utilisation peut compliquer la famille de lois. L'exemple le plus simple de statistique exhaustive est la moyenne empirique des  $T(Y_i)$  où les  $Y_i$  sont i.i.d issues d'une famille exponentielle.

Le **théorème de Neyman-Fisher** [16] assure que  $S$  est une statistique exhaustive si et seulement s'il existe une factorisation de  $p_\theta$  sous la forme

$$p_\theta(y) = q_\theta(S(y))p_0(y),$$

valide pour  $\mu$ -presque tout  $y$ , où  $p_0$  ne dépend pas de  $\theta$ . En particulier, si  $S$  est exhaustive, l'estimateur du maximum de vraisemblance ne dépend que de  $S$ .

Si  $\hat{\theta}$  est un estimateur, alors  $\check{\theta} = E_{\theta_*}[\hat{\theta}|S]$  est un meilleur estimateur de  $\theta$  au sens de l'erreur quadratique (exercice 7 p. 60). C'est le **théorème de Rao-Blackwell** (Rao 1945, Blackwell 1947).

1. Pour tout ensemble mesurable  $A$ , il existe une fonction mesurable  $\varphi$  t.q. pour tout  $\theta$  :  $P_\theta(Y \in A|S(Y)) = \varphi(S(Y))$   $P_\theta$ -p.s.

Noter qu'en raison de l'exhaustivité,  $\check{\theta}$  est effectivement calculable, c.-à-d. sans connaître  $\theta_*$  (ce n'est malheureusement pas le cas de  $E_{\theta_*}[\hat{\theta}]$  !); l'amélioration de  $\hat{\theta}$  par  $\check{\theta}$  est parfois appelée *Rao-Blackwellisation*; ce peut être une opération difficile à mettre en pratique, mais pas toujours (cf. exercice 5 p. 60).

Une statistique exhaustive  $S$  est dite **minimale** si elle est fonction de toute autre statistique exhaustive<sup>2</sup>. Si une statistique exhaustive ne dépend que des rapports  $p_{\theta}(y)/p_{\theta_0}(y)$  pour  $\theta$  appartenant à une famille dénombrable et  $\theta_0$  donné, alors elle est minimale (car en vertu du théorème de factorisation elle est fonction de toute autre statistique exhaustive). Une statistique exhaustive  $S$  minimale existe en général<sup>3</sup>.

Une statistique  $S$  est dite **complète** (ou totale) si toute fonction réelle  $g$  telle que  $E_{\theta}[g(S)^2] < \infty$  et  $E_{\theta}[g(S)] = 0$  pour tout  $\theta \in \Theta$  satisfait forcément  $g(S) = 0$   $P_{\theta}$ -presque sûrement pour tout  $\theta$ ; en d'autres termes, la connaissance de la fonction  $\theta \mapsto E_{\theta}[g(S)]$  suffit à remonter à  $g$ . Cette condition n'est pas forcément très difficile à vérifier (exercice 3 p. 60) mais elle peut très bien être non-satisfaite dans des situations simples. Si une statistique exhaustive minimale n'est pas complète, aucune ne l'est (puisqu'elles sont toutes fonction l'une de l'autre). On peut montrer qu'une statistique exhaustive complète est minimale. La réciproque est fautive et c'est typiquement le cas si la dimension de la statistique minimale est supérieure à la dimension du paramètre (exercices 6 et 12 p. 61).

Le théorème suivant assure alors l'existence d'un estimateur UMR unique pour le risque quadratique parmi les estimateurs non-biaisés; on l'exprime ici dans le cas de l'estimation d'une fonction du paramètre :

20 - THÉORÈME (Lehmann-Scheffé, 1950)

Soit  $S$  une statistique exhaustive et complète et  $g(\theta)$  une fonction de  $\theta$ . Soit  $T$  un estimateur non-biaisé de variance finie de  $g(\theta)$ . Alors  $T^* = E_{\theta}[T|S]$  ne dépend ni du choix de  $\theta$  (en raison de l'exhaustivité) ni du choix de  $T$  (en raison de la complétude) et est de variance minimale parmi les estimateurs non-biaisés (UMVU=Uniformly Minimum Variance Unbiased).

En particulier, pour toute fonction  $f$ ,  $f(S)$  est un estimateur UMVU de son espérance. La démonstration est assez simple (exercice 8 p. 60). Une statistique exhaustive et complète existe très souvent, mais pas nécessairement. Le problème de trouver un estimateur optimal est donc résolu dans le cadre suivant :

1. Estimateurs sans biais
2. Risque quadratique
3. Existence d'une statistique exhaustive et complète.

Par exemple, dans le cadre d'une famille exponentielle régulière canonique complète non-nécessairement standard, cf. § I.2.1, on obtient que  $T(Y)$  est un estimateur UMVU de  $\nabla Z(\theta_*)$  (exercice 13 p. 61). Cependant, les inconvénients majeurs de cette approche sont les suivants :

1. Concernant les estimateurs sans biais :
  - (a) L'estimateur du maximum de vraisemblance et les estimateurs bayésiens sont généralement biaisés (exercice 5 p. 84).
  - (b) Il existe des estimateurs biaisés qui ont un risque quadratique plus faible que l'UMVU, comme l'estimateur de James et Stein (exercice 10 p. 60). L'intérêt des estimateurs biaisés augmente si  $\theta$  est de grande dimension car on s'approche du cadre non paramétrique (p.ex. estimateurs à troncature pour le problème des *normal means*), ceci sort du cadre de ce cours.
  - (c) Il peut ne pas exister d'estimateur non-biaisé (exercice 2 p. 60).
2. Il existe de nombreuses situations où il n'y a pas de statistique complète (exercices 6, 12 et 14).

Voyant là des difficultés insurmontables, les statisticiens se sont alors intéressés aux propriétés asymptotiques des estimateurs, question que l'on abordera progressivement dans la suite, ce qui a finalement donné naissance à la théorie de l'efficacité présentée au § III.3.3.

2. Si  $T$  en est une autre, il existe  $\varphi$  telle que pour tout  $\theta$ ,  $p_{\theta}$ -p.s.,  $S = \varphi(T)$ .

3. On la produit théoriquement ainsi : on dit qu'une tribu  $\mathcal{F}$  est exhaustive si  $P_{\theta}(Y \in \cdot | \mathcal{F})$  ne dépend pas de  $\theta$ ; l'intersection de toutes ces tribus est la tribu exhaustive minimale, et une statistique qui l'engendre est minimale sous des conditions générales. Ce raisonnement réclame quelques précautions, voir [102].

### III.1.3 Exercices et compléments

**Exercice 1.** Montrer qu'un estimateur admissible de risque constant est minimax.

*Indication :* Raisonner par l'absurde.

**Exercice 2 (Un cas simple où tout estimateur est biaisé).** Soit  $Y$  une variable binomiale avec  $P(Y = 1) = p$ . Montrer qu'il n'existe pas d'estimateur non-biaisé de  $\frac{p}{1-p}$ .

*Indication :* N'utiliser aucun théorème, mais s'intéresser à l'équation  $E[\hat{\theta}(X_1, \dots, X_n)] = \frac{p}{1-p}$  pour  $p$  proche de 1 (commencer par le cas  $n = 1$ ).

**Exercice 3 (Exhaustivité et complétude pour la loi de Poisson).** On considère la famille  $(P_\theta)_{\theta > 0}$  des distributions de Poisson :  $P_\theta(k) = e^{-\theta}\theta^k/k!$ . Soit un échantillon  $Y_1, \dots, Y_n$  tiré selon une de ces lois. Montrer que la moyenne empirique est une statistique exhaustive et complète.

*Indication :* On pourra utiliser la propriété suivante : Si  $a_i$  est une suite de réels telle que  $\sum_{i \geq 0} |a_i| < \infty$  et si pour tout  $x$  d'un intervalle non vide de  $]0, 1[$  on a  $\sum_{i \geq 0} a_i x^i = 0$ , alors les  $a_i$  sont tous nuls.

**Exercice 4 (Statistique exhaustive en régression).** Montrer que dans le modèle (I.7) p. 9,  $\hat{\theta}$ , éq. (I.10), est une statistique exhaustive.

*Indication :* Noter que le vecteur  $Y - X\hat{\theta}$  est orthogonal à toutes les colonnes de  $X$ .

Montrer que dans l'exemple (I.8) p. 9,  $\sum Y_i X_i$  est une statistique exhaustive minimale (attention, les  $X_i$  ne sont pas des variables aléatoires).

*Indication :* Se souvenir que si  $U \sim \mathcal{B}(1, p)$ , alors sa probabilité de réalisation est  $p^U(1-p)^{1-U}$ .

**Exercice 5 (Rao-Blackwellisation).** Soit  $(X_1, \dots, X_n)$  une suite de variables iid de Poisson de paramètre  $\lambda$ . On cherche à estimer  $\theta = e^{-\lambda}$ . En vertu de l'exercice précédent, la statistique  $S_n = \sum X_i$  est exhaustive et complète.

1. Montrer que  $\hat{\theta} = 1_{X_1=0}$  est un estimateur non-biaisé de  $\theta$ .
2. Calculer  $P(X_1 = 0 | S_n = k)$  pour tout  $k \geq 0$ .
3. Proposer un estimateur meilleur que  $\hat{\theta}$  par Rao-Blackwellisation.

On pourra comparer l'estimateur obtenu à l'estimateur au maximum de vraisemblance.

**Exercice 6 (Une statistique non-complète).** Soit  $p$  la densité du mélange de  $\mathcal{N}(-1/2, 1)$  et  $\mathcal{N}(1/2, 1)$  avec les poids  $1/2$  et  $1/2$ . Soit  $p_\theta(x) = p(x - \theta)$ . On observe une réalisation  $X$  sous  $p_\theta$ .

1. Vérifier que  $p$  a un maximum unique. *Indication :*  $|\tanh x| \leq |x|$ .
2. Montrer que  $X$  est une statistique exhaustive minimale.  
*Indication :* On pourra s'intéresser au maximum en  $\theta$  rationnel de  $p_\theta(x)/p_0(x)$ .
3. Montrer que  $X$  est une statistique non-complète. *Indication :* Calculer  $E[e^{i\omega X}]$ .

**Exercice 7 (Théorème de Rao-Blackwell).** Soit  $X$  une variable et  $\mathcal{F}$  une tribu, justifier l'inégalité  $E[E[X|\mathcal{F}]^2] \leq E[X^2]$ . En déduire le théorème de Rao-Blackwell.

**Exercice 8.** Démontrer le théorème de Lehmann-Scheffé : Utiliser le théorème de Rao-Blackwell (exercice précédent) ; prouver que  $T^*$  ne dépend pas de  $T$  en utilisant la complétude.

**Exercice 9.** Montrer que si  $\Theta$  est fini,  $\Theta = \{1, \dots, k\}$ , et si l'on pose  $q(Y) = \sum_{j \in \Theta} p_j(Y)$  alors  $S(Y) = (p_i(Y)/q(Y))_{i \in \Theta}$  est une statistique exhaustive minimale.

**Exercice 10 (Estimateur biaisé de James et Stein<sup>4</sup>, 1961).** Soit  $Y$  une observation de loi  $\mathcal{N}(\mu, Id)$ .

1. Montrer que l'estimateur au maximum de vraisemblance de  $\mu$  est  $Y$ . En vertu de l'exercice 13  $Y$  est une statistique exhaustive minimale et complète.
2. On se place en dimension  $d \geq 3$  et l'on considère l'estimateur de James et Stein :

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

---

4. On trouvera des compléments intéressants au § 1 de [6] et au § 7.6 de [23], ainsi qu'une application au baseball et à la toxoplasmose dans [71]

Montrer que son risque vaut

$$E[\|\hat{\mu}(Y) - \mu\|^2] = d - (d-2)^2 E[\|Y\|^{-2}].$$

*Indication* : Commencer par calculer  $E[\|Y\|^{-2}\langle Y, Y - \mu \rangle]$  avec une IPP (i.e., sous des conditions standard, pour une fonction vectorielle  $F$  et une fonction scalaire  $g : \int \langle F, \nabla g \rangle = - \int \operatorname{div}(F)g$ ).

Il est donc inférieur à  $d$ , le risque l'estimateur au maximum de vraisemblance; noter en particulier l'amélioration si  $\mu$  est petit et  $d$  grand.

3. Cet estimateur peut visiblement être amélioré; comment? (On ne démontrera rien)

Pour le cas  $d = 1$ , voir l'exercice 1 p. 76.

**Exercice 11 (SURE-Shrinkage [65]).** On a vu, à l'exercice 10, que pour l'estimation de  $\mu$  dans le modèle  $Y \sim N(\mu, \sigma^2 Id)$ , si la dimension est élevée, un estimateur biaisé qui contracte l'estimateur non-biaisé peut être meilleur. C'est particulièrement vrai si l'on soupçonne que de nombreux paramètres sont très petits devant  $\sigma$ , voire nuls, cas où 0 est plus proche de  $\mu_i$  que  $Y_i$ . Soit l'estimateur

$$\hat{\mu} = \arg \min_{\mu} \frac{1}{2} \sum_i (Y_i - \hat{\mu}_i)^2 + \lambda \sum_i |\hat{\mu}_i|$$

où  $\lambda$  est un seuil à choisir. C'est un cas particulier de l'estimateur lasso (III.37). Montrer que

$$\hat{\mu}_i(Y) = 1_{|Y_i| > \lambda} (Y_i - \lambda \operatorname{sign}(Y_i)).$$

Au vu de l'exercice 1 p. 16, on se propose de choisir  $\lambda$  qui minimise  $\hat{R}(\lambda)$  donné par (I.25). Montrer qu'ici

$$\hat{R}(\lambda) = \sum_i Y_i^2 \wedge \lambda^2 + 2\sigma^2 \sum_i 1_{|Y_i| > \lambda} - n\sigma^2.$$

**Exercice 12 (Statistiques exhaustives, minimales).** Dans les cas suivants, montrer que  $S$  est exhaustive minimale. Seule la minimalité peut être délicate à montrer, d'où les \*.

1.  $(X_i, Y_i)$  est une suite iid de variables de loi de densité  $e^{-\theta x - y/\theta}$ ,  $S = (\sum X_i, \sum Y_i)$ .
2. L'observation est  $(X_1, \dots, X_N)$  où  $X_i$  est une suite iid de variables exponentielles de paramètre  $\theta$ , et  $N$  est une variable de loi connue,  $S = (\sum X_i/N, N)$ . Montrer également que  $S$  est non complète.
- \*3. (Basu)  $X_1, \dots, X_n$  sont iid valant 1, 2, 3 ou 4 avec probabilité  $(1 - \theta)/6$ ,  $(1 + \theta)/6$ ,  $(2 - \theta)/6$  et  $(2 + \theta)/6$ .  $N_i$  est le nombre de fois que la valeur  $i$  est apparue et  $S = (N_1, N_2, N_3)$ .  
Pour la minimalité on pourra se dispenser de détailler.
- \*4.  $Y_1, \dots, Y_n$  est une suite d'observations iid de loi de Laplace  $e^{-|y-\theta|}/2$ ,  $S$  est la statistique d'ordre  $Y_{(1)}, \dots, Y_{(n)}$ .
- \*5. Soit  $(Z_i)$  une suite iid uniforme sur  $[\theta - 1/2, \theta + 1/2]$ ,  $S = (Z_{(1)}, Z_{(n)})$ .

*Indication* : Utiliser la définition de la minimalité.

**\*Exercice 13 (Familles exponentielles).** Soit une famille exponentielle régulière canonique complète non-nécessairement standard, cf. § I.2.1. Montrer que  $T(Y)$  est une statistique minimale et complète. En déduire que si  $Y_1, \dots, Y_n$  sont iid de même loi  $P_\theta$  de cette famille, la moyenne empirique de  $T(Y)$  est un estimateur UMVU de  $\nabla Z(\theta_*)$ .

*Indications pour démontrer la complétude.* On se restreindra au cas où  $\theta$  est scalaire; on se donne une fonction  $g$  telle que  $E_{\theta'}[g(T(Y))] = 0$  pour  $|\theta' - \theta| < \varepsilon$ .

1. Montrer que si  $E_{\theta'}[g^2(T(Y))] < \infty$  pour  $|\theta' - \theta| < \varepsilon$ , alors

$$\int |g(T(y))| e^{\theta T(y)} \cosh(hT(y)) \mu(dy) < \infty$$

pour  $|h| < \varepsilon$ ,

2. En déduire que pour  $|\operatorname{Re}(z)| + |h| < \varepsilon$ , l'application  $f(z) = \int g(T(y)) e^{(\theta+z)T(y)} \mu(dy)$  satisfait

$$f(z+h) = \sum \frac{h^n}{n!} \int g(T(y)) e^{(\theta+z)T(y)} T(y)^n \mu(dy)$$

en utilisant le théorème de Fubini pour permuter  $\int$  et  $\sum$ .

- En déduire que  $f$  est analytique sur  $|\operatorname{Re}(z)| < \varepsilon$ <sup>5</sup>.
- Comme  $f$  est nulle sur  $] - \varepsilon, \varepsilon[$ , elle est nulle sur la bande  $] - \varepsilon, \varepsilon[ + i\mathbb{R}$ ; en déduire que pour tout  $\omega \in \mathbb{R}$  on a

$$\int g_+(T(y))e^{\theta T(y) + i\omega T(y)} \mu(dy) = \int g_-(T(y))e^{\theta T(y) + i\omega T(y)} \mu(dy)$$

où  $g = g_+ - g_-$ . Conclure en remarquant que ceci implique l'identité de deux lois.

**Exercice 14 (Statistiques libres).** Une statistique  $N$  est dite **libre** (ancillary) si sa loi sous  $P_\theta$  ne dépend pas de  $\theta$ . Une conséquence des résultats montrés ici est que toute statistique minimale  $S$  de la forme  $S = (S_0, N)$  où  $N$  est libre et non-triviale n'est pas complète. Ceci s'applique souvent aux statistiques minimales de dimension supérieure à  $\theta$ , et  $S_0$  aura la dimension de  $\theta$ .

- Montrer que les statistiques  $N$ ,  $(\sum X_i)(\sum Y_i)$ ,  $(N_1 + N_2, N_3 + N_4)$ ,  $(N_1 + N_4, N_2 + N_3)$ ,  $(Y_{(2)} - Y_{(1)}, \dots, Y_{(n)} - Y_{(n-1)})$  et  $Z_{(n)} - Z_{(1)}$  de l'exercice 12 sont libres.
- Montrer que si  $S$  est exhaustive minimale et s'il existe une fonction  $\varphi$  telle que  $\varphi(S)$  soit une statistique libre non-constante, alors il n'existe pas de statistique exhaustive et complète.
- En déduire qu'il n'existe pas de statistique exhaustive et complète dans les exemples de l'exercice 12.
- (Théorème de Basu)** Soit  $S$  et  $N$  deux statistiques. Montrer que si  $S$  est exhaustive et complète et si  $N$  est libre, alors  $S$  et  $N$  sont deux variables indépendantes (pour tout  $\theta$ ).  
*Indication* : Noter que pour toute fonction  $h$  on a  $E_\theta[E_\theta[h(N)|S] - E_\theta[h(N)]] = 0$ .
- En déduire que si les  $Y_i$  sont iid et suivent la loi  $\mathcal{N}(0, \sigma^2)$ , alors la variance empirique est indépendante de la moyenne empirique (considérer la famille  $\mathcal{N}(\theta, \sigma^2)$ ).

**Exercice 15 (Estimateur de Barankin [29]).** Soit  $p_\theta(x)\mu(dx)$  un modèle paramétrique; on ne fait pas l'hypothèse de complétude. On s'intéresse à l'existence, l'unicité et la construction de l'estimateur  $T(x)$  non-biaisé de  $g(\theta)$  dont la variance est minimale sous une mesure de probabilité de la forme  $Q(dx) = q(x)\mu(dx)$  (Barankin considère le cas  $q = p_{\theta_0}$  pour un certain  $\theta_0 \in \Theta$ ).

Soit la condition (C) : les fonctions  $p_\theta(x)/q(x)$  appartiennent à  $L_2(Q)$ . Soit  $H$  le sous-espace de  $L_2(Q)$  engendré par ces fonctions. Soit  $T_0 \in L_2(Q)$  un estimateur non-biaisé de  $g(\theta)$  et  $T$  sa projection orthogonale sur  $H$ . Montrer que sous (C)

- $T \in L_2(Q)$ ,  $T \in L_1(P_\theta)$  pour tout  $\theta \in \Theta$ , puis que  $T$  est non-biaisé
- $T$  est indépendant du choix de  $T_0$  (vérifier que si  $T'$  en est un autre, alors  $T - T' \perp H$ )
- $T$  est l'unique estimateur non-biaisé de  $g(\theta)$  de variance minimale sous  $Q$ .

Si le modèle est complet, l'estimateur  $T^*$  du théorème 20 appartient donc pour tout  $\theta_0$  à l'espace de  $L_2(P_{\theta_0})$  engendré par les fonctions  $p_\theta(x)/p_{\theta_0}(x)$ .

Le défaut de cette approche est l'absence de garantie concernant la variance de l'estimateur sous  $P_\theta$  si  $\theta \neq \theta_0$ . L'exercice 10 p. 85 donnera un argument en faveur de cet estimateur.

**\*Exercice 16 (Estimateur de Barankin : calcul explicite dans deux cas).** On conserve les notations de l'exercice précédent et l'on se met dans le cas  $q = p_{\theta_0}$  pour un certain  $\theta_0 \in \Theta$  (en pratique  $\theta_0$  sera un estimateur préliminaire). On se propose de vérifier que l'estimateur  $T$  présenté est le bon en vérifiant qu'il appartient à  $H$  et est non-biaisé.

- Vérifier que pour l'exemple 3 de l'exercice 12 avec  $g(\theta) = \theta$  on trouve

$$T = \frac{n^{-1}}{2 - \theta_0^2} \left( (\theta_0^2 - 2\theta_0 - 4)N_1 - (\theta_0^2 + 2\theta_0 - 4)N_2 + (2\theta_0^2 + \theta_0 - 2)N_3 - (2\theta_0^2 - \theta_0 - 2)N_4 \right)$$

- On considère le modèle de l'exercice 6 avec une seule observation. On se limite ici au cas  $\theta_0 = 0$ , ce qui n'est pas une restriction en raison des propriétés d'invariance par translation.
  - Soit  $H^\perp$  l'orthogonal de  $H$  dans  $L_2(P_0)$ . Montrer que  $g \in H^\perp$  si et seulement si  $g \in L_2(P_0)$  et  $g(x+1) + g(x) = 0$  (on notera que  $\theta \mapsto u(\theta) = \int g(x)e^{-x^2/2 - \theta x} dx$  est analytique<sup>6</sup>).

5. On aurait pu sinon vérifier l'équation de Cauchy-Riemann  $\partial_v f(z) = i\partial_u f(z)$ ,  $z = u + iv$ , en utilisant le théorème de dérivation sous le signe intégral.

6. On montre facilement que pour  $|h|$  assez petit on a  $u(\theta+h) = \sum \frac{h^n}{n!} \int g(x)e^{-x^2/2 - \theta x} (-x)^n dx$  en utilisant le théorème de Fubini pour permuter  $\int$  et  $\sum$ .

(b) En déduire que les fonctions  $e^{i(2n+1)\pi x}$  forment une partie totale de  $H^\perp$ .

(c) Montrer que

$$T(x) = x - \frac{1}{2} \frac{\sum_n (-1)^n e^{-(x-n-1/2)^2/2}}{\sum_k e^{-(x-k-1/2)^2/2}}$$

*Indication* : Le caractère non-biaisé est immédiat car  $T - x \in H^\perp$  et  $x$  est non-biaisé. Il reste à vérifier que  $T$  est bien orthogonal à  $H^\perp$ .

## III.2 L'estimateur au maximum de vraisemblance

C'est celui qui maximise  $\mathcal{L}(\theta, Y)$ . Il également la propriété importante d'être invariant par changement de variables sur  $\theta$ . C'est un bon choix par défaut mais ses performances peuvent être assez décevantes lorsque l'échantillon est petit, particulièrement lorsque  $\theta$  est vectoriel.

On se placera dans le cas d'un échantillon de v.a.iid  $Y = (Y_1, \dots, Y_n)$  de loi  $P_{\theta_*}$  et

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \mathcal{L}(\theta, Y_i). \quad (\text{III.2})$$

Noter que  $\mathcal{L}$  désigne dans cette partie la log-vraisemblance d'un seul échantillon. Comme cette équation n'a pas forcément une unique solution, on parlera d'*un estimateur au maximum de vraisemblance* pour désigner une solution mesurable de cette équation.

### III.2.1 Convergence presque sûre

Elle est simplement traitée à l'aide des théorèmes sur le minimum de contraste. On donne une variante (hypothèses (a') et (a'')) pour prendre en compte les cas où  $\theta \mapsto p_\theta(y)$  peut s'annuler ce qui rend  $\mathcal{L}(\theta, y)$  infini (cf. la discussion précédant le théorème 11 p. 40) :

#### 21 - THÉORÈME

On suppose  $\Theta$  compact, les  $Y_i$  iid de loi  $P_{\theta_*}$ , et

(a) La fonction  $\mathcal{L}(\theta, y) = \log(p_\theta(y))$  satisfait (GNU) (la loi de  $Y$  est  $P_{\theta_*}$ )

(b) L'application  $\theta \rightarrow P_\theta$  est injective

alors tout estimateur au maximum de vraisemblance (III.2) converge p.s. vers  $\theta_*$ . On a la même conclusion si remplace (a) par :

(a') Pour tout  $A \leq 0$ , la fonction  $(\theta, y) \mapsto \max(A, \log(p_\theta(y)))$  satisfait (GNU)

(a'')  $E_{\theta_*} [|\log(p_{\theta_*}(Y))|] < \infty$ .

On peut également remplacer dans cet énoncé (GNU) par (GNU').

Notons, avant de démontrer ce résultat, que le maximum de vraisemblance est un minimum de contraste avec  $K(\theta, y) = -\log(p_\theta(y))$ ; le contraste moyen est ici  $k(\theta) = D(P_{\theta_*} \| P_\theta) + k(\theta_*)$  où  $D(\cdot \| \cdot)$  est la divergence de Kullback

$$D(Q \| P) = \int \log(q(x)/p(x))q(x)\mu(dx), \quad (\text{III.3})$$

$p$  et  $q$  désignant les densités de ces mesures par rapport à une mesure commune  $\mu$  (p. ex.  $\mu = (P+Q)/2$ ). Il est classique que  $D(Q \| P) \geq 0$ , avec égalité si et seulement si  $P = Q$  (conséquence de l'inégalité de Jensen), ce qui justifie la validité du contraste.

*Démonstration.* Par application des théorèmes 10 et 11 avec  $K(\theta, y) = -\log(p_\theta(y))$ , il ne suffit de vérifier que l'unicité de l'extremum (point (b) du théorème 11) qui va provenir de l'hypothèse (b); en effet, comme  $k(\theta) = D(P_{\theta_*} \| P_\theta) + k(\theta_*)$ , on a bien  $k(\theta) = k(\theta_*)$  seulement si  $P_\theta = P_{\theta_*}$  et l'hypothèse (b) permet de conclure. ■

**Exemple.** Soit le modèle de translation sur  $\mathbb{R}^d$  :  $P_\theta(dy) = p(y-\theta)dy$ . Alors (a') sera satisfait si l'ensemble de discontinuité de  $p$  est de mesure de Lebesgue nulle et si  $p$  est borné.

### III.2.2 Normalité asymptotique

On montrera qu'en dehors des pathologies, sous des hypothèses raisonnables, l'estimateur au maximum de vraisemblance est asymptotiquement optimal avec normalité asymptotique :  $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow \mathcal{N}(0, I(\theta_*)^{-1})$  où  $I(\theta)$  est la matrice d'information de Fisher.

**La matrice d'information de Fisher** est une mesure de la sensibilité de la distribution au paramètre, définie par

$$I_{ij}(\theta) = E_\theta \left[ \frac{\partial \mathcal{L}(\theta, y)}{\partial \theta_i} \frac{\partial \mathcal{L}(\theta, y)}{\partial \theta_j} \right] \quad (\text{III.4})$$

en tout point  $\theta$  intérieur à  $\Theta$  tel que la fonction  $\theta \mapsto \mathcal{L}(\theta, y)$  soit  $P_\theta$ -p.s. différentiable. Si l'on pense au maximum de vraisemblance, c'est la matrice  $I$  du théorème 9.

Le calcul pratique de  $I(\theta)$  fera intervenir deux points essentiels :

1. La formule (III.6), bien que parfois (III.4) soit plus pratique.
2. Le fait que l'information de Fisher calculée sur la base de  $n$  échantillons iid vaut  $n$  fois celle calculée sur la base d'un seul (point 3 après le lemme 24 plus bas).

#### 22 - PROPOSITION

Soit un ouvert  $\mathcal{V} \subset \Theta$ . On suppose que  $\mu$ -p.s. la fonction  $\theta \mapsto \mathcal{L}(\theta, y)$  y est deux fois dérivable (et donc  $\theta \mapsto p_\theta(y)$  également) avec

$$\int \sup_{\theta \in \mathcal{V}} \|\nabla_\theta^2 p_\theta(y)\| \mu(dy) < \infty. \quad (\text{III.5})$$

Alors, pour tout  $\theta_* \in \mathcal{V}$  tel que  $I(\theta_*) < \infty$ ,

$$I(\theta_*) = -E_{\theta_*}[\nabla_{\theta_*}^2 \mathcal{L}(\theta_*, Y)]. \quad (\text{III.6})$$

*Démonstration.* La stratégie est de prendre l'espérance sous  $P_{\theta_*}$  dans l'expression suivante

$$\nabla_{\theta_*}^2 \mathcal{L}(\theta_*, y) = p_{\theta_*}(y)^{-1} \nabla_{\theta_*}^2 p_{\theta_*}(y) - \nabla_{\theta_*} \mathcal{L}(\theta_*, y) \nabla_{\theta_*} \mathcal{L}(\theta_*, y)^T. \quad (\text{III.7})$$

Il suffit de montrer que  $\int \nabla_{\theta_*}^2 p_{\theta_*}(y) \mu(dy) = 0$ . Pour montrer cela, on va utiliser deux fois le théorème de dérivation sous le signe intégral afin de sortir le  $\nabla_{\theta_*}^2$ . L'inégalité de Cauchy-Schwartz conduit à

$$\int \|\nabla_{\theta_*} p_{\theta_*}(y)\| \mu(dy) \leq \left( \int \frac{\|\nabla_{\theta_*} p_{\theta_*}(y)\|^2}{p_{\theta_*}(y)} \mu(dy) \right)^{1/2} \left( \int p_{\theta_*}(y) \mu(dy) \right)^{1/2} = \text{Trace}(I(\theta_*))^{1/2} < \infty$$

et donc, avec (III.5), par le théorème de dérivation sous le signe intégral :

$$\int \nabla_{\theta_*}^2 p_{\theta_*}(y) \mu(dy) = \nabla_{\theta_*} \int \nabla_{\theta_*} p_{\theta_*}(y) \mu(dy). \quad (\text{III.8})$$

Comme pour  $\varepsilon$  petit on a

$$\sup_{\|\theta - \theta_*\| \leq \varepsilon} \|\nabla p_\theta(y)\| \leq \|\nabla p_{\theta_*}(y)\| + \varepsilon \sup_{\theta \in \mathcal{V}} \|\nabla^2 p_\theta(y)\|, \quad (\text{III.9})$$

on peut appliquer de nouveau le théorème de dérivation sous le signe intégral et obtenir que le membre de droite de (III.8) est nul. ■

**La condition de R-régularité.** Il est temps d'introduire cette condition technique [1, 2] qui permet en particulier de permuter dérivation et intégrale dans les démonstrations; elle utilise une définition plus générale de l'information de Fisher :

**23 - DÉFINITION** (R-Régularité)

Une expérience est dite R-régulière si les conditions suivantes sont satisfaites :

- (a) Il existe une fonction  $\nabla_{\theta} p_{\theta}(y)$ , nulle si  $p_{\theta}(y) = 0$ , qui satisfait :
- (i) Pour tout  $\theta_0 \in \Theta$ , avec  $\mu$ -mesure pleine : la fonction  $\theta \mapsto p_{\theta}(y)$  est continument différentiable au voisinage de  $\theta$ , avec gradient  $\nabla_{\theta} p_{\theta}(y)$ , et la fonction  $\theta \mapsto \nabla_{\theta} p_{\theta}(y)$  est continue au point  $\theta_0$ .
  - (ii) Pour  $\mu$ -presque tout  $y$  : pour tous  $\theta$  et  $h$  tels que le segment  $[\theta, \theta + h]$  soit inclus dans  $\Theta$ , alors  $\int_0^1 |\langle h, \nabla_{\theta} p_{\theta+th}(y) \rangle| dt < \infty$  et :

$$p_{\theta+h}(y) - p_{\theta}(y) = \int_0^1 \langle h, \nabla_{\theta} p_{\theta+th}(y) \rangle dt.$$

- (b)  $I(\theta) = \int \nabla_{\theta} p_{\theta}(y) \nabla_{\theta} p_{\theta}(y)^T p_{\theta}(y)^{-1} 1_{p_{\theta}(y) > 0} \mu(dy)$  est finie et continue sur  $\Theta$ .

Si pour  $\mu$ -presque tout  $y$  la fonction  $\theta \mapsto p_{\theta}(y)$  est  $C^1$ , alors (a) est bien entendu satisfait, mais ce jeu d'hypothèses recouvre également des situations  $C^1$  par morceaux; par exemple

$$P_{\theta}(dy) = \frac{1}{2} e^{-|y-\theta|} dy$$

forme une expérience R-régulière avec  $I(\theta) = 1$ ; l'ensemble de mesure pleine du (ii) est  $\mathbb{R} \setminus \{\theta_0\}$ . De même l'expérience  $P_{\theta}(dy) = \frac{3}{2} (1 - |y - \theta|)_+^2 dy$  est R-régulière.

Il ne faut cependant pas croire que  $C^1$  par morceaux suffise, cf. exercice 6 p. 71.

La définition montre que la R-régularité pour un paramètre vectoriel est la R-régularité monodimensionnelle dans chaque direction  $h$  avec une dérivée de la forme  $\langle h, \nabla_{\theta} p_{\theta}(y) \rangle$ .

Le lemme qui suit peut paraître technique mais sera d'un grand secours dans la suite :

**24 - LEMME**

On suppose l'expérience R-régulière. Soit  $f$  une fonction mesurable, et  $U$  un ouvert de  $\mathbb{R}^d$  inclus dans  $\Theta$  tels que

$$E_{\theta}[f(Y)^2] < c < \infty, \quad \theta \in U,$$

alors

$$\nabla_{\theta} E_{\theta}[f(Y)] = \int f(y) \nabla_{\theta} p_{\theta}(y) \mu(dy)$$

et cette quantité est une fonction continue de  $\theta \in U$ .

Ce lemme est une conséquence des résultats du § A.1 (équation (A.9)); noter qu'une expérience R-régulière est DMQU en vertu des théorèmes 57 et 59). Donnons-en une démonstration plus directe.

*Démonstration.* On va montrer d'abord la dérivabilité de  $\theta \mapsto E_{\theta}[f(Y)]$  dans toute direction  $h$  avec la formule attendue. Soit donc  $\theta \in \Theta$ ,  $h$  une direction, et  $t > 0$

$$E_{\theta+th}[f(Y)] = E_{\theta}[f(Y)] + \int f(y) \left( \int_0^t \langle h, \nabla_{\theta} p_{\theta+sh}(y) \rangle ds \right) \mu(dy).$$

On va appliquer le théorème de Fubini ; noter qu'en utilisant l'inégalité de Cauchy-Schwartz,

$$\int_0^t \int |f(y)| \|\nabla p_{\theta+sh}(y)\| \mu(dy) ds \leq \int_0^t E_{\theta+sh}[f(Y)^2]^{1/2} \|I(s)\|^{1/2} ds < \infty.$$

On a donc

$$E_{\theta+th}[f(Y)] = E_{\theta}[f(Y)] + \int_0^t \left( \int f(y) \langle h, \nabla p_{\theta+sh}(y) \rangle \mu(dy) \right) ds$$

et il ne reste plus qu'à montrer la continuité de la fonction

$$\theta \mapsto \int f(y) \nabla p_{\theta}(y) \mu(dy).$$

Soit  $\theta_n$  une suite convergente et les fonctions  $w_n(y) = f(y) \nabla p_{\theta_n}(y)$ . Décomposons  $w_n$  comme produit des fonctions

$$u_n(y) = f(y) \sqrt{p_{\theta_n}(y)}, \quad v_n(y) = 1_{p_{\theta_n}(y) > 0} \frac{\nabla p_{\theta_n}(y)}{\sqrt{p_{\theta_n}(y)}},$$

alors  $u_n(y)$  est borné dans  $L_2$  et  $v_n(y)$  converge dans  $L_2$  en raison point (ii) du lemme 51. Il s'ensuit que  $\int w_n(y) \mu(dy)$  converge bien vers la limite attendue en raison point (i) du lemme 51.

On a donc montré que  $\theta \mapsto E_{\theta}[f(Y)]$  est dérivable dans toutes les directions *avec dérivée continue*, ce qui fait qu'elle est effectivement  $C^1$ . ■

En particulier, on en déduit que si l'expérience est R-régulière :

1.  $E_{\theta}[\nabla_{\theta} \mathcal{L}(\theta, Y)] = 0$  (faire  $f = 1$  dans le lemme).
2. Donc l'information de Fisher est la variance du score  $\nabla_{\theta} \mathcal{L}(\theta, Y)$ .
3. L'expérience sur  $n$  échantillons indépendants  $(p_{\theta}(y_1) \dots p_{\theta}(y_n))_{\theta \in \Theta}$  est également R-régulière. Son information de Fisher est  $n$  fois celle correspondant à un seul échantillon (car la variance de la somme vaut la somme des variances ; les détails sont laissés en exercice).

**La normalité asymptotique** peut se montrer simplement avec le théorème 9, mais un jeu d'hypothèses plus faible est obtenu avec un peu plus d'efforts en utilisant le théorème 12 :

#### 25 - THÉORÈME

Ici  $\Theta \subset \mathbb{R}^d$ . On fait les hypothèses suivantes :

- (i) L'expérience est R-régulière, et  $I(\theta_*)$  est inversible.
- (ii) L'estimateur au maximum de vraisemblance  $\hat{\theta}_n$  converge en probabilité vers  $\theta_*$  intérieur à  $\Theta$ .
- (iii) Et de plus

$$E_{\theta_*} \left[ \sup_{\theta} \frac{\|\nabla p_{\theta}(y)\|^2}{p_{\theta}(y)^2} \right] < \infty. \tag{III.10}$$

Alors on a la convergence en loi

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{Loi} \mathcal{N}(0, I(\theta_*)^{-1}).$$

*Démonstration.* Le résultat s'obtient en appliquant le théorème 12 à

$$K(\theta, Y) = -\mathcal{L}(\theta, Y)$$

$$\dot{K}(y) = \sup_{\theta} \|\nabla p_{\theta}(y)\|^2 p_{\theta}(y)^{-1}.$$

Il ne reste qu'à vérifier (II.31) avec  $J = -I(\theta_*)$ . Ce point un peu technique est une conséquence de théorèmes 54 et 59 de l'appendice A. ■

Noter que  $I(\theta_*)$  intervient à la fois comme mesure de sensibilité ( $J = -I(\theta_*)$ ) et comme mesure de bruit ; dans cette combinaison, la sensibilité l'emporte et l'on a intérêt à avoir  $I(\theta_*)$  la plus grande possible.

On a convergence de tous les moments sous des hypothèses supplémentaires raisonnables (théorème 26 plus bas).

**Calcul pratique de la variance asymptotique.** On préfère généralement l'estimateur sandwich [141] pour la variance de  $\hat{\theta}_n$ ,  $\hat{V}_n = \hat{J}^{-1} \hat{I} \hat{J}^{-1}$ , soit :

$$\hat{V}_n = \left( \sum \nabla^2 \mathcal{L}(\hat{\theta}_n, Y_i) \right)^{-1} \left( \sum \nabla \mathcal{L}(\hat{\theta}_n, Y_i) \nabla \mathcal{L}(\hat{\theta}_n, Y_i)^T \right) \left( \sum \nabla^2 \mathcal{L}(\hat{\theta}_n, Y_i) \right)^{-1}. \quad (\text{III.11})$$

qui ne présuppose pas la véracité du modèle paramétrique mais se base simplement sur le fait que  $\mathcal{L}(\theta_*, Y_i)$  est une fonction de contraste. Voir l'exercice 4 p. 71 pour le cas de la régression linéaire.

**Le cas non-compact.** Si  $\Theta$  n'est pas compact, en vertu des remarques faites au § II.2.2 les théorèmes restent vrais si leurs hypothèses sont satisfaites pour tout compact  $\Theta_0 \subset \Theta$  et que presque sûrement la suite  $\hat{\theta}_n$  reste confinée dans un certain compact pouvant dépendre de la réalisation  $\omega$ .

**Démonstration directe.** On peut aussi en donner une démonstration directe simple : En appliquant la formule de Taylor à l'ordre 2 à la fonction  $\sum_{i=1}^n \mathcal{L}(\theta, Y_i)$ , au point  $\hat{\theta}_n$ , et en exploitant que sa dérivée s'y annule, il vient, pour un certain  $\theta'_n \in [\hat{\theta}_n, \theta_*]$  et sur l'ensemble  $\Omega_n = \{\omega : \hat{\theta}_n \in \mathcal{V}\}$  :

$$\begin{aligned} 2 \sum_{i=1}^n \mathcal{L}(\hat{\theta}_n, Y_i) - \mathcal{L}(\theta_*, Y_i) &= -(\hat{\theta}_n - \theta_*)^T \left( \sum_{i=1}^n \nabla_{\theta'}^2 \mathcal{L}(\theta'_n, Y_i) \right) (\hat{\theta}_n - \theta_*) \\ &= \sqrt{n}(\hat{\theta}_n - \theta_*)^T (I(\theta_*) + r_n) \sqrt{n}(\hat{\theta}_n - \theta_*) \end{aligned}$$

et le reste  $r_n$  tend en loi vers 0 en raison de la proposition 5. La convergence en loi de  $\sqrt{n}(\hat{\theta}_n - \theta_*)$  vers  $\mathcal{N}(0, I(\theta_*)^{-1})$  implique alors le résultat car  $P(\Omega_n) \rightarrow 1$ .

**Remarque : Exhaustivité asymptotique.** On a même montré que

$$p_{\theta_*}(Y_1, \dots, Y_n) = e^{-\frac{\varepsilon_n}{2}(\hat{\theta}_n - \theta_*)^T I(\theta_*)(\hat{\theta}_n - \theta_*) + \varepsilon_n} p_{\hat{\theta}_n}(Y_1, \dots, Y_n),$$

où  $\varepsilon_n$  converge en probabilité vers 0 quand  $n$  tend vers l'infini (et si les  $Y_i$  sont tirés selon  $p_{\theta_*}$ ) ; ceci peut s'interpréter comme de l'exhaustivité asymptotique de  $\hat{\theta}_n$ . Plus spécifiquement, si  $\varepsilon_n(\theta_*, Y)$  était identiquement nul, on aurait  $p_{\theta}(y) = e^{-\frac{\varepsilon}{2}(T(y) - \theta)^T I(\theta)(T(y) - \theta)} q(y)$ , avec  $y = (y_1, \dots, y_n)$  et  $T(y) = \hat{\theta}(y)$ , qui est simplement une famille exponentielle.

**Le théorème de Wilks.** Le théorème 17 permet de tester l'hypothèse  $H_0 : \langle g(\theta_*) = 0 \rangle$  (c.f. § IV.2), en utilisant la statistique de rapport de vraisemblances de l'équation (III.12) : sous réserve de satisfaction des hypothèses, si  $g(\theta_*) = 0$ , alors l'estimateur sous contrainte

$$\hat{\theta}_n^c = \arg \min_{g(\theta)=0} \sum_{i=1}^n \mathcal{L}(\hat{\theta}_n, Y_i)$$

satisfait

$$2 \sum_{i=1}^n \mathcal{L}(\hat{\theta}_n, Y_i) - \mathcal{L}(\hat{\theta}_n^c, Y_i) \longrightarrow \chi_m^2 \quad (\text{III.12})$$

où  $m$  est la dimension de  $g$ , c.-à-d. la différence de dimension effective entre  $\theta$  et  $\theta^c$ .

**Exemple : Test de «  $\theta_* = \theta_0$  », région de confiance.** On a en particulier

$$2 \sum_{i=1}^n \mathcal{L}(\hat{\theta}_n, Y_i) - \mathcal{L}(\theta_*, Y_i) \xrightarrow{P} \chi_d^2. \quad (\text{III.13})$$

On pourra tester au niveau  $\alpha$  l'hypothèse  $H_0 : \langle \theta_* = \theta_0 \rangle$  (c.f. § IV.2) par réfutation de cette dernière si  $2 \sum_{i=1}^n \mathcal{L}(\hat{\theta}_n, Y_i) - \mathcal{L}(\hat{\theta}_0, Y_i)$  est supérieur au quantile  $\chi_d^2(1 - \alpha)$ . L'ensemble

$$\mathcal{R}_\alpha = \left\{ \theta : 2 \sum_{i=1}^n \mathcal{L}(\hat{\theta}_n, Y_i) - \mathcal{L}(\theta, Y_i) \leq \chi_d^2(1 - \alpha) \right\}$$

est une région de confiance pour  $\theta_*$  de niveau asymptotique  $\alpha$ , qui s'ajoute aux régions (II.17) et (II.18).

**Exemple : Test de sphéricité.** Soient des v.a.i.d.d.  $Y_1, \dots, Y_n$  de loi  $\mathcal{N}(\mu, \theta)$ ,  $\theta \in \mathbb{R}^{p \times p}$ . On veut tester si  $\theta = \sigma^2 Id$  pour un certain  $\sigma$  (hypothèse  $H_0$ ). On a classiquement  $\hat{\theta} = \frac{1}{n} \sum (Y_i - \bar{Y})(Y_i - \bar{Y})^T$ ,  $\bar{Y} = \frac{1}{n} \sum Y_i$ , et  $\hat{\theta}^c = \hat{\sigma}^2 Id$ ,  $\hat{\sigma}^2 = Tr(\hat{\theta})/p$ . Le membre de gauche de (III.12) vaut  $S = np \ln(Tr(\hat{\theta})/p) - n \ln(\det(\hat{\theta}))$ , qui est asymptotiquement un  $\chi_{p(p+1)/2-1}^2$  sous  $H_0$ . Le test de niveau 5% réfutera  $H_0$  si  $S$  est supérieur au quantile à 0,95 de ce  $\chi^2$ , i.e.  $q$  t.q.  $P(\chi_{p(p+1)/2-1}^2 > q) = 0,95$ .

**Correction de Bartlett (1937).** Un raffinement de (III.12) conduit à une convergence à vitesse  $n^{-2}$  (au sens d'une certaine mesure d'écart entre les densités des deux membres) si l'on modifie le membre de gauche en le divisant par son espérance et en le multipliant par  $d$  [34]. En dehors du cas du test de Bartlett d'identité de variances, l'utilisation pratique de cette correction est délicate car il faut être capable de calculer l'espérance en question avec une précision suffisante.

**Application : Critère d'Akaike pour la sélection de modèle en vue de la prédiction.** Nous reprenons ici la discussion du § II.3.4 où il s'agissait de sélectionner un modèle dans une *famille emboîtée* en évitant le surajustement, mais ici dans le cas du contraste de vraisemblance. Le  $\chi_d^2$  qui apparaît dans l'équation III.13 n'est autre que le terme  $(\hat{\theta}_n - \theta_*)^T J(\hat{\theta}_n - \theta_*)$  de (II.37). Le critère  $TIC(d)$  devient ici le critère d'Akaike, qui est une estimée du contraste moyen en  $\hat{\theta}$ ,  $k(\hat{\theta})$ ,  $k(\theta) = -2E_{\theta_*}[\mathcal{L}(\theta, Y)]$ , estimée consistant à oublier des termes d'ordre inférieur et à remplacer ce  $\chi_d^2$  par son espérance

$$\boxed{AIC(d) = -2 \sum_i \mathcal{L}(\hat{\theta}_n^d, Y_i) + 2d}$$

Le terme négligé  $2(\chi_d^2 - d)$  est d'ordre  $\sqrt{d}$ , et ses fluctuations vont rendre le résultat aléatoire<sup>7</sup>, le minimiseur  $\hat{d}_n$  de AIC étant typiquement sensiblement plus grand que  $d_*$  (rappelons que  $AIC(d) - AIC(d_*)$  est d'ordre  $n$  pour  $d < d_*$ ).

En raison de ces fluctuations, en vue l'estimation de  $d_*$ , d'autres pénalisations (au lieu de  $2d$ ) ont été proposées, la plus connue étant  $d \log n$ , correspondant au critère BIC (cf. appendice E) dont on montre qu'elle garantit la consistance de  $\hat{d}_n$  lorsque  $n$  tend vers l'infini<sup>8</sup>. En somme, *il faut distinguer le problème de prédiction (maximiser  $l(\hat{\theta}_n^d)$ ,  $l(\theta) = E[\mathcal{L}(\theta, Y_1)]$ ) et celui de la sélection de variables (trouver  $d_*$  s'il existe)*. AIC est donc conçu pour la prédiction et l'estimation tandis que BIC l'est pour la sélection<sup>9</sup>. On trouvera des compléments dans [18].

Pour les situations où le nombre de modèles comparés à dimension fixe croît (modèles non-emboîtés) — p.ex. en sélection de variables pour la régression linéaire avec  $\binom{p}{d}$  modèles de dimension  $d$  — la pénalisation à choisir est plus grande<sup>10</sup>; on trouvera des compléments intéressants dans [9], où le terme de pénalisation proposé pour la prédiction est d'ordre  $2d + 2 \log \nu_d$ , où  $\nu_d$  est le nombre de modèles considérés dont la dimension est  $d$ ; un analogue de BIC y est également proposé :  $d \log n + 2 \log \nu_d$ , voir appendice E.

Mentionnons en guise de conclusion que *d'une part le problème de prédiction est généralement plus simple que celui de la sélection, et d'autre part que l'intérêt pratique de ces procédures de pénalisation apparaît particulièrement pour les situations où la validation croisée<sup>11</sup> est difficile à mettre en place, comme dans le cas d'observations non-indépendantes (séries temporelles), ou de la classification non supervisée. Elles présentent également un intérêt pour les procédures en ligne (p. ex. en analyse de séquences d'image), en raison de leur rapidité en comparaison de la validation croisée.*

7. Voir [132] où ces fluctuations et leurs effets sont étudiées avec précision.

8. Il est montré dans [86] qu'une pénalisation d'ordre  $d \log \log n$  suffit à obtenir la convergence forte de  $\hat{d}_n$  vers  $d_*$ .

9. Voir aussi [147] pour une discussion plus précise et un peu différente.

10. Leo Breiman qualifiait l'usage de AIC dans ce contexte de «quiet scandal in the statistical community» [39].

11. Sur CV et ses variantes, voir p.ex. [47, 40, 131]; voir [96] pour une comparaison avec les méthodes de pénalisation.

### III.2.3 Bornes exponentielles. Convergence des moments

On s'intéresse ici à obtenir des théorèmes plus forts que la simple convergence en loi.

#### 26 - THÉORÈME

En plus de la compacité de  $\Theta \subset \mathbb{R}^d$ , de l'injectivité de  $\theta \rightarrow P_\theta$ , et de la R-régularité de l'expérience, on suppose que  $\Theta$  est la fermeture de son intérieur, lequel possède la propriété de cône (p. 54) et que pour un  $q > d$

$$\sup_{\theta} \int \left( \frac{\|\nabla_{\theta} p_{\theta}(y)\|}{p_{\theta}(y)} \right)^q p_{\theta}(y) \mu(dy) < \infty \quad (\text{III.14})$$

(conséquence de la régularité si  $d = 1$ ). Alors, si  $I(\theta_*)$  est inversible, il existe  $C$  et  $\sigma$  tels que pour tous  $a$  et  $n > 0$

$$P_{\theta_*}(\sqrt{n}\|\hat{\theta}_n - \theta_*\| \geq a) \leq C e^{-a^2/2\sigma^2}. \quad (\text{III.15})$$

Les constantes  $C$  et  $\sigma$  peuvent être choisies indépendantes de  $\theta_*$  du moment que  $\theta_*$  est restreint à un compact intérieur à  $\Theta$  où  $I$  est inversible.

En particulier la suite  $\sqrt{n}\|\hat{\theta}_n - \theta_*\|/\sqrt{\log n}$  est bornée avec probabilité 1.

De plus sous les hypothèses du théorème 25, on a convergence de tout les moments de  $\sqrt{n}(\hat{\theta}_n - \theta_*)$  vers ceux de la gaussienne limite.

*Démonstration.* On va utiliser le théorème 19. Il s'agit de vérifier les deux hypothèses, avec ici

$$K_n(\theta, \omega) = - \sum_{i=1}^n \mathcal{L}(\theta, Y_i).$$

Vérifions (II.64) pour  $\eta = 1/2$ , et sans prendre la partie positive, ce qui se réduit exactement à

$$\int \sqrt{p_{\theta}(y)} \sqrt{p_{\theta_*}(y)} \mu(dy) \leq e^{-\|\theta - \theta_*\|^2/2\sigma^2} \quad (\text{III.16})$$

soit

$$1 - d_H(P_{\theta}, P_{\theta_*})/2 \leq e^{-\|\theta - \theta_*\|^2/2\sigma^2} \quad (\text{III.17})$$

où  $d_H$  désigne la distance de Hellinger

$$d_H(P_{\theta}, P_{\theta'}) = \int (\sqrt{p_{\theta'}(y)} - \sqrt{p_{\theta}(y)})^2 \mu(dy).$$

Il se trouve que la R-régularité implique que  $d_H$  satisfait

$$d_H(P_{\theta}, P_{\theta+h}) = \frac{1}{4} h^T [I(\theta) + \varepsilon_{\theta}(h)] h \quad (\text{III.18})$$

avec pour tout compact  $K$  intérieur à  $\Theta$

$$\limsup_{h \rightarrow 0, \theta \in K} |\varepsilon_{\theta}(h)| = 0. \quad (\text{III.19})$$

La démonstration de ce point est faite à l'appendice A.1 (équation (A.8); noter qu'une expérience R-régulière est DMQU en vertu des théorèmes 57 et 59).

Comme l'application  $\theta \rightarrow d_H(P_{\theta}, P_{\theta_*})$  est continue, non nulle pour  $\theta \neq \theta_*$ , et que  $\Theta$  est compact, (III.18) implique que pour un certain  $\varepsilon$  on a  $d_H(P_{\theta}, P_{\theta_*})/2 \geq \varepsilon \|\theta - \theta_*\|^2$ , d'où (III.17) pour un certain  $\sigma$ . L'uniformité en  $\theta$  dans (III.19) fait que  $\sigma$  peut être rendu indépendant de  $\theta \in K$ .

Passons maintenant à (II.63). Le membre de gauche s'estime comme suit en considérant un exposant  $d < q_0 < q$  et deux exposants conjugués  $p$  et  $r$  avec  $p = q/q_0$ , puis en choisissant  $\eta = 1/2r$  (l'équation (II.64) reste vraie pour cette valeur de  $\eta$ , mais  $\sigma$  est changé) :

$$\begin{aligned} E_\theta \left[ \left\| \sum \nabla \mathcal{L}(\theta, Y_i) \right\|^{q_0} \left( \frac{p_{\theta_*}(Y)}{p_\theta(Y)} \right)^\eta \right] &\leq E_\theta \left[ \left\| \sum \nabla \mathcal{L}(\theta, Y_i) \right\|^{q_1} \right]^{1/p} E_\theta \left[ \left( \frac{p_{\theta_*}(Y)}{p_\theta(Y)} \right)^{\eta r} \right]^{1/r} \\ &\leq C n^{q_0/2} e^{-\|\theta - \theta_*\|^2 \eta / \sigma^2} \end{aligned}$$

la majoration du premier terme venant des théorèmes sur les sommes de variables aléatoires indépendantes, et celle du second venant de (III.16). Le théorème 19 s'applique donc avec  $q_0$  (au lieu de  $q$ ).

Concernant la dernière affirmation, noter que (III.15) implique que tout moment de  $\sqrt{n} \|\widehat{\theta}_n - \theta_*\|$  reste borné quand  $n$  tend vers l'infini. Ceci implique que tout moment de  $\sqrt{n}(\widehat{\theta}_n - \theta_*)$  forme une famille uniformément intégrable (quand  $n$  varie), et par conséquent la convergence en loi implique la convergence des moments. ■

### III.2.4 Quelques exemples

**La famille exponentielle** du § I.2.1. On a :  $I(\theta) = \text{Cov}_\theta(Y) = \nabla_\theta^2 Z$ .

**Exemple de non-unicité.** On observe  $Y_i, i = 1, \dots, n$ , uniformes sur  $[\theta_* - 1/2, \theta_* + 1/2]$ . La vraisemblance vaut 1 sur  $[\max Y_i - 1/2, \min Y_i + 1/2]$  et 0 ailleurs. Tous les points de cet intervalle sont des estimateurs consistants avec une variance d'ordre  $1/n^2$  et le milieu donne une variance  $\simeq 1/(2n^2)$ . Il est super-efficace, comme celui de l'exercice 5 p. 7.

**Échec du maximum de vraisemblance.** Soit des observations  $(Y_1, \dots, Y_n)$  provenant d'un mélange de deux gaussiennes :

$$p_\theta(y) = \frac{1}{2} g_{m, \sigma}(y) + \frac{1}{2} g_{m', \sigma'}(y),$$

où  $g_{m, \sigma}$  est la densité gaussienne de moyenne  $m$  et de variance  $\sigma^2$ , et  $\theta = (m, \sigma, m', \sigma')$ . Dans ce cas, le maximum de vraisemblance n'est pas consistant : la vraisemblance de  $\theta = (Y_1, \sigma, 0, 1)$  tend vers l'infini quand  $\sigma$  tend vers 0.  $\Theta$  n'est pas compact.

La tentative de compactifier  $\Theta$  en autorisant  $\sigma = 0$  échoue car alors il ne s'agit plus d'une expérience statistique : les masses de Dirac font que  $\mu(dy)$  ne peut exister (en étant  $\sigma$ -finie). Si l'on restreint  $\Theta$  à  $\sigma \geq \varepsilon$  tout va bien, sauf que l'on vérifie que sur les petits échantillons le même type de problème apparaîtra avec un maximum de vraisemblance pour  $\sigma = \varepsilon$ .

Voir aussi l'exercice 4 p. 7 où le problème vient du trop grand nombre de paramètres. Pour une discussion générale et des exemples classiques, on pourra consulter la contribution de Le Cam dans [32].

### III.2.5 Exercices et compléments

**Exercice 1 (Loi gamma).** On considère des variables iid  $T_1, \dots, T_n$  de distribution gamma :  $p(t) = \frac{1}{\Gamma(a)} b^{-a} t^{a-1} e^{-t/b}$ . La moyenne et la variance s'expriment  $\mu = ab, \sigma^2 = ab^2$ .

Montrer la normalité asymptotique de l'estimateur au maximum de vraisemblance de  $(a, b)$  et calculer sa variance asymptotique (on posera  $\varphi(a) = \partial_{aa} \log(\Gamma(a))$ ). On pourra soit admettre (III.5), soit le démontrer en utilisant le point 3 suivant le lemme 24 et l'inégalité  $\|\nabla_\theta^2 p_\theta(y)\| \leq p_\theta(y) (\|\nabla_\theta^2 \mathcal{L}(\theta, y)\| + \|\nabla_\theta \mathcal{L}(\theta, y)\|^2)$ .

En déduire, à l'aide du théorème 1, la variance asymptotique de  $\widehat{\mu}_n$  en fonction de  $(\mu, \sigma)$ .

**Exercice 2.** Soit des variables  $(X_i)_{1 \leq i \leq n}$  indépendantes de loi  $\mathcal{B}(1, p)$ . Pour chaque  $X_i$  valant 0, on observe une variable exponentielle  $Y_i$  d'espérance  $1-p$ . Le paramètre inconnu est  $p$ . On notera  $S = \sum X_i$  et  $T = \sum Y_i$ .

1. Calculer  $E[S]$  et  $E[T]$ .
2. Écrire la vraisemblance de l'ensemble (Si l'on veut être rigoureux, le plus simple est considérer que  $Y_i = 0$  si  $X_i = 1$ ). Donner l'estimateur au maximum de vraisemblance.

3. Calculer l'information de Fisher (on supposera les hypothèses nécessaires satisfaites).

**Exercice 3.** En utilisant le lemme 24 montrer que, si  $I(\theta)$  est finie,  $E_\theta[\nabla_\theta \mathcal{L}(\theta, y)] = 0$ .

Montrer également que l'information de Fisher correspondant à l'observation de  $n$  échantillons indépendants (c.-à-d. à la mesure produit sur  $\mathbb{R}^n$ ), est  $n$  fois celle d'un seul.

**Exercice 4 (Estimateur sandwich en régression linéaire).** On considère le modèle de régression linéaire (I.7) avec  $\hat{\theta}_n$  donné par (I.10). On pose  $\hat{u} = Y - X\hat{\theta}_n$ , vecteur des erreurs de prédiction. Montrer que l'estimateur sandwich (III.11) p. 67 vaut ici

$$\hat{V}_n = (X^T X)^{-1} \left( \sum_i \hat{u}_i^2 X_i^T X_i \right) (X^T X)^{-1}.$$

**Exercice 5 (Régression logistique).** On considère le modèle (I.8) p.9. On suppose que la suite  $(X_i, Y_i)$  est iid. Étudier la normalité asymptotique de l'estimateur au maximum de vraisemblance.

**Exercice 6 (Une expérience non régulière<sup>12</sup>).** Soit le modèle  $P_\theta = \mathcal{B}(1, \max(\theta, 2\theta - \frac{1}{2}))$ ,  $\theta \in (0, \frac{3}{4})$ . Vérifier que  $\theta \mapsto p_\theta(x)$  est, pour tout  $x$ ,  $C^1$  par morceaux, mais que l'information de Fischer n'est pas correctement définie en  $\theta = \frac{1}{2}$ . Vérifier que le lemme 24 ne s'applique pas si  $f(x) = 1_{x=1}$ .

**Exercice 7 (Maximum de vraisemblance partielle).** On reprend l'exercice 14 p. 28. Les paires  $(X_i, Y_i)$  sont désormais iid de densité  $p_{\theta_*}(x, y)$ . Exprimer la variance asymptotique de  $\hat{\theta}$  en fonction seulement des informations de Fisher  $I_{X,Y}(\theta_*)$  et  $I_Y(\theta_*)$  de  $p_{\theta_*}(x, y)$  et  $p_{\theta_*}(y)$ , et comparer à celle de l'estimateur au maximum de vraisemblance.

**\*Exercice 8 (Vraisemblance, conditionnement, et monotonie de l'information).** Soit  $(P_\theta = p_\theta(y)\mu(dy))_{\theta \in \Theta}$  un modèle d'expérience statistique R-régulière pour une variable  $Y$ . On suppose que  $\mu$  est une probabilité. Soit  $\varphi$  une fonction mesurable et  $Z = \varphi(Y)$ .

1. Soit  $\nu$  la loi image de  $\mu$  par  $\varphi$  et la fonction

$$q_\theta(Z) = E_\mu[p_\theta(Y)|Z].$$

Montrer que si  $Y$  suit la loi  $P_\theta$ , alors  $Z$  suit la loi  $q_\theta(z)\nu(dz)$ .

*Indication :* Montrer que pour toute fonction borélienne bornée  $\int f(z)q_\theta(z)\nu(dz) = E_\theta[f(Z)]$ , par exemple en ramenant les deux termes en des intégrales en  $\mu(dy)$  puis en exploitant la définition de  $q_\theta$ .

2. Montrer que pour toute fonction mesurable bornée  $f$  on a

$$E_\theta[f(Y)|Z] = q_\theta(Z)^{-1} E_\mu[f(Y)p_\theta(Y)|Z].$$

3. Dédire des deux points précédents que si  $\mathcal{L}_p(\theta, Y) = \log p_\theta(Y)$  et  $\mathcal{L}_q(\theta, Z) = \log q_\theta(Z)$  on a

$$\nabla_\theta \mathcal{L}_q(\theta, Z) = E_\theta[\nabla_\theta \mathcal{L}_p(\theta, Y)|Z].$$

4. En déduire que si  $J(\theta)$  désigne l'information de Fisher liée au deuxième modèle on a  $J(\theta) \leq I(\theta)$ .

**Exercice 9 (Statistiques libres et information).** Soit  $N$  une statistique libre (exercice 14 p. 62).

1. Montrer que l'estimateur au maximum de vraisemblance est inchangé si l'on considère la loi conditionnelle à  $N$  au lieu de la loi globale (c.-à-d.  $E_\theta[. | N]$  au lieu de  $E_\theta[.]$ ). Montrer que  $E_\theta[\nabla_\theta \mathcal{L}(\theta, Y) | N] = 0$ .

2. Que penser de l'information de Fisher conditionnelle  $-E_\theta[\nabla_\theta^2 \mathcal{L}(\theta, Y) | N]$  pour estimer la variance? Comparer dans le cas du premier exemple de l'exercice 12 p. 61.

3. Étudier également la question de la variance conditionnelle de  $\hat{\theta}$  dans le cas d'une suite  $Y_i$  iid  $\mathcal{U}([\theta - 1, \theta + 1])$  avec  $\hat{\theta} = \frac{1}{2}(M + m)$ ,  $N = M - m$ ,  $M = \max(Y_i)$ ,  $m = \min(Y_i)$ .

4. Dans le cas où plusieurs statistiques libres existent (voir exercice 14 p. 62 point 1), le choix de  $N$  peut poser problème et Cox [56] propose de choisir celle pour laquelle l'information conditionnelle est la plus grande; pourquoi?

5. Si le nombre d'échantillons est grand, on observe que l'information conditionnelle est proche de l'information moyenne; vérifier ceci sur l'exemple 1 de l'exercice 12 p. 61.

12. Exemple proposé par Arnaud Guyader.

### III.3 Borne de Cramér-Rao. Efficacité

Le but de ce paragraphe est d'approfondir les questions engagées au § III.1, en vue de définir ce qu'est le meilleur estimateur. Il est intéressant de commencer par la borne de Cramér-Rao <sup>13</sup>.

#### III.3.1 Borne de Cramér-Rao

27 - THÉORÈME

On suppose l'expérience R-régulière avec  $I > 0$  sur  $\Theta$ ; alors pour tout estimateur non-biaisé  $\hat{g}$  d'une fonction  $C^1$  vectorielle  $g(\theta)$ , tel que  $E_\theta[|\hat{g}|^2] < c < \infty$ ,  $\theta \in \Theta$ , la variance d'estimation est minorée comme suit :

$$E_\theta \left[ (\hat{g} - g(\theta))(\hat{g} - g(\theta))^T \right] \geq \nabla g(\theta) I(\theta)^{-1} \nabla g(\theta)^T$$

où chaque ligne de  $\nabla g$  contient le gradient de la coordonnée correspondante.

*Remarques.* Cette borne implique que dans le cas de  $n$  échantillons indépendants, on a pour tout estimateur  $\hat{\theta}_n$  de  $\theta$  :

$$nE_\theta \left[ (\hat{\theta}_n - \theta)(\hat{\theta}_n - \theta)^T \right] \geq I(\theta)^{-1}$$

où l'information de Fischer est calculée sur la base d'un seul échantillon.

On observe alors que si l'estimateur  $\hat{\theta}_n$  est optimal au sens où le membre de gauche converge vers  $I(\theta)^{-1}$  quand  $n \rightarrow \infty$ , alors  $g(\hat{\theta}_n)$  le sera également (cf. la delta-method p. 18).

*Démonstration.* En dérivant en  $\theta$  l'identité

$$g(\theta) = \int \hat{g}(y) p_\theta(y) \mu(dy),$$

il vient, en vertu du lemme 24, et du point 1 qui le suit :

$$\nabla g(\theta) = \int \hat{g}(y) \nabla p_\theta(y) \mu(dy) = E_\theta \left[ (\hat{g}(Y) - g(\theta)) \frac{\nabla p_\theta(Y)}{p_\theta(Y)} \right]. \quad (\text{III.20})$$

On peut alors directement conclure en utilisant l'inégalité de Cauchy-Schwarz matricielle

$$E[XZ^T] E[ZZ^T]^{-1} E[ZX^T] \leq E[XX^T] \quad (\text{III.21})$$

avec  $X = \hat{g}(Y) - g(\theta)$  et  $Z = \frac{\nabla p_\theta(Y)}{p_\theta(Y)}$  (vecteur colonne). L'équation (III.21) se démontre en notant que la différence des deux membres de (III.21) est  $E(UU^T)$  où  $U = X - E[XZ^T]E[ZZ^T]^{-1}Z$ . ■

**Exemple : estimation d'une fréquence dans du bruit.** Soit les observations

$$y_k = a \cos(k\omega + \varphi) + u_k, \quad u_k \sim \mathcal{N}(0, \sigma^2), \quad k = 1, \dots, n.$$

$\theta = (\omega, a, \varphi, \sigma^2)$  est inconnu, et  $0 < \omega < \pi$ . La log-vraisemblance vaut :

$$\mathcal{L}(Y, \theta) = -\frac{1}{2\sigma^2} \sum_k (y_k - a \cos(\omega k + \varphi))^2 - n \log(\sigma).$$

Supposons pour simplifier qu'il s'agisse d'un problème de modulation de fréquence pour lequel seul  $\omega$  est inconnu, c.-à-d.  $\theta = \omega$ . On a sous  $P_\omega$

$$\frac{d}{d\omega} \mathcal{L}(Y, \omega) = -\frac{1}{\sigma^2} \sum_k k a \sin(\omega k + \varphi) (y_k - a \cos(\omega k + \varphi)) = -\frac{a}{\sigma^2} \sum_k k \sin(\omega k + \varphi) u_k.$$

13. Publiée par Cramér en 1946 et par Rao en 1945, elle se trouve déjà dans le cours de statistique mathématique à l'Institut Henri Poincaré fait par Fréchet pendant l'hiver 1939-1940, publié en 1943 [77].

Puis

$$E[(\partial_\omega \mathcal{L}(Y, \omega))^2] = \frac{a^2}{\sigma^2} \sum_k k^2 \sin^2(k\omega + \varphi) = \frac{a^2}{2\sigma^2} \sum_k k^2 (1 - \cos(2k\omega + 2\varphi)).$$

En exploitant que  $4 \sum k^2 \cos(2k\omega + 2\varphi) = -\partial_{\omega\omega} \sum \cos(2k\omega + 2\varphi)$  et les identités classiques, on obtient une formule explicite assez compliquée. Pour avoir une idée de la borne, on peut faire l'approximation correspondant à la valeur moyenne si  $\varphi$  est uniforme sur  $[0, 2\pi]$  :

$$E[(\partial_\omega \mathcal{L}(Y, \omega))^2] \simeq \frac{a^2}{2\sigma^2} \sum k^2 \simeq \frac{a^2}{6\sigma^2} n^3.$$

On voit que  $\omega$  n'intervient pas, contrairement au rapport signal sur bruit  $a/\sigma$ , et que la dépendance en  $n$  est  $n^3$ .

**Un exemple d'estimateur super-efficace** (la borne est violée) est donné à l'exercice 5 p. 7. La condition de régularité de l'expérience n'y est pas vérifiée

### III.3.2 Le cas biaisé : Inégalité de van Trees

Il existe une borne dans le cas biaisé, obtenue au prix d'une moyennisation en  $\theta$  sur un ouvert arbitrairement petit, c'est l'inégalité de van Trees<sup>14</sup>. On va se donner une densité de probabilité  $\pi$  sur  $\Theta$  et l'on considèrera la mesure de probabilité ainsi construite sur l'espace produit :

$$E[f(Y, \theta)] = \iint f(y, \theta) p_\theta(y) \mu(dy) \pi(\theta) d\theta = \iint E_\theta[f(Y, \theta)] \pi(\theta) d\theta$$

(noter que  $E$  n'a pas d'indice). Le risque sera alors calculé sur la base de cette espérance  $E[r(\theta, \hat{\theta}(Y))] = \int R_\theta(\hat{\theta}) \pi(d\theta)$  (cf. équation (III.22)) ; il peut être vu comme un risque bayésien mais ce n'est pas vraiment l'objet ici : cette espérance sera utilisée comme une borne inférieure pour le maximum de  $R_\theta(\hat{\theta})$  sur le support (choisi petit) de  $\pi$ .

L'inégalité de van Trees est très importante car elle permet également d'obtenir des bornes utiles en estimation non-paramétrique en observant que la borne inférieure pour un problème non-paramétrique est supérieure à celle obtenue pour tout problème paramétrique contenu dans la classe considérée. Un exemple est donné en exercice 5 p. 77.

#### 28 - THÉORÈME (INÉGALITÉ DE VAN TREES)

On suppose l'expérience  $\mathbb{R}$ -régulière sur  $\Theta \subset \mathbb{R}^d$ . Soit  $\pi$  une densité  $C^1$  sur  $\mathbb{R}^d$  à support compact inclus dans  $\Theta$ . Soit  $g$  une fonction vectorielle  $C^1$  définie sur  $\Theta$ . On pose

$$\begin{aligned} \bar{I} &= \int I(\theta) \pi(\theta) d\theta \\ I_\pi &= \int \frac{\nabla \pi(\theta) \nabla \pi(\theta)^T}{\pi(\theta)} d\theta \\ G &= \int \nabla g(\theta) \pi(\theta) d\theta \end{aligned}$$

(chaque ligne de  $\nabla g$  contient le gradient de la coordonnée correspondante). Soit  $y \mapsto \hat{g}(y)$  un estimateur. On suppose que les quantités ci-dessus sont finies et que

$$E[\|\hat{g}(Y)\|^2 + \|g(\theta)\|^2] < \infty.$$

Alors

$$E[(\hat{g}(Y) - g(\theta))(\hat{g}(Y) - g(\theta))^T] \geq G(\bar{I} + I_\pi)^{-1} G^T. \quad (\text{III.22})$$

14. Pour une discussion des extensions possibles consulter [80].

*Remarques.* Dans le cas de  $n$  observations,  $\bar{I}$  doit simplement être multipliée par  $n$  et le reste est inchangé ; le terme  $I_\pi$  a alors une contribution très faible. Si  $g(\theta) = \theta$  alors  $G = Id$ .

*Démonstration.* Tout va se déduire de l'identité suivante où  $\nabla$  désigne le gradient en  $\theta$  (vecteur ligne) :

$$\iint (\hat{g}(y) - g(\theta)) \left( \nabla p_\theta(y) \pi(\theta) + p_\theta(y) \nabla \pi(\theta) \right) \mu(dy) d\theta = \int \nabla g(\theta) \pi(\theta) d\theta. \quad (\text{III.23})$$

Cette identité est immédiate sous des hypothèses plus fortes de régularité par une simple intégration par parties dans l'intégrale en  $\theta$  puisque le deuxième terme de l'intégrande est le gradient de  $p_\theta(y) \pi(\theta)$ . Pour traiter le cas général, notons avant tout que l'on a bien intégrabilité, comme conséquence immédiate des hypothèses et de l'inégalité de Cauchy-Schwarz, ce qui va permettre d'utiliser le théorème de Fubini-Tonelli. On va maintenant développer l'intégrande, ce qui donne quatre termes ; le premier vaut, en vertu du lemme 24

$$\begin{aligned} \iint \hat{g}(y) \nabla p_\theta(y) \pi(\theta) \mu(dy) d\theta &= \int \left( \nabla \int \hat{g}(y) p_\theta(y) \mu(dy) \right) \pi(\theta) d\theta \\ &= - \int \left( \int \hat{g}(y) p_\theta(y) \mu(dy) \right) \nabla \pi(\theta) d\theta \end{aligned}$$

qui est l'opposé de l'autre terme mettant en jeu  $\hat{g}$  ; cette partie de l'intégrale est donc nulle. Concernant le terme avec  $g(\theta)$ , il conduit bien au membre de droite de (III.23) puisque  $\int \nabla p_\theta(y) \mu(dy) = 0$  et  $\int p_\theta(y) \mu(dy) = 1$ . (III.23) est bien démontré.

On va maintenant appliquer l'inégalité de Cauchy-Schwarz matricielle (III.21) sous la mesure  $\nu(dy, d\theta) = p_\theta(y) \pi(\theta) \mu(dy) d\theta$  avec

$$X = \hat{g}(y) - g(\theta), \quad Z = \frac{\nabla p_\theta(y)}{p_\theta(y)} + \frac{\nabla \pi(\theta)}{\pi(\theta)}.$$

On a donc

$$\nu(XZ^T) \nu(ZZ^T)^{-1} \nu(ZX^T) \leq \nu(XX^T). \quad (\text{III.24})$$

Comme sur l'ensemble  $A_\theta$  de régularité de  $p_\theta$  (cf. Définition 23)  $\nabla p_\theta(y) = 0$  si  $p_\theta(y) = 0$ , en convenant que  $\frac{0}{0} = 0$  dans la définition de  $Z$ , on a bien que le membre de gauche de (III.23) vaut  $\nu(XZ^T)$ . Calculons  $\nu(ZZ^T)$ , on obtient quatre termes en développant :

$$\begin{aligned} \nu(ZZ^T) &= \int \left( \frac{\nabla p_\theta(y) \nabla p_\theta(y)^T}{p_\theta(y)} \pi(\theta) + p_\theta(y) \frac{\nabla \pi(\theta) \nabla \pi(\theta)^T}{\pi(\theta)} \right. \\ &\quad \left. - \nabla \pi(\theta) \nabla p_\theta(y)^T - \nabla p_\theta(y) \nabla \pi(\theta)^T \right) \mu(dy) d\theta \\ &= \bar{I} + I_\pi - 0 - 0 \end{aligned}$$

car les deux derniers termes sont nuls en vertu du lemme 24 ; on obtient alors le résultat en remplaçant dans (III.24).  $\blacksquare$

**Borne minimax asymptotique.** On a donc dans le cas scalaire, pour  $n$  observations indépendantes

$$\int E_\theta [n(\hat{\theta}_n - \theta)^2] \pi(\theta) d\theta \geq \left( \int I(\theta) \pi(\theta) d\theta + \frac{1}{n} I_\pi \right)^{-1}. \quad (\text{III.25})$$

En raison du facteur  $\frac{1}{n}$  et du caractère arbitraire de  $\pi$ , cette propriété peut servir de justification à l'utilisation brute de la matrice de Fisher comme point de comparaison, même pour les estimateurs biaisés ; nous allons voir comment cette affirmation peut être rendue plus précise. Si  $\pi$  a son support dans un petit ouvert  $U \subset \Theta$  on obtient en majorant trivialement le membre de gauche de (III.25)

$$\sup_{\theta \in U} E_\theta [n(\hat{\theta}_n - \theta)^2] \geq \left( \int I(\theta) \pi(\theta) d\theta + \frac{1}{n} I_\pi \right)^{-1}.$$

En prenant la liminf sur  $n$  puis en faisant tendre  $\pi$  vers une masse de Dirac au point de plus mauvaise information de Fisher de  $U$ , il vient la borne minimax asymptotique

$$\liminf_n \sup_{\theta \in U} E_{\theta} [n(\widehat{\theta}_n - \theta)^2] \geq \sup_{\theta \in U} I(\theta)^{-1}. \quad (\text{III.26})$$

Attention, cette borne est valide pour tout ouvert  $U$ , aussi petit soit-il, mais malheureusement pas pour un point.

### III.3.3 Efficacité.

Une première définition d'efficacité d'un estimateur  $\widehat{\theta}$  serait de dire que  $\widehat{\theta}$  est efficace si

$$E_{\theta} [(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T] = I(\theta)^{-1} \quad \text{pour tout } \theta \in \Theta.$$

Un théorème classique (exercice 3 p. 77) implique que tout estimateur efficace non-biaisé n'existe que dans le cadre d'un modèle exponentiel et pour  $\theta = E[X]$ . Il est en particulier obtenu par maximum de vraisemblance. Les estimateurs efficaces en ce sens n'existent donc qu'exceptionnellement ; même un UMVU généralement n'atteindra pas cette borne. En revanche des estimateurs biaisés peuvent parfois faire mieux.

Cette définition n'est donc pas très satisfaisante. On s'intéresse alors aux estimateurs asymptotiquement efficaces : une possibilité serait de considérer qu'une suite d'estimateurs  $\widehat{\theta}_n = \widehat{\theta}_n(Y_1, \dots, Y_n)$  est asymptotiquement efficace si

$$\rho_n(\theta) = nE_{\theta} [(\widehat{\theta}_n - \theta)(\widehat{\theta}_n - \theta)^T] \longrightarrow I(\theta)^{-1} \quad \text{pour tout } \theta \in \Theta. \quad (\text{III.27})$$

Une variante est de demander la convergence en loi

$$\sqrt{n}(\widehat{\theta}_n - \theta) \longrightarrow \mathcal{N}(0, I(\theta)^{-1}) \quad \text{pour tout } \theta \in \Theta. \quad (\text{III.28})$$

Revenons à l'exemple de l'observation d'une variable  $\mathcal{B}(n, \theta)$  évoqué au § III.1.1. L'estimateur minimax admissible de l'introduction,  $\widehat{\theta}_{MA}$  a un risque constant égal à  $(1 + \sqrt{n})^{-2}/4$ , risque à comparer avec  $n^{-1}\theta(1 - \theta)$  pour le maximum de vraisemblance  $\widehat{\theta}_{MV}$  qui satisfait (III.27) contrairement à  $\widehat{\theta}_{MA}$ . Cette dernière mesure d'efficacité conduit donc à un choix plus raisonnable que l'approche minimax admissible.

Hodges a alors mis en évidence un estimateur biaisé qui fera strictement mieux que (III.27) ou (III.28) au point  $\theta = 0$  qu'un estimateur  $\widehat{\theta}_n$  général asymptotiquement normal ; cet estimateur consiste à remplacer  $\widehat{\theta}_n$  par 0 si  $|\widehat{\theta}_n| < n^{-1/4}$ , cf. exercice 2 p. 77 ; le défaut de cet estimateur est que la convergence n'est plus uniforme en  $\theta$  : si  $\theta_*$  est proche de 0,  $\widehat{\theta}_n$  sera longtemps projeté à tort sur zéro, introduisant une erreur d'ordre  $n^{-1/4}$  pour un nombre arbitraire de valeurs de  $n$ .

On a donc alors cherché des définitions plus compliquées qui permettent d'éliminer l'estimateur de Hodges, considéré comme non satisfaisant, bien que super-efficace au sens de la définition précédente, et de conduire à  $I(\theta)^{-1}$  comme borne inférieure absolue atteignable dans les cas réguliers. Les statisticiens sont alors arrivés à la conclusion que, si l'on se base sur (III.27), il faut définir l'efficacité de  $\widehat{\theta}_n$  comme la convergence *uniforme* de  $\rho_n(\cdot)$  vers  $I(\cdot)$ . En effet, une conséquence de (III.26) est que si l'expérience est régulière, alors une convergence uniforme ne peut se faire que vers une limite partout supérieure à  $I(\cdot)$ . On peut montrer que sous des hypothèses générales plus fortes que la R-régularité, l'estimateur au maximum de vraisemblance est efficace, et que les estimateurs efficaces sont équivalents à celui du maximum de vraisemblance au sens où  $\sqrt{n}(\widehat{\theta}_n - \widehat{\theta}_n^{MV}) \xrightarrow{P} 0$  [13], ainsi qu'aux estimateurs bayésiens.

Ces idées ont été développées par Hajek et Le Cam dans les années 1960-1970<sup>15</sup>. Elles les ont conduits à la *théorie asymptotique locale*, qui porte davantage sur les limites en loi, dont un des résultats importants est le suivant, démontré au § A.2 (théorème 61), obtenu lorsque l'on se restreint à la classe des estimateurs

15. On en trouvera une description approfondie dans les chapitres 7 et 8 de [21], ou le chapitre 2 de [1], ou le chapitre I.10 de [13]

réguliers au sens donné dans l'énoncé :

29 - THÉORÈME (THÉORÈME DE CONVOLUTION DE HAJEK. 1970)

On suppose l'expérience R-régulière. Si  $\hat{\theta}_n$  est une suite d'estimateurs régulière au point  $\theta_*$  au sens où pour tout  $h \in \mathbb{R}^d$

$$\sqrt{n} \left( \hat{\theta}_n - \theta_* - \frac{h}{\sqrt{n}} \right) \xrightarrow{Loi} X, \text{ sous } P_{\theta_* + h/\sqrt{n}} \quad (\text{III.29})$$

où la loi de  $X$  ne dépend pas de  $h$ . Alors la variance de  $X$  est supérieure à  $I(\theta_*)^{-1}$ . Plus précisément, en désignant le score  $S_n(\theta) = \frac{1}{\sqrt{n}} \sum_k \frac{\nabla p_\theta(Y_i)}{p_\theta(Y_i)}$ , on a

$$(\sqrt{n}I(\theta_*)(\hat{\theta}_n - \theta_*) - S_n(\theta_*), S_n(\theta_*)) \xrightarrow{Loi} (Y, Z)$$

où  $Y$  et  $Z$  sont indépendantes et  $Z \sim \mathcal{N}(0, I(\theta_*))$ . La variance de  $X = I(\theta_*)^{-1}(Y + Z)$  est donc supérieure à  $I(\theta_*)^{-1}$  et elle vaut  $I(\theta_*)^{-1}$  si et seulement si

$$\sqrt{n}I(\theta_*)(\hat{\theta}_n - \theta_*) - S_n(\theta_*) \xrightarrow{Loi} 0. \quad (\text{III.30})$$

Pour l'existence d'un estimateur satisfaisant (III.30), voir le théorème 63, dû à Le Cam. On montre sans trop de difficulté que sous les hypothèses du théorème 25, l'estimateur au maximum de vraisemblance satisfait (III.29) (utiliser le théorème 44 avec  $T_n = \hat{\theta}_n - \theta_*$ ; noter que LAN est satisfait en raison de la R-régularité (théorème 60)).

L'idée de Le Cam a donc été de considérer comme efficace tout estimateur convergeant en loi vers  $\mathcal{N}(0, I(\theta_*)^{-1})$  mais en se restreignant à la classe des estimateurs satisfaisant (III.29). La force de ce théorème est qu'il est vrai dans le cadre général de l'hypothèse ULAN p. 116 (théorème 61), i.e. sans se restreindre au cas des suites d'observations indépendantes. Des résultats du même type se trouvent montrés plus directement au chapitre 6.2 de [10].

Un bon bilan des aspects théoriques de ces questions se trouve au chapitre 2 de [1].

**Développements récents.** Il apparaît qu'en vue de l'estimation d'un paramètre  $\theta$  de grande dimension (ce qui revient à dire que l'on dispose, en proportion, de peu d'observations), par exemple l'inverse de la matrice de covariance d'un modèle gaussien [112], les estimateurs biaisés sont sensiblement meilleurs (lasso, ridge...); ceci est à rapprocher de l'effet James-Stein, et de l'estimation non-paramétrique classique. Dans ces situations de grande dimension, l'estimateur au maximum de vraisemblance est toujours insatisfaisant.

Il se trouve aussi que des estimateurs à la Hodge sont parfois préférés en pratique (SCAD, Adaptive lasso...), car, avec grande probabilité, ils estiment à zéro les paramètres nuls, pour un prix à payer (dû à l'annulation, par l'estimation, de paramètres proches de zéro) considéré comme faible<sup>16</sup>. Il est démontré que forcément, ces estimateurs ne sont pas efficaces au sens minimax local [146, 104], même si leur variance asymptotique peut être satisfaisante<sup>17</sup> comme l'est celle de l'estimateur de Hodge (sur cet effet trompeur, voir [104] qui montre bien l'intérêt de l'approche minimax locale). Dans ce fait que des estimateurs différents doivent être utilisés pour des objectifs différents (estimer zéro les paramètres nuls ou minimiser le risque), on retrouve un phénomène plus anciennement remarqué expérimentalement qui est l'opposition AIC/BIC mentionnée p. 68.

### III.3.4 Exercices et compléments

**Exercice 1.** Soit la famille paramétrique  $\mathcal{N}(\theta, \sigma^2)$ ,  $\sigma$  est connu,  $\theta \in \mathbb{R}$ . On dispose d'une seule observation  $Y$ . Montrer en utilisant la borne de Van Trees que pour tout estimateur  $\hat{\theta} = g(Y)$  on a  $E[(\hat{\theta} - \theta)^2] \geq \sigma^2$ .

<sup>16</sup>. Attention, le prix est même nul si l'on suppose que les paramètres non nuls sont distants de zéro; c'est la beta-min condition [46]. Tout le problème ici réside dans les tout petits coefficients.

<sup>17</sup>. C'est le cas de SCAD et de *Adaptive lasso*. Concernant ce dernier voir [148].

*Indication* : On pourra choisir pour  $\pi$  une loi gaussienne centrée.

**Exercice 2 (L'estimateur de Hodges).** On considère un estimateur  $\widehat{\theta}_n$  qui est asymptotiquement efficace, au sens de la convergence en loi de  $\sqrt{n}(\widehat{\theta}_n - \theta)$  vers  $\mathcal{N}(0, I(\theta)^{-1})$ . Montrer que l'estimateur qui vaut 0 si  $|\widehat{\theta}_n| \leq n^{-1/4}$  et sinon  $\widehat{\theta}_n$ , a des performances asymptotiques meilleures si  $\theta_* = 0$  et identiques sinon.

**Exercice 3 (Borne de Cramér-Rao et modèles exponentiels).** On considère le modèle exponentiel (I.3) et  $\widehat{\theta}$  désigne l'estimateur du maximum de vraisemblance. On s'intéresse d'abord au cas  $n = 1$ . Démontrer que  $\nabla Z(\widehat{\theta})$  est un estimateur de  $\nabla Z(\theta) = E_\theta[Y]$  qui atteint la borne de Cramér-Rao. Montrer ensuite pourquoi le cas  $n > 1$  peut être vu comme un cas particulier du cas  $n = 1$ , et déduire le résultat pour  $n$  quelconque.

Considérons la réciproque : *Tout estimateur non-biaisé qui atteint la borne de Cramér-Rao est associé à une famille exponentielle non nécessairement standard (cf. p.8) et estime  $E[T(Y)]$ .* On se placera en dimension 1 pour simplifier et l'on fera facilement une démonstration informelle en regardant ce qu'implique le cas d'égalité dans les inégalités de la démonstration de la borne de Cramér-Rao<sup>18</sup>.

**Exercice 4 (Borne de Barankin [29]).** Soit  $\theta_1, \dots, \theta_n \in \Theta$  et  $c_1, \dots, c_n \in \mathbb{R}$ . Utiliser la relation  $Var(T) \geq Cov(T, V)^2 / Var(V)$ , avec  $V = p_\theta(X)^{-1} \sum_{i=1}^n c_i p_{\theta_i}(X)$  pour montrer que si  $T$  est un estimateur non-biaisé de  $g(\theta)$  alors

$$Var_\theta(T) \geq \frac{|\sum c_i (g(\theta_i) - g(\theta))|^2}{Var_\theta(V)} \quad (\text{III.31})$$

(on pourra vérifier que la borne de Cramér-Rao s'obtient de même avec  $V = \partial_\theta \log(p_\theta)$ ). En déduire la borne de Chapman-Robbins

$$Var_\theta(T) \geq \sup_h \frac{(g(\theta + h) - g(\theta))^2}{\int \frac{p_{\theta+h}^2(x)}{p_\theta(x)} \mu(dx) - 1}.$$

En déduire la borne de Cramér-Rao dans le cas monodimensionnel.

Barankin montre que le sup du membre de droite de (III.31) sur tous les  $c_i$  et les  $\theta_i$  correspond à la variance de l'estimateur de l'exercice 15 p. 62 avec  $q = p_\theta$ , si la condition (C) est satisfaite. Si ce sup est infini, il n'existe pas d'estimateur non-biaisé de variance finie en  $\theta$ .

**Exercice 5 (Inégalité de van Trees et bornes inférieures en estimation non-paramétrique).** Soit le modèle paramétrique de densité sur  $J = [-1/2, 1/2]$  :

$$p_\theta(x) = \theta w(x) + 1_J(x)$$

où  $w$  est une fonction donnée, à support dans  $J$ , d'intégrale nulle telle que  $|w(x)| \leq 1/2$ . Soit un estimateur  $\widehat{g}$  de  $g(\theta) = p_\theta(0)$ , basé sur  $n$  échantillons indépendants tirés sous  $p_\theta$ . Soit  $\pi(\theta)$  une fonction de classe  $C^1$ , positive, de support borné contenu dans  $\Theta = [0, 1]$ .

1. Montrer, en utilisant l'inégalité de van Trees, que

$$\sup_{\theta \in \Theta} E_\theta [(\widehat{g}(Y) - p_\theta(0))^2] \geq w(0)^2 \left( 2n \|w\|_2^2 + I_\pi \right)^{-1}, \quad \|w\|_2^2 = \int_{-1/2}^{-1/2} w(x)^2 dx.$$

*Indication* : On minorera le sup par l'espérance sous  $\pi$ .

2. Montrer que si  $w$  prend la forme  $w(x) = \frac{1}{M} \varphi(Mx)$ , où  $M > 1$  et  $\varphi$  satisfait les mêmes conditions que  $w$  précédemment, alors la borne devient

$$\sup_{\theta \in \Theta} E_\theta [(\widehat{g}(Y) - p_\theta(0))^2] \geq \varphi(0)^2 \left( 2n \|\varphi\|_2^2 / M + M^2 I_\pi \right)^{-1}.$$

Un calcul de variations montre qu'un bon choix pour minimiser  $I_\pi$  est de prendre  $\pi(\theta) = 2 \sin^2(\pi\theta) 1_{0 \leq \theta \leq 1}$ . Vérifier que dans ce cas  $I_\pi = 4$ .

18. Voir [16], ou [14] Th.2.5.12, pour des compléments.

3. Soit l'ensemble de densités sur  $[-1/2, 1/2]$

$$\mathcal{E} = \left\{ p : |p(x) - p(y)| \leq |x - y|, p \geq 0, \int p(x) dx = 1 \right\}.$$

En remarquant que  $\mathcal{E}$  contient la famille paramétrique  $p_\theta(x) = \frac{\theta}{M} \varphi(Mx) + 1_J(x)$  où  $M = \frac{1}{6}n^{1/3}$ ,  $\varphi(x) = |x| - \frac{1}{4}$ , montrer que pour tout estimateur non-paramétrique  $\hat{g}(Y)$  de la densité en 0, on a

$$\sup_{p \in \mathcal{E}} E_p [(\hat{g}(Y) - p(0))^2] \geq \frac{1}{6}n^{-2/3}.$$

Les estimateurs à noyau atteignent cette borne à une constante près.

## III.4 La méthode bayésienne

### III.4.1 Estimateurs bayésiens

**Loi a priori, distribution a posteriori.** Thomas Bayes considérait le paramètre  $\theta$  comme étant lui-même aléatoire, ce qui, on va le voir, donne une forme de cohérence d'ensemble à la théorie, puisque l'estimée est une variable aléatoire. En tête de son essai (1763), il formulait ainsi le problème :

#### P R O B L E M.

*Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.*

Noter l'expression «the chance that the probability».

On dirait en français actuel : Etant donné  $Y \sim \mathcal{B}(n, \theta)$ , que vaut  $P(\theta_1 < \theta < \theta_2 | Y)$ ?

Tout ceci suppose l'introduction d'une distribution pour  $\theta$ . On se donne donc une probabilité a priori  $\pi(d\theta)$  sur  $\theta$ , et la paire  $(\theta, Y)$  est probabilisée avec la loi

$$p_\theta(y) \mu(dy) \pi(d\theta).$$

Tout se passe comme si  $\theta$  était d'abord tiré avec la loi  $\pi$ , puis  $y$  tiré conditionnellement à  $\theta$  avec la loi de densité  $p_\theta(y) = p(y|\theta)$ . On verra que, dans les faits, la distribution  $\pi$  représente davantage un apport d'information qu'une réelle distribution du paramètre.

On en déduit la loi de  $\theta$  sachant  $Y$  (probabilité a posteriori) :

$$p(d\theta|Y) = \frac{p_\theta(Y) \pi(d\theta)}{\int p_\tau(Y) \pi(d\tau)} \quad (\text{III.32})$$

dont le tracé (densité...) est considéré en pratique comme un indicateur intéressant. La distribution de  $Y$ , de densité  $\int p_\theta(y) \pi(d\theta)$  par rapport à  $\mu$ , est parfois appelée *distribution inconditionnelle* pour bien la distinguer de  $p_\theta(y)$ .

Lorsqu'on estime la loi d'une composante  $X$  (correspondant ici à  $\theta$ ) d'un vecteur gaussien centré  $(X, Y)$  conditionnellement à  $Y$  comme étant  $N(R_{XY} R_{YY}^{-1} Y, R_{XY} R_{YY}^{-1} R_{YX})$ , on ne fait rien de différent ; la méthode bayésienne est donc une forme de filtrage.

On va voir que l'approche bayésienne présente deux intérêts différents :

1. Travailler sur la base de la loi a posteriori.
2. Proposer de nouveaux estimateurs lorsque la variance d'estimation du maximum de vraisemblance est grande (faible nombre d'observations<sup>19</sup>, colinéarité en régression, etc.). Le rôle de  $\pi$  comme apport d'information supplémentaire, même vague, est ici de stabiliser l'estimation. L'estimateur ridge en régression linéaire (exercice 4 p. 84) est un exemple classique.

<sup>19</sup>. P.ex. la prédiction des votes Carter/Ford (présidentielles 1976) à l'aide de 6 variables socio-économiques et géographiques pour chaque état dans [25] § 5.2.

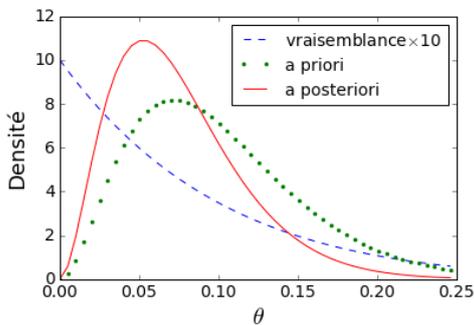
Elle permet également de donner une interprétation parlante aux estimateurs pénalisés (p.ex. 4 p. 84).

Mentionnons finalement qu'il arrive, dans les cas non-réguliers, qu'un estimateur bayésien soit asymptotiquement strictement meilleur que celui du maximum de vraisemblance (exercice 8 p. 85).

**Exemple.** On s'intéresse au taux de mortalité  $\theta$  dans un certain hôpital qui se lance dans un nouveau type d'opération. On sait que le taux de mortalité moyen dans les hôpitaux suite à cette opération est de 10% mais qu'il varie essentiellement entre 3% et 20%. L'hôpital réussit avec succès ses 10 premières opérations. Qu'en déduire sur  $\theta$  ?

Notons d'abord que la présence de répétitions (plusieurs hôpitaux) est un argument pour considérer  $\theta$  comme une variable aléatoire. Comme c'est un paramètre compris entre 0 et 1, il est naturel de lui assigner une distribution Bêta, c.-à-d. de densité  $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ . La distribution Bêta(3,27) a pour moyenne 0.1 et  $P(0.03 < \theta < 0.2) = 0.9$ ; c'est un choix raisonnable au vu des hypothèses. Comme le nombre de succès au terme de dix opérations suit une loi  $\mathcal{B}(10, \theta)$ , on obtient la distribution Bêta(3,37) :

$$p(\theta|y) \propto (\theta^2(1-\theta)^{26})(1-\theta)^{10} = \theta^2(1-\theta)^{36}.$$



$$\begin{aligned} \text{Vraisemblance} &: (1-\theta)^{10} \\ \text{A priori} &: \frac{1}{B(3,27)} \theta^2(1-\theta)^{26} \\ \text{A posteriori} &: \frac{1}{B(3,37)} \theta^2(1-\theta)^{36}. \end{aligned}$$

**Crédibilité vs confiance, bayésien vs fréquentiste.** Une région  $\mathcal{R}$  telle que  $p(\theta \in \mathcal{R}|Y) = 95\%$  est appelée *région de crédibilité à 95%*. Il ne s'agit pas d'une région de confiance et sa dépendance en  $\pi$  rend son interprétation sujette à discussions homériques et récurrentes<sup>20</sup>; car bien entendu  $\pi$  sera toujours erronée. Tout dépend donc de l'importance de cette dépendance et du crédit que l'on prête à  $\pi$ .

Dans les deux cas (crédibilité ou confiance), la variable  $1_{\theta \in \mathcal{R}}$  est une  $\mathcal{B}(1, 95\%)$ . L'intervalle de confiance semble plus astucieux puisqu'il s'affranchit de  $\pi$ . L'intervalle de crédibilité est plus simple à exprimer car  $\theta$  a effectivement une probabilité 95% d'appartenir à  $\mathcal{R}$ . Dans le cas de l'intervalle de confiance, on peut dire que 5% des intervalles à 95% produits dans le monde ne contiennent pas le vrai paramètre. Dans les deux cas, une réponse fautive ( $\theta \notin \mathcal{R}$  en vérité) s'explique par une malchance statistique.

Dans l'exemple qui suit, la méthode des intervalles de confiance aurait pu être employée mais les auteurs ont préféré comparer des distributions à postériori.

**Exemple [52].** Pour chaque individu arrêté dans le comté de Los Angeles en 1990 on s'intéresse au nombre  $Y_i$  d'arrestations dans le passé, la dernière non-incluse. Ces données contiennent le type d'infraction (*trafic de drogue, possession de drogue, cambriolage, vol avec agression*), l'emprisonnement (*oui, non*). Il s'agit en particulier de tirer des conclusions suite à une politique de fort emprisonnement des trafiquants et revendeurs de drogue dans les années 1986-1990.

Les auteurs postulent un modèle binomial négatif de paramètre  $\theta = (\alpha, \beta)$  :

$$P(Y_i = k) = \frac{\Gamma(k + \alpha)}{\Gamma(\alpha)k!} p^\alpha (1-p)^k, \quad p = \frac{1}{1 + t_i \beta}.$$

où  $t_i$  est l'âge du sujet moins 18 ans.  $Y_i$  a pour moyenne  $\alpha\beta t_i$ ; le paramètre  $\mu = \alpha\beta$  représente donc le nombre moyen d'arrestations par an après 18 ans. On se donne une loi a priori sur les paramètres :

$$\pi(\alpha, \beta) \propto \{\alpha(1 + \beta)^2\}^{-1}.$$

20. Cf. [76] et aussi la discussion de Larry Wasserman, qui suit l'article.

C'est la mesure uniforme pour  $(\log \mu, 1/(1+\beta))$ <sup>21</sup>. La loi de  $(\alpha, \beta)$  sachant les observations s'obtient par la formule de Bayes

$$p(\alpha, \beta|Y) \propto \frac{1}{\alpha(1+\beta)^2} \prod_{i=1}^n \frac{\beta^{Y_i} \Gamma(Y_i + \alpha)}{(1+t_i\beta)^{\alpha+Y_i} \Gamma(\alpha)}. \quad (\text{III.33})$$

Bien que  $\pi(\alpha, \beta)$  ne soit pas d'intégrale finie, la mesure conditionnelle ci-dessus est bien finie<sup>22</sup>. Les auteurs peuvent ainsi obtenir par simulation la distribution a posteriori de  $\mu$  dans différents sous-groupes liés aux types d'infraction et à la période. Ils observent en particulier que les trafiquants de drogue ayant subi une peine de prison ont un taux d'arrestation pour des délits non liés à la drogue significativement inférieur aux autres.

**L'estimateur bayésien.** La méthode bayésienne consiste à minimiser en  $\hat{\theta}$  le risque calculé sur la base de la probabilité totale :

$$R(\hat{\theta}) = \int R_{\theta}(\hat{\theta})\pi(d\theta) = \iint r(\theta, \hat{\theta}(y))p_{\theta}(y)\pi(d\theta)\mu(dy)$$

puis à définir les estimateurs bayésiens par la propriété :

$$\hat{\theta}_b \text{ est bayésien si pour tout estimateur } \hat{\theta} \text{ on a } R(\hat{\theta}_b) \leq R(\hat{\theta}).$$

Ceci donne

$$\hat{\theta}_b(y) = \arg \min_{\theta'} \int r(\theta, \theta')p_{\theta}(y)\pi(d\theta) = \arg \min_{\theta'} E[r(\theta, \theta')|Y = y] \quad \mu - p.p. \quad (\text{III.34})$$

(au moins lorsque cette fonction est mesurable). Deux observations sont en théorie très importantes :

- [Unicité implique admissibilité] Si  $\mu(\hat{\theta}_b(y) \neq \hat{\theta}'_b(y)) = 0$  pour tout autre estimateur  $\pi$ -bayésien  $\hat{\theta}'_b$ , alors  $\hat{\theta}_b$  est admissible  
En effet, si  $\hat{\theta}_b$  n'est pas admissible, alors il existe un autre estimateur  $\hat{\theta}$  strictement meilleur ;  $\hat{\theta}$  est forcément bayésien et comme pour un certain  $\theta$  on a  $E_{\theta}[r(\theta, \hat{\theta})] < E_{\theta}[r(\theta, \hat{\theta}_b)]$ , nécessairement  $P_{\theta}(\hat{\theta}(y) \neq \hat{\theta}_b(y)) > 0$ .
- [Risque constant implique minimax] S'il existe  $c$  tel que  $\pi(R_{\theta}(\hat{\theta}_b) = c) = 1$  et  $\sup_{\theta} R_{\theta}(\hat{\theta}_b) \leq c$ , alors  $\hat{\theta}_b$  est minimax (cf. exercice 11 p. 86 (1) ; résultat dû à Hodge et Lehmann, 1951).

Ceci donne donc une méthode pour fabriquer des estimateurs admissibles minimax : *Trouver une distribution  $\pi$  qui satisfasse ces deux conditions.* Ceci n'est pas forcément simple ; il faut noter qu'une telle distribution  $\pi$  est **la plus défavorable** au sens où elle maximise le risque bayésien (exercice 11 p. 86) ; dans cet exercice on observe que dans le cas de l'estimateur minimax de  $p$  pour l'observation de  $n$  variables de Bernoulli  $\mathcal{B}(n, p)$  indépendantes,  $\pi$  se concentre autour de  $\frac{1}{2}$  ce qui conduit à un estimateur dont le risque est fortement détérioré ailleurs : le risque constant obtenu égal à  $\frac{1}{4(1+\sqrt{n})^2}$  est à comparer avec  $n^{-1}p(1-p)$  pour le maximum de vraisemblance, qui est asymptotiquement efficace.

Dans le cas du risque quadratique, on peut être plus explicite :

**30 - THÉORÈME** (Moyenne a posteriori)

On suppose que  $r(\theta, \theta') = \|\theta - \theta'\|^2$  et que  $\int \|\theta\|^2 \pi(d\theta) < \infty$ , alors

$$\hat{\theta}_b = E[\theta|Y] = \frac{\int \theta p_{\theta}(Y)\pi(d\theta)}{\int p_{\theta}(Y)\pi(d\theta)}. \quad (\text{III.35})$$

De plus, si  $Z(y) = \int p_{\theta}(y)\pi(d\theta)$  est  $\mu$ -p.s. non-nul,  $\hat{\theta}_b$  est admissible.

21. Le fait que cette loi ne soit pas une probabilité est un défaut couramment accepté en théorie bayésienne (cf. l'estimateur de Pitman plus bas).

22. Nous ne détaillons pas cet exercice : Remarquer que si  $Y_i > 0$ ,  $\Gamma(\alpha + Y_i) = (\alpha + Y_i - 1) \dots (\alpha + 1) \alpha \Gamma(\alpha)$ , ce qui permet alors de majorer (III.33) par  $C\beta^p(1+\beta)^{-p-n\alpha-2}(\alpha^{p-1} + 1)$  où  $p$  est la somme des  $Y_i$  et  $C$  une constante dépendant des  $Y_i$  et des  $t_i$ . Il ne reste plus qu'à intégrer en  $\alpha$  puis en  $\beta$ .

*Démonstration.* Pour tout autre estimateur  $\widehat{\theta}'$ , on a

$$\begin{aligned} R(\widehat{\theta}') &= E[\|\theta - \widehat{\theta}'(Y)\|^2] \\ &= E[\|\theta - E[\theta|Y] + \widehat{\theta}_b(Y) - \widehat{\theta}'(Y)\|^2] \\ &= E[\|\theta - E[\theta|Y]\|^2] + E[\|\widehat{\theta}_b(Y) - \widehat{\theta}'(Y)\|^2] \quad (\text{Pythagore}) \\ &= R(\widehat{\theta}_b) + \int \|\widehat{\theta}_b(y) - \widehat{\theta}'(y)\|^2 Z(y) \mu(dy). \end{aligned}$$

Donc  $\widehat{\theta}_b$  est bien bayésien. Si  $\widehat{\theta}'$  est toujours meilleur que  $\widehat{\theta}_b$  (i.e.  $\forall \theta, R_\theta(\widehat{\theta}') \leq R_\theta(\widehat{\theta}_b)$ ), alors le deuxième terme est nécessairement nul, ce qui implique que  $\mu(\{\widehat{\theta}_b(y) \neq \widehat{\theta}'(y)\}) = 0$ . Donc  $R_\theta(\widehat{\theta}') = R_\theta(\widehat{\theta}_b)$  pour tout  $\theta$  ce qui empêche  $\widehat{\theta}'$  d'être strictement meilleur que  $\widehat{\theta}_b$ . ■

Sous des hypothèses générales, dans une asymptotique où le nombre d'échantillons tend vers l'infini mais  $\pi$  reste fixe, les estimateurs bayésiens pour le risque quadratique sont asymptotiquement confondus avec ceux du maximum de vraisemblance (la différence est  $\ll 1/\sqrt{n}$ ; cf. [2] Th. 36.9, ou [13] Th. 8.1 et Th. 8.2). Ceci vient essentiellement de ce que dans le numérateur de (III.32), le premier terme est l'exponentielle d'une somme de  $n$  termes qui concentre la mesure exponentiellement autour du maximum de vraisemblance, si bien que l'effet de la loi a priori est réduit asymptotiquement à néant.

**Estimation de  $p_{\theta_*}$ . Densité a posteriori des observations.** La loi  $\pi(d\theta)$  induit, par l'application  $\theta \mapsto p_\theta(\cdot)$ , une mesure sur l'espace des densités de probabilité par rapport à  $\mu$ . Sur cet espace on peut considérer le risque (divergence de Kullback)

$$r(\widehat{p}, p_\theta) = D(p_\theta \| \widehat{p}) = \int \log \left( \frac{p_\theta(y)}{\widehat{p}(y)} \right) p_\theta(y) \mu(dy).$$

Le cadre n'est plus rigoureusement paramétrique car le paramètre est la densité, mais la méthode bayésienne reste opérationnelle. On montre de manière analogue, en résolvant (III.34) sans oublier la contrainte  $\int \widehat{p}(y) \mu(dy) = 1$ , que sous des hypothèses raisonnables l'estimateur

$$\widehat{p}(x) = \frac{\int p_\theta(x) p_\theta(Y) \pi(d\theta)}{\int p_\theta(Y) \pi(d\theta)}$$

est bayésien pour ce risque. C'est la densité conditionnelle d'une nouvelle observation sachant  $Y$ . C'est sans doute une meilleure estimée de  $p_{\theta_*}$  que  $p_{\widehat{\theta}}$ .

On trouve le même résultat avec le risque  $r(\widehat{p}, p_\theta) = \int (p_\theta(y) - \widehat{p}(y))^2 \mu(dy)$ .

**Estimateur de Pitman.** On cherche à estimer un paramètre de translation :  $p_\theta(y) = p_0(y - \theta)$ ,  $\theta \in \mathbb{R}$ . Si l'on considère la mesure  $\pi_k$  uniforme sur  $[-k, k]$  on obtient l'estimateur bayésien  $\widehat{\theta}_b^k$  pour le risque quadratique

$$\widehat{\theta}_p^k(Y) = \frac{\int_{-k}^k \theta p_0(Y - \theta) d\theta}{\int_{-k}^k p_0(Y - \theta) d\theta}, \quad p_0(Y - \theta) = \prod_{i=1}^n p_0(Y_i - \theta).$$

L'estimateur de Pitman s'obtient en faisant tendre  $k$  vers l'infini

$$\widehat{\theta}_p(Y) = \frac{\int \theta p_0(Y - \theta) d\theta}{\int p_0(Y - \theta) d\theta}. \quad (\text{III.36})$$

Il faut bien entendu des hypothèses adéquates d'intégrabilité. On montre que cet estimateur (de risque constant) est minimax sous certaines hypothèses (exercice 11 p. 86 pour le cas où  $p_0$  est supposé à support compact, et [14] Th.5.3.6 pour le cas général).

Notons que si  $\pi$  est une mesure finie,  $\widehat{\theta}_b$  est forcément biaisé;  $\widehat{\theta}_p$  en revanche peut être non-biaisé (exercice 5 p. 84).

**Familles exponentielles et densités conjuguées.** On dit que les distributions sont conjuguées si  $\pi(\theta)$  et  $p(\theta|Y)$  appartiennent à une même famille paramétrique.

Si l'on observe des  $Y_i$  indépendants ayant chacun pour loi

$$p_\theta(y) = e^{T(\theta,y) - Z(\theta)}$$

alors le choix

$$\pi(d\theta) = \pi_{\alpha,\beta}(\theta)d\theta = c_{\alpha,\beta}e^{(T(\theta),\alpha) - \beta Z(\theta)}d\theta$$

où  $\alpha$  et  $\beta$  sont des paramètres à choisir et  $c_{\alpha,\beta}$  est la constante de normalisation, conduit à la loi a posteriori (loi de  $\theta$  sachant les observations)  $p(\theta|Y) = \pi_{\alpha+\sum Y_i, \beta+n}(\theta)$ . Noter que  $p(\theta|Y)$  est la densité de l'échantillon auquel on a ajouté  $\beta$  **observations fictives** de valeur  $\alpha/\beta$ .

Si la loi de  $Y_i$  est normale, gamma, Poisson ou binomiale, on obtient respectivement pour  $\theta$  une loi normale, gamma, gamma ou bêta.

**Convergence.** Il a été remarqué depuis longtemps par Doob que la convergence de l'estimateur bayésien  $E[\theta|Y]$  est une conséquence de la théorie des martingales [66]. En effet, si l'on suppose que  $\theta$  est une fonction mesurable de la suite iid  $(Y_i)_{i \geq 1}$ <sup>23</sup>, on a pour toute fonction borélienne bornée  $f$

$$E[f(\theta)|Y_1, \dots, Y_n] \longrightarrow E[f(\theta)|Y_1, \dots] = f(\theta) \quad p.s.$$

Moyennant une hypothèse d'intégrabilité de  $\theta$ , en particulier si  $\Theta$  est borné, ceci règle le cas du risque quadratique en prenant  $f(\theta) = \theta$ <sup>24</sup>.

**Bilan.** La question du choix de  $\pi$  n'admet guère de réponse générale satisfaisante. Il y a deux tendances : Les mesures non-informatives qui sont plutôt utilisées lorsque la méthode bayésienne sert non pas à améliorer l'estimation mais à travailler avec la densité a posteriori en intervenant le moins possible dans le modèle, et les mesures informatives qui ont pour objet d'intervenir afin améliorer effectivement l'estimation (cf. les points 1 et 2 de l'introduction). Plusieurs types de solutions ont émergé :

- Un choix adhoc : une densité conjuguée.
- Un choix « naturel » à la Pitman, comme la mesure de Lebesgue (invariante par translation) pour les paramètres de translation et la mesure  $d\sigma/\sigma$  (invariante par multiplication) pour un paramètre d'échelle ; c'est un choix raisonnable. Bien que peu informatif, ce choix donne un bon résultat à l'exercice 9 p. 85. Noter que ces lois sont impropres (mesure infinie).
- La mesure de Jeffreys  $\pi(d\theta) = \det(I(\theta))^{1/2}d\theta$ . Elle est, pour le calcul de la loi a posteriori, invariante par changement de variable sur  $\theta$ , ce qui est son principal atout : le niveau de crédibilité d'une région ne change pas avec la paramétrisation ; en revanche, elle semble peu satisfaisante en estimation<sup>25</sup>. C'est que cette forme de neutralité due à l'invariance devient exactement contre-productive si l'on considère des situations non-invariantes, comme à présupposer le vecteur  $\theta$  parcimonieux (peu de coordonnées non-nulles), ou à considérer ses coordonnées comme a priori indépendantes, ce qui recouvre beaucoup de cas rencontrés en pratique.
- Empirical Bayes : Se donner une famille paramétrée de lois a priori  $\pi_\beta$ , et estimer le paramètre  $\beta$  en maximisant la probabilité marginale de l'échantillon  $\int p_\theta(Y)\pi_\beta(d\theta)$ . C'est une piste assez bonne en pratique (possiblement compliquée à mettre en œuvre), et bonne en théorie également. Voir aussi l'exercice 3 p. 84 pour un exemple où  $\beta$  est estimé différemment.

Voici deux méthodes de Bayes empirique qui ont suscité un intérêt particulier pour le traitement de cas de nature plutôt non-paramétrique où  $\theta$  est un vecteur de grande dimension dont les composantes non-nulles sont a priori *sans lien entre elles et homogènes* :

- Bayes empirique paramétrique : Pour modéliser le fait que  $\theta$  est un vecteur ayant beaucoup de paramètres nuls, Johnstone et Silverman [95] choisissent pour chaque composante le même mélange d'une masse de Dirac en 0 et d'une loi exponentielle :

$$\pi_{\alpha,w}(d\theta) = \psi_{\alpha,w}(d\theta_1) \dots \psi_{\alpha,w}(d\theta_d), \quad \psi_{\alpha,w}(dx) = (1-w)\delta_0(dx) + \frac{1}{2}w\alpha e^{-\alpha|x|}dx.$$

23. Ceci n'a rien d'évident en toute généralité. C'est le cas si l'estimateur du maximum de vraisemblance est consistant. Sinon, on doit au moins supposer que les  $p_\theta$  sont toutes différentes ; alors  $\theta$  est l'argument du maximum en  $\tau$  de  $\int \log p_\tau(y)p_\theta(dy) = \lim n^{-1} \sum_{i=1}^n \log p_\tau(Y_i)$  ; si pour tout  $\theta$  cette fonction est continue, pour avoir le maximum il suffit de la calculer en un nombre dénombrable de points et l'on a bien mesurabilité.

24. On trouvera des compléments dans [79].

25. Il est à noter que cette mesure est asymptotiquement la plus défavorable (cf. plus haut p. 80) pour le risque de Kullback [51]. Il semble que ce résultat asymptotique ne soit d'aucun secours pour les échantillons de taille modeste, qui sont précisément l'objet de l'intérêt suscité par la méthode bayésienne, cf. [7] § 2.9.

Ici  $\beta = (a, w)$ . En dépit du caractère très simpliste de la famille  $\psi_\beta$ , les résultats sont très bons.

- Bayes empirique non-paramétrique : Il s'agit de calculer  $\pi$  au maximum de vraisemblance *sous la contrainte* que les  $\theta_j$  sont iid sous  $\pi$  (i.e.  $\pi(d\theta) = \psi(d\theta_1) \dots \psi(d\theta_d)$ )<sup>26</sup>. Ici  $\beta = \psi$ . L'estimée  $\hat{\psi}$  sera typiquement une somme de masses de Dirac [94], solution d'un problème assez lourd à mettre en œuvre<sup>27</sup>.

Voir aussi les exercices 6 et 7 où (III.35) est estimé directement sans passer par  $\pi$ !

Si  $\theta$  est de dimension 1, la contrainte ne joue plus et l'on retrouve le maximum de vraisemblance.

Nous renvoyons à [7] pour des traitements d'exemples et des approfondissements.

### III.4.2 Le maximum a posteriori (MAP).

L'estimation de  $\theta$  se fait ici par maximisation de  $p(\theta, Y)$ , c'est-à-dire qu'on choisit le mode de la probabilité a posteriori  $p(\theta|Y) = p(\theta, Y)/p(Y)$ ; le facteur  $1/p(Y)$  étant constant<sup>28</sup>, il vient :

$$\hat{\theta}_{MAP} = \arg \max_{\theta} \pi(\theta)p_{\theta}(Y).$$

Cet estimateur correspond également à l'estimateur bayésien limite  $\varepsilon \rightarrow \infty$  quand  $r(\theta, \theta') = \varphi(\|\theta - \theta'\|)$  où  $\varphi$  est continue, nulle sur  $[0, \varepsilon]$ , 1 sur  $[2\varepsilon, +\infty[$ , et monotone  $[\varepsilon, 2\varepsilon]$ .

Cette remarque s'interprète simplement dans le cas où la variable  $\theta$  prend des valeurs discrètes : ce choix minimise la probabilité d'erreur  $P(\hat{\theta} \neq \theta|Y)$  (a fortiori, cet estimateur minimise  $P(\hat{\theta} \neq \theta)$ ). En effet si une variable  $\theta$  a une loi  $q(\theta)$  (ici  $q(\theta) = p(\theta|Y)$ ) alors  $\hat{\theta} = \arg \max_{\theta} q(\theta)$  minimise en  $\theta$  la probabilité d'erreur  $1 - q(\theta)$ .

L'estimateur lasso pour le modèle linéaire  $Y \sim \mathcal{N}(X\beta, \sigma^2 Id)$

$$\hat{\beta}_L = \arg \min_{\beta} \frac{1}{2} \|Y - X\beta\|^2 + \lambda \sum_j |\beta_j| \tag{III.37}$$

est un estimateur MAP pour une distribution exponentielle des coefficients. L'estimateur ridge est à la fois MAP et bayésien (exercice 4 p. 84).

Le MAP est utilisé quand la méthode bayésienne est difficile à mettre en œuvre en raison des calculs d'intégrales compliqués qu'elle impose. Il a l'avantage de fournir une interprétation parlante aux estimateurs au maximum de vraisemblance pénalisé. Il n'est pas particulièrement recommandé mais se comporte asymptotiquement comme le maximum de vraisemblance puisque le terme  $\pi(\theta)$  a un poids relatif qui diminue lorsque  $n$  tend vers l'infini. Pour les petits échantillons, il est plus habituel de s'intéresser plus précisément à la probabilité a posteriori, le MAP étant considéré comme une information trop succincte.

### III.4.3 Exercices et compléments.

**Exercice 1.** Soit  $Y_1, \dots, Y_n \sim \mathcal{E}(\theta)$  iid et la loi a priori  $\theta \sim \mathcal{E}(1)$ . Calculer l'estimateur bayésien de  $\theta$  pour le risque quadratique. Montrer sa consistance. Faire le même calcul pour la distribution a priori (impropre)  $\pi(d\theta) = 1_{\theta > 0} \theta^{-1} d\theta$ .

**Exercice 2.** Soit  $Y \sim \mathcal{N}(\theta, 1)$  et une certaine loi a priori  $\pi$  pour  $\theta$ . On désigne par  $f(y)$  la loi inconditionnelle de  $Y$  (mélange de gaussiennes). On supposera  $\pi$  à support compact ce qui rend  $f$  très régulière et rapidement décroissante à l'infini. Exprimer l'estimateur bayésien  $\hat{\theta}(Y)$  en fonction de  $Y$ ,  $f(Y)$  et  $f'(Y)$  seulement. Montrer que, en toute généralité,  $R(\hat{\theta}) = E[\theta^2] - E[\hat{\theta}^2]$  (utiliser  $R(\hat{\theta}) = E[\|\theta - E[\theta|Y]\|^2]$ ). En déduire que le risque vaut

$$R(\hat{\theta}) = 1 - \int \frac{f'(y)^2}{f(y)} dy.$$

26. Une idée de Kiefer et Wolfowitz (1956), voir l'exercice 12 p. 86, mais qu'ils n'utilisent pas pour estimer  $\theta$  ensuite.

27. Une mise en œuvre pour des modèles spécifiques est réalisée dans le programme R REBayes [82].

28. On a bien entendu  $p(y) = \int p(\theta, y) d\theta = \int p_{\theta}(y) \pi(\theta) d\theta$ .

**Exercice 3 (Estimateur bayésien et estimateur de James-Stein).** Soit  $Y \sim \mathcal{N}(\theta, \sigma^2 I_d)$  (observation unique) et la loi a priori pour  $\theta \sim \mathcal{N}(0, \tau^2 I_d)$ .  $\sigma$  est connu.

1. Exprimer  $(\theta, Y)$  comme fonction linéaire d'une autre paire  $(U, V) \sim \mathcal{N}(0, I_{2d})$ . En déduire que  $(\theta, Y)$  est un vecteur gaussien dont on donnera la matrice de covariance.
2. Calculer l'estimateur bayésien  $\hat{\theta}_\tau$  de  $\theta$  correspondant.  
*Indication* : On rappelle que si  $(U, V)$  est un vecteur gaussien *centré*, on a  $E[V|U] = R_{VU}R_{UU}^{-1}U$ .
3. En déduire que pour tout  $0 < a < 1$ ,  $aY$  est un estimateur admissible de  $\theta$ .
4. Quelle est la loi de  $Y$  (i.e. inconditionnelle à  $\theta$ ) ?
5. On admet que si  $X \sim \mathcal{N}(0, s^2 I_d)$ , alors  $(d-2)\|X\|^{-2}$  est un estimateur non biaisé de  $s^{-2}$  (Ceci peut se vérifier par une IPP sur le rayon en coordonnées polaires). En déduire un estimateur de  $(\sigma^2 + \tau^2)^{-1}$ , puis un nouvel estimateur de  $\theta$  pour le cas où  $\tau$  est inconnu ( $\sigma$  est toujours connu). Vérifier qu'il s'agit de l'estimateur de James-Stein de l'exercice 10 p. 60 (on pourra supposer  $\sigma = 1$ ). L'exercice suivant le retrouvera également, d'une manière un peu différente.

**Exercice 4 (L'estimateur ridge est bayésien).** Soit le modèle linéaire gaussien (I.7), soit

$$Y = X\beta_* + u, \quad Y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n \times p}, \quad u \sim \mathcal{N}(0, \sigma_*^2 I_d_n). \quad (\text{III.38})$$

L'estimateur au maximum de vraisemblance  $\hat{\beta}_{MV} = (X^T X)^{-1} X^T Y$  pose des problèmes si le nombre de variables  $p$  est grand car la matrice  $X^T X$  risque d'être mal conditionnée; il arrive même que  $p > n$ . On a proposé l'estimateur pénalisé, appelé estimateur ridge :

$$\hat{\beta}_R = \arg \min_{\beta} \frac{1}{2} \|Y - X\beta\|^2 + \lambda \|\beta\|^2. \quad (\text{III.39})$$

où  $\lambda > 0$  est un paramètre à choisir (possiblement par validation croisée).

1. Vérifier que

$$\hat{\beta}_R = (X^T X + \lambda I_d_p)^{-1} X^T Y \quad (\text{III.40})$$

2. Soit la loi a priori  $\beta_* \sim \mathcal{N}(0, \frac{\sigma_*^2}{\lambda} I_d_p)$ . Il est clair que  $\hat{\beta}_R$  est l'estimateur MAP. Vérifier que c'est également l'estimateur bayésien  $\hat{\beta}_R = E[\beta_* | Y]$ .  
*Indication* : On rappelle que si  $(U, V)$  est un vecteur gaussien *centré*, on a  $E[V|U] = R_{VU}R_{UU}^{-1}U$ .
3. On suppose ici  $X = Id$ . Vérifier que dans ce cas, la validation croisée pour le choix de  $\lambda$  ne peut fonctionner. On décide de choisir  $\lambda$  par minimisation du risque estimé par SURE,  $\hat{R} = \hat{R}(\lambda)$  de l'équation (I.25) p. 17. Montrer que l'on retrouve alors quasiment l'estimateur de James-Stein de l'exercice 10 p. 60.

**Exercice 5 (Un estimateur bayésien est toujours biaisé).** Pour le  $\hat{\theta}$  du § III.4.1 avec risque quadratique, montrer qu'on a

$$R(\hat{\theta}) = - \int \theta^T b(\theta) \pi(d\theta)$$

où  $b(\theta) = E[\hat{\theta} | \theta] - \theta$  est le biais. On pourra utiliser (et justifier) que  $R(\hat{\theta}) = E[\theta^T (\theta - \hat{\theta})]$ .

Attention, ceci ne s'applique que si la mesure  $\pi$  est finie : vérifier que l'estimateur de Pitman est non biaisé si  $p_0$  est symétrique ( $p_0(x) = p_0(-x)$ ) ; commencer par le cas  $n = 1$ .

**Exercice 6 (Estimateur de Robbins-Turing [45, 69]).** On observe  $Y_i \sim \mathcal{P}(\theta_i)$ ,  $i = 1, \dots, n$ , p. ex.  $Y_i$  est le nombre d'occurrences du  $i$ -ième mot du dictionnaire dans les œuvres de Victor Hugo. Soit une distribution  $\pi$  pour les paramètres  $\theta_i$  indépendants; le paramètre est donc  $\theta = (\theta_i)_{1 \leq i \leq n}$ . Vérifier que  $E[\theta_i | Y] = E[\theta_i | Y_i]$ , et que (III.35) conduit à

$$E[\theta_i | Y_i] = (Y_i + 1) \frac{P(Y_i + 1)}{P(Y_i)}$$

où  $P$  est la distribution de  $Y_i$  résultant du modèle bayésien. Cette identité a motivé Robbins en 1956 pour proposer l'estimateur :

$$\hat{\theta}_i = (Y_i + 1) \frac{\hat{P}(Y_i + 1)}{\hat{P}(Y_i)}, \quad \hat{P}(k) = \text{Card}(\{j : Y_j = k\})/n.$$

Discuter brièvement du cas  $n$  petit.

**Exercice 7 (Dans le même style).**  $\mathcal{E}(\lambda)$  désigne la loi exponentielle d'espérance  $1/\lambda$ . On observe une suite indépendante  $Y_i \sim \mathcal{E}(\theta_i)$ ,  $1 \leq i \leq n$ , chaque  $\theta_i$  ayant été tiré indépendamment selon une loi de densité  $\pi(\theta)$ . Ce modèle est vu comme un modèle bayésien.

1. Calculer la densité  $p$  de la loi inconditionnelle commune aux  $Y_i$  en fonction de  $\pi$ .
2. Calculer l'estimateur bayésien  $E[\theta_i|Y_i]$ . Il dépend de  $\pi$ .
3. L'exprimer comme fonction de  $p$  et  $Y_i$ , mais sans faire apparaître  $\pi$ . Proposer alors une procédure d'estimation de  $E[\theta_i|Y_i]$  basée seulement sur une estimée (non paramétrique) de  $p$ .
4. Faire de même si  $Y_i \sim \mathcal{N}(\theta_i, 1)$  (la procédure d'estimation est étudiée dans [44]).
5. (Un petit supplément) Reprendre le point 3 pour le calcul de  $E[\theta_i^{-1}|Y_i]$ . Quel est l'avantage ?

**Exercice 8 (Bayésien supérieur au maximum de vraisemblance).** On considère l'exemple de l'exercice 5 p. 7. Soit l'estimateur bayésien pour la mesure  $\pi(dx) = x^{-2}dx$  (Cette distribution impropre est ici choisie pour la simplicité du calcul auquel elle conduit). Calculer cet estimateur et montrer qu'il est asymptotiquement deux fois meilleur que l'estimateur au maximum de vraisemblance.

Montrer que si maintenant  $\pi(dx) = dx$ , le nouvel estimateur bayésien obtenu est à distance inférieure à  $2(n-2)^{-2}\theta_*$  du précédent, et qu'il est donc asymptotiquement également deux fois meilleur que l'estimateur au maximum de vraisemblance.

Une étude plus poussée montre que le résultat reste du moment que  $\pi(dx) = f(x)dx$  avec  $f$  continue bornée  $> 0$ .

**Exercice 9.** On considère l'exemple de l'exercice 4 p. 7 avec  $p > n$ . On note  $Z = (X, Y)$ . Vérifier que :

$$\mathcal{L}(Z, \theta) = - \sum_{i=1}^n \frac{(X_i - \mu_i)^2}{2\sigma^2} - \sigma \sum_{j=1}^p Y_j + (p-n) \log \sigma \quad (\text{III.41})$$

$$I(\theta) = \sigma^{-2} \begin{pmatrix} Id & 0 \\ 0 & 2n+p \end{pmatrix}. \quad (\text{III.42})$$

1. Calculer l'estimateur bayésien de  $\theta = (\mu_1, \dots, \mu_n, \sigma)$  si  $\pi$  est la mesure  $\lambda^n \times \lambda|_{\mathbb{R}_+}$  (cette mesure n'est pas finie, mais on peut justifier ce choix en procédant comme pour l'estimateur de Pitman).
2. Essayer en remplaçant la mesure  $d\sigma$  restreinte à  $\mathbb{R}_+$  par  $\frac{d\sigma}{\sigma}$  (cf. le bilan du § III.4.1).
3. Calculer l'estimateur bayésien de  $\sigma$  pour la mesure de Jeffreys  $\pi(d\theta) = \det(I(\theta))^{1/2}d\theta$ , et montrer que l'on retrouve le problème original.
4. On a proposé pour la sélection de modèle le critère SIC, variante de BIC (cf. p. 68 et l'appendice E)

$$SIC(d) = -\log p_{\hat{\theta}^d}(Z) + \frac{1}{2} \log \det(I(\hat{\theta}^d)), \quad (\text{III.43})$$

$I(\theta)$  pouvant être remplacé par son estimée  $\hat{I}(\theta) = -\nabla^2 \log p_\theta(Z)$ . Ce critère peut suggérer l'estimateur

$$\hat{\theta} = \arg \max_{\theta} \log p_\theta(Z) - \frac{1}{2} \log \det(I(\theta)). \quad (\text{III.44})$$

Calculer  $\hat{\theta}$  et vérifier qu'il se comporte ici correctement.

**Exercice 10 (Estimateur de Barankin. Estimateur bayésien non-biaisé).** Soit l'estimateur de l'exercice 15 p. 62, c'est-à-dire l'estimateur de variance minimale sous  $q$  sous la contrainte d'être non biaisé. C'est en résumé celui qui minimise

$$\int (\hat{\theta}(x) - \theta_q)^2 q(x) \mu(dx), \quad \theta_q = \int \hat{\theta}(x) q(x) \mu(dx)$$

sous la contrainte

$$\theta = \int \widehat{\theta}(x) p_{\theta}(x) \mu(dx), \quad \theta \in \Theta. \quad (\text{III.45})$$

On choisit  $q$  de la forme  $q(x) = \int p_{\theta}(x) \pi(d\theta)$  pour une certaine mesure  $\pi$ . Démontrer que  $\widehat{\theta}(x)$  est l'estimateur  $\pi$ -bayésien pour le risque quadratique dans la classe des estimateurs non biaisés au sens où il minimise  $\int (\widehat{\theta}(x) - \theta)^2 p_{\theta}(x) \mu(dx) \pi(d\theta)$  sous (III.45). *Indication* :  $\widehat{\theta}(x) - \theta = (\widehat{\theta}(x) - \theta_q) + (\theta_q - \theta)$ .

**Exercice 11 (Estimateurs bayésiens et estimateurs minimax).** Soit  $\pi(d\theta)$  une distribution sur  $\Theta$  et  $\widehat{\theta}_b$  un estimateur bayésien associé. On se propose de justifier la construction proposée au § III.4.1.

1. Soit  $\widehat{\theta}$  tel que pour tout  $\theta' : R_{\theta'}(\widehat{\theta}) \leq \int R_{\theta'}(\widehat{\theta}_b) \pi(d\theta)$ 
  - (i) Montrer que  $\widehat{\theta}$  est minimax, que  $R_{\theta}(\widehat{\theta})$  est  $\pi$ -presque sûrement égal au risque minimax, partout inférieur à ce dernier, et que  $\widehat{\theta}$  est  $\pi$ -bayésien.  
*Indication* : Montrer directement que pour tout estimateur  $\widehat{\theta}'$  on a  $\sup_{\theta} R_{\theta}(\widehat{\theta}) \leq \sup_{\theta} R_{\theta}(\widehat{\theta}')$ , puis conclure.
  - (ii) Vérifier que ceci démontre le point 2 de la page 80 en considérant  $\widehat{\theta} = \widehat{\theta}_b$  (immédiat).
  - (iii) Vérifier que tout estimateur  $\pi$ -bayésien  $\widehat{\theta}_b$  satisfaisant les hypothèses du point 2 satisfait  $R_{\theta'}(\widehat{\theta}) \leq \int R_{\theta'}(\widehat{\theta}_b) \pi(d\theta)$  pour tout  $\widehat{\theta}$  et tout  $\theta'$  (immédiat en utilisant le point 2).
  - (iv) Montrer si  $\widehat{\theta}_b$  est un estimateur  $\pi$ -bayésien satisfaisant les hypothèses du point 2, alors son risque  $\int R_{\theta}(\widehat{\theta}_b) \pi(d\theta)$  est maximal parmi tous les choix possibles de paires  $(\pi', \widehat{\theta}'_b)$ . C'est pour cela que l'on dit que  $\pi$  est la distribution *la plus défavorable*.

On va voir dans la suite comment l'égalisation du risque se fait par un choix de  $\pi$  qui donne un poids élevé aux points où l'estimation est plus difficile ( $\widehat{\theta}_b$  est biaisé vers ces points).  $\pi$  dépend alors de  $n$ .

2. On observe une suite de variables de Bernoulli de longueur  $n$ ,  $(Y_1, \dots, Y_n)$ . On cherche à estimer  $\theta = p = P(Y = 1)$ . On se propose de choisir  $\widehat{\theta}$  de la forme

$$\widehat{\theta} = \frac{\bar{Y} + a}{1 + b}, \quad \bar{Y} = \frac{1}{n} \sum Y_i.$$

- (i) Trouver  $a$  et  $b$  tel que le risque quadratique soit indépendant de  $p$  (on doit trouver un risque de  $1/4(1 + \sqrt{n})^2$ ).
- (ii) Soit  $B_{u,v}$  la distribution bêta sur  $[0, 1]$ , de densité  $\frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} t^{u-1}(1-t)^{v-1}$ . Sa moyenne est  $u/(u+v)$ . Calculer l'estimateur bayésien pour  $\pi = B_{u,v}$  (remarquer que c'est l'espérance d'une certaine  $B_{u',v'}$ ). Montrer qu'il coïncide avec  $\widehat{\theta}$  si  $u = v = \sqrt{n}/2$ .
- (iii) En déduire que, pour ce choix,  $\widehat{\theta}$  est minimax. Utiliser le théorème 30 pour montrer l'admissibilité. Calculer l'information de Fisher et vérifier que lorsque  $n$  tend vers l'infini, la distribution  $\pi$  se concentre sur  $1/2$ , point d'information minimale.
- (iv) Calculer l'estimateur MAP avec  $\pi = B_{u+1,v+1}$ .
3. Étendre le résultat du premier point au cas où  $R_{\theta}(\widehat{\theta}) \leq \sup_k \int R_{\theta}(\widehat{\theta}_b^k) \pi_k(d\theta)$  pour une suite infinie d'estimateurs bayésiens  $\widehat{\theta}_b^k$ .

Application : On considère l'estimateur de Pitman avec les notations de la page 81. On suppose que  $p_0$  est à support inclus dans  $[-A, A]$ .

- (a) Montrer que pour tout  $\theta$  et tout  $k$ ,  $R_{\theta}(\widehat{\theta}_p^k) \leq 4A^2$  et vaut  $R_{\theta}(\widehat{\theta}_p)$  si  $k \geq |\theta| + 2A$ .
- (b) En déduire que  $R_{\theta}(\widehat{\theta}_p) = \lim_k \int R_{\theta}(\widehat{\theta}_b^k) \pi_k(d\theta)$ . Il s'ensuit que  $\widehat{\theta}_p$  est minimax. Vérifier qu'il est de risque constant.

**\*Exercice 12 (NPMLE : Bayes empirique non-paramétrique<sup>29</sup>).** On se donne le modèle suivant pour une suite d'observations indépendantes à valeurs dans  $\mathbb{R}^d$  :

$$Y_k \sim f(y; \beta, \alpha_k) dy, \quad k = 1 \dots n$$

29. Complètement inspiré de [97]. On trouvera des compléments intéressants par exemple dans [83] et [61], et une mise en œuvre dans le programme R REBayes [82].

où pour tous  $\alpha \in A, \beta \in B$ , la fonction  $y \mapsto f(y; \beta, \alpha)$  est une densité de probabilité sur  $\mathbb{R}^d$ . On suppose que  $A \subset \mathbb{R}^p$  et  $B \subset \mathbb{R}^q$  sont compacts. Par exemple des données de la forme

$$Y_{ij} = \beta + \alpha_i e_{ij}, \quad 1 \leq j \leq n_0, \quad 1 \leq i \leq n \quad (\text{III.46})$$

où  $n_0$  est fixe et les  $e_{ij}$  sont iid  $\mathcal{N}(0, 1)$ ; ou encore

$$Y_{ij} = \alpha_i + \beta e_{ij}, \quad 1 \leq j \leq n_0, \quad 1 \leq i \leq n. \quad (\text{III.47})$$

L'estimateur au maximum de vraisemblance pour  $(\alpha_1, \dots, \alpha_n, \beta)$  a toutes chances d'échouer car il y a trop de paramètres. Kiefer et Wolfowitz [97] proposent de postuler que les  $\alpha_k$  forment une suite iid de distribution  $\mu$  inconnue (*indépendance et homogénéité* des paramètres), puis d'estimer  $\theta = (\beta, \mu)$  au maximum de vraisemblance. On a donc le contraste

$$K(y, \theta) = -\log \int f(y; \beta, \alpha) \mu(d\alpha), \quad k(\theta) = -E \left[ \log \int f(Y_1; \beta, \alpha) \mu(d\alpha) \right].$$

On se donne pour  $\Theta$  l'ensemble  $B \times \mathcal{M}$  où  $\mathcal{M}$  est l'ensemble des mesures de probabilité sur  $A$ . Soit  $d_P$  la métrique de Lévy-Prokhorov sur  $\mathcal{M}$ ; il suffira ici de savoir que cette métrique est associée à la topologie de la convergence faible ([36] Th. 6.8), ce qui signifie

$$\left( d_P(\mu, \mu_n) \rightarrow 0 \right) \iff \left( \forall f \text{ continue bornée sur } A, \mu_n(f) \rightarrow \mu(f) \right).$$

La compacité de  $A$  fait que  $\mathcal{M}$  est compact pour cette métrique ([36] Th. 5.1). Soit la distance sur  $\Theta$

$$d(\theta, \theta') = |\beta - \beta'| + d_P(\mu, \mu').$$

On considère les hypothèses (améliorables) :

(H1)  $\|f\|_\infty < \infty$

(H2) Pour tout  $y$ , l'application  $(\alpha, \beta) \mapsto f(y, \alpha, \beta)$  est continue.

(H3) Il existe un couple  $(\mu_*, \beta_*)$  unique tel que la loi de  $Y_1$  soit  $\int f(y, \alpha, \beta_*) \mu_*(d\alpha)$ .

Montrer les points suivants :

1.  $(\Theta, d)$  est compact.
2. Sous (H1), (H2) et (H3), les hypothèses (a,b,c) du théorème 11 sont satisfaites.

*Indication* : Si  $\mu_n$  et  $\nu_n$  sont deux suites de mesures convergeant faiblement vers  $\mu$  et  $\nu$ , alors  $\mu_n \times \nu_n$  converge faiblement vers  $\mu \times \nu$  ([36] Th. 2.8).

Montrer que :

1. L'exemple (III.46) satisfait les hypothèses si  $A$  est de la forme  $[a, b]$  avec  $a > 0$ .
2. L'exemple (III.47) satisfait les hypothèses si  $n_0 > 1$ , et  $B = [a, b]$ ,  $a > 0$ . Expliquer pourquoi  $n_0 > 1$  est nécessaire.

**Exercice 13 (Agrégation avec inégalité oracle).** Soit  $Y_i$  une suite iid,  $\Theta$  un espace vectoriel de paramètres, et  $\pi$  une mesure de probabilité sur  $\Theta$ . Soit l'estimateur basé sur des observations indépendantes du contraste  $K$

$$\hat{\theta}_n = \frac{\int \theta e^{-K_n(\theta)} \pi(d\theta)}{\int e^{-K_n(\theta)} \pi(d\theta)}, \quad K_j(\theta) = K(\theta, Y_1) + \dots + K(\theta, Y_n), \quad K_0(\theta) = 0. \quad (\text{III.48})$$

La mesure  $\pi$  peut être la mesure uniforme sur une famille de candidats présélectionnés (estimateurs sur un autre ensemble de données).

On suppose que la fonction  $\theta \mapsto e^{-K(\theta, y)}$  est concave, ce qui est assez contraignant car  $\Theta$  ne peut être tout  $\mathbb{R}^d$ , mais sera satisfait dans l'application présentée en dernière question. De plus, si  $\Theta$  est compact, et si  $K$  est  $C^2$  strictement convexe, alors pour  $\varepsilon > 0$  assez petit la fonction  $\theta \mapsto e^{-\varepsilon K(\theta, y)}$  est concave. La borne d'erreur (III.50) montre que l'on devrait avoir intérêt à prendre  $\varepsilon$  le plus grand possible.

Il est élémentaire de vérifier que cette condition de concavité implique la convexité de  $K$ .

1. Montrer, en utilisant l'inégalité de Jensen, que pour tout  $j$

$$K(\widehat{\theta}_j, Y_{j+1}) \leq -\log \frac{\int e^{-K_{j+1}(\theta)} \pi(d\theta)}{\int e^{-K_j(\theta)} \pi(d\theta)}.$$

2. Montrer l'inégalité générale pour un processus  $X(\theta, \omega)$

$$E \left[ \log \int e^{X(\theta, \omega)} \pi(d\theta) \right] \geq \log \int e^{E[X(\theta, \omega)]} \pi(d\theta).$$

*Indication* : Poser  $\pi'(d\theta) = e^{E[X(\theta, \omega)]} \pi(d\theta) / \int e^{E[X(\theta, \omega)]} \pi(d\theta)$ , puis utiliser l'inégalité de Jensen.

3. On pose  $k(\theta) = E[K(\theta, Y_1)]$ . Dédurre des deux points précédents que

$$\sum_{j=0}^n E[k(\widehat{\theta}_j)] \leq -\log \int e^{-(n+1)k(\theta)} \pi(d\theta).$$

4. En déduire que si  $\pi$  est une mesure sur une partie finie  $\{\theta_1, \dots, \theta_m\} \subset \Theta$ , on a pour tout  $1 \leq l \leq m$ ,

$$\frac{1}{n+1} \sum_{j=0}^n E[k(\widehat{\theta}_j)] \leq k(\theta_l) - \frac{\ln \pi_l}{n+1}. \quad (\text{III.49})$$

Le membre de gauche est la performance de l'estimateur randomisé basé sur les  $\widehat{\theta}_j$  avec loi uniforme. On aimerait bien en déduire de (III.49) la même inégalité avec seulement  $E[k(\widehat{\theta}_n)]$  à gauche, sous l'argument que  $\widehat{\theta}_n$  est le meilleur de tous, mais cet argument ne tient pas si  $\pi$  est particulièrement clairvoyante, par exemple si  $\theta_0$ , moyenne sous  $\pi$ , réalise le minimum de la fonction  $k$ .

5. Dédurre de la question précédente que, sous les mêmes hypothèses, avec  $\pi(\theta_l) = \frac{1}{m}$ ,  $1 \leq l \leq m$ , l'estimateur  $\bar{\theta}_n = \frac{1}{n+1} \sum_{j=0}^n \widehat{\theta}_j$  satisfait :

$$E[k(\bar{\theta}_n)] \leq \min_l k(\theta_l) + \frac{\ln m}{n+1}. \quad (\text{III.50})$$

L'estimateur  $\bar{\theta}_n$  a donc un risque quasiment aussi bon que le meilleur des candidats  $\theta_l$ , ceci dans un cadre non asymptotique, même si  $m$  est grand.

6. Application (Catoni 1997 [48]). On possède plusieurs candidats  $p_1, \dots, p_m$  pour une densité  $p_*(x)$  (par exemple divers estimateurs basés sur un échantillon antérieur analogue). On cherche à estimer  $p_*(x)$  à l'aide d'un échantillon  $Y = (Y_1, \dots, Y_n)$  de  $n$  variables iid de loi  $p_*$ , comme combinaison linéaire des  $p_j$ . On pose

$$\Theta = \left\{ \theta \in \mathbb{R}_+^m : \sum \theta_i = 1 \right\}, \quad p_\theta = \sum_j \theta_j p_j$$

$$K(\theta, y) = -\ln p_\theta(y), \quad k(\theta) = -\int \ln(p_\theta(x)) p_*(x) \mu(dx).$$

Un minimiseur  $\theta_*$  de  $k$  minimise donc la divergence de Kullback entre  $p_\theta$  et  $p_*$ .

- (a) Vérifier que l'estimateur (III.48) conduit à

$$\widehat{p}_{\widehat{\theta}_n}(x) = \frac{\int p_\theta(x) p_\theta(Y^n) \pi(d\theta)}{\int p_\theta(Y^n) \pi(d\theta)}, \quad p_\theta(Y^n) = \prod_{i=1}^n p_\theta(Y_i).$$

- (b) On suppose que  $\pi$  est la mesure uniforme sur les  $m$  éléments de la base canonique. Montrer que  $\bar{p}_n = \frac{1}{n+1} \sum_{j=0}^n \widehat{p}_j$  satisfait

$$E[r(\bar{p}_n, p_*)] \leq \min_k r(p_k, p_*) + \frac{\log m}{n+1}.$$

- (c) Montrer que  $\widehat{\theta}$  est  $\pi$ -bayésien pour le risque de Kullback  $r(\theta, \theta') = \int \ln(p_\theta(x)/p_{\theta'}(x)) p_\theta(x) \mu(dx)$ . *Indication* : Utiliser la formule (III.34). On montrera directement que pour tout  $\theta'$ ,  $\int \{r(\theta, \theta') - r(\theta, \widehat{\theta})\} p_\theta(y) \pi(d\theta) \leq 0$  (faire apparaître une divergence de Kullback).

**Exercice 14 (Inégalité oracle en agrégation de modèles linéaires).** On observe un vecteur gaussien d'espérance inconnue  $y \sim \mathcal{N}(\mu, \sigma^2 Id)$ . On considère pour ces données une famille d'estimateurs linéaires  $\hat{\mu}_m(y) = P_m y$ ,  $m = 1, \dots, M$ , basé chacun sur un modèle de régression de matrice de design  $X_m$  (cf. p. 9; pour chacun de ces modèles,  $P_m$  sera donc la matrice de projection sur l'espace image de  $X_m$ ,  $P_m = X_m(X_m^T X_m)^{-1} X_m^T$  si  $X_m$  est de rang plein). L'idée est qu'un de ces estimateurs est bon, on ne sait lequel; on voudrait faire aussi bien que s'il était connu. Ces estimateurs peuvent être assez nombreux, par exemple tous les estimateurs linéaires associés chacun à une sous-matrice  $X_m$  d'une certaine matrice de design  $X$ .

Plutôt que d'essayer de choisir le meilleur estimateur par une procédure à la Mallows (cf. exercice 1 p. 16), Barron et Leung [105] proposent de réaliser une combinaison linéaire  $\hat{\mu}$  de ces derniers pour laquelle le risque  $\|\hat{\mu} - \mu\|^2$  sera réduit. L'estimateur se présente ainsi

$$\begin{aligned}\hat{\mu} &= \sum_m w_m \hat{\mu}_m = \sum_m w_m P_m y \\ w_m &= \frac{e^{-\hat{r}_m/4\sigma^2} \pi_m}{\sum_{m'} e^{-\hat{r}_{m'}/4\sigma^2} \pi_{m'}} \\ \hat{r}_m &= \|y - \hat{\mu}_m\|^2 + 2\sigma^2 \text{Tr}(P_m) - n\sigma^2.\end{aligned}$$

Les  $\pi_m$  sont des poids à choisir librement, dont la somme fait 1. On notera que  $\hat{r}_m$  est un estimateur sans biais de  $E[\|\mu - \hat{\mu}_m\|^2]$  (cf. exercice 1 p. 16); on aura également besoin de l'équation I.25 appliquée à  $\hat{\mu}$ :

$$\|\mu - \hat{\mu}\|^2 = \|y - \hat{\mu}\|^2 + 2\sigma^2 \text{Tr}(\nabla \hat{\mu}) - n\sigma^2 + \xi$$

où  $\xi$  est une variable centrée.

1. Montrer que  $2\sigma^2 \nabla w_m = w_m(\hat{\mu}_m - \hat{\mu})$ , puis en déduire

$$2\sigma^2 \text{Tr}(\nabla \hat{\mu}) = -\|\hat{\mu} - y\|^2 + \sum w_m \hat{r}_m + n\sigma^2.$$

2. Montrer que pour tout  $m_0$

$$\sum_m w_m \hat{r}_m = \hat{r}_{m_0} + 4\sigma^2 \left( \log \frac{w_{m_0}}{\pi_{m_0}} - D(w|\pi) \right), \quad D(w|\pi) = \sum_m w_m \log \frac{w_m}{\pi_m}.$$

Il est classique que la divergence  $D(w|\pi)$  est  $\geq 0$ .

3. Dédurre des deux points précédents que pour tout  $m_0$

$$E[\|\mu - \hat{\mu}\|^2] \leq E[\|\mu - \hat{\mu}_{m_0}\|^2] - 4\sigma^2 \log \pi_{m_0}. \quad (\text{III.51})$$

Le premier terme à droite de (III.51) vaut  $\sigma^2 \text{Tr}(P_{m_0}) + \|Q_{m_0} \mu\|^2$ . L'estimateur fusionné fait donc aussi bien que le meilleur de tous, à un terme additionnel près. Pour des compléments sur l'optimalité de cet estimateur dans un contexte de régression en grande dimension, et dans le cas où l'on considère tous les sous-modèles associés à une certaine matrice de design  $X$ , on pourra consulter [124]; en particulier il est favorable de choisir des  $\pi_m$  qui décroissent vite avec le rang de  $X_m$ . S'il y a beaucoup de variables au départ (disons  $> 100$ ), le calcul explicite de l'estimateur devient impossible, mais il se présente comme une moyenne que l'on peut estimer par des méthodes de Monte Carlo.



# IV

---

## TESTS

---

---

Ce chapitre se concentre sur les tests décidant de l'appartenance d'un paramètre à un certain ensemble, et sur le rôle des résultats des chapitres précédents pour les réalisations de ces tests. Il ne faut pas y voir un exposé général de la théorie des tests. Pour des compléments, nous renvoyons à [15] ou [16], et [21] pour la théorie asymptotique.

### IV.1 Théorie non-asymptotique

#### IV.1.1 Introduction

##### 31 - DÉFINITION

On appelle hypothèse une famille de lois de probabilité sur les observables. Une hypothèse  $H$  est dite simple si elle ne contient qu'un seul élément ; sinon elle est composite.

Par exemple les deux hypothèses  $H_0$  : *La suite  $Y = (Y_1, \dots, Y_n)$  est iid de moyenne nulle*, et  $H_1$  : *La suite  $Y = (Y_1, \dots, Y_n)$  est iid de moyenne non-nulle* sont composites.

Dans la situation paramétrique,  $H$  est simplement une partie de  $\Theta$ . On considère en général deux hypothèses notées traditionnellement  $H_0$  et  $H_1$  et appelées *hypothèse nulle* et *hypothèse alternative*. Par exemple  $H_0$  : *La température de la terre reste constante en moyenne*,  $H_1$  : *La température de la terre est en augmentation*.  $H_0$  représentera une absence de changement (ou d'effet d'un médicament ...) tandis que  $H_1$  représentera la présence d'un certain effet. La dissymétrie introduite vient de ce que l'on cherchera à minimiser d'abord l'erreur de première espèce (décider  $H_1$  alors que  $H_0$  est vraie) qui a en pratique un rôle très différent de celui de l'erreur de seconde espèce (décider  $H_0$  alors que  $H_1$  est vraie) : il est prioritaire de ne pas prétendre voir un effet significatif en absence d'effet.

On ne s'intéressera dans la suite qu'aux cas paramétriques ou semi-paramétriques, où  $H_i$  s'exprime par  $\theta_* \in \Theta_i$ .

##### 32 - DÉFINITION

Un test randomisé est une application  $\varphi(Y)$  de  $\Omega$  dans  $[0, 1]$ , qui ne dépend que de l'échantillon. Il est dit pur s'il est à valeurs dans  $\{0, 1\}$ .

En général, on considère des tests purs, et la valeur de  $\varphi(Y)$  indique l'hypothèse choisie. Si le test n'est pas pur, la procédure complète consiste à choisir  $H_0$  ou  $H_1$  aléatoirement avec probabilités  $1 - \varphi(Y)$  et  $\varphi(Y)$ . L'intérêt des tests randomisés est purement théorique : c'est d'obtenir tout les niveaux possibles dans le théorème de Neyman-Pearson énoncé plus bas.

En pratique, on choisit une statistique  $S(Y)$  dont on pense qu'elle sera petite sous  $H_0$  et grande sous  $H_1$  et le test aura la forme  $\varphi = 1_{S > \lambda}$  pour un certain seuil  $\lambda$  à choisir en fonction de l'erreur de première espèce que l'on s'autorise, sans l'augmenter de trop pour éviter de décider  $H_0$  systématiquement. Par exemple, si  $Y = (Y_1, \dots, Y_n)$  est une suite gaussienne i.i.d, un test raisonnable pour décider si la moyenne des  $Y_i$  est nulle ( $H_0$ ) ou pas ( $H_1$ ) est :  $\varphi(Y) = 0$  si  $|\hat{\mu}| < 2,6 \hat{\sigma}/\sqrt{n}$  et 1 sinon ( $\hat{\mu}$  et  $\hat{\sigma}^2$  sont les moyenne et variance empiriques). Le seuil  $\lambda = 2,6$  limite la probabilité d'erreur de première espèce à environ 1%, car  $\hat{\mu}\sqrt{n}/\hat{\sigma}$  est approximativement gaussienne sous  $H_0$  (c'est en fait une loi de Student).

Dans un cadre paramétrique, la probabilité de choisir  $H_1$  si  $\theta \in \Theta_0$  (i.e. sous  $H_0$ ) n'est autre que  $E_\theta[\varphi]$ ; cette observation conduit naturellement à la

**33 - DÉFINITION**

Dans le cadre paramétrique, les fonctions

$$\begin{aligned} \alpha(\theta) &= E_\theta[\varphi], & \theta \in \Theta_0 \\ \beta(\theta) &= 1 - E_\theta[\varphi], & \theta \in \Theta_1 \end{aligned}$$

sont appelées risque de première espèce (ou niveau de risque, ang. size) et risque de seconde espèce. La fonction  $1 - \beta(\theta)$  est la puissance.

Plus généralement, on appelle risque (ou niveau de risque) de  $\varphi$  et puissance les nombres

$$\begin{aligned} \alpha &= \sup_{H_0} E[\varphi] \\ 1 - \beta &= \inf_{H_1} E[\varphi]. \end{aligned}$$

Attention, on parle également du *niveau de confiance* qui est  $1 - \alpha$ . Si bien que niveau (tout court) peut désigner  $1 - \alpha$  ou, plus rarement,  $\alpha$ .

### IV.1.2 Approche UPP

**34 - DÉFINITION**

Un test  $\varphi$  est dit uniformément plus puissant (UPP) au seuil  $\alpha$  si son risque est inférieur à  $\alpha$  et si pour tout test  $\varphi'$  de risque inférieur ou égal à  $\alpha$ , on a  $E[\varphi] \geq E[\varphi']$  sous  $H_1$ .

Si l'on réduit  $H_0$  ou  $H_1$  le test reste UPP avec un niveau et une puissance améliorés.

Comme dans la théorie des estimateurs UMR, cette définition est très bonne mais la difficulté est de montrer qu'il existe de tels tests; c'est généralement faux.

Dans le cas de deux hypothèses simples, le concept UPP se réduit à PP. Introduisons les tests de Neyman, qui consistent à mettre un seuil sur le rapport de vraisemblance :

**35 - DÉFINITION**

On appelle test de Neyman un test tel que  $\mu$ -presque sûrement :

$$\begin{cases} \varphi(y) = 0 & \text{si } p_0(y) > \lambda p_1(y) \\ \varphi(y) = 1 & \text{si } p_0(y) < \lambda p_1(y) \end{cases}$$

où  $\lambda$  est un réel positif.

*Remarque.* Il est immédiat de vérifier que tous les niveaux sont possibles par un bon choix de  $\lambda$  et d'une valeur constante  $\gamma$  de  $\varphi$  sur  $A = \{y : p_0(y) = \lambda p_1(y)\}$  puisque le niveau est

$$\alpha = P_0(p_0(Y) > \lambda p_1(Y)) + \gamma P_0(p_0(Y) = \lambda p_1(Y))$$

et que si la fonction  $f(\lambda) = P_0(p_0(Y) > \lambda p_1(Y))$  ne prend par la valeur  $\alpha$ , alors elle traverse cette valeur en un point où  $P_0(A) \neq 0$  et  $f(\lambda) < \alpha$ .

Généralement  $A$  est de  $\mu$ -mesure nulle et l'on peut prendre  $\gamma = 0$  ou  $\gamma = 1$  sans rien changer ; dans le cas contraire, le meilleur test n'est pas un test pur. On a le résultat fondamental :

**36 - THÉORÈME** (NEYMAN-PEARSON, 1933)

On se place dans le cadre de deux hypothèses simples. Les tests PP sont exactement les tests de Neyman.

*Démonstration.* Soit  $\varphi$  un test de Neyman et  $\psi$  un autre test de même niveau  $\lambda$ , alors la puissance vaut

$$\int \psi(y)p_1(y)\mu(dy) = \int \varphi(y)p_1(y)\mu(dy) - \int (\varphi(y) - \psi(y)) \left( p_1(y) - \frac{p_0(y)}{\lambda} \right) \mu(dy)$$

Comme  $0 \leq \psi \leq 1$ , il est immédiat de remarquer que l'intégrande du dernier terme est  $\geq 0$  (considérer les trois ensembles  $\{p_0 = \lambda p_1\}$ ,  $\{p_0 < \lambda p_1\}$  et  $\{p_0 > \lambda p_1\}$ ), et n'est nul que si le  $\psi$  est un test de Neyman. ■

Ce théorème a trouvé une généralisation pour le cas composite du test de  $H_0 : \theta \leq \theta_0$  contre  $H_1 : \theta > \theta_0$  ; elle requiert une hypothèse très forte :

**37 - DÉFINITION**

On dit que la famille  $p_\theta(y)$  est à vraisemblance croissante s'il existe une fonction  $T(y)$  telle que pour tous  $\theta_1 > \theta_0$ , le rapport  $p_{\theta_1}(y)/p_{\theta_0}(y)$  soit fonction strictement croissante de  $T(y)$ .

C'est le cas des familles exponentielles de la forme

$$p_\theta(y) = \exp\{Q(\theta)T(y) - Z(\theta)\} \tag{IV.1}$$

pour lesquelles  $Q$  est strictement croissante (si  $Q$  est décroissante, remplacer  $(Q, T)$  par  $(-Q, -T)$ ).

**38 - THÉORÈME** (LEHMANN, 1955)

Si la famille  $p_\theta(y)$  est à vraisemblance croissante, les tests de la forme

$$\begin{cases} \varphi(y) = 0 & \text{si } T(y) < \lambda \\ \varphi(y) = \gamma & \text{si } T(y) = \lambda \\ \varphi(y) = 1 & \text{si } T(y) > \lambda. \end{cases}$$

sont UPP pour tester  $H_0 : \theta \leq \theta_0$  contre son contraire.

*Démonstration.* Notons avant tout que pour  $\theta_1 > \theta_0$ , il s'agit d'un test de Neymann de niveau

$$\alpha = P_{\theta_0}(T(y) > \lambda) + \gamma P_{\theta_0}(T(y) = \lambda)$$

de  $\theta = \theta_0$  contre  $\theta = \theta_1$ . Montrons que pour tout  $\theta < \theta_0$ , le niveau est encore  $\leq \alpha$  :

$$\begin{aligned} P_\theta(\text{décider } H_1) &= P_\theta(T(y) > \lambda) + \gamma P_\theta(T(y) = \lambda) \\ &= (1 - \gamma)E_{\theta_0} \left[ 1_{T > \lambda} \frac{p_\theta}{p_{\theta_0}} \right] + \gamma E_{\theta_0} \left[ 1_{T \geq \lambda} \frac{p_\theta}{p_{\theta_0}} \right]. \end{aligned}$$

Pour montrer que cette quantité est  $\leq \alpha$ , il suffit d'utiliser que pour toute paire de variables  $X$  (ici  $1_{T > \lambda}$ ) et  $Y$  (ici  $-\frac{p_\theta}{p_{\theta_0}}$ ) fonction croissante l'une de l'autre (c.-à-d.  $(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \geq 0$ )

avec  $P \times P$ -probabilité 1) on a  $E[XY] \geq E[X]E[Y]$  (pour le montrer, prendre l'espérance dans le produit ci-dessus).

Soit  $\theta_1 > \theta_0$ , comme  $\varphi$  est un test de Neyman  $\theta_0$  contre  $\theta_1$ , on a  $E_{\theta_1}[\varphi] \geq E_{\theta_0}[\varphi]$  pour tout autre test de risque  $\alpha$  d'hypothèse nulle  $\theta_0$ , et donc en particulier pour tout test de risque  $\alpha$  d'hypothèse nulle  $H_0$ . ■

Il existe un résultat analogue au théorème précédent pour le test de  $H_0 : \theta \leq \theta_1$  ou  $\theta \geq \theta_2$  contre  $H_1 : \theta_1 < \theta < \theta_2$  dans le cas d'une famille exponentielle de la forme (IV.1) avec  $Var_{\theta}(T(y)) > 0$  pour tout  $\theta$ , et le test prend la forme attendue [16] :

$$\varphi(y) = \gamma_1 1_{T(y)=\lambda_1} + \gamma_2 1_{T(y)=\lambda_2} + 1_{\lambda_1 < T(y) < \lambda_2}.$$

On vient de voir l'essentiel des situations pour lesquelles il existe un test UPP. En particulier, dans le cas d'hypothèses du type  $H_0 : \theta_1 \leq \theta \leq \theta_2$  contre son contraire (en particulier  $H_0 : \theta = \theta_1$ ), on montre qu'il n'y a pas de test UPP. La solution est alors, comme en estimation, de restreindre la classe des tests considérés :

**39 - DÉFINITION**

Un test  $\varphi$  est dit *sans biais* si sa puissance est toujours supérieure à son niveau.

C'est-à-dire que décider  $H_1$  est toujours plus vraisemblable sous  $H_1$  que sous  $H_0$ . Le test biaisé type est le test (dissymétrique) de comparaison de  $|\hat{\mu} - a|$  à un seuil, avec  $a \neq 0$ , alors que  $H_0$  est la nullité de la moyenne de la suite  $(Y_i)$  iid  $\mathcal{N}(\mu, 1)$  (vérifier !)<sup>1</sup>.

Si l'on se restreint aux tests non-biaisés, on trouve alors des tests UPPS (UPP sans biais) pour  $H_0 : \theta_1 \leq \theta \leq \theta_2$  dans le cas des familles exponentielles décrites ci-dessus [15]. Par exemple, la comparaison de la valeur absolue de la moyenne empirique  $\hat{\mu}$  à un seuil donne un test UPPS pour décider si la moyenne d'une gaussienne de variance 1 est nulle (vérifier que ce test n'est pas UPP !).

Si le problème possède des propriétés de symétrie, la classe des tests à considérer peut être également réduite par des arguments d'**invariance** : par exemple, la statistique  $|\hat{\mu}|/\hat{\sigma}$  est bien invariante par multiplication des données par une constante de signe arbitraire. De même, si  $Y_i$  est multidimensionnel, pour tester si la moyenne est nulle, on cherchera un test invariant par rotation des observations, à moins que certaines directions soient privilégiées.

**Aspects pratiques.** En pratique, l'approche la plus utilisée est celle du test du **rapport de vraisemblance généralisé** (Generalised Likelihood Ratio Test), qui peut être vu comme une extension du test de Neyman

$$T(y) = \frac{\sup_{\theta \in \Theta_0} p_{\theta}(y)}{\sup_{\theta \in \Theta_1} p_{\theta}(y)} \geq \lambda.$$

Dans le cas de deux ensembles  $\Theta_0$  et  $\Theta_1$  compacts séparés, on peut mettre en évidence des propriétés d'optimalité asymptotique de ce test [2].

Si l'on veut tester contre toute autre hypothèse  $H_1$  on peut choisir  $T(y) = \sup_{\theta \in \Theta_0} p_{\theta}(y)$ .

Si  $H_0$  n'est pas simple, une autre approche consiste à utiliser une statistique  $T(y)$  **libre** sous  $H_0$ , c.-à-d. dont la loi ne dépend pas de  $\theta \in H_0$ . Supposons par exemple que les  $Y_i$  soient iid de densité  $h(y - \theta)$

1. Le lien avec la définition des estimateurs sans biais est le suivant. Si l'on considère la fonction de perte  $r(H_i, \varphi) = a_i 1_{\varphi \neq i}$ , où  $a_0$  et  $a_1$  sont deux réels positifs, l'équation (III.1) devient

$$a_i P_{H_i}(\varphi = i) \geq a_{1-i} P_{H_i}(\varphi = 1 - i), \quad i = 0, 1$$

soit

$$P_{H_0}(\varphi = 1) \leq \frac{a_1}{a_0 + a_1} \leq P_{H_1}(\varphi = 1).$$

Nous renvoyons à [16] pour plus de détails.

sur  $\mathbb{R}^2$ , avec  $\int y h(y) dy = 0$ . Si  $H_0$  est  $\theta_1 = a$ , on pourra utiliser la statistique  $T(Y) = |\bar{Y}_1 - a|$ , où  $\bar{Y}_1$  est la moyenne empirique de la première coordonnée. La loi de  $T(Y)$  est bien la même pour toutes les distributions de  $H_0$ , ce qui facilite le calcul du niveau. En analyse de variance, le rapport de vraisemblance généralisé est une statistique libre pour les hypothèses linéaires (test de Fisher).

Le calcul explicite des seuils pose souvent problème, mais on peut les estimer par simulation (sous  $H_0$ ). Dans le cas où  $H_0$  n'est pas simple et que la statistique n'est pas libre, on commence par estimer  $\theta$  sous la contrainte  $\theta \in \Theta_0$  et l'on simule sous la loi correspondante; il est recommandé que la statistique soit asymptotiquement libre. Les niveaux asymptotiques pour les tests de significativité sont connus dans de nombreux cas, c'est l'objet du § IV.2.

**Approche minimax.**  $H_0$  et  $H_1$  sont ici composites et séparées. Il s'agit de maximiser la puissance à niveau fixé. C'est un cadre moins exigeant que l'approche UPP. Cette situation a été traitée dans certains articles difficiles [37, 91, 59], le résultat étant essentiellement que si  $H_0$  et  $H_1$  sont convexes, le test minimax est un test de Neyman d'une certaine hypothèse simple  $P \in H_0$  contre une autre  $Q \in H_1$  ces deux probabilités étant en un sens les plus proches sous la contrainte d'appartenir l'une à  $H_0$  et l'autre à  $H_1$ .

Donnons un énoncé partiel, car les questions d'existence sont évitées, mais qui présente le premier avantage d'explicitier le critère de proximité entre  $P$  et  $Q$  (c.-à-d.  $\int |p - \lambda q| d\mu$  pour un  $\lambda$  réglant du niveau) ce qui permet la construction effective du test, et le second de se démontrer simplement :

40 - THÉORÈME

On suppose que  $H_0$  et  $H_1$  sont deux familles de densités par rapport à une mesure  $\mu(dx)$ , et que  $H_0$  comme  $H_1$  est convexe. Soit  $\lambda > 0$  et la norme  $L_1 : \|f\|_1 = \int |f(x)| \mu(dx)$ .  
On suppose qu'existent  $p \in H_0$ ,  $q \in H_1$  tels que

$$\forall p' \in H_0, \forall q' \in H_1, \|p - \lambda q\|_1 \leq \|p' - \lambda q'\|_1$$

et que

$$\mu(x : p(x) = \lambda q(x)) = 0.$$

Alors le test défini par

$$\varphi = 1_{p(x) < \lambda q(x)}$$

minimise la pondération des risques de première et seconde espèce

$$\sup_{H_0} P(\varphi = 1) + \lambda \sup_{H_1} P(\varphi = 0).$$

Tout autre test de même niveau  $\sup_{H_0} P(\varphi = 1)$  a une puissance détériorée.

*Démonstration.* Dans cette démonstration, on note  $\int f d\mu$  pour  $\int f(x) \mu(dx)$ . Interprétons la première inégalité. Si  $p' \in H_0$ , alors pour tout  $0 < \varepsilon < 1$ ,  $p + \varepsilon(p' - p) \in H_0$  et donc

$$\int |p - \lambda q| d\mu \leq \int |p + \varepsilon(p' - p) - \lambda q| d\mu.$$

En utilisant que pour tout  $a \neq 0$  on a  $\varepsilon^{-1}(|a + \varepsilon b| - |a|)$  converge vers  $\text{sgn}(a)b$  quand  $\varepsilon \rightarrow 0$ , on peut faire la différence des deux membres divisée par  $\varepsilon$  et appliquer le théorème de convergence dominée, ce qui donne

$$\int \text{sgn}(p - \lambda q)(p' - p) d\mu \geq 0$$

et comme le signe vaut  $1 - 2\varphi$  on obtient finalement

$$\int (p - p') \varphi d\mu \geq 0. \tag{IV.2}$$

En raisonnant de même sur  $q$  on obtient pour tout  $q' \in H_1$  :

$$\int (q - q')(1 - \varphi) d\mu \geq 0. \quad (\text{IV.3})$$

Soit maintenant  $\psi$  un autre test, on a en notant  $P_{H_0}(\psi = 1)$  pour  $\sup_{H_0} P(\psi = 1)$

$$\begin{aligned} P_{H_0}(\psi = 1) + \lambda P_{H_1}(\psi = 0) &\geq P_p(\psi = 1) + \lambda P_q(\psi = 0) \\ &= \int \psi p d\mu + \lambda \int (1 - \psi)q d\mu \\ &= \int (\psi - \frac{1}{2})(p - \lambda q) d\mu + \frac{\lambda + 1}{2} \\ &\geq \frac{1}{2}(\lambda + 1 - \|p - \lambda q\|_1). \end{aligned}$$

Montrons qu'on a égalité dans le cas où  $\psi = \varphi$ , pour cela on utilise les relations (IV.2) et (IV.3)

$$\begin{aligned} P_{H_0}(\varphi = 1) + \lambda P_{H_1}(\varphi = 0) &= \max_{p' \in H_0} \int \varphi p' d\mu + \lambda \max_{q' \in H_1} \int (1 - \varphi)q' d\mu \\ &\leq \int p\varphi d\mu + \lambda \int (1 - \varphi)q d\mu \\ &= \frac{1}{2}(\lambda + 1 - \|p - \lambda q\|_1). \end{aligned}$$

Le test minimise donc bien l'erreur pondérée. Il reste à montrer que si  $\psi$  a un niveau meilleur la puissance se détériore; mais c'est évident car si  $\psi$  avait une puissance strictement meilleure, son risque pondéré serait strictement meilleur que celui de  $\varphi$ . ■

Si pour  $k = 0, 1$ ,  $H_k$  est de la forme  $\{p : p(x) = \sum_{i=1}^d \theta_i p_i(x), \theta \in \Theta_k\}$  où  $\Theta_0$  et  $\Theta_1$  sont des parties convexes de  $\mathbb{R}^d$  et  $p_i$  une famille de densités, alors la fabrication du test nécessitera de résoudre

$$\min_{\theta_0 \in \Theta_0, \theta_1 \in \Theta_1} \int \left| \sum_i (\theta_{0i} - \lambda \theta_{1i}) p_i(x) \right| \mu(dx)$$

ce qui n'est pas simple mais reste un problème convexe en dimension finie.

### IV.1.3 Dualité tests/régions de confiance.

#### 41 - DÉFINITION

On appelle région de confiance de niveau de confiance  $1 - \alpha$  tout domaine aléatoire  $\mathcal{R} = \mathcal{R}(Y) \subset \Theta$  tel que pour tout  $\theta$ ,  $P_\theta(\theta \in \mathcal{R}) \geq 1 - \alpha$ .

Un exemple typique est  $\mathcal{R} = \{\theta : \|\theta - \hat{\theta}\| \leq \delta\}$ . Un autre exemple est  $\mathcal{R} = \{\theta, S(\theta, Y) \in \mathcal{S}\}$  où  $S(\theta, Y)$  est une fonction pivotale (sa loi ne dépend pas de  $\theta$ ; c'est comme une statistique libre sauf que c'est également une fonction de  $\theta$ ) auquel cas  $P_\theta(\theta \in \mathcal{R})$  est indépendant de  $\theta$ .

C'est un simple exercice de vérifier que si  $\mathcal{R}$  est une région de confiance de risque  $\alpha$ , alors le test qui accepte l'hypothèse  $\theta_* = \theta_0$  ssi  $\theta_0 \in \mathcal{R}$  a un risque inférieur à  $\alpha$ . Il est également simple de vérifier que réciproquement, si l'on a pour tout  $\theta_0 \in \Theta$  un test  $\varphi(Y, \theta_0)$  de risque  $\alpha$  pour décider  $H_0 : \theta_* = \theta_0$  contre son contraire, alors la région aléatoire suivante est une région de confiance de risque  $\alpha$  :

$$\mathcal{R} = \{\theta_0 : H_0 \text{ est acceptée}\} = \{\theta_0 : \varphi(Y, \theta_0) = 0\}.$$

## IV.1.4 Approche bayésienne

Comme dans le cas de l'estimation, cf. § III.4.1, on va voir que la méthode bayésienne donne une solution simple au problème de trouver le meilleur test, mais a le défaut de requérir une probabilité  $\pi(d\theta)$  sur  $\Theta$ .

On a donc une loi globale pour la paire  $(\theta, Y) : p_\theta(y)\mu(dy)\pi(d\theta)$ .  $Y$  a une loi marginale de densité  $p(y) = \int p_\theta(y)\pi(d\theta)$ . La probabilité d'erreur globale est l'espérance de la variable aléatoire valant  $\varphi$  sur  $H_0$  et  $1 - \varphi$  sur  $H_1$  :

$$\begin{aligned} R(\varphi) &= E[1_{\theta \in H_0}\varphi + 1_{\theta \in H_1}(1 - \varphi)] \\ &= \int (\varphi(y)q_0(y) + (1 - \varphi(y))q_1(y))\mu(dy), \quad q_i(y) = \int_{H_i} p_\theta(y)\pi(d\theta) = P(H_i|y)p(y) \\ &= \int (q_0(y) - q_1(y))\varphi(y)\mu(dy) + \pi(H_1). \end{aligned}$$

Par conséquent les tests qui minimisent  $R(\varphi)$  sont exactement ceux qui satisfont

$$\begin{cases} \varphi(y) = 0 & \text{si } \frac{q_0(y)}{q_1(y)} > 1 \\ \varphi(y) = 1 & \text{si } \frac{q_0(y)}{q_1(y)} < 1 \end{cases}$$

la valeur étant arbitraire en cas d'égalité. Noter que ce test revient également à comparer à 1 le rapport de vraisemblance des hypothèses sachant les observations :  $P(H_0|Y)/P(H_1|Y)$ . C'est aussi un rapport de vraisemblance généralisé où les sup sont remplacés par des intégrales.

Noter que si  $\pi$  est non-atomique et  $H_0$  est simple, alors  $q_0(y) = 0$  et le test choisit toujours  $H_1$  (puisque'on a aucune chance d'être sous  $H_0$ ).

## IV.2 Seuils asymptotiques

La normalité asymptotique permet de construire des tests qui sont asymptotiquement optimaux. On ne traitera pas de l'optimalité de ces tests (en ce qui concerne le sujet délicat de la comparaison des tests par leur propriétés asymptotiques, on pourra consulter le chapitre 10 de [20]) mais on s'intéressera essentiellement à leur niveau asymptotique, dans le cas d'une hypothèse simple  $g(\theta) = 0$  contre son contraire.

$$H_0 : g(\theta_*) = 0, \quad H_1 : g(\theta_*) \neq 0$$

pour une certaine fonction  $g$  à valeurs dans  $\mathbb{R}^q$ . On notera  $\hat{\theta}^c$  l'estimateur sous la contrainte  $g(\theta) = 0$ .

### IV.2.1 Cas paramétrique

Notons la vraisemblance  $\mathcal{L}(\theta) = \mathcal{L}(\theta, Y_1, \dots, Y_n)$  et omettons l'indice  $n$  qui reste sous-entendu. Soit  $\hat{\theta}^c = \hat{\theta}_n^c$  l'estimateur au maximum de vraisemblance de  $\theta$  sous la contrainte  $g(\theta) = 0$ . On montre sans problème à l'aide du théorème 16 p. 50, comme on l'a déjà fait pour démontrer le théorème de Wilks (p. 67), que l'on a sous  $H_0$  la convergence en loi de  $S$  vers un  $\chi_q^2$  pour chacun des quatre choix suivants de la statistique  $S$  :

$$S = \begin{cases} 2(\mathcal{L}(\hat{\theta}) - \mathcal{L}(\hat{\theta}^c)) & \text{(statistique du rapport de vraisemblance)} \\ \nabla \mathcal{L}(\hat{\theta}^c) \hat{I}^{-1} \nabla \mathcal{L}(\hat{\theta}^c) & \text{(statistique des scores, ou de Rao)} \\ (\hat{\theta} - \hat{\theta}^c) \hat{I} (\hat{\theta} - \hat{\theta}^c) & \text{(statistique d'Hausman)} \\ g(\hat{\theta})^T (\hat{G} \hat{I}^{-1} \hat{G}^T)^{-1} g(\hat{\theta}) & \text{(statistique de Wald)} \end{cases}$$

où  $\hat{I}$  est l'estimée (ou la vraie valeur) de la matrice d'information non-normalisée (i.e. correspondant à l'échantillon complet), par exemple  $\hat{I} = -\nabla^2 \mathcal{L}(\hat{\theta}^c)$  et  $\hat{G}$  l'estimée de  $\nabla g(\theta_*)$ , par exemple  $\hat{G} = \nabla g(\hat{\theta}^c)$ . Ces quatre statistiques sont donc **asymptotiquement libres**. D'où les tests :

$$\boxed{\text{Rejeter } H_0 \text{ si } S \geq \chi_q^2(1 - \alpha)}$$

où  $\chi_q^2(\cdot)$  désigne la fonction quantile du  $\chi_q^2$ . Ces statistiques sont asymptotiquement égales sous  $H_0$ . Un exemple est le test de sphéricité de la page 68.

L'intérêt de ce calcul n'est pas simplement d'obtenir le seuil du test pour un niveau donné, car on argumentera qu'en pratique il pourra être plus prudent de l'estimer par simulation (le seuil théorique n'est qu'asymptotique); il est surtout de montrer que  $S$  est asymptotiquement libre, ce qui explique son intérêt et valide (asymptotiquement!) la méthode par simulation, et par bootstrap [31, 85].

## IV.2.2 Cas semi-paramétrique

### Tests passant par un estimateur sous contrainte.

► CONTRASTE QUADRATIQUE. On considère l'estimateur en deux temps  $\hat{\theta}_c$  de la page 51. On a le test des scores (ou d'Hausman puisque le contraste est quadratique)

$$\boxed{\text{Rejeter } H_0 \text{ si } (\hat{\theta}^c - \hat{\theta})^T V^{-1} (\hat{\theta}^c - \hat{\theta}) \geq \chi_q^2(1 - \alpha)}$$

où  $V$  est la variance de  $\hat{\theta} - \theta_*$  (ou une estimée). La validité de ce test est une conséquence de (II.53).

► CONTRASTE GÉNÉRAL. On est ici dans le cadre du § II.5 avec les  $Y_i$  iid. Notons  $H = \nabla K$ ,  $\hat{G} = \nabla g(\hat{\theta}^c)$ ,  $\hat{J} = \sum_{i=1}^n \nabla H(\hat{\theta}^c, Y_i)$  et la matrice d'information non-normalisée estimée  $\hat{I} = \sum_{i=1}^n H(\hat{\theta}^c, Y_i) H(\hat{\theta}^c, Y_i)^T$ . On peut aussi les estimer en  $\hat{\theta}$ . Le test de Wald devient :

$$\boxed{\text{Rejeter } H_0 \text{ si } g(\hat{\theta})^T (\hat{G} \hat{J}^{-1} \hat{I} \hat{J}^{-T} \hat{G}^T)^{-1} g(\hat{\theta}) \geq \chi_q^2(1 - \alpha)}$$

Dans le cas d'un minimum de contraste, le test du multiplicateur de Lagrange se déduit de la normalité de ce dernier car  $\hat{\lambda} = -\frac{1}{n} (\hat{G} \hat{G}^T)^{-1} \hat{G} \sum_{i=1}^n \nabla K(\hat{\theta}^c, Y_i)^T$  (cf. théorème 16) :

$$\boxed{\text{Rejeter } H_0 \text{ si } \hat{\lambda} \hat{G} \hat{J}^{-1} \hat{G}^T (\hat{G} \hat{J}^{-1} \hat{I} \hat{J}^{-1} \hat{G}^T)^{-1} \hat{G} \hat{J}^{-1} \hat{G}^T \hat{\lambda} \geq \chi_q^2(1 - \alpha)}$$

Notons finalement que dans le cadre d'un estimateur à minimum de contraste, nous avons vu que si  $I = J$ , le théorème de Wilks (théorème 17) s'applique (cf. l'estimateur (II.57), ou encore (II.51)) ce qui conduit à

$$\boxed{\text{Rejeter } H_0 \text{ si } 2 \sum_i K(\hat{\theta}^c) - K(\hat{\theta}) \geq \chi_q^2(1 - \alpha)} \quad (\text{cas } I = J).$$

### Tests passant par un Z-estimateur.

On s'intéresse ici au test de l'hypothèse simple  $\theta = \theta_0$ , qui par sa simplicité peut se faire dans un cadre assez général, *sans recourir à une estimation* ce qui est important dans les problèmes pratiques où l'on est prêt à investir une grande énergie dans l'estimation de  $\theta_0$  une fois pour toutes mais où des tests réguliers (journaliers...) doivent être faits de manière simple et robuste<sup>2</sup>. De même, toutes les matrices intervenant dans les statistiques seront calculées une fois pour toutes. C'est le problème de la **surveillance** (monitoring). On va également aborder le problème du **diagnostic** qui est de comprendre, si  $\theta = \theta_0$  est refusé, dans quelle direction le paramètre a dévié de sa valeur nominale  $\theta_0$ . *On va dans ce paragraphe normaliser les quantités, contrairement au paragraphe précédent.*

► SURVEILLANCE. Plaçons-nous dans la situation paramétrique<sup>3</sup>, mais plus nécessairement sous l'hypothèse d'indépendance des  $Y_i$ , p.ex. un processus ARMA; on utilise une fonction d'estimation  $H(\theta, y)$  générale. On a alors moyen d'estimer  $I$  et  $J$  avec une précision arbitraire par simulation. Comme  $\hat{\theta}^c = \theta_0$

2. Nous reprenons ici les idées exposées dans [30]. Voir les références pour des applications aux structures mécaniques soumises à des vibrations.

3. Ce n'est en fait pas nécessaire car on utilise essentiellement que  $I$  et  $J$  puissent être estimés.

il est naturel de s'intéresser aux tests ne nécessitant aucune estimation. On a le test des scores qui se déduit simplement de l'hypothèse (II.14)

$$\boxed{\begin{aligned} H_n(\theta_0)^T I^{-1} H_n(\theta_0) &\leq \chi_d^2(1 - \alpha) \\ H_n(\theta) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n H(\theta, Y_i) \\ I &= \text{Var}_{\theta_0}(H_n(\theta_0)) \end{aligned}} \quad (\text{IV.4})$$

► APPROCHE LOCALE. CAS D'UN SURPLUS DE FONCTIONS D'ESTIMATION. L'heuristique présentée ici sera précisée au § IV.3. Pour étudier la puissance considérons l'alternative asymptotique locale  $\theta = \theta_0 + \delta/\sqrt{n}$ . Sous cette alternative on devine que l'on a, sous des hypothèses adéquates, la convergence en loi

$$H_n(\theta) = H_n(\theta_0) - H_n(\theta_0 + \delta/\sqrt{n}) + H_n(\theta_0 + \delta/\sqrt{n}) \xrightarrow{\text{Loi}} \mathcal{N}(-J\delta, I).$$

Si l'on a  $p$  fonctions d'estimation (dimension de  $H$ ) et  $d$  paramètres, c.-à-d. que  $J \in \mathbb{R}^{p \times d}$ , la statistique utilisée dans (IV.4), qui doit maintenant être comparée à un  $\chi_p^2(1 - \alpha)$ , est donc asymptotiquement sous  $H_\delta$  de la forme  $\|I^{-1/2}J\delta + X_p\|^2$ ,  $X_p \sim \mathcal{N}(0, Id_p)$ . En revanche, la statistique du test

$$\boxed{H_n(\theta_0)^T I^{-1} J(J^T I^{-1} J)^{-1} J^T I^{-1} H_n(\theta_0) \leq \chi_d^2(1 - \alpha) \quad (\text{cas } p > d)}$$

(c'est le test précédent appliqué à la fonction d'estimation fusionnée  $H' = J^T I^{-1} H$ , cf. le § II.2.4) est maintenant sous  $H_\delta$  de la forme  $\|(J^T I^{-1} J)^{1/2} \delta + X_d\|^2$ ,  $X_d \sim \mathcal{N}(0, Id_d)$ . Elle est donc plus discriminante, car l'écart de l'espérance de la statistique de test entre  $H_0$  et  $H_\delta$  est inchangé (il vaut  $\delta J^T I^{-1} J \delta$ ) mais sa variance a été réduite.

► DIAGNOSTIC. On peut le faire en testant  $g(\theta) = 0$  pour certaines fonctions  $g$  (p. ex.  $g(\theta) = \theta_i - \theta_{0i}$ ). Dans le cas où  $J$  est carrée inversible, on a les équivalents  $g(\hat{\theta}) \simeq G(\hat{\theta} - \theta_0) \simeq -GJ^{-1}H_n(\theta_0)/\sqrt{n}$  (cf. la formule (II.5)). Sous  $H_0$  la variance du terme ci-dessus est explicite et l'on obtient le test de Wald

$$\boxed{H_n(\theta_0)^T J^{-T} G^T (GJ^{-1} I J^{-T} G^T)^{-1} GJ^{-1} H_n(\theta_0) \leq \chi_q^2(1 - \alpha), \quad G = \nabla g(\theta_0)}$$

où  $q$  est la dimension de  $g$ . Dans le cas d'un surplus de fonctions d'estimation ( $p > d$ ), on procède comme précédemment en appliquant ce dernier test à la nouvelle fonction d'estimation  $H' = J^T I^{-1} H$  et l'on obtient le test de Wald

$$\boxed{\begin{aligned} H_n(\theta_0)^T I^{-1} J_1 (J_1^T I^{-1} J_1)^{-1} J_1^T I^{-1} H_n(\theta_0) &\leq \chi_q^2(1 - \alpha) \\ J_1 &= J(J^T I^{-1} J)^{-1} G^T \end{aligned}}$$

L'espérance de la statistique sous  $H_\delta$  est (asymptotiquement)  $q + \delta^T G^T (G(J^T I^{-1} J)^{-1} G^T)^{-1} G \delta$ .

### IV.3 Puissance asymptotique

On s'intéresse ici au comportement des statistiques de test sous  $H_1$ . Il y a deux types d'approches asymptotiques possibles : soit on fixe  $H_0$  et  $H_1$  et l'on fait tendre le nombre d'échantillons vers l'infini, auquel cas  $\beta$  s'effondre à une vitesse contrôlée par des résultats de type *grandes déviations*, soit on rapproche  $H_1$  de  $H_0$  à une vitesse adéquate de sorte à avoir une puissance toujours d'ordre 1. Cette dernière solution est le point de vue de la théorie asymptotique locale que nous discutons ici. On aura génériquement  $H_0 : \theta_* = \theta_0$  et  $H_1 : \theta_* = \theta_0 + h/\sqrt{n}$  où  $n$  est le nombre d'échantillons ; la raison essentielle est que le rapport de vraisemblance reste alors d'ordre 1. Lucien Le Cam a montré qu'un cadre adapté pour traiter ce problème dans le cas d'une suite générale d'expériences statistiques (cf. I.3.1) est l'hypothèse suivante dite de *normalité asymptotique locale* :

(LAN)  $P_\theta^n$  est une suite d'expériences statistiques. La vraisemblance satisfait

$$\log \left( \frac{P_{\theta+h/\sqrt{n}}^n(Y)}{P_\theta^n(Y)} \right) = h^T S_n(\theta) - \frac{1}{2} h^T I(\theta) h + \varepsilon_n(\theta, h) \quad (\text{IV.5})$$

avec pour tout  $M < \infty$  et tout  $\alpha > 0$

$$\lim_{n \rightarrow \infty} \sup_{|h| \leq M} P_\theta^n(|\varepsilon_n(\theta, h)| > \alpha) = 0.$$

De plus  $I(\theta)$  est déterministe et  $S_n(\theta)$  converge en loi sous  $P_\theta^n$  vers  $\mathcal{N}(0, I(\theta))$ .

Cette hypothèse exprime que le score est asymptotiquement localement une statistique suffisante. Notons que cette approche peut être vue comme une façon de mesurer l'écart entre les lois  $P_\theta$  sur les processus, bien que les mesures correspondantes soient étrangères. Énonçons tout de suite le résultat suivant qui montre que ces hypothèses sont vérifiées dans le cas régulier indépendant :

42 - THÉORÈME

On suppose que  $P_\theta(dy) = p_\theta(y)\mu(dy)$  définit une expérience R-régulière. Soit  $Y_i$  des variables indépendantes de loi commune  $P_\theta(dy)$ . Alors l'hypothèse LAN est satisfaite avec pour  $I(\theta)$  l'information de Fisher et

$$S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\nabla_\theta p_\theta}{p_\theta}(Y_i).$$

La démonstration, difficile, se trouve à l'appendice A.2. La propriété LAN est également vérifiée pour des chaînes de Markov générales [90] et même pour des processus à source markovienne cachée [35]. Il est simple de vérifier qu'elle est satisfaite pour le modèle AR(1) gaussien avec variance connue.

43 - LEMME (LE CAM, 1960)

Soit  $(\Omega, \mathcal{B})$  un espace d'événements et  $\mathcal{B}_n$  une suite croissante de sous-tribus de  $\mathcal{B}$ . Soit  $\{P_n\}$  et  $\{Q_n\}$  deux suites de lois définies sur  $\mathcal{B}_n$ . Soit  $\Lambda_n$  la densité de  $Q_n$  par rapport à  $P_n$  (i.e.  $Q_n = \Lambda_n P_n + R_n$  où  $R_n$  est une mesure positive singulière par rapport à  $P_n$ ) et  $T_n$  une suite de variables aléatoires  $\mathcal{B}_n$ -mesurables. Si

$$(T_n, \Lambda_n) \xrightarrow{d} (T, \Lambda) \quad \text{sous } P_n$$

avec  $E[\Lambda] = 1$ , alors

$$T_n \xrightarrow{d} T' \quad \text{sous } Q_n$$

où la loi de  $T'$  est donnée par

$$E[f(T')] = E[f(T)\Lambda] \quad (\text{IV.6})$$

pour toute  $f$  borélienne bornée.

*Démonstration.* Pour simplifier, on note  $P_n[\cdot]$ ,  $Q_n[\cdot]$ ,  $R_n[\cdot]$ , les espérances sous les lois correspondantes. Montrons avant tout que  $R_n[\Omega]$  tend vers 0 :

$$\begin{aligned} R_n[\Omega] &= Q_n[\Omega] - P_n[\Lambda_n] \\ &\leq 1 - P_n[\Lambda_n 1_{\Lambda_n < c}] \\ \overline{\lim}_n R_n[\Omega] &\leq 1 - E[\Lambda 1_{\Lambda < c}] \end{aligned}$$

qui est arbitrairement petit. Soit maintenant  $f$  une fonction positive continue bornée, et  $c$  un point de continuité de la distribution de  $\Lambda$  :

$$\begin{aligned}
Q_n[f(T_n)] &= P_n[f(T_n)\Lambda_n] + R_n[f(T_n)] \\
&= P_n[f(T_n)\Lambda_n 1_{\Lambda_n < c}] + P_n[f(T_n)\Lambda_n 1_{\Lambda_n \geq c}] + O(R_n[\Omega]) \\
&= E[f(T)\Lambda 1_{\Lambda < c}] + o_c(1) + O(1)P_n[\Lambda_n 1_{\Lambda_n \geq c}] + o(1) \\
&= E[f(T)\Lambda 1_{\Lambda < c}] + O(1)(1 - P_n[\Lambda_n 1_{\Lambda_n < c}]) + o_c(1) \\
&= E[f(T)\Lambda 1_{\Lambda < c}] + O(1)E[\Lambda 1_{\Lambda \geq c}] + o_c(1) \\
&= E[f(T)\Lambda] + O(1)E[\Lambda 1_{\Lambda \geq c}] + o_c(1)
\end{aligned}$$

qui est arbitrairement proche de  $E[f(T)\Lambda]$  pour  $n$  grand pour un choix adéquat de  $c$ . ■

#### 44 - THÉORÈME

On se place sous les hypothèses LAN.  $\theta$  et  $h$  sont fixés. Soit  $T_n = T_n(Y_1, \dots, Y_n)$  une suite de statistiques vectorielles. Si l'on a la convergence en loi

$$\begin{pmatrix} T_n \\ S_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Sigma^2 & J^T \\ J & I \end{pmatrix}\right) \quad \text{sous } P_\theta,$$

pour un certain  $(J, \Sigma^2)$ , alors

$$T_n \xrightarrow{d} \mathcal{N}(J^T h, \Sigma^2) \quad \text{sous } P_{\theta+h/\sqrt{n}}.$$

De plus, si  $E_{\theta'}[\|T_n\|^2] < c$  pour un certain  $c$  et  $\theta'$  au voisinage de  $\theta$ , alors

$$J = \lim_n \frac{1}{\sqrt{n}} \nabla_\theta E[T_n].$$

*Remarque.* La dernière affirmation dit que  $J$  est la sensibilité moyenne du test au paramètre.

*Démonstration.* Le premier point est l'application directe du lemme 43. Pour le deuxième point, sous  $P_\theta$ , le théorème de représentation de Skorohod implique l'existence de variables  $(T'_n, S'_n)$  et  $(T', S')$  telles que  $(T'_n, S'_n) \sim (T_n, S_n)$  et  $(T', S') \sim \mathcal{N}(0, \begin{pmatrix} \Sigma^2 & J^T \\ J & I \end{pmatrix})$ , avec  $(T'_n, S'_n) \rightarrow (T', S')$  presque sûrement ; le point (ii) du lemme 51 implique alors la convergence dans  $L_2$  de la suite  $S'_n$ , et le point (i) permet ensuite de conclure que  $E[T'_n S'_n] \rightarrow E[T' S'] = J$  ; en utilisant alors l'équation (A.9) avec  $\psi = T_n$  on obtient

$$J = \lim_n E[T'_n S'_n] = \lim_n E_\theta[T_n S_n] = \lim_n \frac{1}{\sqrt{n}} \nabla_\theta E_\theta[T_n]. \quad \blacksquare$$

**Application aux tests.** Si l'on veut tester l'hypothèse  $\theta$  contre  $\theta + h/\sqrt{n}$  à partir d'une statistique  $T_n$  centrée sous  $P_\theta$ , le théorème 44 indique que le rapport de vraisemblance asymptotique (terme de droite)

$$\log \frac{P_{\theta+h/\sqrt{n}}(T_n)}{P_\theta(T_n)} \simeq h^T J \Sigma^{-2} T_n - \frac{|\Sigma^{-1} J^T h|^2}{2} \quad (\text{IV.7})$$

a une variance  $v = |\Sigma^{-1} J^T h|^2$  sous les deux hypothèses et possède une moyenne de  $\pm v/2$  selon le cas ; ceci conduit à utiliser la statistique de test normalisée

$$\frac{h^T J \Sigma^{-2} T_n}{|\Sigma^{-1} J^T h|}$$

qui a une variance 1 sous les deux hypothèses et dont la moyenne est 0 ou  $|\Sigma^{-1} J^T h|$  selon le cas. La mise en place du test suppose que l'on est capable d'estimer  $J$ .

Noter que si  $T_n = \sqrt{n}(\hat{\theta}_n - \theta)$  où  $\hat{\theta}_n$  est l'estimateur au maximum de vraisemblance, alors sous des hypothèses assez générales,  $J = Id$  et  $\Sigma^{-2} = I(\theta)$ .

Si maintenant on veut tester  $\theta$  contre son contraire, il est naturel de remplacer  $h$  par sa valeur la plus probable dans (IV.7),  $J^T h = T_n$ , ce qui conduit à la statistique

$$|\Sigma^{-1}T_n|^2$$

qui est asymptotiquement un  $\chi^2$  à  $d$  degrés de liberté sous l'hypothèse nulle et possède sous l'alternative une excentricité de  $|\Sigma^{-1}J^T h|^2$ .

## IV.4 Exercices

**Exercice 1.** Soit  $X_1, \dots, X_n$  iid tirées selon la loi exponentielle d'espérance  $\theta$ , et  $\theta_0 < \theta_1$ .

1. Trouver le test U.P.P. de niveau  $\alpha$  pour tester  $H_0 : \theta = \theta_0$  contre  $H_1 : \theta = \theta_1$ .
2. Même question avec  $H_0 : \theta = \theta_0$  contre  $H_1 : \theta > \theta_0$ .
3. Même question avec  $H_0 : \theta \leq \theta_0$  contre  $H_1 : \theta > \theta_1$ .
4. Même question avec  $X_i \sim \mathcal{U}([0, \theta])$ ,  $H_0 : \theta \leq \theta_0$  et  $H_1 : \theta > \theta_0$ .

**Exercice 2.** On reprend l'exercice 1 p. 70. On y a montré que l'estimateur au maximum de vraisemblance  $(\hat{\mu}_n, \hat{\sigma}_n^2)$  converge p.s. vers  $(\mu, \sigma^2)$  avec normalité asymptotique et  $\sqrt{n}(\hat{\mu}_n - \mu) \rightarrow \mathcal{N}(0, \sigma^2)$ . On possède un autre jeu de données indépendant  $T'_1, \dots, T'_n$  de loi gamma et l'on s'intéresse à l'hypothèse  $H_0 : \mu = \mu'$ .

1. L'hypothèse  $H_0$  est-elle simple?
2. Quelle est la limite en loi de  $\hat{\sigma}_n^{-1} \sqrt{n}(\hat{\mu}_n - \mu)$ ?
3. Quelle est la limite en loi sous  $H_0$  de  $\sqrt{n}(\hat{\mu}_n - \hat{\mu}'_n) / (\hat{\sigma}_n^2 + \hat{\sigma}'_n{}^2)^{1/2}$ ?
4. Proposer un test de niveau de confiance asymptotique 95% de  $H_0$  contre son contraire.

**Exercice 3.** Vérifier que l'hypothèse LAN est satisfaite pour le modèle AR(1) gaussien avec variance connue.

# V

---

## L'APPROCHE MARTINGALE

---

---

Dans le chapitre III on a considéré des jeux d'hypothèses pour lesquelles l'indépendance des observations est le cadre où les calculs asymptotiques sur les exemples se font bien. On s'intéresse ici spécifiquement au cas dépendant ou non-stationnaire, la suite  $Y_n$  peut même par exemple diverger, converger... Ceci nous rapproche, dans les cas stationnaires au moins, du chapitre II, mais on va voir une particularité du cas paramétrique qui permet d'exploiter fructueusement la théorie des martingales. Nous traiterons également de processus instables et de régression linéaire. Une partie des résultats mathématiques utiles sont exposés et illustrés dans [10], dont deux théorèmes sont rappelés en appendice D.

### V.1 Le maximum de vraisemblance en situation générale

#### V.1.1 Théorie

On se donne une suite  $(Y_0, Y_1, \dots)$  dont la loi dépend de  $\theta \in \mathbb{R}^d$ . On note  $\hat{\theta}_n$  l'estimateur au maximum de vraisemblance basé sur l'échantillon  $(Y_0, Y_1, \dots, Y_n)$ . On supposera que  $\hat{\theta}_n$  converge presque sûrement vers le vrai paramètre  $\theta_*$  (il n'existe pas de condition générale très pratique pour le vérifier et si (GNU) n'est pas satisfait on devra passer par le théorème 2 p. 23; voir [10] § 6.2 pour plus de détails). On notera

$$\begin{aligned} L_n(\theta) &= \log p_\theta(Y_0, \dots, Y_n) - \log p_\theta(Y_0) \\ L'_n(\theta) &= \nabla_\theta L_n(\theta) \quad (\text{vecteur colonne}) \\ L''_n(\theta) &= \nabla_\theta^2 L_n(\theta) \quad (\text{matrice}) \end{aligned}$$

où  $p_\theta(y_0, \dots, y_n)$  est la densité de la loi par rapport à la mesure  $\mu(dy_0) \dots \mu(dy_n)$ . On a retranché un terme à la vraisemblance (il s'agit donc de la vraisemblance conditionnelle à  $Y_0$ ) de sorte à pouvoir ne travailler que sur les lois conditionnelles ce qui simplifiera les hypothèses (dans le cas indépendant cela ne change rien car  $Y_0$  s'élimine).

On supposera que pour tout  $n$  ces quantités existent et que pour tout  $n$  avec probabilité 1 la densité  $p_\theta^n(y) = p_\theta(Y_n = y | Y_0, \dots, Y_{n-1})$  est R-régulière au sens de la définition p. 65.

La propriété essentielle est que  $L'_n(\theta_*)$  est une martingale sous  $P_{\theta_*}$  : en effet (on pose  $Y^n = (Y_0, \dots, Y_n)$ ) comme  $L_n(\theta) - L_{n-1}(\theta) = \log p_\theta(Y_n | Y^{n-1})$ , on a

$$\begin{aligned} E_{\theta_*} [L'_n(\theta_*) - L'_{n-1}(\theta_*) | Y^{n-1}] &= \int \nabla_\theta \log p_{\theta_*}(y_n | Y^{n-1}) p_{\theta_*}(y_n | Y^{n-1}) \mu(dy_n) \\ &= \int \nabla_\theta p_{\theta_*}(y_n | Y^{n-1}) \mu(dy_n) \\ &= 0. \end{aligned}$$

On a utilisé à la fin le lemme 24. Posons

$$\begin{aligned} U_n &= L'_n(\theta_*) - L'_{n-1}(\theta_*) = \nabla_{\theta} \log p_{\theta_*}(Y_n | Y_0, \dots, Y_{n-1}) \quad (\text{vecteur colonne}) \\ \mathcal{F}_i &= \sigma(Y_0, \dots, Y_i) \\ I_n &= \sum_{i=1}^n E[U_i U_i^T | \mathcal{F}_{i-1}]. \end{aligned}$$

Il faut bien voir que dans le cas iid, on a  $I_n = nI$ .

Considérons par exemple le modèle ARCH

$$\sigma_n^2 = \theta_0 + \sum_{i=1}^q \theta_i Y_{n-i}^2 \quad (\text{V.1})$$

$$Y_n = \sigma_n \varepsilon_n \quad (\text{V.2})$$

où  $\varepsilon_n$  est une suite i.i.d  $\mathcal{N}(0, 1)$  et les paramètres sont  $\geq 0$  avec  $\sum_i \theta_i < 1$ ; on vérifie simplement que

$$\begin{aligned} \log p_{\theta}(y_n | y_0, \dots, y_{n-1}) &= -\frac{y_n^2}{2\sigma_n^2} - \frac{1}{2} \log(\sigma_n^2), \quad \sigma_n^2 = \theta_0 + \sum_{i=1}^q \theta_i y_{n-i}^2 \\ U_n &= \frac{Y_n^2 - \sigma_n^2}{2\sigma_n^4} \nabla_{\theta} \sigma_n^2 = \frac{\varepsilon_n^2 - 1}{2\sigma_n^2} (1, Y_{n-1}^2, \dots, Y_{n-q}^2) \\ I_n &= \sum_{i=1}^n \frac{1}{2\sigma_i^4} (1, Y_{i-1}^2, \dots, Y_{i-q}^2) (1, Y_{i-1}^2, \dots, Y_{i-q}^2)^T. \end{aligned}$$

La vérification des hypothèses de la proposition qui vient pour ce modèle sont laissées en exercice 2 p. 106.

On a alors sous ces notations :

45 - PROPOSITION (Normalité asymptotique des scores)

On suppose que pour tout  $n$  avec probabilité 1 sous  $P_{\theta_*}$  la densité  $p_{\theta_*}^n(y) = p_{\theta}(Y_n = y | Y_0, \dots, Y_{n-1})$  est  $\mathbb{R}$ -régulière au sens de la définition p. 65. S'il existe une matrice p.s. inversible  $I$  telle que

$$n^{-1} I_n \xrightarrow{P} I \quad (\text{V.3})$$

$$n^{-1} \sum_{i=1}^n E[\|U_i\|^2 1_{\|U_{ni}\| > \varepsilon \sqrt{n}} | \mathcal{F}_{i-1}] \xrightarrow{P} 0, \quad \text{pour tout } \varepsilon > 0. \quad (\text{V.4})$$

Alors la suite  $n^{-1/2} L'_n(\theta_*)$  converge en loi, conditionnellement à  $I$ , vers  $\mathcal{N}(0, I)$ <sup>1</sup>.

*Remarque.* La dernière condition est impliquée par  $n^{-1-\eta} \sum_{i=1}^n E[\|U_i\|^{2+2\eta}] \rightarrow 0$  pour un  $\eta > 0$ .

*Démonstration.* On a  $n^{-1/2} L'_n(\theta_*) = \sum X_{ni}$ ,  $X_{ni} = U_i / \sqrt{n}$ . Les hypothèses du théorème 69 p. 125 sont satisfaites puisque

$$\sum_{i=1}^n E[X_{ni} X_{ni}^T | \mathcal{F}_{n,i-1}] = n^{-1} I_n \xrightarrow{P} I.$$

On a donc convergence en loi de  $\sum X_{ni}$  vers  $\mathcal{N}(0, I)$ ; d'où le résultat. ■

1. Ceci signifie que pour toute fonction  $\varphi$  continue bornée, on a la convergence

$$E[\varphi(I, n^{-1/2} L'_n(\theta_*))] \rightarrow \int E[\varphi(I, y) e^{-\frac{1}{2} y^T I^{-1} y} \det(I)^{-1/2}] (2\pi)^{d/2} dy.$$

Pour avoir la normalité asymptotique de  $\widehat{\theta}_n$ , il faut faire comme dans la démonstration du théorème 6 p. 31 :

$$0 = L'_n(\widehat{\theta}_n) = L'_n(\theta_*) + L''_n(\theta'_n)(\widehat{\theta}_n - \theta_*)$$

avec  $\theta'_n \in [\widehat{\theta}_n, \theta_*]$  (dans le cas vectoriel,  $\theta'_n$  change d'une ligne à l'autre de la matrice  $L''_n$ ) ; d'où

$$n^{1/2}(\widehat{\theta}_n - \theta_*) = -\left\{nL''_n(\theta'_n)^{-1}I\right\}\left\{n^{-1/2}I^{-1}L'_n(\theta_*)\right\}$$

On vérifie facilement que  $I_n + L''_n(\theta_*)$  est, sous de faibles conditions, une martingale, car en raison de (III.6), on a que  $E[L''_n - L''_{n-1} + U_n U_n^T | \mathcal{F}_{n-1}] = 0$ . Il est alors naturel de supposer que  $n^{-1}L''_n(\theta_n)$  tend vers  $-I$ , ce qui est l'analogie de la quatrième hypothèse du théorème 6. D'où le théorème :

#### 46 - THÉORÈME

En plus des hypothèses de la proposition 45, on suppose que les fonctions  $L_n(\theta)$  sont, avec probabilité 1, de classe  $C^2$  sur un voisinage fixe de  $\theta_*$  et que  $n^{-1}L''_n(\theta_n)$  converge en probabilité vers  $-I$  pour toute suite de variables aléatoires  $\theta_n$  convergeant presque sûrement vers  $\theta_*$ . Alors tout estimateur  $\widehat{\theta}_n$  tel que  $L'_n(\widehat{\theta}_n) = 0$  et tel que  $\widehat{\theta}_n$  converge vers  $\theta_*$  satisfait

$$n^{1/2}(\widehat{\theta}_n - \theta_*) \longrightarrow \mathcal{N}(0, I^{-1}) \text{ en loi conditionnellement à } I.$$

Ce théorème s'applique sans difficulté au processus AR de l'équation (I.12) (avec  $q = 0$ ) ;  $I$  est alors la matrice de covariance du processus.

Concernant le modèle ARCH de (V.1, V.2), on obtient

$$L''_n(\theta) = \sum_{i=1}^n \frac{1}{2\sigma_i(\theta)^6} (\sigma_i^2(\theta) - 2Y_i^2) M_{i-1}$$

$$M_{i-1} = (1, Y_{i-1}^2, \dots, Y_{i-q}^2)(1, Y_{i-1}^2, \dots, Y_{i-q}^2)^T.$$

Comme  $Y_i = \sigma_i(\theta_*)\varepsilon_i$ , il vient, en notant  $\sigma_{i*} = \sigma_i(\theta_*)$

$$\begin{aligned} L''_n(\theta) + I_n &= (L''_n(\theta) - L''_n(\theta_*)) + (L''_n(\theta_*) + I_n) \\ &= \sum_{i=1}^n \left( \frac{1}{2\sigma_i(\theta)^4} - \frac{1}{2\sigma_{i*}^4} - \left( \frac{1}{\sigma_i(\theta)^6} - \frac{1}{\sigma_{i*}^6} \right) \sigma_{i*}^2 \varepsilon_i^2 \right) M_{i-1} + \sum_{i=1}^n \frac{1 - \varepsilon_i^2}{\sigma_{i*}^4} M_{i-1}. \end{aligned} \quad (\text{V.5})$$

La convergence en probabilité de  $n^{-1}(L''_n(\theta_n) + I_n)$  vers 0 se montre alors sans grande difficulté, en remarquant que  $\sigma_i^2(\theta) > \theta_{*0}/2$  pour  $\theta$  assez proche de  $\theta_*$ , et que le deuxième terme est une martingale (exercice 2 p. 106).

### V.1.2 Exemple : Chaînes de Markov stationnaires

Supposons que  $Y_n, n \geq 0$  est une chaîne de Markov stationnaire ergodique de mesure invariante  $\pi_{\theta_*}(y)\mu(dy)$  et de probabilité de transition  $P(Y_n \in dy | Y_{n-1}) = q_{\theta_*}(y | Y_{n-1})\mu(dy)$ . Pour la convergence p.s. noter que l'on a un estimateur à minimum de contraste assez simple car

$$L_n(\theta) = \sum_{i=1}^n \log(q_{\theta}(Y_i | Y_{i-1})) = \sum_{i=1}^n l(\theta, Z_i), \quad Z_i = (Y_i, Y_{i-1}).$$

Après division par  $n$  le théorème 10 de convergence de l'estimateur à minimum de contraste s'applique à condition que  $l(\theta, Z_i)$  satisfasse (GNU) et à condition que le contraste moyen ait  $\theta_*$  comme unique minimum ; il vaut

$$\begin{aligned} k(\theta) &= -E[\log(q_{\theta}(Y_i | Y_{i-1}))] = -\int \int \log(q_{\theta}(y | x)) q_{\theta_*}(y | x) \pi_{\theta_*}(x) \mu(dy) \mu(dx) \\ &= \int D(Q_{\theta_*}^x | Q_{\theta}^x) \pi_{\theta_*}(x) \mu(dx) + k(\theta_*), \quad Q_{\theta}^x(dy) = q_{\theta}(y | x) \mu(dy). \end{aligned}$$

Le premier terme est  $\geq 0$  et ne s'annule que si  $\pi_{\theta_*}(x)\mu(dx)$ -p.s.,  $Q_{\theta_*}^x = Q_\theta^x$ . Il faut donc que pour tout  $\theta$ , il existe un ensemble  $A$  de  $\pi_{\theta_*}$ -probabilité non nulle tel que pour tout  $x \in A$  les probabilités de transition  $Q_{\theta_*}^x$  et  $Q_\theta^x$  diffèrent (identifiabilité).

La condition de régularité imposée à  $p_\theta^n(y)$  (proposition 45) est satisfaite si pour  $\mu$ -presque tout  $y_0$  la densité  $q_\theta(y|y_0)$  satisfait l'hypothèse de régularité habituelle.

On obtient ensuite

$$\begin{aligned} U_n &= \nabla_\theta \log(q_{\theta_*}(Y_n|Y_{n-1})), \\ E[U_n U_n^T | \mathcal{F}_{n-1}] &= \int \nabla_\theta \log(q_{\theta_*}(y|Y_{n-1})) \nabla_\theta \log(q_{\theta_*}(y|Y_{n-1}))^T q_{\theta_*}(y|Y_{n-1}) \mu(dy) \\ L_n''(\theta) - L_{n-1}''(\theta) &= \nabla^2 \log(q_\theta(Y_n|Y_{n-1})) = \nabla_\theta^2 l(\theta, Z_n). \end{aligned}$$

L'ergodicité implique que  $I_n/n$  converge p.s. vers  $E[I_1]$  dès que  $E_{\theta_*}[\|U_1\|^2] < \infty$ .

L'hypothèse (V.4) est satisfaite si par exemple  $E_{\theta_*}[\|U_1\|^3] < \infty$ .

Il ne reste plus qu'à vérifier l'hypothèse du théorème 46. En vertu de la proposition 5, si  $\nabla^2 l(\theta, Z_n)$  satisfait (GNU) au voisinage de  $\theta_*$ , la suite  $L_n''(\theta_n)/n$  converge vers  $E[\nabla_\theta^2 l(\theta_*, Z)]$ , quantité qui vaut  $-E[I_1]$  si pour  $\mu$ -presque tout  $x$ ,  $\int \nabla_\theta^2 q_\theta(y|x) \mu(dy) = 0$  (cf. la proposition 22 p. 64 appliquée ici à la loi conditionnelle à  $Y_{n-1}$ ), ce qui est le cas si  $\nabla^2 l(\theta, Z_n)$  satisfait (GNU).

Alors  $\sqrt{n}(\hat{\theta}_n - \theta_*)$  a pour variance asymptotique  $E[U_n U_n^T]^{-1}$  si cette quantité est finie. Rappelons les hypothèses :

1.  $\hat{\theta}_n$  reste dans un compact de  $\Theta$  et  $\theta_*$  est intérieur à  $\Theta$ .
2.  $\int |q_\theta(y|x) - q_{\theta_*}(y|x)| \pi_{\theta_*}(x) \mu(dy) \mu(dx) = 0$  implique  $\theta = \theta_*$ .
3.  $l(\theta, Z_i)$  satisfait (GNU).
4. Il existe un voisinage de  $\theta_*$  où presque sûrement  $\nabla_\theta^2 l(\theta, Z_i)$  existe et satisfait (GNU).
5. Pour  $\mu$ -presque tout  $y_0$  la densité  $q_\theta(y|y_0)$  satisfait l'hypothèse de régularité (cf. p. 65), et  $\int \nabla_\theta^2 q_\theta(y|y_0) \mu(dy) = 0$ .
6.  $E_{\theta_*}[\|\nabla_\theta l(\theta, Z_i)\|^3] < \infty$ .

### V.1.3 Exercices et compléments

**Exercice 1.** Soit  $(Y_n)$  une chaîne de Markov telle que conditionnellement à  $Y_n$ ,  $Y_{n+1} \sim \mathcal{P}(\theta_* Y_n + \theta_*)$ . Etudier la convergence du maximum de vraisemblance sous l'hypothèse de stationnarité de la suite.

**Exercice 2.** On considère le processus ARCH  $Y_n$  (V.1, V.2) que l'on suppose stationnaire ergodique<sup>2</sup>. On suppose que  $\theta_i^* > 0$  pour tout  $i > 0$ .

1. Montrer que pour une certaine constante  $C(\theta_*)$  on a  $\|U_i\| \leq C(\theta_*)|\varepsilon_i - 1|$ .
2. Vérifier les hypothèses de la proposition 45 (utiliser la remarque qui suit).
3. Vérifier l'hypothèse additionnelle du théorème 46 (exploiter (V.5)).

**Exercice 3 (Processus de branchement<sup>3</sup>).** Soit  $\{Z_0 = 1, Z_1, Z_2, \dots\}$  un processus de Bienaymé-Galton-Watson avec distribution de renouvellement  $p_j$ ,  $j = 0, 1, 2, \dots$ . On peut réaliser ce processus de la façon suivante : Partir d'un tableau de variables iid  $X_{nk}$  de distribution  $P(X_{nk} = j) = p_j$ , et poser

$$Z_{n+1} = \sum_{k=1}^{Z_n} X_{nk}.$$

2. Par un procédé analogue à celui qui a mené à (I.14),  $Y_n^2$  peut être construit comme  $Y_n^2 = \varphi(\varepsilon_n^2, \varepsilon_{n-1}^2, \dots)$  pour une certaine fonction  $\varphi$ , et donc la loi du processus  $Y_n$  est celle de  $S_n \varphi(\varepsilon_n^2, \varepsilon_{n-1}^2, \dots)$  où  $S_n$  est une suite iid à valeurs 0 ou 1, indépendante des  $\varepsilon_i$ . Ceci implique la stationnarité et l'ergodicité car la suite  $(\varepsilon_n^2, S_n)$  est stationnaire ergodique.

3. On trouvera des compléments dans [10] p.178.

1.  $Z_n$  représente l'effectif d'une génération au temps  $n$ . Interpréter les  $p_j$ .
2. Justifier les relations

$$E[Z_{n+1}|Z_n] = mZ_n$$

$$E[(Z_{n+1} - mZ_n)^2|Z_n] = \sigma^2 Z_n.$$

où  $m = E[Z_1]$  et  $\sigma^2 = Var(Z_1)$ .

3. On suppose  $m, \sigma < \infty$ . Démontrer la convergence presque sûre de  $m^{-n}Z_n$  vers une certaine variable aléatoire  $W$ .
4. On suppose que  $W > 0$  avec probabilité 1, en déduire la consistance de l'estimateur  $\hat{m} = Z_{n+1}/Z_n$ . Proposer par un calcul heuristique, un ordre de grandeur de la variance d'estimation.

**Exercice 4.** On considère le processus  $(Y_i)_{i \geq 0}$  tel que  $Y_0 = 1$  et

$$Y_{i+1} = \begin{cases} Y_i & \text{avec probabilité } e^{-\theta_* - \sum_{j=1}^i Y_j} \\ Y_i/2 & \text{sinon.} \end{cases}$$

pour un  $\theta_* > 0$ . On notera  $p_i(\theta) = e^{-\theta - \sum_{j=1}^i Y_j}$  et  $\chi_{i+1} = 1_{Y_{i+1}=Y_i}$ .

1. Exprimer les quantités  $L'_n(\theta)$ ,  $I_n$ ,  $X_{ni}$ . On précisera également la mesure  $\mu(dy)$ .
2. Démontrer que  $E[\sum_n Y_n] < \infty$ .
3. Vérifier que les équations (V.3, V.4) sont bien satisfaites.
4. Donner un équivalent de  $I_n$  puis montrer que

$$I_n^{-1} L''_n(\theta_n) \xrightarrow{P} -1$$

pour toute suite  $\theta_n$  convergeant vers  $\theta_*$ . On pourra montrer et utiliser que

$$n^{-1} \sum_{i=0}^{n-1} u_i(\chi_{i+1} - p_i(\theta_*))$$

converge dans  $L_2$  vers 0 si la suite  $u_i$  est bornée et adaptée (à une suite de tribus bien choisies).

5. On admet que  $\hat{\theta}_n$  converge vers  $\theta_*$ ; démontrer la normalité asymptotique de l'estimateur.

## V.2 Processus autorégressif instable

Nous donnons ici un exemple où la distribution asymptotique n'est pas gaussienne. Considérons le modèle autorégressif scalaire d'ordre 1

$$Y_i = \theta_* Y_{i-1} + \varepsilon_i$$

où  $Y_0$  est donné et les  $\varepsilon_i$  sont iid  $\mathcal{N}(0, \sigma^2)$ . On s'intéresse ici à la situation  $|\theta_*| > 1$ , cas étudié par White en 1958 [142]. On obtient l'estimateur au maximum de vraisemblance :

$$\hat{\theta}_n = \frac{\sum_{i=1}^n Y_i Y_{i-1}}{\sum_{i=1}^n Y_{i-1}^2} = \theta_* + \frac{\sum_{i=1}^n \varepsilon_i Y_{i-1}}{\sum_{i=1}^n Y_{i-1}^2}.$$

Notons que

$$Y_i = \theta_*^i Z + r_i$$

avec

$$Z = Y_0 + \theta_*^{-1} \varepsilon_1 + \theta_*^{-2} \varepsilon_2 + \dots = Y_0 + T$$

$$r_i = -\theta_*^{-1} \varepsilon_{i+1} - \theta_*^{-2} \varepsilon_{i+2} + \dots$$

et la suite  $r_i$  est stationnaire gaussienne. On peut vérifier alors simplement que

$$\theta_*^n(\hat{\theta}_n - \theta_*) = \frac{Z \sum_{i=1}^n \theta_*^{-n+i-1} \varepsilon_i + u_n}{Z^2 \sum_{i=1}^n \theta_*^{-2n+2i-2} + v_n}$$

où  $u_n$  et  $v_n$  convergent dans  $L_1$  vers 0. Comme  $Z = Y_0 + T$ , ceci se réécrit

$$\theta_*^n(\hat{\theta}_n - \theta_*) = \left( \sum_{i=1}^n \theta_*^{-2n+2i-2} + v_n \right)^{-1} \left( \frac{T'_n}{Y_0 + T} + Z^{-1} u_n \right), \quad T'_n = \sum_{i=1}^n \theta_*^{-n+i-1} \varepsilon_i.$$

Le premier facteur converge en probabilité vers  $\theta_*^2 - 1$  et  $Z^{-1} u_n$  converge vers 0. Le vecteur gaussien  $(T, T'_n)$  convergeant en loi vers  $\mathcal{N}(0, \sigma^2(\theta_*^2 - 1)^{-1} Id)$ , on a finalement

$$\theta_*^n(\hat{\theta}_n - \theta_*) \longrightarrow (\theta_*^2 - 1) \frac{X'}{X + \sigma^{-1} \sqrt{\theta_*^2 - 1} Y_0}, \quad (X, X') \sim \mathcal{N}(0, Id).$$

Si  $Y_0 = 0$ , le rapport suit une distribution de Cauchy.

**Exercice.** Quelle est la limite et la distribution asymptotique de l'estimateur :

$$\hat{\theta}_n = \frac{\sum_{i=1}^n Y_i Y_{i-1}}{\sum_{i=0}^n Y_i^2} ?$$

### V.3 Régression linéaire

Soit le modèle de régression linéaire classique

$$Y_i = X_i \beta^* + u_i$$

où  $X_i$  est un vecteur ligne, pour l'instant déterministe, et les  $u_i$  sont iid scalaires de variance  $\sigma_*^2$ . La suite  $Y_i$  n'est pas iid.

#### 47 - DÉFINITION

L'estimateur de  $\beta^*$  aux moindres carrés ordinaires (OLS, ordinary least squares) est

$$\hat{\beta} = \arg \min_{\beta} \sum_i (Y_i - X_i \beta)^2.$$

C'est l'estimateur de  $\beta^*$  au maximum de vraisemblance sous l'hypothèse de normalité de  $u$ . On a le résultat élémentaire

#### 48 - PROPOSITION

$\hat{\beta}$  est un estimateur sans biais avec, en notant  $R = \sum_i X_i^T X_i$  :

$$\begin{aligned} \hat{\beta} &= R^{-1} \sum X_i^T Y_i = \beta^* + R^{-1} \sum X_i^T u_i \\ \text{Var}(\hat{\beta}) &= \sigma_*^2 R^{-1}. \end{aligned}$$

On s'intéresse maintenant à l'asymptotique quand le nombre d'observations  $n$  tend vers l'infini. Notons

$$\hat{\beta}_n = R_n^{-1} \sum_{i=1}^n X_i^T Y_i.$$

Il est intéressant de considérer le cas où les  $X_i$  sont aléatoires et les  $u_i$  sont seulement des accroissements de martingale, par exemple pour traiter de l'estimation des systèmes dynamiques (le cas typique étant le

processus autorégressif d'ordre  $p$  avec  $X_i = (Y_{i-1} \dots Y_{i-p})$ , cf. [101]). Noter que dans la suite,  $X_i$  devra être bien entendu  $\mathcal{F}_{i-1}$ -mesurable. On a le théorème suivant :

49 - THÉORÈME

Soit une suite de tribus croissante  $\mathcal{F}_n$  telle que  $X_i$  soit  $\mathcal{F}_{i-1}$ -mesurable et telle que les  $u_i$  forment une suite d'accroissements de martingale avec :

$$\begin{aligned} E[u_i | \mathcal{F}_{i-1}] &= 0, \\ \sup_i E[u_i^2 | \mathcal{F}_{i-1}] &< \infty, \quad p.s. \end{aligned}$$

Si presque sûrement  $\sup_n \|R_n\| \|R_n^{-1}\| < \infty$  et  $R_n^{-1} \rightarrow 0$ , alors  $\hat{\beta}_n \rightarrow \beta^*$  presque sûrement. S'il existe une matrice  $Q$  déterministe telle que p.s.  $\frac{1}{n} R_n \rightarrow Q > 0$ ,  $E[u_i^2 | \mathcal{F}_{i-1}] \rightarrow \sigma_*^2$  p.s., et si

$$\sup_i E[|u_i|^\eta | \mathcal{F}_{i-1}] < \infty, \quad p.s.$$

pour un  $\eta > 2$  alors

$$\sqrt{n}(\hat{\beta}_n - \beta) \longrightarrow \mathcal{N}(0, \sigma_*^2 Q^{-1}) \quad \text{en loi.}$$

*Démonstration.* Commençons par la convergence presque sûre. On a besoin du lemme suivant dont la démonstration est faite à l'appendice B :

50 - LEMME (de Kronecker)

Soit  $R_n$  une suite croissante de matrices symétriques ( $R_n - R_{n-1} \geq 0$ ) telle que  $R_n^{-1}$  tend vers 0, et  $\sum M_n$  une série convergente de matrices (pas nécessairement absolument convergente), alors la suite

$$\|R_n\|^{-1} \sum_{i=1}^n R_i M_i$$

tend vers zéro.

On a

$$\hat{\beta}_n - \beta^* = R_n^{-1} \sum_{i=1}^n X_i^T u_i = (R_n^{-1} \|R_n\|) \left( \|R_n\|^{-1} \sum_{i=1}^n R_i (R_i^{-1} X_i^T u_i) \right).$$

Comme  $\|R_n^{-1}\| \cdot \|R_n\|$  est borné, il suffit en vertu du lemme précédent de montrer que la martingale

$$\sum_{i=l}^n R_i^{-1} X_i^T u_i$$

converge presque sûrement ( $l$  est le premier indice tel que  $R_{l-1}^{-1}$  est définie). Il suffit d'avoir, en raison du théorème 67

$$\sum_{i \geq l} \|R_i^{-1} X_i^T\|^2 E[u_i^2 | \mathcal{F}_{i-1}] < \infty$$

ce qui est réalisé si

$$\text{Tr} \sum_{i \geq l} R_i^{-1} X_i^T X_i R_i^{-1} < \infty. \tag{V.6}$$

Mais

$$\sum_{i=l}^n R_i^{-1} X_i^T X_i R_i^{-1} = \sum_{i=l}^n R_i^{-1} (R_i - R_{i-1}) R_i^{-1}.$$

Notons que pour deux matrices symétriques  $A$  et  $B$ ,  $B$  étant définie positive, on a

$$A(A^{-1} - B^{-1})A \leq B - A.$$

Ceci vient simplement de ce que la différence des deux membres vaut

$$B - A - A + AB^{-1}A = (B - A)B^{-1}(B - A) \geq 0.$$

On a donc

$$\sum_{i=l}^n R_i^{-1} X_i^T X_i R_i^{-1} \leq \sum_{i=l}^n R_{i-1}^{-1} - R_i^{-1} = R_{l-1}^{-1} - R_n^{-1} \leq R_{l-1}^{-1}$$

ce qui implique bien (V.6). La convergence presque sûre est donc démontrée. Pour le théorème-limite central, il suffit de montrer que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i u_i \longrightarrow \mathcal{N}(0, \sigma_*^2 Q^{-1}) \quad \text{en loi.}$$

Il s'agit donc de vérifier les hypothèses du théorème 68 avec  $X'_{ni} = X_i u_i / \sqrt{n}$ . Les conditions (D.1) et (D.2) sont satisfaites si

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n E[\|X_i u_i\|^2 1_{\|X_i u_i\| > \varepsilon \sqrt{n}} | \mathcal{F}_{i-1}] &\longrightarrow 0, \quad \text{pour tout } \varepsilon > 0 \\ \frac{1}{n} \sum_{i=1}^n X_i^T X_i E[u_i^2 | \mathcal{F}_{i-1}] &\xrightarrow{P} \sigma_*^2 Q. \end{aligned}$$

La seconde relation est élémentaire. Pour la première,

$$\begin{aligned} \frac{1}{n} \sum_{i \leq n} E[\|X_i u_i\|^2 1_{\|X_i u_i\| > \varepsilon \sqrt{n}} | \mathcal{F}_{i-1}] &\leq \frac{1}{n} \sum_{i \leq n} E[\|X_i u_i\|^\eta | \mathcal{F}_{i-1}] (\varepsilon \sqrt{n})^{2-\eta} \\ &\leq \left( \frac{1}{n} \sum_{i \leq n} \|X_i\|^2 \right) \sup_{i \leq n} \|X_i\| / \varepsilon \sqrt{n}^{\eta-2} \sup_{i \leq n} E[|u_i|^\eta | \mathcal{F}_{i-1}] \\ &\leq C(\omega) \sup_{i \leq n} (\|X_i\| / \varepsilon \sqrt{n})^{\eta-2}. \end{aligned}$$

Il ne suffit plus que de montrer que  $\sup_{i \leq n} \|X_i\| / \sqrt{n}$  tend vers zéro. Mais la convergence de  $n^{-1} R_n$  implique que  $\|X_i\| / \sqrt{i}$  tend vers zéro car

$$\frac{\|X_i\|^2}{i} = \text{Tr} \left( \frac{R_i}{i} - \frac{R_{i-1}}{i} \right) \rightarrow 0$$

et donc

$$\sup_{i \leq n} \|X_i\| / \sqrt{n} \leq \sup_{i \leq n_0} \|X_i\| / \sqrt{n} + \sup_{n_0 \leq i \leq n} \|X_i\| / \sqrt{i}.$$

Le deuxième terme peut être rendu arbitrairement petit par un choix adéquat de  $n_0$  et le premier tend vers zéro. ■

# A

## COMPLÉMENTS SUR LES EXPÉRIENCES RÉGULIÈRES

On va montrer que la R-régularité (cf. p. 65) implique l'hypothèse plus générale de différentiabilité en moyenne quadratique uniforme (DMQU). Les statisticiens du XX<sup>e</sup> siècle ont découvert que cette dernière hypothèse est mathématiquement plus adaptée à l'étude de l'efficacité (théorème de convolution de Hajek p. 76). Sous l'hypothèse légèrement plus faible de différentiabilité en moyenne quadratique (DMQ), on démontrera le lemme 24. L'hypothèse DMQU est compliquée et ne se vérifie en pratique qu'au travers de la R-régularité, c'est pourquoi nous ne l'avons pas introduite auparavant.

L'hypothèse DMQ est la différentiabilité de l'application  $\theta \mapsto \sqrt{p_\theta}$  considérée comme application à valeurs dans  $L_2(\mu)$ , DMQU ajoutant une condition d'uniformité. Cette application a donc une dérivée notée dans la suite  $R_\theta(y)$ , qui ne peut que coïncider avec  $\nabla p_\theta / \sqrt{p_\theta}$  lorsque cette quantité existe. Une troisième hypothèse, également classiquement considérée, est la régularité, qui demande en plus de DMQ la continuité de  $\theta \mapsto R_\theta$  dans  $L_2(\mu)$ . Les hypothèses discutées jusqu'à présent concernent les suites d'observations indépendantes, chacune suivant la loi de densité  $p_\theta$ . Un autre jeu d'hypothèses est adapté aux suites d'expériences statistiques plus générales, c'est l'hypothèse LAN (déjà rencontrée p. 100, et théorème 44), et sa version uniforme ULAN (cf. p. 116). On va montrer les implications suivantes

$$\begin{array}{ccccccc}
 \text{R-régularité} & \implies & \text{Régularité} & \implies & \text{DMQU} & \implies & \text{DMQ} & \text{(Cas i.i.d)} \\
 & & & & \downarrow & & \downarrow & \\
 & & & & \text{ULAN} & \implies & \text{LAN} & \text{(Cas général)}
 \end{array}$$

Dans le paragraphe A.2 nous démontrerons le théorème de convolution de Hajek sous l'hypothèse générale ULAN (théorème 61) ce qui permet d'aborder le cas d'une suite générale d'expériences statistiques (cf. I.3.1), en particulier le cas dépendant, cf. la remarque suivant le théorème 42 p. 100. On démontrera également de théorème d'existence d'un estimateur efficace de Le Cam sous l'hypothèse LAN (théorème 63).

### A.1 Liens entre les jeux d'hypothèses

Le référence pour cette section est [1] (chapitre 2, ou appendice A.5) ou encore [21]. On démontrera au passage le lemme 24 (à nouveau) et la formule (III.18). Il faut commencer par le lemme suivant :

51 - LEMME

Soit un espace  $\Omega$  muni d'une mesure  $\mu(dx)$ , et des fonctions  $u_n, u, v_n, v, w_n, w$  de  $L_2(\mu)$ .

- (i) Si  $u_n$  converge vers  $u$   $\mu$ -presque partout,  $\sup_n \|u_n\|_2^2 < \infty$  et si  $\|v_n - v\|_2 \rightarrow 0$ , alors  $\mu(u_n v_n)$  converge vers  $\mu(uv)$ .
- (ii) Si  $w_n 1_{|w|>0}$  converge  $\mu$ -presque partout vers  $w$  et si  $\overline{\lim}_n \|w_n\|_2^2 \leq \|w\|_2^2 < \infty$ , alors  $\|w_n - w\|_2 \rightarrow 0$ .

*Démonstration.* Commençons par le premier point. On pose  $\tilde{u}_n = u_n - u$  et  $\tilde{v}_n = v_n - v$ ; noter qu'en vertu du lemme de Fatou  $\|u\|_2 < \infty$ ; on a pour tout  $A > 0$

$$\begin{aligned} |\mu(u_n v_n) - \mu(uv)| &= |\mu(u_n \tilde{v}_n) + \mu(\tilde{u}_n v)| \\ &\leq \|u_n\|_2 \|\tilde{v}_n\|_2 + |\mu(\tilde{u}_n v 1_{|\tilde{u}_n| \leq A|v|})| + |\mu(\tilde{u}_n v 1_{|\tilde{u}_n| > A|v|})| \\ &\leq \|u_n\|_2 \|\tilde{v}_n\|_2 + |\mu(\tilde{u}_n v 1_{|\tilde{u}_n| \leq A|v|})| + A^{-1} \mu(\tilde{u}_n^2). \end{aligned}$$

Le théorème de convergence dominée s'applique au deuxième terme car  $|\tilde{u}_n v| 1_{|\tilde{u}_n| \leq A|v|} \leq Av^2$ , et par conséquent

$$\overline{\lim}_n |\mu(u_n v_n) - \mu(uv)| \leq A^{-1} \sup_n \mu(\tilde{u}_n^2)$$

qui peut être rendu arbitrairement petit prenant  $A$  grand.

Concernant le deuxième point :

$$\|w_n - w\|_2^2 = 2\mu(w(w - w_n 1_{|w|>0})) + \mu(w_n^2) - \mu(w^2)$$

En utilisant le premier point avec  $v_n = w$  et  $u_n = w - w_n 1_{|w|>0}$ , on obtient la convergence du premier terme vers 0, et le second a une limite supérieure  $\leq 0$ . ■

#### 52 - DÉFINITION (DIFFÉRENTIABILITÉ EN MOYENNE QUADRATIQUE)

L'expérience est dite différentiable en moyenne quadratique uniformément sur  $K$  compact intérieur à  $\Theta$  (DMQU), s'il existe une fonction  $R_\theta$  telle que pour tout  $\theta \in K$  :

$$\sqrt{p_{\theta+h}}(y) = \sqrt{p_\theta}(y) + \frac{1}{2} h^T R_\theta(y) + \varphi_{\theta,h}(y) \quad (\text{A.1})$$

avec

$$\sup_{\theta \in K} \int \varphi_{\theta,h}(y)^2 \mu(dy) = o(|h|^2). \quad (\text{A.2})$$

Si  $K$  est un singleton,  $K = \{\theta\}$ , on dit simplement que l'expérience est différentiable en moyenne quadratique (DMQ) en  $\theta$ . L'expérience est différentiable en moyenne quadratique (DMQ) sur  $\Theta$  si elle l'est en tout point de  $\Theta$ .

Si  $\sqrt{p_\theta}$  est différentiable,  $R_\theta$  est simplement  $\nabla_\theta p_\theta / \sqrt{p_\theta}$ . Dans la suite on posera

$$\nabla_\theta p_\theta = R_\theta \sqrt{p_\theta} \quad (\text{A.3})$$

$$S_\theta = \frac{R_\theta}{\sqrt{p_\theta}} 1_{p_\theta > 0} = \frac{\nabla_\theta p_\theta}{p_\theta} 1_{p_\theta > 0} \quad (\text{le score}) \quad (\text{A.4})$$

$$I(\theta) = \int R_\theta(y) R_\theta(y)^T \mu(dy) = E_\theta[S_\theta(Y) S_\theta(Y)^T]. \quad (\text{A.5})$$

#### 53 - LEMME

Si l'expérience est DMQ en tout point du segment  $[\theta, \theta + h]$  de  $\mathbb{R}^d$  et si  $I(\cdot)$  y est borné, alors, pour presque tout  $y$ ,

$$\sqrt{p_{\theta+h}}(y) = \sqrt{p_\theta}(y) + \frac{1}{2} \int_0^1 h^T R_{\theta+th}(y) dt \quad (\text{A.6})$$

(cet ensemble de  $y$  de  $\mu$ -mesure pleine dépend de  $h$ ).

*Démonstration.* L'équation (A.1) implique que pour toute fonction  $\psi \in L_2(\mu)$  l'intégrale  $\int \psi(y) \sqrt{p_{\theta+th}(y)} \mu(dy)$  est une fonction dérivable de  $t$  avec pour dérivée  $\frac{1}{2} \int \psi(y) h^T R_{\theta+th}(y) \mu(dy)$  par conséquent

$$\int \psi(y) (\sqrt{p_{\theta+h}(y)} - \sqrt{p_{\theta}(y)}) \mu(dy) = \frac{1}{2} \int \int_0^1 \psi(y) h^T R_{\theta+th}(y) \mu(dy)$$

où l'on a utilisé le théorème de Fubini pour permuter les intégrales. En faisant la différence des deux membres on obtient que la différence des deux membres de (A.6) est orthogonale à  $\psi$ , donc à toute fonction de  $L_2(\mu)$ . ■

Nous passons maintenant à un résultat qui sera utilisé pour démontrer le théorème 25.

54 - THÉORÈME

Si l'expérience est DMQ en  $\theta_*$  avec  $I(\theta_*) < \infty$ , alors, pour tout  $\theta_*$  intérieur à  $\Theta$ , la fonction  $\theta \mapsto l(\theta) = E_{\theta_*}[\ln p_{\theta}(Y)]$  admet le développement de Taylor en  $\theta_*$  à l'ordre deux

$$l(\theta_* + h) - l(\theta_*) = -\frac{1}{2} h^T I(\theta_*) h + o(|h|^2).$$

*Démonstration.* On a pour  $x \geq 0$

$$\ln(1+x) - x + 2(\sqrt{1+x} - 1)^2 = (\sqrt{1+x} - 1)^2 \varepsilon(x)$$

où la fonction  $\varepsilon$  ainsi définie est bornée et tend vers 0 quand  $x \rightarrow 0$ . Avec  $x = p_{\theta_*+h}(y)/p_{\theta_*}(y) - 1$  et en prenant l'espérance sous  $p_{\theta_*}$ , il vient

$$\begin{aligned} E_{\theta_*}[\ln(p_{\theta_*+h}(y)) - \ln(p_{\theta_*}(y))] + 2 \int (\sqrt{p_{\theta_*+h}(y)} - \sqrt{p_{\theta_*}(y)})^2 \mu(dy) \\ = \int (\sqrt{p_{\theta_*+h}(y)} - \sqrt{p_{\theta_*}(y)})^2 \varepsilon(\sqrt{p_{\theta_*+h}(y)} - \sqrt{p_{\theta_*}(y)}) \mu(dy). \end{aligned}$$

Posons  $\delta_h(y) = \sqrt{p_{\theta_*+h}(y)} - \sqrt{p_{\theta_*}(y)}$ . Par application de l'équation (A.1)  $\int \delta_h(y)^2 \mu(dy) = \frac{1}{4} h^T I(\theta_*) h + o(|h|^2)$ , il suffit donc de montrer que

$$\int \delta_h(y)^2 \varepsilon(\delta_h(y)) \mu(dy) = o(|h|^2).$$

L'équation (A.1) implique

$$\begin{aligned} \int \delta_h(y)^2 \varepsilon(\delta_h(y)) \mu(dy) &\leq \int \frac{\langle h, \nabla p_{\theta_*}(y) \rangle^2}{p_{\theta_*}(y)} \varepsilon(\delta_h(y)) \mu(dy) + 2\|\varepsilon\|_{\infty} \int \varphi_{\theta_*,h}(y)^2 \mu(dy) \\ &\leq |h|^2 \int \frac{|\nabla p_{\theta_*}(y)|^2}{p_{\theta_*}(y)} \varepsilon(\delta_h(y)) \mu(dy) + o(|h|^2) \|\varepsilon\|_{\infty}. \end{aligned} \quad (\text{A.7})$$

Notons que pour tout  $\eta > 0$

$$\begin{aligned} \overline{\lim}_{h \rightarrow 0} \int \frac{|\nabla p_{\theta_*}(y)|^2}{p_{\theta_*}(y)} \varepsilon(\delta_h(y)) \mu(dy) &\leq \|\varepsilon\|_{\infty} \overline{\lim}_{h \rightarrow 0} \int \frac{|\nabla p_{\theta_*}(y)|^2}{p_{\theta_*}(y)} 1_{|\delta_h(y)| \geq \eta} \mu(dy) + I(\theta_*) \sup_{t \leq \eta} \varepsilon(t) \\ &\leq I(\theta_*) \sup_{t \leq \eta} \varepsilon(t) \end{aligned}$$

par application du théorème de Lebesgue (les v.a.  $\frac{|\nabla p_{\theta_*}(y)|^2}{p_{\theta_*}(y)} 1_{|\delta_h(y)| \geq \eta}$  convergent en  $p_{\theta_*}$ -probabilité vers 0 quand  $h \rightarrow 0$ ). La  $\overline{\lim}$  est donc nulle ce qui fait que le membre de droite de (A.7) est bien  $o(|h|^2)$ . ■

On suppose l'expérience DMQU sur  $K$ . Alors

$$\sup_{\theta \in K} I(\theta) < \infty.$$

De plus on a l'estimée des distances de Hellinger

$$d_H(P_\theta, P_{\theta+h}) = \int \left( \sqrt{p_{\theta+h}(y)} - \sqrt{p_\theta(y)} \right)^2 \mu(dy) = \frac{1}{4} h^T I(\theta) h + o(|h|^2) \quad (\text{A.8})$$

le  $o(|h|^2)$  étant uniforme en  $\theta \in K$ .

Pour tout  $\theta \in K$  et toute fonction  $\psi$  telle que  $\sup_{\theta' \in [\theta, \theta+h]} E_{\theta'}[\psi(Y)^2] < \infty$  la fonction  $t \mapsto E_{\theta+th}[\psi(Y)]$  est dérivable en 0 et

$$\frac{d}{dt} E_{\theta+th}[\psi(Y)]|_{t=0} = E_\theta[\psi(Y) h^T S_\theta(Y)]. \quad (\text{A.9})$$

*Démonstration.* L'équation (A.1) se réécrit

$$\frac{1}{2} h^T R_\theta = \sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \varphi_{\theta,h} \quad (\text{A.10})$$

Chaque terme du membre de droite appartient à  $L_2(\mu)$ , donc le membre de gauche également. Ceci implique que  $I(\theta) < \infty$ , et l'uniformité est immédiate.

L'équation (A.8) est alors une conséquence immédiate de (A.1).

Passons à (A.9). En remplaçant dans (A.10)  $h$  par  $\varepsilon h$  et en multipliant par  $2\sqrt{p_\theta}$  il vient

$$\begin{aligned} \varepsilon h^T S_\theta p_\theta &= 2 \left( \sqrt{p_{\theta+\varepsilon h}} \sqrt{p_\theta} - p_\theta - \varphi_{\theta,\varepsilon h} \sqrt{p_\theta} \right) \\ &= -(\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta})^2 - p_\theta + p_{\theta+\varepsilon h} - 2\varphi_{\theta,\varepsilon h} \sqrt{p_\theta} \end{aligned}$$

et en multipliant par  $\psi(y)$  puis en intégrant sous  $\mu$  il vient

$$E_{\theta+\varepsilon h}[\psi(Y)] - E_\theta[\psi(Y)] - \varepsilon E_\theta[\psi(Y) h^T S_\theta(Y)] = \int \left( (\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta})^2 + 2\varphi_{\theta,\varepsilon h} \sqrt{p_\theta} \right) \psi d\mu.$$

Il suffit donc de montrer que

$$\int (\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta})^2 \psi d\mu = o(\varepsilon) \quad (\text{A.11})$$

car la majoration du dernier terme est immédiate. Mais

$$\varepsilon^{-1} (\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta})^2 \psi \leq |R_\theta| |\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta}| \psi + \varepsilon^{-1} |\varphi_{\theta,\varepsilon h}| |\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta}| \psi = u_\varepsilon + v_\varepsilon$$

Comme pour toute suite  $\varepsilon_n$  tendant vers 0 on a en vertu de (A.6)

$$\sqrt{p_{\theta+\varepsilon_n h}}(y) - \sqrt{p_\theta}(y) = \frac{1}{2} \int_0^{\varepsilon_n} h^T R_{\theta+th}(y) dt$$

(l'ensemble de  $\mu$ -mesure pleine peut être choisi commun à tous les  $\varepsilon_n$ ),  $u_{\varepsilon_n}$  tend presque sûrement vers 0; de plus, comme  $u_{\varepsilon_n}^2 \leq 2(p_{\theta+\varepsilon_n h} + p_\theta) \psi^2 + 2|R_\theta|^2$ , cette suite est bornée dans  $L_2(\mu)$ ; le premier point du lemme 51, avec  $v_n = |R_\theta|$  et  $u_n = |\sqrt{p_{\theta+\varepsilon_n h}} - \sqrt{p_\theta}| \psi$ , permet donc de conclure à la convergence vers 0 de  $\mu(u_{\varepsilon_n})$ ;  $\mu(v_{\varepsilon_n})$  tend également vers 0 en raison de l'inégalité de Cauchy-Schwarz et de (A.2). ■

## 56 - DÉFINITION (RÉGULARITÉ)

L'expérience est dite régulière si elle est DMQ en tout  $\theta \in \Theta$  et si l'application  $\theta \mapsto R_\theta$  est continue de  $\Theta$  dans  $L_2(\mu)$ .

57 - THÉORÈME

Une expérience régulière est DMQU sur tout compact intérieur à  $\Theta$ .

*Démonstration.* La fonction  $\varphi_{\theta,h}$  définie par (A.1) satisfait presque sûrement en raison de (A.6)

$$\varphi_{\theta,h}(y) = \frac{1}{2} \int_0^1 \langle h, R_{\theta+th}(y) - R_{\theta}(y) \rangle dt$$

d'où en utilisant l'inégalité de Cauchy-Schwarz et le théorème de Fubini-Tonelli

$$\begin{aligned} \int \varphi_{\theta,h}(y)^2 \mu(dy) &\leq \frac{|h|^2}{4} \int_0^1 \int |R_{\theta+th}(y) - R_{\theta}(y)|^2 \mu(dy) dt \\ &\leq \frac{|h|^2}{4} \sup_{|\theta-\theta'| \leq |h|} \int |R_{\theta'}(y) - R_{\theta}(y)|^2 \mu(dy). \end{aligned}$$

D'où le résultat. ■

58 - LEMME

Soit un modèle R-régulier au sens de la définition 23 p. 65, alors pour toute fonction  $\psi$  de régularité  $C^1$ , et  $\mu$ -presque tout  $y$ , pour tous  $\theta$  et  $h$  tels que le segment  $[\theta, \theta_h]$  soit contenu dans  $\Theta$ ,

$$\psi(p_{\theta+h}(y)) - \psi(p_{\theta}(y)) = \int_0^1 \psi'(p_{\theta+th}(y)) \langle h, \nabla_{\theta} p_{\theta+th}(y) \rangle dt,$$

*Démonstration.* Posons  $F(t) = p_{\theta+th}(y)$  et  $f(t) = \langle h, \nabla_{\theta} p_{\theta+th}(y) \rangle$ . La R-régularité implique  $F(t) - F(s) = \int_s^t f(u) du$ . Pour toute fonction  $\psi$  de régularité  $C^2$ , on a pour tout  $n > 0$

$$\begin{aligned} \psi(F(1)) - \psi(F(0)) &= \sum_{k=1}^n \psi(F(k/n)) - \psi(F((k-1)/n)) \\ &= \sum_{k=1}^n \psi'(c_k) \int_{(k-1)/n}^{k/n} f(t) dt, \quad F(k/n) < c_k < F((k-1)/n) \\ &= \sum_{k=1}^n \int_{(k-1)/n}^{k/n} \psi'(F(t)) f(t) dt + O(\|\psi''\|_{\infty}) \int_{(k-1)/n}^{k/n} |F(t) - c_k| |f(t)| dt \\ &= \int_0^1 \psi'(F(t)) f(t) dt + O(\|\psi''\|_{\infty}) \sup_{|t-s| < \frac{1}{n}} |F(t) - F(s)| \int_0^1 |f(t)| dt \end{aligned}$$

La continuité de  $F$  fait que le dernier terme peut être rendu arbitrairement petit, d'où la formule

$$\psi(F(1)) - \psi(F(0)) = \int_0^1 \psi'(F(t)) f(t) dt,$$

qui s'étend à  $\psi$  de régularité  $C^1$  par approximation et passage à la limite. ■

59 - THÉORÈME (RÉGULARITÉ DE L'EXPÉRIENCE R-RÉGULIÈRE)

Soit un modèle R-régulier au sens de la définition 23 p. 65, où  $I(\theta)$  est donné par (A.5) avec

$$R_\theta = \frac{\nabla p_\theta}{\sqrt{p_\theta}} 1_{p_\theta > 0}.$$

Alors l'expérience est régulière.

*Démonstration.* Notons d'abord que le deuxième point du lemme 51 implique la convergence de  $R_{\theta'}$  vers  $R_\theta$  dans  $L_2(\mu)$  lorsque  $\theta'$  tend vers  $\theta$ .

On définit  $\varphi_{\theta,h}$  par (A.1) et il ne reste qu'à démontrer (A.2). Posons  $F(t) = p_{\theta+th}(y)$  et  $f(t) = \langle h, \nabla_\theta p_{\theta+th}(y) \rangle$ . En vertu du lemme 58 appliqué à  $\psi(x) = \sqrt{x} \vee \varepsilon$  (cette fonction n'étant pas  $C^1$ , il faut en fait raisonner par approximation, c'est immédiat), on a

$$\sqrt{p_{\theta+h}(y) \vee \varepsilon} = \sqrt{p_\theta(y) \vee \varepsilon} + \int_0^1 \frac{h^T \nabla p_{\theta+th}(y)}{\sqrt{p_{\theta+th}(y) \vee \varepsilon}} 1_{p_{\theta+th}(y) > \varepsilon} dt.$$

Le fait que  $I(\cdot)$  soit borné implique que  $\mu$ -p.s.  $\int_0^1 |R_{\theta+th}(y)| dt$  est fini et donc, sur cet ensemble de  $y$ , on a par application du théorème de Lebesgue en faisant tendre  $\varepsilon$  vers 0

$$\sqrt{p_{\theta+h}(y)} = \sqrt{p_\theta(y)} + \int_0^1 R_{\theta+th}(y) dt.$$

ce qui permet de calculer  $\varphi_{\theta,h}$  comme dans la démonstration du théorème 57. ■

Pour l'énoncé du théorème suivant, nous aurons besoin de la définition de l'hypothèse LAN uniforme :

**(ULAN)** LAN (p. 100) est satisfaite et renforcée par : pour tout  $M < \infty$ , tout  $\alpha > 0$  et tout compact  $K$  intérieur à  $\Theta$

$$\lim_{n \rightarrow \infty} \sup_{\theta \in K} \sup_{|h| \leq M} P_\theta(|\varepsilon_n(\theta, h)| > \alpha) = 0. \quad (\text{A.12})$$

60 - THÉORÈME

Sous l'hypothèse de DMQ sur  $\Theta$  et d'indépendance des observations, l'hypothèse LAN (p. 100) est vérifiée. Si la différentiabilité est uniforme sur  $K$  (et donc sous l'hypothèse de régularité de l'expérience) ULAN est satisfait sur  $K$ .

*Démonstration.* On a pour une certaine constante  $c$  et tous  $a, b > 0$ , en vertu de la formule de Taylor pour la fonction log au voisinage de 1 :

$$1_{1/2 \leq b/a \leq 2} \left| \log(b) - \log(a) - 2(\sqrt{b} - \sqrt{a}) \frac{1}{\sqrt{a}} \right| \leq \frac{c}{a} (\sqrt{b} - \sqrt{a})^2$$

(poser  $b = x^2 a$ ). En utilisant cette majoration avec

$$\begin{aligned} a_i &= p_\theta(Y_i) \\ b_i &= p_{\theta+h/\sqrt{n}}(Y_i) \end{aligned}$$

on voit que sur l'ensemble

$$A_n = \{\omega : \forall i, 1/2 \leq b_i/a_i \leq 2\}$$

on peut utiliser (A.1) pour majorer la variable  $\varepsilon_n(\theta, h)$  de la formule (IV.5) par

$$|\varepsilon_n(\theta, h)| \leq \left| \frac{1}{2} h^T I(\theta) h - 2 \sum_{i=1}^n \frac{\varphi_{\theta, h_n}(Y_i)}{\sqrt{a_i}} \right| + \sum_{i=1}^n \frac{c}{a_i} (\sqrt{b_i} - \sqrt{a_i})^2 = |A| + B.$$

On a, en utilisant le théorème 55 puis (A.1, A.2)

$$E_\theta[A] = n o(h^2/n) = o(1)$$

$$E_\theta[A^2] = n o(h^2/n) = o(1)$$

$$E_\theta[B] = n o(h^2/n) = o(1).$$

Ces  $o(1)$  tendent vers 0 quand  $n$  tend vers l'infini uniformément en  $h$ . Donc

$$E[1_{A_n} |\varepsilon_n(\theta, h)|] = o(1).$$

Et par ailleurs

$$P(\Omega \setminus A_n) \leq \sum_{i=1}^n P(b_i/a_i \geq 2 \text{ ou } b_i/a_i \leq 1/2) \leq 12 \sum_{i=1}^n E[(\sqrt{b_i}/\sqrt{a_i} - 1)^2] = \frac{12}{c} E[B] = o(1).$$

La convergence étant bien uniforme en  $h$  et  $\theta \in K$ , le théorème est démontré. ■

## A.2 Démonstration des théorèmes 29, 42. Régularité des scores

L'objet de cette section est de démontrer le théorème 61 (théorème convolution de Hajek, c'est le théorème 29 avec un jeu d'hypothèses légèrement plus général) et de le théorème 63 (théorème d'existence d'un estimateur efficace de Le Cam. Ces deux résultats ont été présentés page 76).

Passons à la démonstration du théorème 29.

<b>61 - THÉORÈME</b>	(THÉORÈME DE CONVOLUTION DE HAJEK)
----------------------	------------------------------------

Sous l'hypothèse ULAN, le théorème 29 est vrai (le score $S_n$ étant donné par (IV.5)).
---

*Démonstration.* On va appliquer le lemme 43 avec

$$T_n = \sqrt{n}(\hat{\theta}_n - \theta)$$

$$P_n = P_{\theta+h/\sqrt{n}}$$

$$Q_n = P_\theta$$

$$\Lambda_n = \frac{p_\theta(Y)}{p_{\theta+h/\sqrt{n}}(Y)}$$

(on pose  $\Lambda_n = 0$  si le dénominateur est nul). Commençons par montrer la convergence en loi de  $\Lambda_n$  sous  $P_n$ . L'équation IV.5 appliquée à  $\theta - h/\sqrt{n}$  donne

$$\log \left( \frac{P_\theta^n(Y)}{P_{\theta-h/\sqrt{n}}^n(Y)} \right) = h^T S_n(\theta - h/\sqrt{n}) - \frac{1}{2} h^T I(\theta - h/\sqrt{n}) h + \varepsilon_n(\theta - h/\sqrt{n}, h)$$

et en vertu de (A.12) le dernier terme tend en probabilité vers 0 sous  $P_{\theta-h/\sqrt{n}}$ . On a donc, après changement de  $h$  en  $-h$ , en vertu de la continuité de  $I$

$$\Lambda_n \xrightarrow{Loi} \Lambda \sim \mathcal{N}\left(-\frac{1}{2} h^T I(\theta) h, h^T I(\theta) h\right), \text{ sous } P_{\theta+h/\sqrt{n}}.$$

On a par hypothèse

$$T_n \xrightarrow{Loi} X + h, \text{ sous } P_{\theta+h/\sqrt{n}}.$$

De toute suite d'entiers, on peut extraire une sous-suite telle que la paire  $(T_{n_k}, \Lambda_{n_k})$  converge en loi vers une certaine limite. On peut alors appliquer le lemme 43 à une telle sous-suite. Pour identifier la limite notons que l'équation (IV.6) avec  $f(x) = e^{ia^T x}$  devient

$$E[e^{ia^T(X+h)}] = E[e^{ia^T X + h^T Z - h^T I(\theta)h/2}]$$

où  $Z \sim \mathcal{N}(0, I(\theta))$ ,  $X$  suit sa loi, et la loi liée est encore inconnue. En posant  $a = I(\theta)b$ , cette identité se réécrit, en posant  $Y = X - Z$  (conformément aux notations du théorème 29)

$$e^{ib^T I(\theta)h} E[e^{ib^T(Y+Z)}] = E[e^{ib^T(Y+Z) + h^T Z - h^T I(\theta)h/2}]$$

soit

$$\begin{aligned} E[e^{ib^T Y + (ib+h)^T Z}] &= e^{ib^T I(\theta)h + h^T I(\theta)h/2} E[e^{ib^T(Y+Z)}] \\ &= e^{(h+ib)^T I(\theta)(h+ib)/2 + b^T I(\theta)b/2} E[e^{ib^T(Y+Z)}] \end{aligned}$$

ce qui prouve bien l'indépendance car on a le produit d'une fonction de  $h + ib$  par une fonction de  $b$ . La distribution obtenue étant indépendante de la sous-suite, on a bien convergence. ■

Poursuivons cette section avec la propriété de régularité des scores dont l'application principale sera la démonstration du théorème 63

## 62 - THÉORÈME (RÉGULARITÉ DES SCORES)

On se place sous l'hypothèse LAN (p. 100). Alors on a pour tout  $\alpha > 0$  :

$$\sup_{|h| \leq M} P_\theta(|S_n(\theta + h/\sqrt{n}) - S_n(\theta) + I(\theta)h| > \alpha) \longrightarrow 0. \quad (\text{A.13})$$

Sous l'hypothèse ULAN, en particulier si l'expérience est R-régulière, pour tout  $K$  compact intérieur à  $\Theta$  la convergence ci-dessus est uniforme en  $\theta \in K$ .

*Démonstration.* Reprenons les notations de l'équation (IV.5). Notons  $L_{h,k} = \mathcal{L}_n(\theta + (h+k)/\sqrt{n}) - \mathcal{L}_n(\theta + h/\sqrt{n})$ ,  $S_h = S_n(\theta + h/\sqrt{n})$ ,  $\varepsilon_{h,k} = \varepsilon_n(\theta + h/\sqrt{n}, k)$ , alors

$$\begin{aligned} L_{0,h} &= h^T S_0 - \frac{1}{2} h^T I(\theta)h + \varepsilon_{0,h} \\ L_{h,h+k} &= k^T S_h - \frac{1}{2} k^T I(\theta)k + \varepsilon_{h,k} - \frac{1}{2} k^T (I(\theta + h/\sqrt{n}) - I(\theta))k \\ L_{0,h+k} &= (h+k)^T S_0 - \frac{1}{2} (h+k)^T I(\theta)(h+k) + \varepsilon_{0,h+k}. \end{aligned}$$

et en retranchant à la troisième équation la somme des deux premières, les termes  $L_{\dots}$  s'éliminent et il vient

$$k^T (S_h - S_0) + k^T I(\theta)h = \varepsilon_{0,h+k} - \varepsilon_{h,k} - \varepsilon_{0,h} + \frac{1}{2} k^T (I(\theta + h/\sqrt{n}) - I(\theta))k.$$

Les quatre restes sauf  $\varepsilon_{h,k}$  tendent clairement vers 0 en probabilité sous la loi  $P_\theta$ , la convergence étant uniforme en  $\theta, h, k$ . Pour  $\varepsilon_{h,k}$ , on écrit avec les notations du lemme 43

$$\begin{aligned} P_\theta(|\varepsilon_{h,k}| > \alpha) &= P_\theta(|\varepsilon_{h,k}| > \alpha, \Lambda_n \leq \eta) + P_\theta(|\varepsilon_{h,k}| > \alpha, \Lambda_n \geq \eta) \\ &\leq P_\theta(\Lambda_n \leq \eta) + \eta^{-1} E_\theta(1_{|\varepsilon_{h,k}| > \alpha} \Lambda_n) \\ &\leq P_\theta(\log(\Lambda_n) \leq A) + \eta^{-1} P_{\theta+h/\sqrt{n}}(|\varepsilon_{h,k}| > \alpha), \quad A = \log(\eta) \\ &\leq P_\theta(|\varepsilon_{0,h}| > A/2) + P_\theta(|h^T S_n(\theta) - \frac{1}{2} h^T I(\theta)h| > A/2) + \eta^{-1} P_{\theta+h/\sqrt{n}}(|\varepsilon_{h,k}| > \alpha). \end{aligned}$$

Si l'expérience est régulière, la suite  $S_n(\theta)$  étant bornée dans  $L_2(P_\theta)$ , uniformément pour  $\theta \in K$ , tous ces termes peuvent être rendus arbitrairement proches de zéro en choisissant  $\eta$  petit ( $A$  grand) puis  $n$  grand, ceci uniformément en  $\theta$  et  $h$  si le modèle est régulier, grâce à (A.12). En absence de régularité on perd le contrôle uniforme du deuxième terme, et donc l'uniformité en  $\theta$ . ■

On se place sous l'hypothèse LAN. On suppose que l'on dispose d'un estimateur  $\theta_n$  qui converge à vitesse  $n^{-1/2}$  vers  $\theta_*$  au sens où

$$\lim_{A \rightarrow \infty} \sup_n P(\sqrt{n}|\theta_n - \theta_*| > A) = 0.$$

Alors l'estimateur

$$\widehat{\theta}_n = [\theta_n] + n^{-1/2}I^{-1}S_n([\theta_n])$$

où  $[\cdot]$  désigne l'arrondi coordonnée par coordonnée à  $1/\sqrt{n}$  près, satisfait

$$n^{1/2}(\widehat{\theta}_n - \theta_*) + I^{-1}S_n(\theta_*) \xrightarrow{P} 0.$$

*Démonstration.* Nous reprenons les arguments de Le Cam [103]. Remarquons que

$$|n^{1/2}(\widehat{\theta}_n - \theta_*) + I^{-1}S_n(\theta_*)| \leq |n^{1/2}([\theta_n] - \theta_*) + I^{-1}S_n([\theta_n]) - I^{-1}S_n(\theta_*)|.$$

On ne peut pas utiliser directement (A.13) en raison du caractère aléatoire de  $[\theta_n]$ , mais si  $|[\theta_n] - \theta_*| < A/n^{1/2}$ ,  $[\theta_n]$  ne peut prendre qu'un nombre fini de valeurs déterminées qui sont les multiples (vectoriels) de  $1/\sqrt{n}$  à distance inférieure à  $A/\sqrt{n}$  de  $\theta_*$ , valeurs que nous notons  $\delta_i$ ; elles sont en nombre inférieur à  $(2A)^d$ . On a donc

$$\begin{aligned} P(|n^{1/2}(\widehat{\theta}_n - \theta_*) + I^{-1}S_n(\theta_*)| \geq \alpha) &\leq P(n^{1/2}|[\theta_n] - \theta_*| > A) + \sum_i P(|n^{1/2}(\delta_i - \theta_*) + I^{-1}S_n(\delta_i) - I^{-1}S_n(\theta_*)| > \alpha) \\ &\leq P(n^{1/2}|[\theta_n] - \theta_*| > A) + (2A)^d \sup_{|h| \leq A} P(|h + I^{-1}S_n(\theta_* + h/\sqrt{n}) - I^{-1}S_n(\theta_*)| > \alpha). \end{aligned}$$

Un choix de  $A$  assez grand permet de rendre le premier terme arbitrairement petit et le second peut alors être réduit en augmentant  $n$ . ■



## B

---

# DÉMONSTRATION DU LEMME 50

---

Tout repose sur une sommation d'Abel (intégration par parties). Posons  $S_n = \sum_{i=n+1}^{\infty} M_i$ , alors

$$\sum_{i=1}^n R_i M_i = \sum_{i=1}^n R_i (S_{i-1} - S_i) = \sum_{i=1}^{n-1} (R_{i+1} - R_i) S_i - R_n S_n + R_1 S_0$$

d'où

$$\|R_n\|^{-1} \sum_{i=1}^n R_i M_i = \|R_n\|^{-1} \sum_{i=1}^{n-1} (R_{i+1} - R_i) S_i - \|R_n\|^{-1} R_n S_n + \|R_n\|^{-1} R_1 S_0.$$

Les deux derniers termes tendent vers 0. Pour le premier, on va couper la somme en deux :

$$\|R_n\|^{-1} \sum_{i=1}^{n-1} (R_{i+1} - R_i) S_i = \|R_n\|^{-1} \sum_{i=1}^{n_0-1} (R_{i+1} - R_i) S_i + \|R_n\|^{-1} \sum_{i=n_0}^{n-1} (R_{i+1} - R_i) S_i.$$

Le premier terme tend vers 0 et le second satisfait

$$\begin{aligned} \|R_n\|^{-1} \left\| \sum_{i=n_0}^{n-1} (R_{i+1} - R_i) S_i \right\| &\leq \|R_n\|^{-1} \sum_{i=n_0}^{n-1} \|R_{i+1} - R_i\| \|S_i\| \\ &\leq \varepsilon(n_0) \|R_n\|^{-1} \sum_{i=n_0}^{n-1} \|R_{i+1} - R_i\|, \quad \varepsilon(n_0) = \sup_{i \geq n_0} \|S_i\| \\ &\leq \varepsilon(n_0) \|R_n\|^{-1} \sum_{i=n_0}^{n-1} \text{Tr}(R_{i+1} - R_i) \\ &\leq \varepsilon(n_0) \|R_n\|^{-1} \text{Tr}(R_n) \\ &\leq \varepsilon(n_0) d. \end{aligned}$$

On a donc montré que

$$\limsup_n \|R_n\|^{-1} \left\| \sum_{i=1}^n R_i M_i \right\| \leq \varepsilon(n_0) d$$

pour tout  $n_0$ . Par conséquent la limite est nulle. ■



# C

## UNE BORNE SUR LES PROCESSUS EMPIRIQUES

Nous présentons ici des résultats un peu techniques utiles à la démonstration du théorème 12. Commençons par une inégalité très importante due à Arkadi Nemirovski (2000) :

**64 - THÉORÈME** Inégalité de Nemirovski

Soit  $X_i \in \mathbb{R}^T$ ,  $T \geq 3$ , une suite de variables indépendantes vectorielles. En notant  $\|x\|_\infty = \sup_{t \in T} |x_t|$  pour  $x \in \mathbb{R}^T$ , et  $S_n = \sum_{i=1}^n X_i$ , on a

$$E \left[ \|S_n - E[S_n]\|_\infty^2 \right] \leq 8 \ln(2T) E \left[ \left\| \sum_{i=1}^n X_i^2 \right\|_\infty \right]. \quad (\text{C.1})$$

Dans la version originale de Nemirovski,  $\left\| \sum X_i^2 \right\|_\infty$  est remplacé par  $\sum \|X_i^2\|_\infty$ , c'est une inégalité beaucoup plus simple à montrer, voir par exemple [67] (où l'on voit que l'inégalité reste vraie pour des accroissements de martingale), qui nous suffira pour la démonstration du théorème qui suit (nous référons à [46] pour la démonstration de (C.1)).

**65 - THÉORÈME**

Soit  $\theta \mapsto L_i(\theta)$  une suite processus indépendants à valeurs réelles définis sur une boule  $\{\theta : \|\theta\| \leq r\} \subset \mathbb{R}^d$ . On suppose qu'il existe des variables indépendantes  $\dot{L}_i$  telles que

$$|L_i(\theta) - L_i(\theta')| \leq \dot{L}_i \|\theta - \theta'\| \quad (\text{C.2})$$

$$E[L_i(\theta)] = 0 \quad (\text{C.3})$$

$$L_i(0) = 0 \quad (\text{C.4})$$

et que les constantes suivantes sont finies

$$C(L) = \sup_i E[\sup_\theta L_i(\theta)^2]^{\frac{1}{2}} \quad (\text{C.5})$$

$$C(\dot{L}) = \max_i E[\dot{L}_i^2]^{\frac{1}{2}}. \quad (\text{C.6})$$

Alors, pour tout  $r > 0$

$$E \left[ \sup_{\|\theta\| < r} n^{-1} \left| \sum_{i=1}^n L_i(\theta) \right|^2 \right]^{\frac{1}{2}} \leq C_0 d C(L) \left( 1 + \log \left( \frac{C(\dot{L})r}{C(L)} \right) \right)^{\frac{1}{2}} \leq C_0 d C(\dot{L})r \quad (\text{C.7})$$

où  $C_0$  est une constante universelle.

*Remarque.* La première inégalité sera utilisée dans un contexte particulier où  $r \rightarrow 0$ , cas où l'on peut avoir  $C(L) \ll C(\dot{L})r$ . Sinon la seconde convient a priori.

*Démonstration.* Pour tout  $x > 0$ , définissons le réseau  $\Theta_x = \{\theta : \|\theta\| \leq r\} \cap ((2x)\mathbb{Z})$ . On note

$$S_n(\theta) = n^{-\frac{1}{2}} \sum_{i=1}^n L_i(\theta), \quad \varphi(x) = E \left[ \sup_{\theta \in \Theta_x} |S_n(\theta)|^2 \right]^{\frac{1}{2}}.$$

En vertu du théorème de Lebesgue, la limite de  $\varphi(x)$  quand  $x$  tend vers 0 est bien  $\varphi(0)$ , la racine carrée du membre de gauche de (C.1). On a pour  $x \leq y$ , en notant  $\hat{\theta}$  le point de  $\Theta_y$  le plus proche de  $\theta$

$$\sup_{\theta \in \Theta_x} |S_n(\theta)| \leq \sup_{\theta \in \Theta_x} |S_n(\hat{\theta})| + \sup_{\theta \in \Theta_x} |S_n(\hat{\theta}) - S_n(\theta)| = \sup_{\theta \in \Theta_y} |S_n(\theta)| + \sup_{\theta \in \Theta_x} |S_n(\hat{\theta}) - S_n(\theta)|.$$

Pour traiter le dernier terme, on va appliquer l'inégalité de Nemirovski avec  $X_i = n^{-\frac{1}{2}}(L_i(\hat{\theta}) - L_i(\theta))_{\theta \in \Theta_x}$ , et  $T = \Theta_x$ . Comme  $\|\hat{\theta} - \theta\|^2 \leq dy^2$ , on a  $\|X_i\|_\infty^2 \leq n^{-1}dy^2\dot{L}_i^2$ , et par ailleurs  $|\Theta_x| \leq (r/x)^d$ . Donc

$$E \left[ \sup_{\theta \in \Theta_x} |S_n(\hat{\theta}) - S_n(\theta)|^2 \right] \leq 8 \ln(2(r/x)^d) n^{-1} dy^2 E \sum \dot{L}_i^2.$$

En rassemblant ces équations on obtient

$$\varphi(x) \leq \varphi(y) + \left( 8d \ln(2(r/x)^d) C(\dot{L})^2 y^2 \right)^{\frac{1}{2}}.$$

Soit  $n_0 \geq 1$ . En faisant  $x = 2^{-n-1}r$  et  $y = 2^{-n}r$  et en sommant sur  $n \geq n_0$ , il vient

$$\begin{aligned} \varphi(0) &\leq \varphi(2^{-n_0}r) + \sum_{n=n_0}^{\infty} \left( 8d \ln(2^{(n+1)d+1}) C(\dot{L})^2 r^2 2^{-2n} \right)^{\frac{1}{2}} \\ &\leq \varphi(2^{-n_0}r) + C_0 d C(\dot{L}) r \sum_{n=n_0}^{\infty} n^{\frac{1}{2}} 2^{-n} \\ &\leq \varphi(2^{-n_0}r) + C_0 d C(\dot{L}) r n_0^{\frac{1}{2}} 2^{-n_0}. \end{aligned}$$

On peut majorer à son tour  $\varphi(2^{-n_0}r)$  en utilisant l'inégalité de Nemirovski avec  $X_i = n^{-\frac{1}{2}}(L_i(\theta))_{\theta \in T}$ , et  $T = \Theta_{2^{-n_0}r}$ . On a  $E[\|X_i\|_\infty^2] \leq n^{-1}C(L)^2$ , d'où

$$\varphi(2^{-n_0}r)^2 \leq 8 \ln(2^{dn_0+1}) C(L)^2 \leq C_0 d n_0 C(L)^2.$$

Donc

$$\varphi(0) \leq C_\alpha n_0^{\frac{1}{2}} d \left( C(L) + C(\dot{L}) r 2^{-n_0} \right).$$

On obtient alors la première inégalité en prenant  $n_0 = \lceil \log_2(C(\dot{L})r/C(L)) \rceil$ . La deuxième inégalité vient de ce que  $C(L) \leq C(\dot{L})r$  et  $x \mapsto (1 + \ln x)/x$  est décroissante pour  $x \geq 1$ . ■

#### 66 - LEMME

Soit  $U$  une variable aléatoire positive telle pour tout  $x > 0$

$$E[U 1_{U \leq x}] \leq C_1 \sqrt{x} + C_2,$$

alors pour tout  $x > 0$

$$P(U > x) \leq \frac{4C_1}{\sqrt{x}} + \frac{2C_2}{x}.$$

*Démonstration.* Posons  $x_n = 4^n x$ , alors

$$P(x_n \leq U \leq x_{n+1}) \leq x_n^{-1} E[U 1_{U \leq x_{n+1}}] \leq C_1 2^{1-n} x^{-\frac{1}{2}} + C_2 x^{-1} 4^{-n}$$

et il ne reste plus qu'à sommer sur  $n \geq 0$ . ■

# D

## THÉORÈMES-LIMITE POUR LES MARTINGALES

Nous énonçons un théorème classique, puis le théorème 2.17 et le corollaire 3.1 de [10].

### 67 - THÉORÈME

Soit  $S_n = \sum_{i=1}^n X_i$  une  $\mathcal{F}_n$ -martingale, alors, sur l'ensemble  $\{\omega : \sum E[X_i^2 | \mathcal{F}_{i-1}] < \infty\}$ ,  $S_n$  converge p.s.

Un tableau de martingales est une famille de variables et de tribus  $\{S_{ni}, \mathcal{F}_{ni}, 1 \leq i \leq n\}$  telles que

$$E[S_{n,i+1} | \mathcal{F}_{ni}] = S_{ni}.$$

C'est-à-dire que pour tout  $n$  fixé, la suite  $S_{ni}, 1 \leq i \leq n$ , est une martingale. On notera  $X_{ni} = S_{ni} - S_{n,i-1}$  les accroissements de martingale. L'exemple le plus simple est  $S_{ni} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$  où  $Y_i$  est une suite iid centrée,  $X_{ni} = \frac{1}{\sqrt{n}} Y_i$ .

### 68 - THÉORÈME

Soit  $\{S_{ni}, \mathcal{F}_{ni}, 1 \leq i \leq n, 1 \leq n\}$  un tableau de martingales vectorielles :

$$E[S_{n,i+1} | \mathcal{F}_{ni}] = S_{ni}.$$

Supposons que les variables  $X_{ni} = S_{ni} - S_{n,i-1}$  satisfont les conditions suivantes

$$\sum_i E[\|X_{ni}\|^2 1_{\|X_{ni}\| > \varepsilon} | \mathcal{F}_{n,i-1}] \xrightarrow{P} 0, \quad \text{pour tout } \varepsilon > 0 \tag{D.1}$$

$$\sum_{i=1}^n E[X_{ni} X_{ni}^T | \mathcal{F}_{n,i-1}] \xrightarrow{P} V. \tag{D.2}$$

pour une certaine matrice  $V$ . Alors

$$S_{nn} \xrightarrow{d} \mathcal{N}(0, V). \tag{D.3}$$

### 69 - THÉORÈME

On se place sous les hypothèses du théorème 68 sauf que l'on autorise  $V$  à être une variable aléatoire. On suppose en outre que  $\mathcal{F}_{ni} \subset \mathcal{F}_{n+1,i}$ . Dans ce cas on a la même conclusion, conditionnellement à  $V$ , au sens où pour tout  $u \in \mathbb{R}^d$  et toute fonction  $\varphi$  continue bornée

$$E[\varphi(V) e^{i\langle u, S_{nn} \rangle}] \rightarrow E[\varphi(V) e^{-\langle Vu, u \rangle / 2}]. \tag{D.4}$$



# E

---

## LES CRITÈRES MML, MDL, ET BIC

---

---

Nous nous proposons ici de donner une brève histoire de ces critères et d'en présenter les idées principales. L'article de Hansen et Yu [87] pourra compléter cette partie.

Un petit rappel est nécessaire : Soit un espace probabilisé fini  $\Omega = \{y_1, \dots, y_p\}$  muni de probabilités  $p_i = p(y_i)$ . Un code est une application  $y_i \mapsto C(y_i) = C_i$  qui associe à chaque  $y_i$  une suite de 0 et 1 différente (séquence de bits, mot binaire). On note  $l_i$  la longueur de  $C_i$ . Alors il existe un code (le code de Shannon) tel que  $l_i = \lceil \log(1/p_i) \rceil$  et dont la longueur moyenne  $\sum_i p_i l_i$  est à distance inférieure à 1 du minimum possible. L'expression de  $l_i$  montre qu'il attribue les séquences de bits les plus courtes aux  $y$  les plus probables ; les détails se trouvent par exemple dans [54]. La quantité  $\lceil \log(1/p_i) \rceil$  représente donc le coût de codage de  $y_i$  par ce code optimisé pour  $p$ . Le code Morse international, qui traduit la lettre  $e$  par un simple point, illustre très bien cela.

Dans le cas continu,  $y \in \mathbb{R}^d$ , la formulation exacte est que le coût de codage de  $y$  avec précision  $\varepsilon$  sur chaque coordonnée est  $-\log(\varepsilon^d p(y))$  (i.e. on s'est ramené au cas discret en attribuant la probabilité  $\varepsilon^d p(y_i)$  à chaque valeur  $\varepsilon$ -discrétisée  $y_i \in \varepsilon \mathbb{Z}^d$ , à un terme d'ordre supérieur près). En raison du caractère additif constant du terme  $-\log(\varepsilon^d)$ , on peut résumer ainsi la situation :

*La quantité  $\log(1/p(y))$  représente le coût de la transmission de  $y$  par un code basé sur la distribution  $p$ . Si  $y$  est aléatoire, le coût de transmission moyen est minimisé pour  $p$  coïncidant avec la distribution de  $y$ .*

Interprétant la vraie distribution comme le meilleur codeur, on est mené à juger de la qualité d'un modèle  $(p_\theta)_{\theta \in \Theta}$  pour des données  $y$  par sa capacité à les résumer par un message contenant un paramètre  $\theta$  bien choisi et  $y$  codé sur la base de  $p_\theta$ . Mais si l'on veut comparer deux modèles  $(p_\theta)_{\theta \in \Theta}$  et  $(p_\theta)_{\theta \in \Theta'}$  à cette aune, p.ex. les AR(2) et les AR(20), il faut aussi prendre en compte les coûts de codage de  $\theta \in \Theta$  et  $\theta' \in \Theta'$  qui peuvent être très différents si ces vecteurs sont de dimensions différentes : si  $\Theta$  augmente, le coût de codage de  $y$  diminuera probablement tandis que celui de  $\theta$  augmentera. On est donc conduit à minimiser le coût de codage du message global

$$C(Y, \theta) = C(Y|\theta) + C(\theta) = -\log(p_\theta(y)) + C(\theta),$$

ces deux quantités se mesurant en nombre de bits. La minimisation se fait modèle par modèle, ce qui permet d'attribuer à chacun un coût qui lui est propre, comme dans les formules (E.2) ou (E.3) plus bas. Rappelons que dans le cas continu le terme additionnel  $-d \log(\varepsilon)$  a été omis car il ne dépend pas de  $\theta$ .

WALLACE ET BOULTON ont lancé en 1968 cette idée de faire intervenir la théorie de l'information, dans un article concernant le choix du nombre de classes en classification : « The "best" classification will result from an optimum compromise between the efficiency of attribute encoding and the length of message needed to specify the distribution » ([139] §2), i.e. *entre avoir des classes petites (réduire  $C(Y|\theta)$ ) et avoir peu de classes (réduire  $C(\theta)$ )*.

L'élégance de cette approche est qu'elle ne nécessite pas que les données soient issues d'un modèle, elle cherche simplement le modèle le plus opérationnel en termes de codage, ce codage étant fait *en deux*

*temps* : coder  $\theta$ , puis  $y$ . Ce principe basé sur le codage en deux temps est généralement associé au terme MML (minimum message length).

Comme pour le maximum de vraisemblance, ou la théorie bayésienne, l'interprétation la moins risquée de cette théorie est qu'elle permet de proposer des méthodes qu'il reste ensuite à valider en théorie et en pratique.

JORMA RISSANEN étend en 1978 [126] le calcul de Wallace et Boulton à un cadre plus général que la classification. Il reprend le critère de coût total du codage des observations  $(Y_i)$  avec précision  $\pm\varepsilon/2$  quand on se base sur la distribution  $p_{\hat{\theta}}(y)$ ; ce coût de codage doit aussi prendre en compte le codage des composantes de  $\hat{\theta}$  qu'il faut également transmettre pour que le récepteur puisse savoir quel décodeur utiliser. La procédure, dont le point reste l'optimisation de la précision  $\delta_k$  avec laquelle on transmet  $\hat{\theta}$ , est bien décrite dans [87], donnons-en les grands traits<sup>1</sup>. On se place dans le cadre d'observations iid, ce qui permet de mieux mettre en valeur l'essence du calcul.

Soit  $\pm\delta_k/2$  la précision (à déterminer) du codage de chaque coordonnée  $\hat{\theta}_k$ ; ce codage sera basé sur une probabilité a priori  $p_d$  sur  $\Theta_d$  (le modèle de dimension  $d$ ) d'entropie finie; le  $\hat{\theta}$  codé sera noté  $\check{\theta}$ . On obtient le coût de codage

$$-\sum_{i=1}^n \log \varepsilon^q p_{\check{\theta}}(Y_i) - \log \left( \left( \prod \delta_k \right) p_d(\hat{\theta}) \right)$$

où  $q$  est la dimension de  $Y_i$ , les logarithmes sont négatifs (on raisonne avec  $\varepsilon$  très petit et on verra que  $\delta_k$  tend vers 0; sinon le calcul serait bien plus compliqué), et le coût de la transmission de  $d$  a été négligé. Rissanen calcule alors la valeur optimale des  $\delta_k$  comme suit (elle sera de l'ordre de la précision sur  $\hat{\theta}$ , c'est-à-dire  $1/\sqrt{n}$ ). Si l'on pose  $\check{\delta} = \check{\theta} - \hat{\theta}$ , l'équation plus haut s'approxime en utilisant la formule de Taylor à l'ordre 2, et en ne gardant que les termes dépendants de  $\delta$

$$-\sum_{i=1}^n \log p_{\check{\theta}}(Y_i) - \sum_k \log \delta_k \simeq -\sum_{i=1}^n \log p_{\hat{\theta}}(Y_i) + \frac{1}{2} \check{\delta}^T H_n(\hat{\theta}) \check{\delta} - \sum_k \log \delta_k$$

$$H_n(\theta) = -\nabla^2 \sum_{i=1}^n \log p_{\theta}(Y_i)$$

(le terme du premier ordre est nul car  $\hat{\theta}$  réalise le maximum).  $H_n(\theta)$  est donc proche de l'information de Fisher *si  $Y$  suit le modèle*. Supposons dans un premier temps que  $H_n(\theta)$  est diagonale. On voit alors que pour minimiser la quantité ci-dessus, chaque coordonnée de  $\theta$  doit être discrétisée à un pas différent  $\delta_k^{-2} \simeq c H_n(\theta)_{kk}$  ( $c$  est une constante d'ordre 1 liée à la distribution effective des  $\check{\delta}_k \in [-\delta_k, \delta_k]$ ) et on obtient le critère (première approximation) :

$$-\sum_{i=1}^n \log p_{\hat{\theta}(Y)}(Y_i) + \frac{d}{2} \log n + \frac{1}{2} \log \det \left( \frac{1}{n} H_n(\hat{\theta}) \right) - \log p_d(\hat{\theta}) + O(d). \quad (\text{E.1})$$

Si  $H_n(\theta)$  n'est pas diagonale, on s'y ramène par une rotation, c.-à-d. que l'on discrétise les coordonnées de  $\hat{\theta}$  dans une base de vecteurs propres de  $H(\hat{\theta})$ , et l'on retrouve (E.1). Si l'on considère  $n$  grand, on obtient le critère

$$BIC(d) = -\log p_{\hat{\theta}(Y)}(Y) + \frac{d}{2} \log n \quad (\text{E.2})$$

(Bayesian Information Criterion de Gideon Schwarz) présenté plus bas<sup>2</sup>. La référence [134] recommande

1. Dans ce qui suit, on s'en tient à l'idée de coder  $\theta$  à l'aide d'une probabilité a priori, probabilité qui on va le voir, n'intervient asymptotiquement pas. Dans [126] ce codage est la partie entière  $\hat{\theta}/\delta_k$ , écrite en base 2, ce qui fait essentiellement un coût de  $\log(|\hat{\theta}|/\delta_k)_+$  bits, qui ne change pas à l'asymptotique.

2. Wallace et Freeman dans [140] obtiennent (E.1) par un calcul différent (§5.1, formule p.245) : Ils cherchent la stratégie  $y \mapsto \hat{\theta}(y)$  qui minimise le coût de codage *moyen* sous l'hypothèse que les données ont pour densité  $\int p_{\theta}(y) \pi(d\theta)$ , où  $\pi$  est une loi sur l'ensemble de tous les paramètres; le développement n'est pas fait en  $\hat{\theta}$  mais en un point (voisin) qui doit minimiser  $-\log p_{\theta}(Y) + \frac{1}{2} \log \det(I(\theta))$  où  $I(\theta)$  désigne l'information de Fisher de l'échantillon complet; l'annulation du terme d'ordre 1 dans le développement vient du fait que  $\delta$  est centré et que les calculs y sont faits en espérance.

Dans un cadre où l'on connaît  $\pi$ , c'est bien entendu plus logique que le codage en deux temps avec  $\hat{\theta}(y)$ .

L'avantage de cette approche est la précision du cadre, c'est aussi son inconvénient car on perd la simplicité de l'approche basée sur le codage de l'échantillon  $Y$ , en imposant une hypothèse de modèle sur la distribution de  $Y$ .

de conserver la forme plus complète

$$SIC(d) = -\log p_{\hat{\theta}(Y)}(Y) + \frac{1}{2} \log \det(H_n(\hat{\theta})), \quad (\text{E.3})$$

la dénomination SIC provenant de [87] éq. (17), où son intérêt est davantage relativisé (cf. § 4.1.2 et 4.1.3 de l'article).

Considérons le cas de la régression linéaire gaussienne  $Y_i = X_i\beta + u_i$ ,  $u_i \sim \mathcal{N}(0, \sigma^2)$ ,  $\beta \in \mathbb{R}^p$ ; il s'agirait ici de sélectionner le nombre des variables parmi un grand nombre  $p_0$  (l'analogue de l'ordre maximal pour les AR(p)). Avec pour paramètres  $\theta = (\beta, \tau = \sigma^2)$ ,  $d = p + 1$ , on trouve

$$H(\hat{\theta}) = \begin{pmatrix} \frac{X^T X}{\tau} & 0 \\ 0 & \frac{n}{2\tau^2} \end{pmatrix} \quad (\text{E.4})$$

$$SIC(d) = \frac{1}{2}(n - p - 2) \log \hat{\sigma}^2 + \frac{p}{2} \log n + \frac{1}{2} \log \det\left(\frac{X^T X}{n}\right) \quad (\text{E.5})$$

$$BIC(d) = \frac{1}{2}n \log \hat{\sigma}^2 + \frac{p+1}{2} \log n. \quad (\text{E.6})$$

LE MDL. On pourrait choisir de coder autrement qu'en deux temps, et il se trouve que ce choix apparaît important dès que l'on veut des résultats théoriques précis; en particulier, le codage en deux temps est sous-optimal puisque redondant du fait que la connaissance de  $\hat{\theta}(y)$  donne une information sur  $y$  permettant réduire encore le coût de codage. L'extension à d'autres codages que le codage en deux temps est généralement associée au terme MDL (minimum description length) proposé par Rissanen<sup>3</sup>.

Ce problème de choix de codage a conduit à une spirale mathématico-philosophique émaillée de discussions byzantines et de calculs sophistiqués dans de nombreux articles, visant à gagner en conceptualisation. Rien de bien clair ne semble s'en dégager en ce qui concerne l'estimation. Finalement, tout dépend du sens précis que l'on donne à *chacun* des mots de la phrase « *un bon modèle est un modèle qui permet de coder efficacement les données* »; la théorie initiale de Wallace et Boulton (MML, codage en deux temps) reste une option sagement opérationnelle.

Le lien entre codage et probabilité a priori incite à comparer avec une approche bayésienne, ce qui nous conduit à l'approche de Schwarz.

GIDEON SCHWARZ a considéré en 1978 une approche Bayésienne (MAP) postulant une probabilité a priori sur les paramètres  $p(\theta) = \sum_d \alpha_d p_d(\theta)$  où  $\alpha_d > 0$  et  $p_d$  est une probabilité sur  $\Theta_d$  [130]; dans cette approche,  $d$  sera obtenu par maximisation de la probabilité a posteriori  $\alpha_d \int p_\theta(Y) p_d(\theta) d\theta$ . Il montre que cette méthode revient asymptotiquement à maximiser  $BIC(d)$ , indépendamment des  $p_d$  et des  $\alpha_d$  (sous certaines conditions raisonnables...). Ici aussi, c'est la forme (E.3) qui apparaît naturellement.

CHRISTOPHE GIRAUD souligne que si pour chaque valeur de  $d$  de nombreux modèles sont proposés, comme en sélection des variables en régression, le calcul de Schwarz doit être modifié, et de la même façon le troisième terme de (E.1) doit tenir compte de ce qu'une certaine probabilité, disons  $\pi_d$  si elle ne dépend que de  $d$ , doit être affectée à chacun de ces modèles ([9] p. 49); ceci ajoute à BIC un terme  $-\log(\pi_d)$ . Si  $\nu_d$  est le nombre de modèles de dimension  $d$ , le choix  $\pi_d^{-1} = d(d+1)\nu_d$  conduit à

$$BIC(d) = -\sum_{i=1}^n \log p_{\hat{\theta}(Y)}(Y_i) + \frac{d}{2} \log n + \log \nu_d$$

(le terme  $\log(d(d+1))$  a été négligé). Dans le cas de la régression, si l'on considère tous les  $\binom{p_0}{d}$  modèles de dimension  $d$  obtenus à partir des  $p_0$  variables de départ, il vient pour  $d \ll p_0$ ,  $\log \nu_d \simeq d \log(p_0/d)$ .

**Exercice (Echec de BIC ou SIC si  $\dim(\theta)$  approche  $n$ ).** On reprend l'exemple de la régression linéaire. Montrer que si  $p_0 > n$  alors BIC choisira toujours un modèle à  $n$  variables ( $\hat{\sigma} = 0$ ). Montrer que si l'on prend maintenant pour paramètre  $\theta = (\beta, \tau = \sigma^{-2})$  dans SIC, alors le facteur  $(n - p - 2)$  dans (E.5) devient  $(n - p + 2)$ , ce qui aboutit au même résultat.

3. P.ex. il propose dans [125] de coder  $Y$  sur la base de la mesure  $p_{\hat{\theta}(y)}(y) / \int p_{\hat{\theta}(z)}(z) dz$ , où ici  $\hat{\theta}$  est l'estimateur au maximum de vraisemblance dans la classe concernée (dans l'exemple du début l'AR(d)). La comparaison des coûts de codage dans les différentes classes permettant ensuite de choisir.



# Bibliographie

- [1] P. J. BICKEL et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, 1993.
- [2] A. BOROVKOV. *Statistique Mathématique*. Mir, 1987.
- [3] A. DASGUPTA. *Asymptotic theory of statistics and probability*. Springer, 2008.
- [4] P. J. DIGGLE. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press, 2014.
- [5] R. DURRETT. *Probability : theory and examples*. Cambridge Univ. Press, 2010.
- [6] B. EFRON. *Large-scale inference*. Cambridge Univ. Press, 2010.
- [7] A. GELMAN et al. *Bayesian data analysis*. CRC Press, 2014.
- [8] E. GINÉ et R. NICKL. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Univ. Press, 2016.
- [9] C. GIRAUD. *Introduction to high-dimensional statistics*. T. 139. CRC Press, 2015.
- [10] P. HALL et C. C. HEYDE. *Martingale limit theory and its application*. Academic Press, 1980.
- [11] T. HASTIE, R. TIBSHIRANI et J. FRIEDMAN. *The elements of statistical learning*. 2<sup>e</sup> éd. Springer, 2009.
- [12] P. J. HUBER et E. M. RONCHETTI. *Robust statistics*. Wiley, 2009.
- [13] I. A. IBRAGIMOV et R. Z. HASMINSKIÏ. *Statistical estimation*. Springer, 1981.
- [14] E. L. LEHMANN et G. CASELLA. *Theory of point estimation*. Springer, 1998.
- [15] E. L. LEHMANN et J. P. ROMANO. *Testing statistical hypotheses*. Springer, 2005.
- [16] A. MONFORT. *Cours de statistique mathématique*. Economica, 1997.
- [17] A. PAKES et D. POLLARD. « Simulation and the asymptotics of optimization estimators ». In : *Econometrica* 57.5 (1989), p. 1027-1057.
- [18] C. R. RAO et Y. WU. « On model selection ». In : *Model selection*. T. 38. Inst. Math. Statist., 2001, p. 1-64.
- [19] C. P. ROBERT. *The Bayesian choice*. 2<sup>e</sup> éd. Springer-Verlag, 2001.
- [20] R. J. SERFLING. *Approximation theorems of mathematical statistics*. Wiley, 1980.
- [21] A. W. van der VAART. *Asymptotic statistics*. Cambridge Univ. Press, 1998.
- [22] A. W. van der VAART et J. A. WELLNER. *Weak convergence and empirical processes*. Springer, 1996.
- [23] L. WASSERMAN. *All of nonparametric statistics*. Springer, 2006.



# Références

- [24] R. ADAMS. *Sobolev spaces*. Academic Press, 1975.
- [25] J. H. ALBERT et S. CHIB. « Bayesian analysis of binary and polychotomous response data ». In : *J. Amer. Statist. Assoc.* 88.422 (1993), p. 669-679.
- [26] Y. ATCHADÉ. *Markov Chain Monte Carlo confidence intervals*. Rapp. tech. 2013.
- [27] A. BADDELEY et R. TURNER. « Practical maximum pseudolikelihood for spatial point patterns (with discussion) ». In : *Aust. N. Z. J. Stat.* 42.3 (2000), p. 283-322.
- [28] A. BADDELEY. « Time-invariance estimating equations ». In : *Bernoulli* 6.5 (2000), p. 783-808.
- [29] E. BARANKIN. « Locally best unbiased estimates ». In : *Ann. Math. Statistics* 20 (1949), p. 477-501.
- [30] M. BASSEVILLE, M. GOURSAT et L. MEVEL. « Statistical model-based damage detection and localization : subspace-based residuals and damage-to-noise sensitivity ratios ». In : *J. of Sound and Vibration* 275 (2004), p. 769-794.
- [31] R. BERAN. « Prepivoting test statistics : a bootstrap view of asymptotic refinements ». In : *J. Amer. Statist. Assoc.* 83.403 (1988), p. 687-697.
- [32] J. BERKSON. « Minimum chi-square, not maximum likelihood ! » In : *Ann. Statist.* 8.3 (1980).
- [33] J. BESAG. « Spatial interaction and the statistical analysis of lattice systems ». In : *J. Roy. Statist. Soc. Ser. B* 36 (1974), p. 192-236.
- [34] P. BICKEL et J. GHOSH. « A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument ». In : *Ann. Statist.* 18.3 (1990), p. 1070-1090.
- [35] P. BICKEL et Y. RITOV. « Inference in hidden Markov models. I. Local asymptotic normality in the stationary case ». In : *Bernoulli* 2.3 (1996), p. 199-228.
- [36] P. BILLINGSLEY. *Convergence of probability measures*. John Wiley & Sons, 1999.
- [37] L. BIRGÉ. « Sur un théorème de minimax et son application aux tests ». In : *Probab. Math. Statist.* 3.2 (1984), p. 259-282.
- [38] O. BJØRNSTAD, B. FINKENSTÄDT et B. GRENFELL. « A stochastic model for extinction and recurrence of epidemics : estimation and inference for measles outbreaks ». In : *Biostatistics* 4.4 (2002), p. 493-510.
- [39] L. BREIMAN. « The little bootstrap and other methods for dimensionality selection in regression : X-fixed prediction error ». In : *J. Amer. Statist. Assoc.* 87.419 (1992), p. 738-754.
- [40] L. BREIMAN et P. SPECTOR. « Submodel Selection and Evaluation in Regression. The X-Random Case ». In : *Int. Statis. Review* 60.3 (1992), p. 291-319.
- [41] P. BRÉMAUD. *Point processes and queues*. Springer, 1981.
- [42] P. BRÉMAUD et J. JACOD. « Processus ponctuels et martingales : résultats récents sur la modélisation et le filtrage ». In : *Advances in Appl. Probability* 9.2 (1977), p. 362-416.
- [43] H. BREZIS. *Analyse fonctionnelle*. Masson, 1983.
- [44] L. D. BROWN et E. GREENSHTEIN. « Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means ». In : *Ann. Statist.* 37.4 (2009), p. 1685-1704.

- [45] L. D. BROWN, E. GREENSHTEIN et Y. RITOV. « The Poisson compound decision problem revisited ». In : *J. Amer. Statist. Assoc.* 108.502 (2013), p. 741-749.
- [46] P. BÜHLMANN et S. van de GEER. *Statistics for high-dimensional data*. Springer, 2011.
- [47] P. BURMAN. « A comparative study of ordinary cross-validation,  $v$ -fold cross-validation and the repeated learning-testing methods ». In : *Biometrika* 76.3 (1989), p. 503-514.
- [48] O. CATONI. *A mixture approach to universal model selection*. Rapp. tech. LMENS-97-30. <http://www.math.ens.fr/EDITION/preprints/1997/resu9730.html>. 1997.
- [49] S. X. CHEN et I. VAN KEILEGOM. « A review on empirical likelihood methods for regression ». In : *TEST* 18.3 (2009), p. 415-447.
- [50] S. N. CHIU et al. *Stochastic geometry and its applications*. Third. John Wiley & Sons, 2013.
- [51] B. CLARKE et A. BARRON. « Jeffreys' prior is asymptotically least favorable under entropy risk ». In : *J. Statist. Plann. Inference* 41.1 (1994), p. 37-60.
- [52] J. COHEN et al. « Hierarchical Bayesian Analysis of Arrest Rates ». In : *J. Amer. Statist. Assoc.* 93.444 (1998). [www.stat.cmu.edu/tr/tr636/tr636.html](http://www.stat.cmu.edu/tr/tr636/tr636.html), p. 1260-1270.
- [53] R. COOK et L. NI. « Sufficient dimension reduction via inverse regression : a minimum discrepancy approach ». In : *J. Amer. Statist. Assoc.* 100.470 (2005), p. 410-428.
- [54] T. M. COVER et J. A. THOMAS. *Elements of information theory*. Second. Wiley, 2006.
- [55] D. COX. « Partial likelihood ». In : *Biometrika* 62.2 (1975), p. 269-276.
- [56] D. COX. « The choice between alternative ancillary statistics ». In : *J. Roy. Statist. Soc. Ser. B* 33 (1971), p. 251-255.
- [57] D. COX et N. REID. « A note on pseudolikelihood constructed from marginal densities ». In : *Biometrika* 91.3 (2004), p. 729-737.
- [58] N. CRESSIE. *Statistics for spatial data*. Wiley, 1991.
- [59] J. CVITANIĆ et I. KARATZAS. « Generalized Neyman-Pearson lemma via convex duality ». In : *Bernoulli* 7.1 (2001), p. 79-97.
- [60] S. DACHIAN et Y. KUTOYANTS. « Hypotheses testing : Poisson versus self-exciting ». In : *Scand. J. Statist.* 33.2 (2006), p. 391-408.
- [61] L. DICKER et S. ZHAO. *Nonparametric empirical Bayes and maximum likelihood estimation for high-dimensional data analysis*. Rapp. tech. 2014.
- [62] O. DIEKMANN, H. HEESTERBEEK et T. BRITTON. *Mathematical tools for understanding infectious disease dynamics*. Princeton University Press, 2013.
- [63] P. DIGGLE. « A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point ». In : *J. R. S. S. A* 153.3 (1990), p. 349-362. Données disponibles sous R, package spatstat.
- [64] P. DIGGLE et B. S. ROWLINGSON. « A Conditional Approach to Point Process Modelling of Elevated Risk ». In : *J. R. S. S. A* 157.3 (1994), p. 433-440.
- [65] D. L. DONOHO et I. M. JOHNSTONE. « Adapting to unknown smoothness via wavelet shrinkage ». In : *J. Amer. Statist. Assoc.* 90.432 (1995), p. 1200-1224.
- [66] J. DOOB. « Application of the theory of martingales ». In : *Le Calcul des Probabilités et ses Applications*. CNRS, 1949, p. 23-27.
- [67] L. DÜMBGEN et al. « Nemirovski's inequalities revisited ». In : *Amer. Math. Monthly* 117.2 (2010), p. 138-160.
- [68] N. DUNFORD et J. SCHWARTZ. *Linear operators. Part I*. Wiley, 1988.
- [69] B. EFRON. « Robbins, empirical Bayes and microarrays ». In : *Ann. Statist.* 31.2 (2003), p. 366-378.
- [70] B. EFRON. « The estimation of prediction error : covariance penalties and cross-validation ». In : *J. Amer. Statist. Assoc.* 99.467 (2004). With comments and a rejoinder by the author, p. 619-642.

- [71] B. EFRON et C. MORRIS. « Data Analysis Using Stein's Estimator and its Generalizations ». In : *J. Amer. Statist. Assoc.* 70.350 (1975), p. 311-319.
- [72] H.-r. FANG et Y. SAAD. « Two classes of multiseccant methods for nonlinear acceleration ». In : *Numer. Linear Algebra Appl.* 16.3 (2009), p. 197-221.
- [73] J. FLEGAL et G. JONES. « Batch means and spectral variance estimators in Markov chain Monte Carlo ». In : *Ann. Statist.* 38.2 (2010), p. 1034-1070.
- [74] T. FLEMING et D. HARRINGTON. *Counting processes and survival analysis*. Wiley, 1991.
- [75] C. FRANCO et J.-M. ZAKOÏAN. « Estimating linear representations of nonlinear processes ». In : *J. Statist. Plann. Inference* 68.1 (1998), p. 145-165.
- [76] D. A. S. FRASER. « Is Bayes posterior just quick and dirty confidence? » In : *Statist. Sci.* 26.3 (2011), p. 299-316.
- [77] M. FRÉCHET. « Sur l'extension de certaines évaluations statistiques au cas de petits échantillons ». In : *Rev. Inst. Intern. Stat.* 11 (1943), p. 182-205.
- [78] C. GAETAN et X. GUYON. *Modélisation et statistique spatiales*. T. 63. Springer-Verlag, 2008.
- [79] S. GHOSAL. « A review of consistency and convergence rates of posterior distribution ». In : *Proc. of Varanashi Symp. in Bayesian Inference*. [www4.stat.ncsu.edu/~sghoshal/papers.html](http://www4.stat.ncsu.edu/~sghoshal/papers.html). 1996.
- [80] R. GILL et B. LEVIT. « Applications of the Van Trees inequality : a Bayesian Cramér-Rao bound ». In : *Bernoulli* 1.1-2 (1995), p. 59-79.
- [81] V. GODAMBE et B. KALE. « Estimating functions : an overview ». In : t. 7. Oxford Univ. Press, 1991, p. 3-20.
- [82] J. GU et R. KOENKER. « REBayes : Empirical Bayes Mixture Methods in R ». In : *Journal of Statistical Software* 82.8 (2015).
- [83] J. GU et R. KOENKER. « Unobserved heterogeneity in income dynamics : an empirical Bayes perspective ». In : *Journal of Business & Economic Statistics* (2015).
- [84] P. HALL. « Cross-validation in density estimation ». In : *Biometrika* 69.2 (1982), p. 383-390.
- [85] P. HALL et S. WILSON. « Two guidelines for bootstrap hypothesis testing ». In : *Biometrics* 47.2 (1991), p. 757-762.
- [86] E. HANNAN. « The estimation of the order of an ARMA process ». In : *Ann. Statist.* 8.5 (1980), p. 1071-1081.
- [87] M. HANSEN et B. YU. « Model selection and the principle of minimum description length ». In : *J. Amer. Statist. Assoc.* 96.454 (2001), p. 746-774.
- [88] A. HAWKES et D. OAKES. « A cluster process representation of a self-exciting process ». In : *J. Appl. Probability* 11 (1974), p. 493-503.
- [89] G. HINTON et R. SALAKHUTDINOV. « Using Deep Belief Nets to Learn Covariance Kernels for Gaussian Processes ». In : *Advances in neural information processing systems 20 (NIPS'07)*. MIT Press, 2008, p. 1249-1256.
- [90] R. HÖPFNER, J. JACOD et L. LADELLI. « Local asymptotic normality and mixed normality for Markov statistical models ». In : *Probab. Theory Related Fields* 86.1 (1990), p. 105-129.
- [91] P. HUBER et V. STRASSEN. « Minimax tests and the Neyman-Pearson lemma for capacities ». In : *Ann. Statist.* 1 (1973), p. 251-263.
- [92] T. INGLOT. « Inequalities for quantiles of the chi-square distribution ». In : *Probab. Math. Statist.* 30.2 (2010), p. 339-351.
- [93] C. JACKSON et al. « Multistate Markov models for disease progression with classification error ». In : *The Statistician* 52.2 (2003), p. 193-209.
- [94] W. JIANG et C.-H. ZHANG. « General maximum likelihood empirical Bayes estimation of normal means ». In : *Ann. Statist.* 37.4 (2009), p. 1647-1684.
- [95] I. JOHNSTONE et B. SILVERMAN. « Needles and straw in haystacks : empirical Bayes estimates of possibly sparse sequences ». In : *Ann. Statist.* 32.4 (2004), p. 1594-1649.

- [96] M. KEARNS et al. « An experimental and theoretical comparison of model selection methods ». In : *Machine Learning* 27 (1997).
- [97] J. KIEFER et J. WOLFOWITZ. « Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters ». In : *Ann. Math. Statist.* 27 (1956), p. 887-906.
- [98] N. KIEFER, T. VOGELSANG et H. BUNZEL. « Simple robust testing of regression hypotheses ». In : *Econometrica* 68.3 (2000), p. 695-714.
- [99] J. F. C. KINGMAN. *Poisson processes*. The Clarendon Press, 1993.
- [100] Y. KITAMURA. *Empirical Likelihood Methods in Econometrics : Theory and Practice*. Rapp. tech. cowles.yale.edu/sites/default/files/files/pub/d15/d1569.pdf. 2006.
- [101] T. L. LAI et C. Z. WEI. « Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems ». In : *Ann. Statist.* 10.1 (1982), p. 154-166.
- [102] D. LANDERS et L. ROGGE. « Minimal sufficient  $\sigma$ -fields and minimal sufficient statistics. Two counterexamples ». In : *Ann. Math. Statist.* 43 (1972), p. 2045-2049.
- [103] L. M. LE CAM. *Théorie asymptotique de la décision statistique*. Séminaire de Mathématiques Supérieures, No. 33 (Été, 1968). Les Presses de l'Université de Montréal, 1969, p. 140.
- [104] H. LEEB et B. PÖTSCHER. « Sparse estimators and the oracle property, or the return of Hodges' estimator ». In : *J. Econometrics* 142.1 (2008), p. 201-211.
- [105] G. LEUNG et A. R. BARRON. « Information theory and mixing least-squares regressions ». In : *IEEE Trans. Inform. Theory* 52.8 (2006).
- [106] M. N. M. van LIESHOUT. *Markov point processes and their applications*. Imperial College, 2000.
- [107] M. LOIZEAUX et I. MCKEAGUE. « Perfect sampling for posterior landmark distributions with an application to the detection of disease clusters ». In : *Selected Proceedings of the Symposium on Inference for Stochastic Processes (Athens, GA, 2000)*. T. 37. Inst. Math. Statist., 2001, p. 321-331.
- [108] C. MALLOWS. « Some Comments on  $C_p$  ». In : *Technometrics* 15.4 (1973), p. 661-675.
- [109] S. MASE, Y. OGATA et M. TANEMURA. « Statistical analysis of mapped point patterns—present condition of theory and application ». In : *Selected papers on analysis, probability, and statistics*. T. 161. Amer. Math. Soc. Transl. Ser. 2. Amer. Math. Soc., 1994, p. 95-108.
- [110] G. MCLACHLAN et D. PEEL. *Finite mixture models*. Wiley, 2000.
- [111] G. J. MCLACHLAN et T. KRISHNAN. *The EM algorithm and extensions*. Wiley, 2008.
- [112] N. MEINSHAUSEN et P. BÜHLMANN. « High-dimensional graphs and variable selection with the lasso ». In : *Ann. Statist.* 34.3 (2006), p. 1436-1462.
- [113] G. MOHLER et al. « Self-exciting point process modeling of crime ». In : *J. Amer. Statist. Assoc.* 106.493 (2011), p. 100-108.
- [114] R. M. NEAL. « Connectionist learning of belief networks ». In : *Econometrica* 56.71 (1992), p. 71-113.
- [115] W. NEWBY. *Introduction à la théorie des bornes d'efficacité semi-paramétriques*. Annales d'économie et de statistique, 1989.
- [116] J. NEYMAN et E. SCOTT. « Statistical approach to problems of cosmology ». In : *J. Roy. Statist. Soc. Ser. B* 20 (1958), p. 1-43.
- [117] Y. OGATA, K. KATSURA et M. TANEMURA. « Modelling heterogeneous space-time occurrences of earthquakes and its residual analysis ». In : *J. Roy. Statist. Soc. Ser. C* 52.4 (2003), p. 499-509.
- [118] Y. OGATA et D. VERE-JONES. « Inference for earthquake models : a self-correcting model ». In : *Stochastic Process. Appl.* 17.2 (1984), p. 337-347.
- [119] C. ONOF et al. « Rainfall modelling using Poisson-cluster processes : a review of developments ». In : *Stoch. Env. Res. and Risk Ass.* 14 (2000), p. 384-411.
- [120] R. PÉREZ-OCÓN, J. RUIZ-CASTRO et L. M. LUZ GÁMIZ-PÉREZ. « A Multivariate Model to Measure the Effect of Treatments in Survival to Breast Cancer ». In : *Biometrical Journal* 6 (1998), p. 703-715.

- [121] J. W. PRATT. « Concavity of the log likelihood ». In : *J. Amer. Statist. Assoc.* 76.373 (1981), p. 103-106.
- [122] S. PRESS et S. WILSON. « Choosing Between Logistic Regression and Discriminant Analysis ». In : *Journal of the American Statistical Association* 73.364 (1978), p. 699-705.
- [123] J. RASMUSSEN et al. « Continuous time modelling of dynamical spatial lattice data observed at sparsely distributed times ». In : *J. R. Stat. Soc. Ser. B* 69.4 (2007), p. 701-713.
- [124] P. RIGOLLET et A. TSYBAKOV. « Exponential screening and optimal rates of sparse estimation ». In : *Ann. Statist.* 39.2 (2011), p. 731-771.
- [125] J. RISSANEN. « Fisher information and stochastic complexity ». In : *IEEE Trans. Inform. Theory* 42.1 (1996), p. 40-47.
- [126] J. RISSANEN. « Modeling by Shortest Data Description ». In : *Automatica* 14 (1978), p. 465-471.
- [127] R. ROCKAFELLAR. *Convex analysis*. Princeton University Press, 1970.
- [128] E. RONCHETTI. « Robust model selection in regression ». In : *Statist. Probab. Lett.* 3.1 (1985), p. 21-23.
- [129] W. RUDIN. *Functional analysis*. McGraw-Hill, 1991.
- [130] G. SCHWARZ. « Estimating the dimension of a model ». In : *Ann. Statist.* 6.2 (1978), p. 461-464.
- [131] J. SHAO. « Linear model selection by cross-validation ». In : *J. Amer. Statist. Assoc.* 88.422 (1993), p. 486-494.
- [132] R. SHIBATA. « Selection of the order of an autoregressive model by Akaike's information criterion ». In : *Biometrika* 63.1 (1976), p. 117-126.
- [133] S. M. STIGLER. « Studies in the History of Probability and Statistics. XXXII : Laplace, Fisher and the Discovery of the Concept of Sufficiency ». In : *Biometrika* 60.3 (1973), p. 439-45.
- [134] P. STOICA et P. BABU. « On the proper forms of BIC for model order selection ». In : *IEEE Trans. Signal Process.* 60.9 (2012), p. 4956-4961.
- [135] C. STONE. « Adaptive maximum likelihood estimators of a location parameter ». In : *Ann. Statist.* 3 (1975), p. 267-284.
- [136] D. STRAUSS et M. IKEDA. « Pseudolikelihood estimation for social networks ». In : *J. Amer. Statist. Assoc.* 85.409 (1990), p. 204-212.
- [137] L. TRUQUET. *A new smoothed QMLE for AR processes with LARCH errors*. Rapp. tech. hal-00284776. 2008.
- [138] A. WALD. « Note on the consistency of the maximum likelihood estimate ». In : *Ann. Math. Statistics* 20 (1949), p. 595-601.
- [139] C. S. WALLACE et D. M. BOULTON. « An information measure for classification ». In : *Computer Journal* 11.2 (1968), p. 185-194.
- [140] C. WALLACE et P. FREEMAN. « Estimation and inference by compact coding ». In : *J. Roy. Statist. Soc. Ser. B* 49.3 (1987), p. 240-265.
- [141] H. WHITE. « Maximum likelihood estimation of misspecified models ». In : *Econometrica* 50.1 (1982), p. 1-25.
- [142] J. S. WHITE. « The limiting distribution of the serial correlation coefficient in the explosive case ». In : 29 (1958), p. 1188-1197.
- [143] R. WILCOX. *Introduction to robust estimation and hypothesis testing*. Elsevier, 2012.
- [144] R. WOLPERT et K. ICKSTADT. « Poisson/gamma random field models for spatial statistics ». In : *Biometrika* 85.2 (1998), p. 251-267.
- [145] J.-H. XUE et D. M. TITTERINGTON. « On the generative-discriminative tradeoff approach : interpretation, asymptotic efficiency and classification performance ». In : *Comput. Statist. Data Anal.* 54.2 (2010), p. 438-451.
- [146] Y. YANG. « Can the strengths of AIC and BIC be shared ? A conflict between model identification and regression estimation ». In : *Biometrika* 92.4 (2005), p. 937-950.

- [147] Y. ZHANG et Y. YANG. « Cross-validation for selecting a model selection procedure ». In : *J. Econometrics* 187.1 (2015), p. 95-112.
- [148] H. ZOU. « The adaptive lasso and its oracle properties ». In : *J. Amer. Statist. Assoc.* 101.476 (2006), p. 1418-1429.
- [149] H. ZOU, T. HASTIE et R. TIBSHIRANI. « On the “degrees of freedom” of the lasso ». In : *Ann. Statist.* 35.5 (2007), p. 2173-2192.

# Index

- admissible (estimateur), 57
- agrégation, 87
- agrégation bayésienne, 88
- agrégation de modèles linéaires, 89
- AIC, 68
- Akaike, 45
- algorithme EM, 27
- ancillary, *voir* libre
- ARCH, 104
- ARMA, 10
  
- Barankin (borne), 77
- Barankin (estimateur), 62, 77, 85
- Bartlett (correction), 68
- Basu (théorème), 62
- bayésienne (méthode), 78
- biais, 18, 58
- BIC, 68, 85, 127
- branchement (processus), 106
  
- Chapman-Robbins (borne), 77
- complète (statistique), 59
- conditionnellement binomial (modèle), 13
- conjuguées (familles), 82
- consistance, 18
- contraintes (estimation), 48
- contraste, 19, 24, 25
- Cox (processus), 16
- $C_p$ -Mallows, 16, 43, 44
- Cramér-Rao, 72, 77
- critère de sélection, 44, 68, 85, 127
  
- delta-method, 18
- discriminative vs generative (training), 6, 20
  
- efficacité, 75
- EM (algorithme), 27
- empirical Bayes, 82, 83
- entropie, 22
- équation d'estimation, 20, 32
- estimateur
  - bayésien, 78, 85
  - Bayes empirique, 82, 83
  - de Barankin, 62, 77, 85
  - de James-Stein, 60, 84
  - de Pitman, 81
  - de Robbins-Turing, 84
  - des moments, 22, 46
  - OLS, 9, 16, 108
  - par pseudo-vraisemblance, 20, 25, 28, 44
  - ridge, 16, 78, 83, 84
  - robuste, 19, 39
  - sandwich, 67, 71
  - sous contraintes, 48
- exhaustivité, 58
- expérience R-régulière, 65
- expérience régulière, 114
  
- familles exponentielles, 7, 59, 77
- fonction d'estimation, 20, 32
  
- generative vs discriminative (training), 6, 20
- Gibbs (processus), 15
- Godambe (matrice d'information), 34
  
- Hawkes (processus), 12
- Hellinger (distance), 69
- Hodges (estimateur), 75
- hypothèses
  - EC, 49
  - FE, 34
  - GNU, 29
  - GNU', 33
  - LAN, 100, 111, 116
  - ME, 41
  - ULAN, 111, 116
  
- inégalité de Cauchy-Schwarz matricielle, 72
- information
  - matrice, 64
  - monotonie, 71
  
- jackknife, 28
- James-Stein (estimateur), 60, 84
  
- Kronecker (lemme), 109
- Kullback, 63
  
- LAN, 100, 111, 116

LARCH, 44  
 lasso, 17, 61, 83  
 Lehmann-Scheffé, 58, 59  
 libre (statistique), 62, 71, 94  
 loi des grands nombres uniforme, 30  
  
 M-estimateur, 19, 39  
 Mallows, 16, 43, 44  
 MAP, 83  
 martingales, 125  
 maximum d'entropie, 22  
 maximum de vraisemblance, 24, 63, 103  
     avec normalisation inconnue, 17  
     conditionnelle, 20, 25, 28, 71  
     dans les familles exponentielles, 8  
     empirique, 38  
     et exhaustivité, 58  
     mis en défaut, 7, 70, 79, 85  
     par algorithme EM, 27  
     partielle, 20, 25, 27, 28, 71  
 MDL, 68, 127  
 mélange de lois, 8  
 minimale (statistique), 62  
 minimax (estimateur), 57, 86  
 minimum de  $\chi^2$ , 46  
 minimum de contraste, 19, 39  
 MML, 68, 127  
 moindres carrés conditionnels, 9  
 moments (méthode), 22, 46  
  
 Neyman-Fisher, 58  
 Neyman-Scott, 15, 16  
 non-biaisé, 18, 58  
 non-paramétrique, 6  
  
 OLS (estimateur), 9, 16, 108  
 oracle, 87, 89  
 overfitting, 43  
  
 paramétrique, 6  
 pénalisation, 16, 43, 61, 68, 85, 127  
 Pitman (estimateur), 81  
 pivotale (fonction), 96  
 Poisson (processus), 11, 15–17, 26  
 ponctuel (processus), 15, 16, 26, 32  
 pseudo-vraisemblance, 20, 25, 27, 28, 44, 71  
  
 quasi-vraisemblance, 44  
  
 R-régulière (expérience), 65, 71  
 Rao-Blackwell, 58, 60  
 Rao-Blackwellisation, 59, 60  
 rapport de vraisemblance généralisé, 94  
 région de confiance, 34, 68  
 régression, 8, 26, 39, 40, 52, 53  
 régulière (expérience), 65, 71  
 ridge (estimateur), 16, 78, 83, 84  
 risque, 57  
 Robbins (estimateur), 84  
 robuste (estimateur), 19, 39  
  
 sandwich (estimateur), 67, 71  
 sensibilité (matrice), 34, 41, 47  
 SIC, 85, 129  
 stabilisation de variance, 28  
 Stein (estimateur SURE), 16, 61  
 Stein (estimateur), 60, 84  
 Strauss (processus), 15, 17, 22  
 suffisant, *voir* exhaustivité  
 surajustement, 43  
 SURE, 16, 61, 84  
  
 Takeuchi (critère), 44, 45, 53, 68  
 test  
     d'Hausman, 97, 98  
     de Wald, 97, 98  
     des scores, 97, 98  
     du multiplicateur de Lagrange, 98  
     du rapport de vraisemblance, 97  
 TIC, 44, 45, 53, 68  
 totale (statistique), 59  
 Turing, 84  
  
 ULAN, 111, 116  
 UMR, 58  
 UMVU, 59  
  
 validation croisée, 25, 68  
     vs SURE, 84  
 vraisemblance  
     et conditionnement, 71  
     partielle, 20, 25, 27, 28, 71  
  
 Wilks, 52, 67  
  
 Z-estimateur, 20, 32