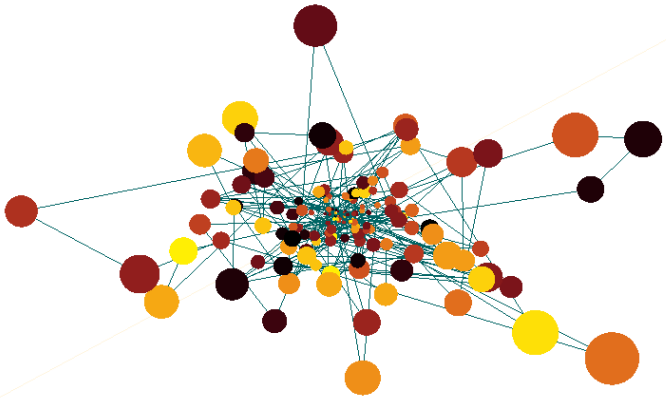
Estimation paramétrique

Cours de Master 2

Bernard Delyon

5 décembre 2024



1. IRMAR, Université Rennes I, Campus de Beaulieu, 35042 Rennes cédex, France.

Table des matières

I Introduction. Définitions **5**

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 . . . 10
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 Premiers résultats de convergence . . . 23
I.3.4 Exercices et compléments . . . 25

II Estimation semi-paramétrique **31**

II.1 Loi des grands nombres uniforme. Score . . . 31
II.2 Fonctions d'estimation (Z-estimateurs) . . . 34
II.2.1 Exemple : la méthode de la variable instrumentale . . . 34
II.2.2 Convergence presque sûre . . . 35
II.2.3 Normalité asymptotique . . . 36
II.2.4 Théorème-limite central pour une fonction d'estimation générale . . . 38
II.2.5 Fusion optimale de fonctions d'estimation . . . 38
II.2.6 Exercices et compléments . . . 40
II.3 Estimateurs à minimum de contraste (M-estimateurs) . . . 41
II.3.1 Exemple : régression non-linéaire . . . 41
II.3.2 Propriétés asymptotiques . . . 42
II.3.3 Surajustement. Débiaisement du contraste . . . 44
II.3.4 Un raffinement dans le cas indépendant . . . 45
II.3.5 Exercices et compléments . . . 47
II.4 Méthode des moments généralisée . . . 49
II.4.1 Un exemple : le minimum de χ^2 . . . 49
II.4.2 Définition et convergence . . . 50
II.4.3 Exercices et compléments . . . 51
II.5 Estimation sous contraintes . . . 52
II.5.1 Exemple : estimation de matrice sous contrainte de rang . . . 52
II.5.2 Convergence . . . 52
II.5.3 Exemple : estimation de matrice sous contrainte de rang. Suite . . . 54
II.5.4 Théorème de Wilks . . . 55
II.5.5 Exercices . . . 56
II.6 Applications aux tests . . . 57

II.6.1 Tests passant par un estimateur sous contrainte. . . . 57
II.6.2 Tests passant par un Z-estimateur. . . . 58
II.7 Bornes de grandes déviations et estimées des moments . . . 59

III Estimation paramétrique **63**
III.1 Comparaison des estimateurs . . . 63
III.1.1 Risque, estimateur admissible et approche minimax. . . . 63
III.1.2 L'approche de Rao-Blackwell-Lehmann-Scheffé . . . 64
III.1.3 Exercices et compléments . . . 66
III.2 L'estimateur au maximum de vraisemblance . . . 69
III.2.1 Convergence presque sûre . . . 69
III.2.2 Normalité asymptotique . . . 70
III.2.3 Bornes exponentielles. Convergence des moments . . . 75
III.2.4 Quelques exemples . . . 76
III.2.5 Exercices et compléments . . . 77
III.3 Borne de Cramér-Rao. Efficacité . . . 78
III.3.1 Borne de Cramér-Rao . . . 78
III.3.2 Le cas biaisé : Inégalité de van Trees . . . 80
III.3.3 Efficacité. . . . 81
III.3.4 Exercices et compléments . . . 83
III.4 La méthode bayésienne . . . 84
III.4.1 Estimateurs bayésiens . . . 84
III.4.2 Le maximum a posteriori (MAP). . . . 90
III.4.3 Exercices et compléments. . . . 90
III.5 Les tests classiques et leur seuil asymptotique . . . 99

IV L'approche martingale **101**
IV.1 Le maximum de vraisemblance en situation générale . . . 101
IV.1.1 Théorie . . . 101
IV.1.2 Exemple : Chaînes de Markov stationnaires . . . 103
IV.1.3 Exercices et compléments . . . 104
IV.2 Processus autorégressif instable . . . 105
IV.3 Régression linéaire . . . 106

A Hypothèse LAN. Théorèmes de Hajek et Le Cam **109**
A.1 Théorème convolution de Hajek . . . 109
A.2 Existence d'un estimateur efficace . . . 112

B Compléments sur les expériences régulières **115**
B.1 Différentiabilité en moyenne quadratique . . . 115
B.2 Liens entre les jeux d'hypothèses . . . 119

C Démonstration du lemme de Kronecker matriciel **123**

D Une inégalité de Sobolev **125**

E Une borne sur les processus empiriques **127**

F Théorèmes-limite pour les martingales **129**

G Les critères MML, MDL, et BIC **131**

I

Introduction. Définitions

Les statisticiens doivent certainement être en train de statistiquer là-dessus. Quelle occasion de savantes controverses !

Rhinocéros, acte III. Ionesco

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 une 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 Θ est une partie de l'espace euclidien. On suppose que ces lois ont une densité par rapport à une mesure positive σ-finie [1] commune $\mu(dy)$, généralement la mesure de Lebesgue ou une

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

mesure de comptage :

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

On dispose d'une observation Y tirée selon P_{θ_*} et l'on cherche à estimer θ_*. La vraisemblance de Y est $p_\theta(Y)$ (cette fonction de θ dépend du choix de μ), et son logarithme sera noté $\mathscr{L}(\theta, Y) = \log(p_\theta(Y))$ [2]. L'estimateur le plus populaire, celui du maximum de vraisemblance, est

$$\widehat{\theta} = \operatorname*{argmax}_{\theta\in\Theta} \mathscr{L}(\theta, Y).$$

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

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

Il se justifie d'une part par le théorème de Neymann Pearson 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 (noter qu'on ne compare que deux valeurs), 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 θ qui conduit à la meilleure prédiction n'est pas forcément le plus vraisemblable pour le modèle considéré, c'est la distinction entre *discriminative training* et *generative training* [162] ; 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 [34] : « 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 $Y = (Y_1, ..Y_n)$ de variables iid gaussiennes. Estimer leur espérance.
2. On observe une suite $Y = (Y_1, ..Y_n)$ de variables iid uniformes sur $[0, a]$, a est inconnu. Estimer leur espérance.
3. On observe une suite $Y = (Y_1, ..Y_n)$ 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 fine, le paramètre est ici la densité commune aux variables.

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

1. On observe une suite $Y = (Y_1, ..Y_n)$ de variables iid. Estimer leur espérance θ.
2. Comme le précédent mais on suppose que la densité est symétrique.
3. On observe une suite $\big((Y_1, X_1), ..(Y_n, X_n)\big)$ 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 θ.

Pour ces problèmes, la loi est caractérisée par une paire (θ, η) où θ est le paramètre d'intérêt, de dimension finie, et η est le paramètre dit de nuisance, de dimension infinie, que l'on ne cherche pas à estimer précisément ; dans l'exemple 3, η 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 η, dans l'exemple 2 ce serait l'estimation de la densité [3]. Nous traiteront dans la suite de problèmes semi-paramétriques mais pas de méthodes semi-paramétriques.

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

3. Ce problème est traité dans [150]. Il existe effectivement un estimateur strictement meilleur que $\widehat{\theta} = n^{-1}\sum Y_i$.

I.1.3 Exercices

Exercice 1 (Données censurées). Soit $(X_i)_{1\leqslant i\leqslant n}$ une suite de variables iid de loi $\mathcal{E}(\theta^{-1})$ (exponentielles d'espérance θ). 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 μ et p_θ, 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 $\widehat{\theta}_n$ (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 $\widehat{\theta}_n$ converge vers θ_*.

Exercice 3 (Un cas un peu spécial). Soit la chaîne de Markov $(Y_i)_{i\geqslant 0}$, où Y_{i+1} vaut Y_i avec probabilité θ 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 σ^{-1}). On s'intéressera particulièrement au cas $n = p$. Le paramètre est ici $\theta = (\sigma, \mu_1, \dots \mu_n)$.

Calculer $\widehat{\sigma}$, l'estimateur au maximum de vraisemblance de σ. 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. 26 et 9 p. 92. L'exercice 12 p. 93 en suggère un dernier bien plus sophistiqué (les μ_k d'ici sont les α_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 $\widehat{\theta}_n = \max Y_i$.
2. Calculer $P(\widehat{\theta}_n > t)$ puis $P(n\theta_*^{-1}(\theta_* - \widehat{\theta}_n) > t)$ pour $t > 0$ et montrer que

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

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

   où $\mathscr{E}$ est une variable exponentielle d'espérance 1.
3. Déduire de (I.1) que l'estimateur $\widehat{\theta}'_n = (n+1)\widehat{\theta}_n/n$ satisfait : $E\big[n^2(\theta_* - \widehat{\theta}'_n)^2\big] \longrightarrow \theta_*^2$.

$\widehat{\theta}'_n$ est donc, en terme d'erreur quadratique moyenne, asymptotiquement deux fois meilleur que l'estimateur au maximum de vraisemblance. L'exercice 2 p. 90 traitera de l'estimation bayésienne de θ_*.

Exercice 6 (Le MV dépend de la version des densités [4]). Soit g la densité gaussienne centrée réduite. On considère $f(x) = g(x)1_{x\neq 1} + 10 \times 1_{x=1}$ puis la famille paramétrique de densités $f(x-\theta)$. C'est donc essentiellement le modèle de translation gaussien. Que vaut le maximum de vraisemblance basé sur une réalisation X_1 ?

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 μ sur $\mathbb{R}^d$, mesure non portée par un hyperplan, est définie par

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

4. Exemple proposé par Arnaud Guyader.

où Z se déduit du reste par la condition d'intégrale 1. L'ensemble Θ où varie θ est :

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

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

La famille est dite **régulière** si Θ est ouvert. La condition d'hyperplan sur μ implique que pour deux θ différents, les mesures P_θ sont différentes. On vérifie facilement les propriétés (du moins si θ est intérieur à Θ) :

$$\nabla Z(\theta) = E_\theta[Y]$$
$$\nabla^2 Z(\theta) = E_\theta[YY^T] - E_\theta[Y]E_\theta[Y]^T = Cov_\theta(Y).$$

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

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

$$\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \nabla Z(\widehat{\theta}). \tag{I.5}$$

Noter que $\widehat{\theta}$ est également un estimateur par méthode de moment puisqu'il fait coïncider la moyenne empirique avec son espérance sous $P_{\widehat{\theta}}$. Cet estimateur peut être difficile à obtenir si l'on ne possède pas d'expression simple pour la fonction Z (exercices 4 p. 17, 7 p. 48).

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 Θ 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. La famille $\mathcal{N}(\theta, 1)$ l'est.

Les modèles présentés dans la suite seront souvent issus de familles exponentielles (p. ex. le modèle autorégressif, le processus de Strauss). 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 $\nabla Z(\widehat{\theta})$ puisque le théorème de Lehmann-Schéffé s'y applique (théorème 23, exercice 13 p. 67) et que la borne de Cramér-Rao est atteinte (exercice 3 p. 83).

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[5].

Si le nombre P de populations est connu, les paramètres sont les (q_j, μ_j, σ_j) ; sinon, on sort du cadre paramétrique.

L'étude du cas simple du n-échantillon d'un mélange de deux gaussiennes d'espérance θ_1 et θ_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

5. On trouvera de nombreux exemples dans [122].

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. 29).

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

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

Les X_i sont considérés comme déterministes et la fonction $n_\theta(x)$ est connue (comme fonction de θ et x) ; θ 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*

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

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

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

X_i et θ 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, etc. Ou Y est le fait qu'une femme travaille et X son nombre d'enfants, le salaire de son mari, etc. La vraisemblance est ici immédiate à écrire ; son expression est compliquée mais son logarithme est une fonction concave de θ (exercice 11 p. 27), ce qui facilite sa maximisation.

L'estimateur le plus classique de θ 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 :

$$\widehat{\theta}_n = \operatorname*{argmin}_\theta \sum \big(Y_i - n_\theta(X_i)\big)^2. \tag{I.9}$$

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

$$\widehat{\theta} = (X^T X)^{-1} X^T Y \tag{I.10}$$

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 $n \to \infty$ une approche commode est de supposer que la suite (X_i) est aléatoire stationnaire (**random design**, par opposition à **fixed design**), considérant que n_θ paramètre l'espérance de Y sachant X, $E[Y_i|X_i] = n_\theta(X_i)$, ce qui revient à reformuler le modèle ainsi :

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

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

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

c.-à-d. le θ 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 θ lui-même (l'objet tomographié), soit d'obtenir un estimateur qui conduise à un bon prédicteur $\widehat{Y}_i = n_{\widehat{\theta}}(X_i)$ sur de nouvelles données. Noter que le modèle aléatoire pour X ne tient pas dans l'exemple de la tomographie ; il est en revanche adéquat pour les problèmes de prédiction car il prend en compte le fait que de nouveaux échantillons n'auront pas un régresseur déjà observé, ce que l'on peut voir comme une propriété d'extrapolation.

I.2.2 Modèles de séries temporelles

Ces modèles sont généralement markoviens, et θ 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 décrit par la formule suivante :

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

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

Partant de conditions initiales, pour l'instant déterministes, pour Y et ε, 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 ε_n par récurrence à l'aide du passé de Y_n par la formule :

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

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

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

Si l'on connaît les conditions initiales $(Y_1, \dots Y_p)$ et $(\varepsilon_p, \dots \varepsilon_{p+1-q})$, alors tous les ε_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

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

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 θ est rendue très compliquée au travers du calcul des $\widehat{\varepsilon}_i$, sans compter le problème des conditions initiales inconnues en pratique pour les ε_k ; dans le cas autorégressif cette complication disparaît et l'estimation est relativement facile.

Si l'on note

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

on a

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

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

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

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 [6]) par les équations de récurrence suivantes :

$$\Lambda_n = a_0 + \sum_{k=1}^{p} a_k Y_{n-k} + \sum_{k=1}^{q} b_k \Lambda_{n-k}$$
$$Loi\big(Y_n | Y_1, \dots Y_{n-1}, \Lambda_1, \dots \Lambda_n\big) = \mathcal{P}(\Lambda_n).$$

On pensera d'abord au cas simple $p = 1, q = 0$. Les coefficients doivent être positifs pour garantir la positivité de Λ_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 λ, 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 $a \leqslant t_1 \leqslant \dots \leqslant t_n \leqslant b$).

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 $[0, A]$, $[A, B]$, etc.

Vraisemblance. Soit $\tau > 0$ fixé, $E_n = \{(t_1, \dots t_n) : 0 \leqslant t_1 \dots \leqslant t_n \leqslant \tau\}$ 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

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

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 $N_t = N^0_{\Lambda_0^t}$ où et 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))

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

(on a omis un terme $1/N!$ car il ne dépend pas de θ) [7]. Notons que l'on a la formule classique suivante qui montre que λ_t s'interprète comme une mesure de la probabilité d'observer un évènement dans un

6. Pour un modèle plus complet, voir les équations (1) à (4) de [40]. 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, $\sum_i p_i \delta_{\tau_i}$, $p_i \leqslant 1$. Interprétation : la simulation doit alors être complétée en ajoutant indépendamment un événement à chaque instant τ_i avec probabilité p_i. Un terme $\prod_i p_i^{\varepsilon_i}(1-p_i)^{1-\varepsilon_i}$, où ε_i vaut 1 si un évènement est observé en τ_i et 0 sinon, doit être ajouté à la vraisemblance. Voir [45].

avenir très proche :

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

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

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 λ_t peut dépendre du passé de la trajectoire (i. e. être une variable $\mathscr{F}_t$-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 [44]

$$\Big(\prod_{T_i\leqslant b} \lambda_{T_i}\Big) e^{-\int_0^b (1-\lambda_t)dt}.$$

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

Un exemple typique est un λ de la forme $\lambda_t = \mu(t) + \sum_{T_i\leqslant t} \gamma(t-T_i)$. Ce sont des processus de Hawkes [95] [8]. Des modèles de ce type sont utilisés pour les tremblements de terre (secousse initiale puis répliques [130, 60]), avec par exemple $\gamma(x) = a(x+b)^{-p}$, $\mu = cste$. Ou encore $\lambda_t = \psi(\theta t - N_t)$ où ψ est une fonction croissante > 0 [131] ; on trouvera dans [136] une application concernant les parasites du pin, et dans [126] 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 λ 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 λ_{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 :

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

La log-vraisemblance d'une trajectoire $(Y_t)_{0\leqslant t\leqslant T}$ 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

$$\frac{A_{e_1e_2}}{-A_{e_1e_1}} \cdots \frac{A_{e_{n-1}e_n}}{-A_{e_{n-1}e_{n-1}}} \big(-A_{e_1e_1} e^{A_{e_1e_1}t_1}\big) \dots \big(-A_{e_{n-1}e_{n-1}} e^{A_{e_{n-1}e_{n-1}}t_{n-1}}\big)\, e^{A_{e_ne_n}t_n}$$

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 μ. Partant de chaque évènement T_i de première génération, simuler ses enfants comme un processus de Poisson d'intensité $\gamma(t-T_i)$ (ce sont donc des évènements postérieurs à T_i). Simuler ensuite la troisième génération de manière analogue en partant de chaque évènement 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.

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

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

où N_{ij} est le nombre de fois que la transition $i \to j$ 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'attentes ; 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, $\mathcal{S} = \{0,1,2,3\}$. 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 [133, 100]. 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 :

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

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 [9]. 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	$(S,I,R) \to (S, I-1, R+1)$	$\lambda_0 I$
Infection	$(S,I,R) \to (S-1, I+1, R)$	$\lambda_1 SI/n$
Sensibilisation	$(S,I,R) \to (S+1, I, R-1)$	$\lambda_2 R.$

La première transition représente une guérison (I→R), le taux est proportionnel à I car on suppose que chaque individu infecté guérit indépendamment du reste en un temps exponentiel de paramètre λ_0. Le risque d'infection pour un individu sain étant proportionnel à la proportion d'individus infectés, le taux de la deuxième transition (S→I) (infection) est choisi proportionnel à SI/n. Si sur des données $(S(t), I(t), R(t))_{t\leqslant T}$ on observe N_0, N_1, N_2 transitions de chaque type, la log-vraisemblance est, à un terme indépendant des paramètres près

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

I.2.3 Modèles spatiaux

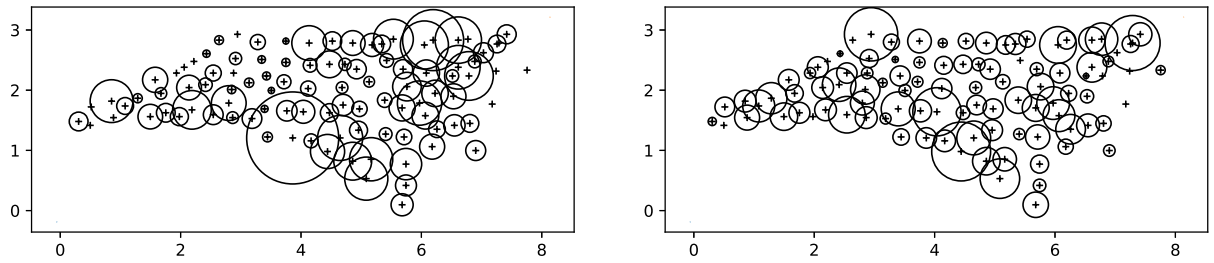
On ne parlera pas ici des champs gaussiens qui sont les modèles les plus utilisés : il s'agit de vecteurs gaussiens particuliers. S'ils sont classiquement utilisés pour modéliser les données spatiales, les réseaux

9. Pour un traité général sur le sujet, recommandons [64].

bayésiens également en sont des cas particuliers.

On trouvera des compléments à cette section dans [81, 51] et dans un article célèbre de Julian Besag [35].

Modèle conditionnellement binomial. Considérons les données de mort subite de nouveau-né dans les 100 comtés de Caroline du Nord [59].



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. Une corrélation spatiale semble apparaître, elle serait dûe à l'existence d'une variable explicative spatialement corrélée (structure de la population).

On considère le modèle suivant (les n_a sont déterministes)

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

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

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

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 :

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

où $d(a,b)$ est la distance entre les comtés a et b. Le modèle n'a plus pour paramètres que α, γ, δ et q. On peut également faire dépendre α 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 α 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 $\theta = (\alpha, \beta, \gamma, q)$, à cause de la dépendance de Z en θ (la méthode proposée à l'exercice 4 p. 17 pour contourner ce problème ne fonctionne ici pas très bien). On choisit plutôt de maximiser la pseudo-vraisemblance, une des méthodes classiques que l'on verra au § I.3.2.

On obtient alors les estimées $\widehat{\delta} = 0,64$ (ce qui fait une moyenne de 7 termes non nuls par ligne de la matrice c), $\widehat{\alpha} = -6.1$, $\widehat{\gamma} = 0.01165$, $\widehat{q} = 0.00188$ et $\widehat{\beta} = -1.64\ 10^{-5}$. 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, $e^{\widehat{\alpha}}/(1 + e^{\widehat{\alpha}}) = 0.0022$, est proche de $\widehat{q}$ ce qui confirme l'interprétation du terme qn_a comme un recentrage.

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 [161], de maladies dans des régions [65, 66] ; voir aussi [3] pour des exemples d'études en biologie. Soit E (généralement une partie compacte de $\mathbb{R}^d$) un espace muni d'une mesure finie μ sans atome : une réalisation du **processus de Poisson spatial** de mesure d'intensité μ 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 $\mu(.)/\mu(E)$ (l'absence d'atome implique que les X_i sont différents). Donc pour toute fonction φ définie sur $\cup_{n\geqslant 0} E^n$ symétrique de ses arguments (fonction d'ensemble), borélienne [10] et bornée

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

Le nombre de points présents dans une partie mesurable D suit une loi de Poisson de paramètre $\mu(D)$ [107]. 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 $\mathcal{P}\big(q_\theta(y)\mu(dy)\big)$ (q_θ est une fonction positive avec $\mu(q_\theta) < \infty$)

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

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 θ 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_θ par rapport au processus de Poisson $\mathcal{P}(\mu)$ lui-même. Par exemple le processus de **Strauss** défini par

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

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 β 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. 28.

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

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

où Φ est une certaine fonction [121].

Une autre famille importante est formée des modèles de **Neyman-Scott**, appelés encore **cluster point process**, originellement utilisés pour modéliser la distribution des galaxies [129] : 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 μ_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 [11]).

10. Au sens où $\varphi(x)1_{|x|=n}$ coïncide avec une fonction borélienne sur $\mathbb{R}^d$, voir p. ex. [118].

11. Par exemple dans [119] : $Y = (Y_1, \dots Y_p)$ 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[12], 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 μ_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 [3].

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 λ 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 [132] ; 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[13] de Stein). Soit un vecteur d'observations $Y = (Y_1, \dots Y_n)$, de variance $\sigma^2 Id$, les paramètres $\mu_i = E[Y_i]$, et un estimateur $\widehat{\mu}_i = \widehat{\mu}_i(Y)$. On peut penser au cas de la régression, linéaire ou non. On note le risque instantané et le risque moyen

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

1. Montrer que

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

2. On se place dans le cas d'un estimateur linéaire, i. e. $\widehat{\mu}(y) = \mu_0 + PY$, $P \in \mathbb{R}^{n\times n}$ (p. ex. l'estimateur OLS (I.10), ridge (III.44)).
   (a) Déduire de ce qui précède que l'estimateur suivant de $\overline{R}$ est non biaisé :

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

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

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

12. 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.

13. Stein's Unbiased Risk Estimate. Nous nous inspirons d'une approche proposée par Efron [72].

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

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

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

L'intention de Mallows [120] était de mettre en place un critère permettant de comparer les performances de modèles linéaires ayant moins de variables explicatives ; σ est traditionnellement calculé sur la base du modèle le plus compliqué. Nous approfondirons p. 44. Une application de (I.25) à l'estimateur lasso, éq. (III.41), se trouve à l'exercice 11 p. 67, et une autre en régression à l'exercice 14 p. 96.

On pourrait voir (I.25) comme résultant de (I.23) et d'une linéarisation, assez abusive, de $y \mapsto \widehat{\mu}(y)$ dans $Cov(\widehat{\mu}_i, Y_i) = E[(\widehat{\mu}_i(Y) - \widehat{\mu}_i(\mu))(Y_i - \mu_i)]$. Ceci, ou plus simplement (I.24), rend défendable son utilisation dans des cas non gaussiens proches de la linéarité.

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

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

Exercice 4 (Vraisemblance non normalisée). On dispose d'une observation y d'une variable Y de densité (par rapport à une mesure fixe $\mu(dy)$) de la forme $p_\theta(y) = e^{Q(y,\theta)}/Z(\theta)$ 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. Le but est de réaliser le maximum de vraisemblance sans calculer $Z(\theta)$. Un alternative sera proposée à l'exercice 7 p. 48.

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

$$\nabla_\theta Q(y, \theta) - E_\theta\big[\nabla_\theta Q(Y, \theta)\big].$$

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

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

où E_ψ est la loi associée à ψ. Noter que cette équation permet de démontrer le premier point.

La première équation permet d'estimer θ 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 θ_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 ψ fixe, proche de la limite (i. e. $\psi = \theta_n$ pour un n grand). Cette méthode est très utilisée [96, 127] 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 est observé. Montrer que le gradient de la log-vraisemblance de y_V vaut

$$E_\theta\big[\nabla_\theta Q(Y, \theta) | Y_V = y_V\big] - E_\theta\big[\nabla_\theta Q(Y, \theta)\big].$$

La première espérance s'obtiendra typiquement par simulation de réalisations de Y_H conditionnellement à $Y_V = y_V$. C'est un algorithme proposé pour les modèles graphiques (équation (17.37) de [9] ; 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 λ_a et sont servis à leur tour avec un temps de service exponentiel de paramètre λ_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

Risque. C'est l'espérance d'un certaine fonction de l'estimateur $\widehat{\theta}$ que l'on cherche à minimiser, typiquement $R(\widehat{\theta}) = E[|\widehat{\theta} - \theta_*|^2]$

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

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 $\widehat{\theta}_n$ vers une limite θ_* désirée, puis la normalité asymptotique de $\sqrt{n}(\widehat{\theta}_n - \theta_*)$. 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_θ est la loi uniforme sur $[0, \theta]$).

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

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

Biais. Le biais d'un estimateur $\widehat{\theta}$ de θ_* est $b(\widehat{\theta}) = E[\widehat{\theta}] - \theta_*$, et s'il est nul $\widehat{\theta}$ est dit non-biaisé[14]. Les estimateurs non-biaisés sont rares en pratique. Ce concept n'est pas invariant par changement de variables, car si $\widehat{\theta}$ est non-biaisé, $f(\widehat{\theta})$ sera généralement un estimateur biaisé de $f(\theta_*)$ (exercice 3 p. 26).

Variance asymptotique. Si l'on a une limite en loi du type $\sqrt{n}(\widehat{\theta}_n - \theta_*) \to \mathcal{N}(0, V)$, on dit que V est la variance asymptotique de l'estimateur. On est souvent capable de calculer V, contrairement à la variance de chaque $\widehat{\theta}_n$. Ceci permet d'avoir une évaluation de l'erreur d'estimation, et surtout de comparer différents estimateurs (en comparant leur variance asymptotique[15]).

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 $g(\widehat{\theta}_n) \simeq g(\theta_*) + \nabla g(\theta_*)(\widehat{\theta}_n - \theta_*)$:

1 - Théorème *(Delta-method)*

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

$$r_n(\widehat{\theta}_n - \theta_*) \longrightarrow X \qquad \text{en loi}$$

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

$$r_n(g(\widehat{\theta}_n) - g(\theta_*)) \longrightarrow \nabla g(\theta_*) X \qquad \text{en loi.}$$

Démonstration. On a

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

14. 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.

15. Ceci a été mis en pratique pour la première fois par Laplace en 1818 pour estimer θ_* dans le modèle de régression $y_i = \theta_* x_i + u_i$, les u_i étant iid centrés de densité symétrique φ. Il compare $\widehat{\theta}_{LS} = \sum_i y_i x_i / \sum x_i^2$ (moindres carrés) à $\widehat{\theta}_{MS} = \arg\min \sum_i |y_i - \theta x_i|$. Il trouve que $\widehat{\theta}_{LS} - \theta_*$ est proche d'une gaussienne de variance $\int u^2 \varphi(u) du / \sum x_i^2$ tandis que $\widehat{\theta}_{MS} - \theta_*$ est proche d'une gaussienne de variance $1/(4\varphi(0)^2 \sum x_i^2)$. La méthode à choisir dépend donc du résultat de la comparaison entre $1/(4\varphi(0)^2)$ et $\int u^2 \varphi(u) du$, indépendamment des x_i. On notera que le cas particulier $x_i \equiv 1$, comparaison de le moyenne et de la médiane, est déjà intéressant. Ce point d'histoire est relaté par Stephen Stigler [148].

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

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

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). ∎

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

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

alors

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

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 $\mathcal{N}(0, Cov_{\theta_*}(Y)^{-1})$.

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 pistes classiquement utilisées pour fabriquer des estimateurs, à l'exception de la méthode bayésienne que nous verrons plus tard. 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 :

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

K est appelé fonction de perte (loss function), ou plus rarement contraste. On suppose que θ_* minimise le contraste moyen :

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

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 qui satisfont

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

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

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

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 ε_i ont une distribution symétrique par rapport à 0. On a les deux estimateurs

$$\widehat{\theta} = \arg\min_\theta \sum_{i=1}^n (Y_i - \theta)^2 = \frac{1}{n} \sum_i Y_i$$
$$\widehat{\theta}_R = \arg\min_\theta \sum_{i=1}^n \psi(Y_i - \theta) \tag{I.30}$$

où $\psi(x) = x^2 1_{|x|\leqslant\alpha} + (2\alpha|x| - \alpha^2)1_{|x|>\alpha}$ pour un certain α. $\widehat{\theta}_R$ 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 [16].

La vraisemblance partielle consiste à remplacer la vraisemblance par un produit de marginales extraites des données ; il est toutefois rare que les marginales aient une expression simple, un exemple est donné à l'exercice 7 p. 26.

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

$$\widehat{\theta} = \operatorname*{argmax}_\theta \prod_{a\in A} P_\theta\big(Y_a | Y_b, b \neq a\big) \tag{I.31}$$

ce qui peut conduire, comme on va le voir, à des formules beaucoup plus simples à traiter. Elle est couramment utilisée en fiabilité ([77] p.10) et en processus spatiaux ([35] § 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) :

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

où $c_{ab} = \gamma 1_{d(a,b)<\delta}$ et $\theta = (\alpha, \beta, \gamma, \delta, q)$. Comme Z dépend de θ, la vraisemblance est difficile à calculer. Strauss et Ikeda [151] [17] 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

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

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

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

et les paramètres sont $(a, \beta, \gamma, -\gamma q)$ (tout du moins si δ 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 δ a été estimé par essais successifs de sorte à maximiser la pseudo-vraisemblance obtenue à δ fixé.

Pour la convergence, voir l'exercice 14 p. 29. D'autres exemples se trouvent aux exercices 7 p. 26, 10 p. 27, 12 p. 28, et également 5 p. 48 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 [18].

16. Pour la terminologie, voir par exemple [56]. Chaque terme est un cas particulier du précédent.
17. Ils travaillent sur les données *monastère* présentées à l'exercice 10 p. 27.
18. Pour des compléments concernant les approches génératives et discriminatives, on peut consulter [162]. Pour la comparaison directe de l'analyse discriminante et de la régression logistique, voir [135].

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

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

ou plus généralement

$$H_n(\widehat{\theta}_n, \omega) \longrightarrow 0 \tag{I.33}$$

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

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

On suppose que la limite existe en probabilité.

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

$$y_i = \alpha_* + \beta_* a_i + e_i. \tag{I.35}$$

où e_i est indépendant de a_i. La méthode des moindres carrés qui consiste à minimiser en α et β

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

conduit à (I.32) avec

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

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, β_*, 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 α_* et β_* peut néanmoins se faire par l'utilisation d'une variable explicative décorrélée de c_i et de e_i, par exemple le salaire des parents s_i. En effet on voit que la paire d'équations

$$E[y_i] = \alpha_* + \beta_* E[a_i]$$
$$E[y_i s_i] = \alpha_* E[s_i] + \beta_* E[s_i a_i]$$

permet d'estimer α_* et β_* 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 :

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

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.

Exemple : Processus de Strauss. Pour ce processus, cf. p. 15, la vraisemblance est difficile à calculer. On montre (exercice 12 p. 28) que la fonction suivante est une fonction d'estimation pour la paire (β, γ) :

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

Méthode des moments. Elle consiste à exprimer θ_* comme fonction de statistiques simples des données (espérance, variance, corrélations...)

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

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

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

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 σ^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

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

où q est le rang de Q, d'où $\widehat{\sigma}^2 = \frac{1}{q}\|Qy\|^2$ (généralement, X est de rang plein et $q = n - p$ où p est le nombre de variables). 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. Comme $q\widehat{\sigma}^2/\sigma^2$ est un χ^2_q, cet estimateur reste bon pour des valeurs de p assez grandes (i. e. si beaucoup de variables sont prises en compte).

Méthode des moments généralisée. Dans le cas paramétrique iid avec observations scalaires et $\Theta \subset \mathbb{R}^d$, on peut par exemple chercher $\widehat{\theta}_n$ 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, θ 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,

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

où $K \geqslant d$ 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 μ^n pour $Y = (Y_1, \dots Y_n)$ qui soit d'entropie maximale sous la contrainte que certains de ses moments coïncident avec les moments empiriques et

éventuellement sous d'autres contraintes (indépendance des Y_i sous π...). Rappelons que l'entropie de π 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 σ^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 Premiers résultats de convergence

Des hypothèses adéquates sur les fonctions $K_n(\theta, \omega)$ et $H_n(\theta, \omega)$ vont permettre d'obtenir la convergence de $\widehat{\theta}_n$ vers θ_*. 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 espace métrique compact Θ 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 θ_n une suite de points de Θ :

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

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 θ_n vient de ce que dans le cas contraire on peut extraire de (θ_n) une suite convergente ayant une limite différente de θ_*, 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 Θ est compact et si $E[d(Y_1, \theta)]$ a un minimum unique $\theta_* \in \Theta$. Si seulement les parties fermées bornées de Θ sont compactes, on peut se ramener au cas compact en notant que la suite θ_n est p.s. bornée car

$$d(0, \theta_n) \leqslant K_n(\theta_n) + K_n(0) \leqslant 2K_n(0)$$

(le compact dans lequel on se place dépend donc de ω, mais peu importe).

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

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 θ). Le théorème suivant donne deux résultats simples permettant de passer de la convergence ponctuelle à la convergence uniforme de fonctions [19] ; 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.

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 $A_\theta = \{\omega : H_n(\theta) \to h(\theta)\}$ et $P(A_\theta) = 1$ pour tout θ, 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. 27).

Un résultat pour la convergence en probabilité. Les théorèmes précédents sont bien adaptés à la convergence presque sûre. Une version plus faible existe pour la convergence en probabilité :

4 - Théorème

Soit Θ un espace métrique compact et $(\theta, \omega) \mapsto H_n(\theta, \omega)$ (resp. $K_n(\theta, \omega)$) une suite de fonctions mesurables définies sur $\Theta \times \Omega$ à valeurs dans un espace vectoriel normé (resp. dans $\mathbb{R}$). On suppose que pour une fonction **continue** $\theta \mapsto h(\theta)$ (resp. $\theta \mapsto k(\theta)$).

$$\sup_{\theta \in \Theta} \|H_n(\theta, \omega) - h(\theta)\| \xrightarrow{P} 0 \qquad \Big(\text{resp.} \quad \sup_{\theta \in \Theta} \|K_n(\theta, \omega) - k(\theta)\| \xrightarrow{P} 0\Big). \tag{I.37}$$

La mesurabilité des sup fait partie des hypothèses.
Soit $\theta_n = \theta_n(\omega)$ une suite de v.a. à valeurs dans Θ.

- Si $H_n(\theta_n, \omega) \xrightarrow{P} 0$ et si l'équation $h(\theta) = 0$ admet une solution unique θ_*, alors $\theta_n \xrightarrow{P} \theta_*$.
- Si $K_n(\theta_n) - \min_\theta K_n(\theta) \xrightarrow{P} 0$ et si le minimum de k est atteint en un unique point θ_*, alors $\theta_n \xrightarrow{P} \theta_*$.

Démonstration. Dans le cas du processus H_n, les hypothèses entraînent que

$$\|h(\theta_n)\| \leqslant \|H_n(\theta_n, \omega)\| + \sup_{\theta \in \Theta} \|H_n(\theta, \omega) - h(\theta)\| \xrightarrow{P} 0.$$

La continuité de h et l'unicité de la solution de $\{h = 0\}$ font que pour tout $\varepsilon > 0$, il existe $\eta(\varepsilon) > 0$ tel que $|h(\theta)| < \eta(\varepsilon) \Longrightarrow d(\theta, \theta_*) \leqslant \varepsilon$, sinon on trouverait une suite x_n telles que $h(x_n) \to 0$ et qui convergerait vers un point différent de θ_* (compacité), annulant h. Donc ayant déjà montré que pour tout $\varepsilon > 0$, $P(\|h(\theta_n)\| > \eta(\varepsilon)) \to 0$, on obtient $P(d(\theta_n, \theta_*) > \varepsilon) \to 0$.

Le raisonnement est similaire avec K_n car les hypothèses impliquent que $k(\theta_n) - \min_\theta k(\theta) \xrightarrow{P} 0$. ∎

19. Le premier est classique. Pour le second, voir [141], Th.I.8.

La condition (I.37) peut être vérifiée à l'aide d'une inégalité de Sobolev, c'est l'objet du résultat suivant (un résultat de la normalité asymptotique du même esprit sera donné au th. 12 p. 38) :

5 - Théorème

Soit Θ une partie bornée de $\mathbb{R}^d$ dont l'intérieur satisfait la propriété de cône (p. ex. convexe bornée, cf. appendice D). Soit $(\theta,\omega)\mapsto H_n(\theta,\omega)$ une suite de fonctions mesurables sur $\Theta\times\Omega$ à valeurs dans $\mathbb{R}^k$. On suppose que pour un $q>d$ et pour une fonction h continue sur Θ, la suite de fonctions

$$Q_n(\theta)=E\big[\|\nabla H_n(\theta,\omega)-\nabla h(\theta)\|^q\big]+E\big[\|H_n(\theta,\omega)-h(\theta)\|^q\big] \tag{I.38}$$

satisfait

$$\sup_{\theta\in\Theta}\sup_n Q_n(\theta)<+\infty \tag{I.39}$$

$$\forall\theta\in\Theta\,,\ Q_n(\theta)\longrightarrow 0. \tag{I.40}$$

Alors

$$E\Big[\sup_{\theta\in\Theta}\|H_n(\theta,\omega)-h(\theta)\|^q\Big]\to 0. \tag{I.41}$$

En particulier (I.37) est satisfait.

Démonstration. L'inégalité de Sobolev présentée en appendice, éq. (D.1) p.125 implique

$$\sup_{\theta\in\Theta}\|H_n(\theta,\omega)-h(\theta)\|^q\leqslant C\int_\Theta\Big(\|\nabla H_n(\theta,\omega)-h(\theta)\|^q+\|H_n(\theta,\omega)-h(\theta)\|^q\Big)d\theta$$

pour une constante C qui ne dépend que de Θ et de q. Il ne reste plus qu'à prendre l'espérance et appliquer le théorème de Fubini-Tonelli puis le théorème de convergence dominée. ∎

Rappelons que pour une suite de v.a. indépendantes X_i et tout $p>1$, on a l'inégalité de Rosenthal

$$E\Big[\Big(\sum_{i=1}^n X_i\Big)^p\Big]\leqslant(18qp^{1/2})^p\Big\{\Big(\sum_{i=1}^n E\big[X_i^2\big]\Big)^{p/2}+\sum_{i=1}^n E\big[|X_i|^p\big]\Big\},$$

inégalité qui s'adapte aux martingales (cf. [8] Th. 2.12). Pour un estimateur du type (I.32), sous des hypothèses adéquates, ceci conduira directement à un terme d'ordre $n^{-p/2}$ pour $Q_n(\theta)$.

Pour le cas où Θ n'est pas borné, mais que les autres hypothèses sont vérifiées dans les théorèmes 4 et 5, une stratégie pour obtenir la convergence est de se placer sur $\Theta_k=\Theta\cap\{x:\|x\|\leqslant k\}$, avec $\theta'_n=\theta_n 1_{\theta_n\in\Theta_k}+\theta_* 1_{\theta_n\notin\Theta_k}$, pour $k>\|\theta_*\|$. Pour tout k la suite θ'_n ainsi construite converge vers θ_*, et si l'on a montré que la suite θ_n est presque sûrement bornée, ceci prouve la convergence de θ_n vers θ_*.

En raison du caractère un peu restrictif des hypothèses de ces théorèmes, on présentera dans la suite le théorème 6 qui est spécifiquement adapté aux suites de fonctions aléatoires qui sont des moyennes empiriques, et exige moins de régularité.

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 λ et pour μ basé sur un n-échantillon $(X_i,Y_i)_{1\leqslant i\leqslant n}$. Calculer $E[(Y-\mu)^2(X-\lambda^{-1})]$. En déduire un estimateur de α. A quel type d'estimateur

a-t-on affaire ? On pourra comparer, en termes de complexité, à l'estimation de α au maximum de vraisemblance (à μ et λ connus).

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

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

1. On suppose $Y_i \sim \mathcal{N}(\mu, \sigma^2)$.
   (a) Quelle est l'estimateur au maximum de vraisemblance $\widehat{\theta}$ de $\theta = (\mu, \sigma^2)$? Que vaut $p_{\widehat{\theta}}(Y)$?
   (b) Soit X observation unique de loi $\Gamma(p, \beta)$, i.e. de densité $(x/\beta)^p e^{-x/\beta} x^{-1}/\Gamma(p)$. Quel est l'estimateur au maximum de vraisemblance de β lorsque p est connu ?
   (c) On sait que S est σ^2 fois un χ^2_{n-1}, c.-à-d. qu'il suit la loi $\Gamma(p, \beta)$ avec $\beta = 2\sigma^2$, $p = (n-1)/2$. En déduire un autre estimateur au maximum de vraisemblance de σ^2, basé 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 σ 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 $\Gamma(n, 1/\lambda)$.

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

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

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 θ_*. Vérifier que θ_* minimise bien k, ce qui justifie l'utilisation de ce contraste.
   (b) Quelle sera la limite de la suite $\widehat{\theta}_n$ si la loi de Y_i a une densité q qui n'appartient pas au modèle ? On demande juste 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 $u \geqslant 0$. Le meilleur u est calculé à l'exercice 4 p. 47.
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 θ_* ?

Exercice 7 (Pseudo-vraisemblance pour un modèle de v.a. de Bernoulli corrélées [58]). Soit $X \sim \mathcal{N}(0, R)$, $R_{ii} = 1$, $R_{ij} = \theta$, $i \neq j$. On observe $Y_i = 1_{X_i > 0}$ et l'on cherche à estimer θ. On pourra admettre que pour $i \neq j$, si θ_* 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

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

et donner l'équation dont $\widehat{\theta}$ est solution. θ_* 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)). Le paramètre θ est a_1 de (I.12). Par simplicité, on suppose σ connu, ce qui en fait ne change rien à la suite. 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) $\widehat{\theta}_1$. En déduire également la loi de Y_i sachant $(Y_j)_{j\neq i}$, puis l'estimateur au maximum de vraisemblance conditionnelle associé $\widehat{\theta}_2$ (équation (I.31) avec $A = \{2, 3, \dots n-1\}$). Montrer que si $(Y_i)_{i\geqslant 1}$ est stationnaire ergodique (i. e. la loi des grands nombres s'applique) alors $\widehat{\theta}_1$ et $\widehat{\theta}_2$ 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) [20]**.** Soit (Y_i) une suite iid dont on cherche à estimer la densité. On a classiquement une famille d'estimateurs $\widehat{p}_h(x)$ indexée par un paramètre h qu'il va falloir choisir (largeur de fenêtre de l'estimateur à noyau, etc).

La famille $\widehat{p}_h$ 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 :

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

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. Social network.). Strauss et Ikeda [151] 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 est le suivant

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

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 ; θ_2 mesure la réciprocité des relations. Proposer une pseudo-vraisemblance basée sur un maximum de vraisemblance conditionnelle.

Exercice 11 (Régression catégorielle ordonnée [134]). La réponse Y est à valeurs dans $\{1, ...m\}$, p. ex. {guérison, élimination des symptômes, aucun effet, aggravation}, noter le caractère ordonné ; 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 $\mu_0 = -\infty < \mu_1 \cdots < \mu_m < \mu_{m+1} = +\infty$, de $\beta \in \mathbb{R}^p$, et d'une variable Z, dite latente car non-observée telle que

$$Y = k \quad \text{si} \quad \mu_k \leqslant Z + X\beta < \mu_{k+1}.$$

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

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

20. Inspiré de [91]. Une discussion générale sur le choix de h par validation croisée se trouve dans [2] p. 540 et suivantes.

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

On dit qu'une fonction $\varphi \geqslant 0$ est log-concave si l'ensemble $\{\varphi > 0\}$ est convexe et que $\log \varphi$ y est concave. On vérifiera que F_l et F_g sont log-concaves, 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 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 et un M-estimateur pour les processus de Poisson ponctuels).** Soit $(S, \mathscr{A}, \mu)$ un espace mesuré, la mesure μ est supposée finie et sans atome, et soit $\mathcal{P}_\mu$ la loi du processus de Poisson ponctuel d'intensité μ 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 θ

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

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 ξ la fonction $X \mapsto \varphi(X, \xi)$ soit symétrique en ses variables (fonction d'ensemble). Montrer que

$$E_\mu E_\theta \left[\varphi(X, \xi)\frac{f_\theta(X \cup \{\xi\})}{f_\theta(X)}\right] = \frac{1}{\mu(S)} E_\theta \left[\sum_{x\in X} \varphi(X\backslash\{x\}, x)\right] \tag{I.42}$$

   où E_μ désigne l'espérance sous $\mu(d\xi)/\mu(S)$ et E_θ est l'espérance sous P_θ.
   Indication : Commencer par le cas $f_\theta = 1$ en utilisant (I.21) ; terminer utilisant la définition de E_θ et en jouant sur φ.

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 β règle l'intensité du Poisson où r est connu et $\theta = (\beta, \gamma)$. Utiliser les deux fonctions

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

   pour justifier la validité du Z-estimateur basé sur p réalisations du processus, donnant (β, γ) comme solution du système

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

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

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

   On peut l'interpréter comme un estimateur de maximum de vraisemblance partielle (voir [27], ou [118] 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 θ on a

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

On utilisera la formule (I.42) 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 $\log u \leqslant u - 1$.
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, Z) de densité $p_{\theta_*}(x, z)$ par rapport à une mesure de référence produit notée par simplicité $dxdz$. On est dans le cadre paramétrique habituel. On observe seulement X ; Z est appelée variable cachée, ou latente. L'estimateur du maximum de vraisemblance $\widehat{\theta}$ satisfait

$$\int p_\theta(X, z) dz \leqslant \int p_{\widehat{\theta}}(X, z) dz \tag{I.43}$$

pour tout θ. Malheureusement cette maximisation est difficile à réaliser en pratique.

(i) On rappelle que pour toutes densités p et q, on a $\int \log(q(z)/p(z)) p(z) dz \leqslant 0$ (l'intégrale pouvant valoir $-\infty$; cf. la divergence de Kullback-Leibler). Montrer que pour tout θ

$$\int \log p_\theta(X, z) p_{\widehat{\theta}}(X, z) dz \leqslant \int \log p_{\widehat{\theta}}(X, z) p_{\widehat{\theta}}(X, z) dz. \tag{I.44}$$

$\widehat{\theta}$ réalise donc le maximum en θ du membre de gauche de (I.44).

(ii) Vérifier que (I.44) se réécrit

$$E_{\widehat{\theta}}[\log p_\theta(X, Z) | X] \leqslant E_{\widehat{\theta}}[\log p_{\widehat{\theta}}(X, Z) | X]. \tag{I.45}$$

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

$$\widehat{\theta} = F(\widehat{\theta}), \qquad F(\theta) = \arg\max_{\theta'} E_\theta[\log p_{\theta'}(X, Z) | X]. \tag{I.46}$$

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

$$\theta_{n+1} = F(\theta_n). \tag{I.47}$$

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.43) augmente d'une étape à la suivante.
Indication : On notera que (I.47) signifie que l'inégalité *inverse* de (I.44) est satisfaite avec $\widehat{\theta} = \theta_n$ et $\theta = \theta_{n+1}$; puis on utilisera l'inégalité de Jensen, prenant garde au fait que $z \to p_\theta(X, z)$ n'est pas une densité.

(iv) Exemple. Soit $X = (X_1, ...X_n)$, 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 $Z = (Z_1, ...Z_n) \in \{1, 2\}^n$ la variable cachée correspondante (cf. p. 8). Vérifier que (I.47) prend une forme simple, tandis que (I.43) est compliqué 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.46) sera calculée soit explicitement comme au point (iv), soit par simulation de réalisations de Z conditionnellement à X sous P_θ (algorithme MCEM), soit en utilisant une approximation dite variationnelle (cf. exercice 13).

La convergence peut être très lente, et sensible à l'initialisation. Il existe des techniques d'accélération, basées sur des méthodes numériques de recherche de point fixe, cf. [123] chap. 4, [75].

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 $Q^i_{\theta_*}$; on pourra commencer par le cas où cette loi est indépendante de i. Soit l'estimateur

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

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 θ_* à moins que pour un autre θ, pour tout i et $Q^i_{\theta_*}$-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_θ 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_θ est incalculable.

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

1. Trouver une relation entre σ et h. On supposera h continuement dérivable en θ_*.
2. Vérifier que si $\widehat{\theta}_n = n^{-1}\sum_{i=1}^n X_i$, 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 iid ?

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

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

On suppose que le biais de $\widehat{\theta}_n$ est de la forme $\sum_{k\geqslant 1} b_k n^{-k}$. 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, et risque de placer finalement $\check{\theta}_n$ en position d'infériorité .

On vérifie facilement que si $\widehat{\theta}_n$ est l'estimateur de la variance $n^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$ alors $\check{\theta}_n$ est l'estimateur non-biaisé de l'exercice 4 p. 26.

II

Estimation semi-paramétrique

II.1 Loi des grands nombres uniforme. Score

Nous allons introduire un résultat permettant d'obtenir la convergence uniforme imposée au théorème 2, ou dans (I.37), dans le cadre de fonctions d'estimations additives, $H_n(\theta,\omega) = n^{-1}\sum_{i=1}^n H(\theta, Y_i)$, possiblement peu régulières. La loi des grands nombres uniforme présentée ici généralise au cas dépendant un théorème classique consacré aux suites de variables indépendantes (p. ex. § 19.8 de [18] ou Th.3.1(vi) de [2]).

On a vu au théorème 5 qu'une inégalité de Sobolev permet de traiter cette question assez directement dans un cadre très général, mais au prix d'une régularité suffisante. Par exemple, dans le cas de processus ponctuels (exercice 12), le nombre de points est aléatoire et la forme additive ci-dessus n'est plus adaptée. Nous y reviendrons au § II.2.4 pour la normalité asymptotique et au § II.7 pour des inégalités de déviations.

Une autre méthode basée sur l'entropie métrique et la dimension de Vapnik-Chervonenkis est plus sophistiquée, très fructueuse, mais essentiellement restreinte aux suites indépendantes ; nous ne l'aborderons pas. On en trouvera une présentation simple et opérationnelle dans [14], qui peut introduire à des exposés plus pointus comme [6] chap.3, [18] chap.19, ou [19].

On considérera le jeu d'hypothèses suivant :

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

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

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

De plus

$$E\Big[\sup_{\theta\in\Theta} \big\|H(\theta, Y)\big\|\Big] \;<\; \infty \tag{II.2}$$

et pour toute boule B de centre θ_0, la fonction $\psi(y) = \sup_{\theta\in B} \big\|H(\theta, y) - H(\theta_0, y)\big\|$ satisfait

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

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 λ, $\lambda(\int f\, d\mu) = \int \lambda(f) d\mu$. Le cas non séparable est plus compliqué. Voir [70], ou encore [143].

(2) *Pour tout $\theta_0 \in \Theta$, presque sûrement, la fonction $\theta \mapsto H(\theta, Y)$ est continue au point θ_0 (cet ensemble de probabilité 1 peut dépendre de θ_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 ([125] § 17.1.3). L'hypothèse (GNU2) accepte certaines situations où la fonction H est discontinue en θ, par exemple si $\Theta = [a, b]$, Y a une loi diffuse et $H(\theta, Y) = 1_{Y \leqslant \theta} f(\theta, Y)$ où f est continue en θ [2].

6 - 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

$$\lim_n \sup_{\theta \in \Theta} \Big\| h(\theta) - \frac{1}{n} \sum_{i=1}^{n} H(\theta, Y_i) \Big\| = 0. \tag{II.4}$$

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

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

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

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

Fixons $\varepsilon > 0$. Pour tout θ_0, il existe $\eta(\theta_0) > 0$ tel que $f_{\theta_0}(\eta(\theta_0)) < \varepsilon$. Les boules ouvertes de centre θ et de rayon $\eta(\theta)$ forment un recouvrement d'ouverts de Θ dont on peut extraire par compacité un sous-recouvrement fini :

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

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 θ dont il s'agit de majorer la norme uniforme, sans oublier que j dépend de θ. Celle du troisième est inférieure à ε ; 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 :

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

mais avec probabilité 1, en vertu de (II.3)

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

On a montré que le membre de gauche de (II.4) est inférieur à 2ε ; comme ε est arbitraire, il est nul. ∎

2. Wald semble être le premier, en 1949, à avoir souligné ce point dans un cadre général [155]

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

7 - Proposition

Soit θ_n une suite de variables aléatoires à valeurs dans Θ (non-nécessairement compact) convergeant presque sûrement vers une limite déterministe θ_*. Si (GNU) est satisfaite sur un voisinage compact Θ_* de θ_*, alors presque sûrement :

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

Démonstration. Soit θ'_n qui vaut θ_n si $\theta_n \in \Theta_*$ et θ_* sinon, alors

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

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 $M_n(\widehat{\theta}_n) = 0$ 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 $M_n(\widehat{\theta}_n) - M_n(\theta_*) \sim \nabla M_n(\theta_*)(\widehat{\theta}_n - \theta_*)$ qui va permettre de déduire la normalité asymptotique de $\widehat{\theta}_n - \theta_*$, car $\nabla M_n(\theta_*)$ converge en vertu de la loi des grands nombres (et $M_n(\widehat{\theta}_n) = 0$). 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 :

8 - 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 $n \geqslant 1$ et presque tout ω 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

$$\begin{aligned}
&\theta_n \xrightarrow{P} \theta_* \\
&\sqrt{n}\ M_n(\theta_n) \xrightarrow{P} 0 \\
&\sqrt{n}\ M_n(\theta_*) \xrightarrow{d} \mathcal{D}, \quad \text{pour une certaine distribution } \mathcal{D} \\
&\sup_{\|\theta-\theta_*\|\leqslant\varepsilon} \big\|\nabla M_n(\theta) - J(\theta)\big\| \xrightarrow{P} 0 \qquad \text{(II.5)}
\end{aligned}$$

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

$$\sqrt{n}(\theta_n - \theta_*) + \sqrt{n}J_*^{-1}M_n(\theta_*) \xrightarrow{P} 0. \qquad \text{(II.6)}$$

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

$$M_n(\theta_n) = M_n(\theta_*) + \nabla M_n(\theta'_n)(\theta_n - \theta_*) = M_n(\theta_*) + (J_* + r_n)(\theta_n - \theta_*) \qquad \text{(II.7)}$$

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 θ'_n est différent d'une ligne à l'autre de $\nabla M_n(\theta'_n)$. L'équation (II.7) équivaut à

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

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 θ_n et la convergence uniforme en probabilité de la dernière hypothèse ; nous y reviendrons, une version générale étant proposée au théorème 12.

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

On considère l'estimateur solution de :

$$\sum_{i=1}^{n} H(\widehat{\theta}_n, Y_i) = 0. \tag{II.8}$$

On va montrer que sous des hypothèses raisonnables, $\widehat{\theta}_n$ converge p.s vers θ_*, avec normalité asymptotique de $\sqrt{n}\ (\widehat{\theta}_n - \theta_*)$. On verra le rôle important joué par la matrice de bruit $I = Var(\sum_{i=1}^{n} H(\theta_*, Y_i))/n$, 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, 86], situations où le maximum de vraisemblance est souvent très difficile à mettre en œuvre (cf. § I.3.2 et les exercices 7 p. 26, 10 p. 27, 14 p. 29). 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

$$y_n = \theta_* y_{n-1} + e_n = \sum_{k \geqslant 0} \theta_*^k e_{n-k} \tag{II.9}$$

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 [3]. Dans ce cas l'estimateur des moindres carrés

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

n'est généralement pas consistant ($\widehat{r}_i$ désigne la i-ième corrélation empirique), car (II.9) implique que $r_1 = \theta_* r_0 + E[y_{n-1} e_n] \neq \theta_* r_0$. En multipliant l'équation (II.9) 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 θ_* par $\widehat{r}_q / \widehat{r}_{q-1}$. C'est le Z-estimateur basé sur la fonction

$$H(\theta, Y_k) = (y_k - \theta y_{k-1}) y_{k-q}, \qquad Y_k = (y_k, \dots y_{k-q}). \tag{II.10}$$

Comme $E[H(\theta, Y_k)] = r_q - \theta r_{q-1}$, θ_* 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,

$$y_n = \sum_{i=1}^{p} \theta_i y_{n-i} + e_n \tag{II.11}$$

3. On a considéré dans ce paragraphe un exemple plus général que l'ARMA, car dans de nombreuses application, 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

$$H(\theta, Y_k)_j = \Big(y_k - \sum_i \theta_i y_{k-i}\Big) y_{k-q+1-j}, \qquad j = 1, \dots p \tag{II.12}$$

$$= e_k(\theta) t_{k-q,j}. \tag{II.13}$$

La dernière écriture suggère la philosophie générale de la méthode instrumentale : les $e_k(\theta)$ sont les erreurs de prédiction sous l'hypothèse θ, 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

9 - Théorème

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

$$\lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{i=1}^{n} H(\widehat{\theta}_n, Y_i) = 0, \quad p.s. \tag{II.14}$$

satisfait p.s.

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

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

Démonstration. Il s'agit d'une conséquence immédiate des théorèmes 6 et 2. ∎

Le cas non-compact. On s'est placé sous l'hypothèse de compacité de Θ. Cette hypothèse est restrictive mais a le mérite de la simplicité. Si Θ 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 Θ_p dont la réunion des intérieurs fait Θ, et sur chacun desquels est satisfait (GNU).*
De plus la suite $\widehat{\theta}_n$ reste presque sûrement dans un compact (pouvant dépendre de ω).

Cette hypothèse se vérifiera facilement si par exemple $\|\widehat{\theta}_n\|$ 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)).

10 - Théorème

Le théorème 9 reste vrai si l'on remplace (GNU) par (GNU').

Démonstration. Notons d'abord que comme les intérieurs des Θ_p recouvrent le compact où la suite $\widehat{\theta}_n$ reste confinée, il existe un $p(\omega)$ tel que $\Theta_{p(\omega)}$ contienne toute la suite. Considérons pour tout p l'estimateur $\widehat{\theta}_n^{(p)}$ qui vaut $\widehat{\theta}_n$ si $\widehat{\theta}_n \in \Theta_p$ et sinon θ_*. Le théorème 9 implique la convergence presque sûre de $\widehat{\theta}_n^{(p)}$ vers θ_*. On a donc convergence de $\widehat{\theta}_n$ vers θ_* sur l'ensemble $A_p = \{\omega : \forall n,\ \widehat{\theta}_n \in \Theta_p\}$. 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 $\mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$. Il existe un voisinage Θ_0 de θ_*, sur lequel, p.s. pour tout i, la fonction $\theta \mapsto H(\theta, Y_i)$ est de classe C^1. Le triplet $(\Theta_0, (Y_i)_{i\geqslant 0}, \nabla H(\theta, y))$ 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 Θ ; il suffit, pour appliquer le théorème qui suit, de lui donner la valeur 0 ailleurs, car en réalité, seul Θ_0 compte désormais.

11 - Théorème

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

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} H(\theta_*, Y_i) \xrightarrow{d} \mathcal{N}(0, I) \tag{II.15}$$

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

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

satisfait :

$$\sqrt{n}\Big(\theta_n - \theta_* + \tfrac{1}{n} J^{-1} \sum_{i=1}^{n} H(\theta_*, Y_i)\Big) \xrightarrow{P} 0 \tag{II.16}$$

$$\sqrt{n}\ (\widehat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, J^{-1} I J^{-T}). \tag{II.17}$$

Remarques. 1/ Par analogie avec la théorie du maximum de vraisemblance, l'inverse de la matrice $J^{-1}IJ^{-T}$ est parfois appelé matrice d'information de Godambe (Godambe information matrix).
2/ La variance de $\widehat{\theta}_n$ 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 $\widehat{\theta}_n$ est asymptotiquement linéaire.

Démonstration. C'est une application directe du théorème 8 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 θ dans un voisinage compact de θ_*. ∎

Régions de confiance. Si l'on dispose d'estimateurs $\widehat{I}_n \xrightarrow{P} I$ (p. ex. formule (II.25) plus bas ; dans le cas dépendant, il peut être difficile de trouver un $\widehat{I}_n$ satisfaisant) alors l'hypothèse (II.15) implique que

$$\mathcal{R}_\alpha = \Big\{\theta : \big\|\widehat{I}_n^{-1/2} \sum_{i=1}^{n} H(\theta, Y_i)\big\|^2 \leqslant n\chi_d^2(1-\alpha)\Big\} \tag{II.18}$$

est une région de confiance pour θ_* de risque asymptotique α (i. e. son risque réel α_n tend vers α. Rappelons que $\chi_d^2(1-\alpha) \simeq d - 2\log\alpha$ [99]. Comme de plus sous les hypothèses du théorème $\widehat{J}_n = n^{-1} \sum \nabla H(\widehat{\theta}_n, Y_i) \xrightarrow{P} J$ on a également la région

$$\mathcal{R}_\alpha = \Big\{\theta : n(\widehat{\theta}_n - \theta)^T \widehat{V}^{-1} (\widehat{\theta}_n - \theta) \leqslant \chi_d^2(1-\alpha)\Big\}, \quad \widehat{V} = \widehat{J}_n^{-1} \widehat{I}_n \widehat{J}_n^{-T}. \tag{II.19}$$

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

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

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) [4] :

$$P\Big(\forall k, X_k^2 \leqslant \sigma_k^2\chi_1^2(1-\alpha_k)\Big) \geqslant \prod_k P\Big(X_k^2 \leqslant \sigma_k^2\chi_1^2(1-\alpha_k)\Big) = \prod_k(1-\alpha_k).$$

Le membre de gauche étant la limite de $P(\theta_* \in \mathcal{R})$, ceci prouve bien le niveau de confiance asymptotique. Si ε est le niveau attendu et que tous les α_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 va tout autrement de 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.11) du § II.2.1, qui est un peu plus général que l'ARMA(p,q), avec l'équation d'estimation (II.12). Pour simplifier les écritures on se restreint à (II.9, II.10), 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

$$\begin{aligned}\frac{1}{n}E\Big[\Big(\sum_{i=1}^n H(\theta_*,Y_i)\Big)^2\Big] &= \frac{2}{n}E\Big[\sum_{i<j} H(\theta_*,Y_i)H(\theta_*,Y_j)\Big] + \frac{1}{n}E\Big[\sum_{i=1}^n H(\theta_*,Y_i)^2\Big] \\ &= \frac{2}{n}E\Big[\sum_{i=1}^n (n-i)H(\theta_*,Y_0)H(\theta_*,Y_i)\Big] + E\big[H(\theta_*,Y_0)^2\big]. \qquad \text{(II.20)}\end{aligned}$$

Comme $H(\theta_*,Y_0)H(\theta_*,Y_i) = e_0e_iy_{-q}y_{i-q}$, son espérance est nulle pour $i \geqslant q$ (noter que la q-décorrélation est ici insuffisante, puisqu'on considère des moments d'ordre 4), et sinon vaut $E[e_qe_{q+i}y_0y_i]$ (stationnarité), et en faisant tendre n vers l'infini on obtient, en supposant que le membre de gauche de (II.20) converge bien vers I,

$$I = 2\sum_{i=1}^q E\big[e_qe_{q+i}y_0y_i\big] + E[e_q^2y_0^2]. \qquad \text{(II.21)}$$

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.20) on voit que, sous des hypothèses raisonnables de décroissance des covariances :

$$I = R(0) + \sum_{i>0} R(i) + R(i)^T \ , \quad R(i) = \mathrm{Cov}(H_1, H_{i+1}), \quad H_i = H(\theta_*,Y_i). \qquad \text{(II.22)}$$

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.22) l'estimateur obtenu n'est pas consistant. Il faut prendre un estimateur fenêtré [76], ce qui n'est pas très facile car il y a un paramètre sensible à régler pour équilibrer biais et variance.

Si l'on veut faire des tests d'adéquation pour comparer des modèles, une autre stratégie consiste à utiliser quand même un estimateur non-consistant, comme par exemple $\check{I} = n^{-2}\sum S^2$, 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 $(n\check{I})^{-1/2}\sum_{i=1}^n H_i$ est néanmoins asymptotiquement pivotale si le modèle est valide, ce qui est parfois le cas [106, 24]. Sa loi peut alors s'estimer par simulation, ce qui permet de l'utiliser pour des tests.

4. Pour tout vecteur gaussien centré $(Y_1,\dots Y_n)$ et des c_k réels, $P\big(|Y_k| \leqslant c_k, k=1\dots n\big) \geqslant \prod_k P\big(|Y_k| \leqslant c_k\big)$. C'est une conséquence de l'inégalité de Royen (Gaussian correlation inequality [112]) : Si E et F sont convexes symétriques, $P(Y \in E \cap F) \geqslant P(Y \in E)P(Y \in F)$.

II.2.4 Théorème-limite central pour une fonction d'estimation générale

Il arrive que la fonction d'estimation ne possède pas la forme additive utilisée jusqu'à présent auquel cas il faut faire appel au théorème 8 plus directement, la difficulté étant de garantir (II.5). Le théorème suivant se propose de surmonter cette difficulté en utilisant une inégalité de Sobolev, comme cela a été fait pour montrer la convergence en probabilité de $\widehat{\theta}_n$ (théorème 5) :

12 - Théorème

Soit $H_n(\theta,\omega)$ des fonctions mesurables de $\mathbb{R}^d\times\Omega$ dans $\mathbb{R}^d$. Soit $\theta_*\in\mathbb{R}^d$ possédant un voisinage $\mathscr{V}=\{\theta:\|\theta-\theta_*\|<\varepsilon\}$ sur lequel pour tout $n\geqslant 1$ et presque tout ω l'application $\theta\mapsto H_n(\theta,\omega)$ est de classe C^1. Soit $\widehat{\theta}_n(\omega)$ une suite d'estimateurs du paramètre $\theta_*\in\mathbb{R}^d$. On suppose que

$$\widehat{\theta}_n \xrightarrow{P} \theta_*$$
$$\sqrt{n}\ H_n(\widehat{\theta}_n) \xrightarrow{P} 0$$
$$\sqrt{n}\ H_n(\theta_*) \xrightarrow{d} \mathcal{D}, \quad \text{pour une certaine distribution } \mathcal{D}$$
$$\forall\theta\in\mathscr{V}, \left\|\nabla H_n(\theta)-J(\theta)\right\| + \left\|\nabla^2 H_n(\theta)-\nabla J(\theta)\right\| \xrightarrow{P} 0$$
$$\sup_{\theta\in\mathscr{V}} E\Big[\left\|\nabla H_n(\theta)\right\|^q + \left\|\nabla^2 H_n(\theta)\right\|^q\Big] < +\infty$$

pour un $q>d$ et pour une certaine fonction $\theta\mapsto J(\theta)$ différentiable sur $\mathscr{V}$. Alors, si $J_*=J(\theta_*)$ est inversible,

$$\sqrt{n}(\theta_n-\theta_*)+\sqrt{n}J_*^{-1}H_n(\theta_*)\xrightarrow{P}0. \tag{II.23}$$

Remarque. Si $\mathcal{D}=\mathcal{N}(0,I)$, $\sqrt{n}(\theta_n-\theta_*)\longrightarrow\mathcal{N}(0,J_*^{-1}IJ_*^{-T})$ en loi.

Démonstration. Il s'agit simplement de vérifier (II.5). Posons

$$G_n(\theta)=\nabla H_n(\theta)-J(\theta).$$

En vertu de l'inégalité de Sobolev présentée plus bas appendice, éq. (D.1) p.125, on a avec probabilité 1

$$\sup_{\theta\in\mathscr{V}}\left\|G_n(\theta)\right\|^q \leqslant C_q\int_{\mathscr{V}}\Big(\|\nabla G_n(\theta)\|^q+\|G_n(\theta)\|^q\Big)d\theta,$$

où C_q ne dépend que de q. Il ne suffit plus que de prendre l'espérance et d'appliquer le théorème de Fubini-Tonelli, puis le théorème de convergence dominée. ∎

II.2.5 Fusion optimale de fonctions d'estimation

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.12) 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 $\theta\in\Theta\subset\mathbb{R}^d$, $d\leqslant m-1$; on a donc m fonctions $p_1(\theta),\dots p_m(\theta)$ dont la somme fait 1. C'est le cas du modèle de Hardy-Weinberg pour la fréquence de génotypes dans le 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

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

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 $\widehat{\theta}$ vaut alors

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

avec I et J définies au théorème 11. Noter que si K dépend de θ, tout en étant de classe C^2 au voisinage de θ_*, 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 non énonçons de façon informelle (cf. l'exercice 5 p. 40) :

> *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 $\widetilde{H}(\theta, y) = K^T H(\theta, y)$ satisfait $\widetilde{I} = \widetilde{J}$. Cette méthode permet d'obtenir de bons estimateurs dans des cadres assez généraux de recherche d'estimateurs efficaces ([128] 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 d'une combinaison des fonctions d'estimation.

Dans certains cas paramétriques, I et J s'expriment directement en fonction de θ_*, auquel cas une estimée préliminaire $\widehat{\theta}$ de θ_* conduira à une matrice $K(\widehat{\theta})$ convenable. Sinon il faut les estimer directement sur les observations, éventuellement à l'aide d'une première estimée $\widehat{\theta}$ de θ_*, par exemple avec les deux premières formules de l'encadré (II.25) dont la convergence se justifie avec les hypothèses (GNU). Si $\widehat{\theta}$ converge à vitesse $n^{-\frac{1}{2}}$, la seconde estimée $\widetilde{\theta}$ pourra se calculer en annulant $\sum \widetilde{H}(\theta, Y_i)$ 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) :

$$\begin{aligned}\widetilde{\theta} &= \widehat{\theta} - \Big(\sum_i \nabla\widetilde{H}(\widehat{\theta}, Y_i)\Big)^{-1}\Big(\sum_i \widetilde{H}(\widehat{\theta}, Y_i)\Big), \quad \widetilde{H}(\theta, y) = \widehat{J}^T\widehat{I}^{-1}H(\theta, y) \\ &= \widehat{\theta} - \big(\widehat{J}^T\widehat{I}^{-1}\widehat{J}\big)^{-1}\Big(\tfrac{1}{n}\sum_i \widetilde{H}(\widehat{\theta}, Y_i)\Big)\end{aligned} \tag{II.24}$$

si $\widehat{J}$ correspond à (II.25). Soit au final

On suppose les Y_i indépendantes.

Calcul de la nouvelle fonction d'estimation :

$$\begin{aligned}&\widehat{J} = \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla H(\widehat{\theta}, Y_i) \\ &\widehat{I} = \tfrac{1}{n}\textstyle\sum_i H(\widehat{\theta}, Y_i)H(\widehat{\theta}, Y_i)^T - \tfrac{1}{n^2}\Big(\sum_i H(\widehat{\theta}, Y_i)\Big)\Big(\sum_i H(\widehat{\theta}, Y_i)\Big)^T \\ &\widetilde{H}(\theta, y) = \widehat{J}^T\widehat{I}^{-1}H(\theta, y).\end{aligned}$$

Calcul de l'estimateur :

$$\widetilde{\theta} = \widehat{\theta} - \Big(\textstyle\sum_i \widetilde{H}(\widehat{\theta}, Y_i)\widetilde{H}(\widehat{\theta}, Y_i)^T\Big)^{-1}\sum_i \widetilde{H}(\widehat{\theta}, Y_i)$$

ou bien $\sum_i \widetilde{H}(\widetilde{\theta}, Y_i) = 0$. (II.25)

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.2.6 Exercices et compléments

Exercice 1. On considère l'exercice 2 p. 25 pour lequel l'estimateur est $\widehat{\lambda} = n/\sum X_i$, $\widehat{\mu} = \sum Y_i/n$ et $\widehat{\alpha} = -2/\widehat{\lambda} + \frac{1}{2}\widehat{\lambda}^2 \sum (Y_i - \widehat{\mu})^2 (X_i - \widehat{\lambda}^{-1})/n$. 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.10). 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 e_1. En déduire une expression de la variance asymptotique de $\widehat{\theta}$ en fonction de θ_* seulement.

Exercice 3 (Fusion de fonctions d'estimation et fusion d'estimateurs). Soient $H_1(\theta, y)$ et $H_2(\theta, y)$ deux fonctions d'estimation satisfaisant les hypothèses du théorème 11, avec les matrices de sensibilité J_1 et J_2. Soit $\widehat{\theta}_1$ et $\widehat{\theta}_2$ les deux estimées correspondantes. Soit P une matrice, et $\widehat{\theta}_3$ l'estimée obtenue avec $H_3 = PJ_1^{-1}H_1 + (I - P)J_2^{-1}H_2$. On suppose que H_3 satisfait également (II.15). Montrer, en utilisant le théorème 11 que $\sqrt{n}(\widehat{\theta}_3 - P\widehat{\theta}_1 - (1 - P)\widehat{\theta}_2)$ 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 θ_* à partir des n premières observations par $\widehat{\theta}_n$ 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 $\widehat{\theta}_n$ vers θ_*.
2. On admet la normalité asymptotique. Calculer la variance asymptotique de $\widehat{\theta}_n$ 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.2.5. On rappelle qu'une matrice symétrique $S \in \mathbb{R}^{d \times d}$ est $\geqslant 0$ si pour tout $x \in \mathbb{R}^d$, $x^T S x \geqslant 0$, et que $S \geqslant T$ si $S - T \geqslant 0$. 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 θ ; 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 $V_{K_0} \leqslant V_K$ pour tout K, c'est-à-dire que

$$(K_0^T I K_0)^{-1} \leqslant (K^T I K_0)^{-1} K^T I K (K^T I K_0)^{-T}. \tag{II.26}$$

1. Montrer que si $J = IK_0$ est de rang plein, K_0 aussi et $K_0^T I K_0$ est bien inversible.
   Indication : Vu les dimensions de J, dire qu'elle est de rang plein revient à dire que ses colonnes sont indépendantes : $Jx = 0 \Rightarrow x = 0$. Pour montrer l'inversibilité on vérifiera que l'équation $x^T K_0^T I K_0 x = 0$ est sans solution non-nulle ; utiliser une racine carrée de I.
2. Montrer que si S et T sont symétriques $\geqslant 0$, et M une matrice rectangulaire, alors $S \leqslant T$ implique $M^T S M \leqslant M^T T M$, si les dimensions sont compatibles. En déduire que (II.26) se réduit à montrer :

   $$I K_0 (K_0^T I K_0)^{-1} K_0^T I \leqslant I. \tag{II.27}$$

   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 $J(J^T J)^{-1} J^T \leqslant Id$.
   Indication. Une matrice P telle que $P^T = P = P^2$ est une matrice de projection orthogonale ; par conséquent pour tout x, $x^T P x \leqslant \|x\|^2$.
4. En déduire (II.27). *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 (signe(0) = 0) :

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

On pourra le faire en utilisant soit le théorème 6 soit le théorème 3.

Exercice 7 (Maximum de vraisemblance empirique (MVE) [5]). On est dans le cas où $\dim(H) > \dim(\theta)$. On considère le vecteur $(\widehat{\theta}, \widehat{p}_1, \dots \widehat{p}_n)$ solution du problème

$$\max \sum_{i=1}^{n} \log p_i \qquad \text{sous} \qquad \sum_{i=1}^{n} p_i H(\theta, Y_i) = 0, \qquad \sum_{i=1}^{n} p_i = 1. \tag{II.28}$$

On cherche donc à réaliser quand même l'annulation 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.28). 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 $\widehat{\theta}$.

1. Vérifier que les conditions du premier ordre pour (II.28), après élimination du multiplicateur de Lagrange associé à $\sum_{i=1}^{n} p_i = 1$, sont ($\gamma \in \mathbb{R}^{\dim(H)}$ est le vecteur des multiplicateurs de Lagrange)

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

   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 (θ, γ). Que vaut γ_* ?
3. Soit

   $$\widehat{J}(\theta) = \sum_i \widehat{p}_i \nabla H(\theta, Y_i), \qquad \widehat{I}(\theta) = \sum_i \widehat{p}_i H(\theta, Y_i) H(\theta, Y_i)^T, \qquad \overline{H}(\theta) = \frac{1}{n} \sum_i H(\theta, Y_i).$$

   Vérifier que $\widehat{\theta}$ est solution de $\widehat{J}(\theta)^T \widehat{I}(\theta)^{-1} \overline{H}(\theta) = 0$ (multiplier l'équation $n\widehat{p}_i + \widehat{p}_i \gamma^T H(\widehat{\theta}, Y_i) = 1$ par $H(\widehat{\theta}, Y_i)^T$ 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.2.5.
4. Invariance. Soit $M(\theta)$ une matrice inversible. Que se passe-t-il d'insatisfaisant dans la méthode du § II.2.5 (i. e. la résolution de $J(\theta_*) I(\theta_*)^{-1} \overline{H}(\theta) = 0$) si l'on remplace la fonction $H(\theta, y)$ par $M(\theta) H(\theta, y)$, mais qui ne 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 θ_* telle que $E[H(\theta_*, Y)] = 0$ (modélisation trop approximative), la solution de (II.28) peut donner des poids très inégaux aux observations et conduire à des solutions insatisfaisantes. Ceci rend la méthode peu robuste.

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

II.3.1 Exemple : régression non-linéaire

Reprenons le modèle de la page 9 :

$$Y_i = n_\theta(X_i) + e_i, \qquad E[e_i] = 0. \tag{II.29}$$

5. On trouvera des compléments dans le livre de Owen ou dans [108], ou encore [50].

L'estimateur des moindres carrés est

$$\widehat{\theta}_n = \operatorname{argmin} \sum (Y_i - n_\theta(X_i))^2. \tag{II.30}$$

L'estimateur robuste de Huber est

$$\widehat{\theta}_n = \operatorname{argmin} \sum \psi(Y_i - n_\theta(X_i)) \tag{II.31}$$

où ρ est une fonction qui commence en parabole, et se prolonge linéairement : $\psi(x) = x^2 1_{|x|\leqslant\alpha} + (2\alpha|x| - \alpha^2)1_{|x|>\alpha}$. Cet estimateur est moins sensible aux données aberrantes ; il a été présenté en détail dans le cas du problème du paramètre de translation éq. (I.30) (c.-à-d. $n_\theta(x) = \theta$), et dans ce cas, 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 α dépend du poids attribué à la contamination dans le mélange, voir [10] § 4 pour des arguments précis, et [160] 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 :

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} K(\theta, Y_i). \tag{II.32}$$

13 - 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 $k(\widehat{\theta}_n)$ converge p.s. vers le minimum de la fonction k.
Si de plus θ_* est le minimum unique de $k(\theta)$ (identifiabilité) alors $\widehat{\theta}_n$ tend p.s. vers θ_*.
Les conclusions restent vraies si l'on remplace (GNU) par (GNU').

Démonstration. Sous (GNU) c'est une conséquence immédiate du théorème 6 et du théorème 2. L'extension à (GNU') se fait comme à la démonstration du théorème 10. ∎

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 favorable pour l'estimation. Pour traiter ces situations, on a le résultat suivant :

14 - Théorème

On considère les hypothèses suivantes :

(a) Pour tout $A \geqslant 0$, la suite $(Y_i)_{i\geqslant 1}$ 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 θ_*. En particulier $k(\theta_*) < +\infty$.

(c) On a p.s. $\frac{1}{n}\sum_{i=1}^{n} K(\theta_*, Y_i) \to k(\theta_*)$.

Sous (a,b,c), $\widehat{\theta}_n$ tend p.s. vers θ_*.
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 $A \geqslant 0$ (valeur qui tendra vers $+\infty$)

$$K_n^A(\theta) = \frac{1}{n}\sum_{i=1}^n \min(A, K(\theta, Y_i))$$
$$k^A(\theta) = E\big[\min(A, K(\theta,Y))\big]$$
$$K_n(\theta) = \frac{1}{n}\sum_{i=1}^n K(\theta, Y_i).$$

Alors pour tout n et tout A

$$K_n^A(\widehat{\theta}_n) \leqslant K_n(\widehat{\theta}_n) \leqslant K_n(\theta_*).$$

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

$$\overline{\lim_n}\, k^A(\widehat{\theta}_n) \leqslant \overline{\lim_n}\left(K_n^A(\widehat{\theta}_n) + \|K_n^A - k^A\|_\infty\right) \leqslant \overline{\lim_n}\, K_n(\theta_*) = k(\theta_*)$$

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

$$k^A(\widehat{\theta}_\infty) \leqslant k(\theta_*).$$

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

L'extension à (GNU') se fait comme dans la démonstration du théorème 10. ∎

Revenons à (II.30), 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
$E\Big[\sup_{\theta\in\Theta} n_\theta(X_1)^2\Big] < \infty$
$E[e_i^2] < \infty$
$E\Big[(f(X_1) - n_\theta(X_1))^2\Big]$ 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 11 à la fonction $H(\theta, y) = \nabla_\theta K(\theta, y)$ car on a :

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

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 II.3.4.

II.3.3 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 $d = 1, \dots 50$, $\widehat{\theta}_n^d \in \mathbb{R}^d$, et de choisir celui qui donne la meilleure vraisemblance. Pour $d \leqslant d_*$ 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.

Supposons donnée une famille croissante d'espaces de paramètres Θ_d ; on voudrait choisir le meilleur des $\widehat{\theta}_n^d$, au sens de la minimisation de $k(\widehat{\theta}_n^d)$. L'interprétation proposée par Akaike (1973) est que pour $d > d_*$, la comparaison des $K_n(\widehat{\theta}_n^d)$ (au lieu de $k(\widehat{\theta}_n^d)$) 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 (cf. exercice 1 p. 16), afin de prendre en compte le surajustement. Le débiaisement introduit sera faible de sorte qu'il affectera peu le comportement pour $d \leqslant d_*$, mais suffira à corriger la diminution du contraste pour $d > d_*$.

On considère l'estimation dans un espace Θ, et θ_* va désigner le minimiseur de k dans cet espace. Comme $\nabla k(\theta_*) = 0$, on a par un développement limité de Taylor-Lagrange en θ_* du contraste k :

$$\begin{aligned} k(\widehat{\theta}_n) &= k(\theta_*) + \tfrac{1}{2}(\widehat{\theta}_n - \theta_*)^T \nabla^2 k(\theta_c)(\widehat{\theta}_n - \theta_*), \qquad \theta_c \in [\theta_*, \widehat{\theta}_n] \\ &= k(\theta_*) + \tfrac{1}{2}(\widehat{\theta}_n - \theta_*)^T J(\widehat{\theta}_n - \theta_*) + \tfrac{1}{n} o_P(1) \end{aligned} \tag{II.33}$$

($o_P(1)$ désigne un terme qui tend vers 0 en probabilité quand $n \to \infty$). Par ailleurs, par un développement limité de Taylor-Lagrange K_n en $\widehat{\theta}_n$, et en appliquant la loi des grands nombres uniforme à la dérivée seconde de K_n, il vient

$$\begin{aligned} K_n(\widehat{\theta}_n) &= K_n(\theta_*) - \tfrac{1}{2}(\widehat{\theta}_n - \theta_*)^T \nabla^2 K_n(\theta'_c)(\widehat{\theta}_n - \theta_*) \\ &= K_n(\theta_*) - \tfrac{1}{2}(\widehat{\theta}_n - \theta_*)^T J(\widehat{\theta}_n - \theta_*) + \tfrac{1}{n} o_P(1) \qquad \text{(surajustement)} \end{aligned} \tag{II.34}$$

si bien que par soustraction de ces deux équations

$$\begin{aligned} &k(\widehat{\theta}_n) = \tau + K_n(\widehat{\theta}_n) + (\widehat{\theta}_n - \theta_*)^T J(\widehat{\theta}_n - \theta_*) + \tfrac{1}{n} o_P(1), \\ &\tau = k(\theta_*) - K_n(\theta_*). \end{aligned} \tag{II.35}$$

La quantité $K_n(\widehat{\theta}_n)$ sous-estime donc le contraste $k(\widehat{\theta}_n)$ par un terme qui, au vu du théorème 11, est asymptotiquement d'espérance $Tr(IJ^{-1})$, auquel s'additionne τ, variable centrée. On a donc l'estimée asymptotiquement non-biaisée de $k(\widehat{\theta}_n)$: $K_n(\widehat{\theta}_n) + \frac{1}{n} Tr(IJ^{-1})$.

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

$$\boxed{\begin{array}{l} \widehat{\theta} = \arg\min_{\theta \in \Theta} Q(\theta), \qquad \theta_* = \arg\min_{\theta \in \Theta} E[Q(\theta)] \\ TIC = 2Q(\widehat{\theta}) + 2\,Tr(\widehat{J}^{-1}\widehat{I}), \qquad \widehat{J} \simeq E[\nabla^2 Q(\theta_*)], \qquad \widehat{I} \simeq Var(\nabla Q(\theta_*)) \end{array}}$$

où le $\simeq$ signale une estimation (cf. la discussion du § II.2.3 sur l'estimation de I). TIC, appelé *critère de Takeuchi* (prononcé Takéouchi), est une estimée corrigée du contraste en $\widehat{\theta}_n$. Dans l'exemple de la famille d'espaces de paramètres Θ_d, TIC sera une fonction de d, $TIC(d)$, et son minimiseur $\widehat{d}$ est un estimateur de d_* pour lequel $k(\widehat{\theta})$ devrait être raisonnablement faible. Dans le cas du critère de vraisemblance, on verra que $I = J$ ce qui fait que $Tr(J^{-1}I) = d$ (du moins si $d > d_*$), 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 décorrélés, 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.21)) [6]. Deux autres exemples sont proposés dans les exercices 8 plus bas et 2 p. 56.

Nous poursuivrons cette discussion page 74.

II.3.4 Un raffinement dans le cas indépendant

Il se trouve que dans le cas indépendant, on peut affaiblir substantiellement les hypothèses pour la normalité asymptotique. On peut alors, par exemple, traiter le cas de la médiane où $K(\theta, y) = |y - \theta|$, ce qui n'est pas possible avec le théorème 11 car l'hypothèse (FE) n'est pas satisfaite par la fonction discontinue $H(\theta, y) = \nabla K(\theta, y)$. Ceci va réclamer de gros efforts. Le jeu d'hypothèses est le suivant :

(ME) *(Θ, d) est un espace métrique compact. $K(\theta, y) : \Theta \times \mathbb{R}^m \to E$ est une fonction telle que pour tout θ, $y \mapsto K(\theta, y)$ soit mesurable. $(Y_i)_{i \geqslant 1}$ est une suite de variables iid. Il existe une fonction $\dot{K}$ et un voisinage $\mathscr{V}_*$ de θ_*, tels que, avec probabilité 1,*

$$\forall \theta, \theta' \in \mathscr{V}_*, \quad |K(\theta, Y_1) - K(\theta', Y_1)| \leqslant \dot{K}(Y_1)\|\theta - \theta'\| \tag{II.36}$$

et

$$E[\dot{K}(Y_1)^2] < \infty. \tag{II.37}$$

Presque sûrement $\theta \mapsto K(\theta, Y_1)$ est différentiable en θ_*. 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 θ_* de la forme

$$k(\theta_* + h) - k(\theta_*) = \frac{1}{2} h^T J h + o(|h|^2) \tag{II.38}$$

et la matrice de sensibilité J est inversible.

15 - Théorème

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

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

Alors

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta_* + \tfrac{1}{n} J^{-1} \sum_{i=1}^n \nabla K(\theta_*, Y_i)\Big) \xrightarrow{P} 0 \tag{II.39}$$

$$\sqrt{n}\ (\widehat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, J^{-1} I J^{-T}), \qquad I = E[\nabla K(\theta_*, Y_1) \nabla K(\theta_*, Y_1)^T]. \tag{II.40}$$

Démonstration. Soit $r_0 > 0$ tel que $B(\theta_*, r_0)$ soit inclus dans un voisinage de θ_* dans Θ ; remarquons tout d'abord que quitte à remplacer $\widehat{\theta}_n$ par $\widehat{\theta}_n 1_{\|\widehat{\theta}_n - \theta_*\| \leqslant r_0} + \theta_* 1_{\|\widehat{\theta}_n - \theta_*\| > r_0}$, on peut supposer que $\widehat{\theta}_n \in B(\theta_*, r_0)$ toujours, car, puisque $\widehat{\theta}_n$ converge en probabilité vers θ_*, ce changement n'affecte pas les équations (II.39) et (II.40).

La trame de la démonstration suit [19] § 3.2. Elle repose ici sur le théorème 57, un résultat délicat démontré à l'appendice E. Appliquons ce théorème avec

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

6. 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 décorrélés mais non indépendants), voir [78].

En remarquant que $\|\nabla K(\theta_*, Y_1)\| < \dot{K}(Y_1)$ p.s., et que, sur $\mathscr{V}$, $|k(\theta) - k(\theta')| \leqslant E[\dot{K}(Y_1)]\|\theta - \theta'\|$, on obtient $C(\dot{L}) \leqslant E[(2\dot{K}(Y_1) + E[\dot{K}(Y_1)])^2]^{1/2}$. Pour le calcul de $C(L)$, qui dépend de $r \leqslant r_0$, on a

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

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

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

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

$$E\Big[\sup_{\|h\|<r} n^{-1}\Big|\sum_{i=1}^n L_i(h)\Big|^2\Big] \leqslant C\log\Big(\frac{1}{\eta(r)}\Big) r^2\eta(r)^2 \leqslant r^2\widetilde{\eta}(r)^2 \tag{II.41}$$

pour une autre fonction $\widetilde{\eta}$ 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 θ tel que $\|\theta - \theta_*\| \leqslant r$:

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

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

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

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

$$\begin{aligned} &\big|M_n(\widehat{\theta}_n) - M_n(\theta_*) - \nabla M_n(\theta_*)(\widehat{\theta}_n-\theta_*) - \tfrac{1}{2}(\widehat{\theta}_n-\theta_*)^T J(\widehat{\theta}_n-\theta_*)\big| \leqslant n^{-\frac{1}{2}} r\widetilde{\eta}(r) X_r + o(\|\widehat{\theta}_n-\theta_*\|^2) \\ &\big|M_n(\theta') - M_n(\theta_*) + \tfrac{1}{2}\nabla M_n(\theta_*)^T J^{-1}\nabla M_n(\theta_*)\big| \leqslant n^{-\frac{1}{2}} r\widetilde{\eta}(r) X_r + o(\|\nabla M_n(\theta_*)\|^2) \end{aligned}$$

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

$$\begin{aligned} -M_n(\widehat{\theta}_n) + M_n(\theta') + \tfrac{1}{2}\Big(\widehat{\theta}_n - \theta_* + J^{-1}\nabla M_n(\theta_*)\Big)^T J\Big(\widehat{\theta}_n - \theta_* + J^{-1}\nabla M_n(\theta_*)\Big) \\ \leqslant 2n^{-\frac{1}{2}} r\widetilde{\eta}(r) X_r + o(\|\widehat{\theta}_n-\theta_*\|^2) + o(\|\nabla M_n(\theta_*)\|^2) \end{aligned}$$

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

$$\begin{aligned} &\tfrac{1}{2}\Big(\widehat{\theta}_n - \theta_* + J^{-1}\nabla M_n(\theta_*)\Big)^T J\Big(\widehat{\theta}_n - \theta_* + J^{-1}\nabla M_n(\theta_*)\Big) \\ &\qquad \leqslant 2n^{-\frac{1}{2}} r\widetilde{\eta}(r) X_r + o(\|\widehat{\theta}_n-\theta_*\|^2) + o(\|\nabla M_n(\theta_*)\|^2) + n^{-1}o_P(1) \\ &\qquad = 2n^{-\frac{1}{2}} r\widetilde{\eta}(r) X_r + o(\|\widehat{\theta}_n-\theta_*\|^2) + n^{-1}o_P(1) \end{aligned}$$

donc, on a finalement obtenu

$$1_{A_n\cap B_n}\|\widehat{\theta}_n - \theta_* + J^{-1}\nabla M_n(\theta_*)\|^2 \leqslant \frac{2}{\lambda_{\min}(J)} n^{-\frac{1}{2}} r\widetilde{\eta}(r) X_r + o(\|\widehat{\theta}_n-\theta_*\|^2) + n^{-1}o_P(1). \tag{II.42}$$

Noter que si l'on savait que $\|\widehat{\theta}_n - \theta_*\|^2 = O_P(1/n)$ [7], la démonstration serait terminée en choisissant $r = r_n = \alpha_n/\sqrt{n}$, où $\alpha_n \to \infty$ est choisi tel que $\alpha_n\widetilde{\eta}(\alpha_n/\sqrt{n})$ tende vers 0 (p. ex. la solution de $\alpha_n\widetilde{\eta}(\alpha_n/\sqrt{n}) = \sqrt{\widetilde{\eta}(1/\sqrt{n})}$ dans l'intervalle $[1, \sqrt{n}]$), car alors après multiplication par n, le membre de droite tend vers 0, tandis que $P(A_n \cap B_n) \to 1$.

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

Il ne reste donc plus qu'à démontrer que $\|\widehat{\theta}_n - \theta_*\|^2 = O_P(1/n)$. Pour simplifier un peu les calculs qui suivent, notons que le même raisonnement que précédemment avec cette fois

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

conduit aux mêmes équations sauf que η et $\widetilde{\eta}$ sont désormais simplement bornés et les termes en ∇M_n disparaissent, si bien que (II.42) devient une équation un peu plus simple (B_n a disparu) qui nous suffira :

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

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

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

Pour tout $R > 0$, la variable $U_n = 1_{|\xi_n| \leqslant R}\rho_n 1_{\rho_n \geqslant 1/2}\|\widehat{\theta}_n - \theta_*\|^2$ satisfait pour $u > 0$

$$E[1_{U_n \leqslant u^2} U_n] \leqslant E[1_{\|\widehat{\theta}_n - \theta_*\|^2 \leqslant 2u^2} U_n] \leqslant c n^{-\frac{1}{2}} u + n^{-1} R.$$

Ceci est a priori valide en fait seulement pour tout $u \leqslant \sqrt{2}\, r_0$, mais reste en fait vrai pour tout u car le terme central ne change pas lorsque u dépasse cette valeur. Par application du lemme 58 avec $C_1 = c n^{-\frac{1}{2}}$, $C_2 = n^{-1}R$ et $x = R^2/n$, on obtient que

$$P(nU_n > R^2) \leqslant (4c + 2)/R$$

et donc

$$P(n\rho_n 1_{\rho_n \geqslant 1/2}\|\theta_n - \theta_*\|^2 > R^2) \leqslant P(nU_n > R^2) + P(\xi_n > R) \leqslant \frac{4c+2}{R} + P(\xi_n > R)$$

qui tend vers 0 quand $R \to \infty$ uniformément en n. Les variables $n 1_{\rho_n \geqslant 1/2}\|\theta_n - \theta_*\|^2$ sont donc $O_P(1)$. Par ailleurs, les variables $n 1_{\rho_n < 1/2}\|\theta_n - \theta_*\|^2$, qui convergent en probabilité vers 0, sont également $O_P(1)$ (cf. note 7), ce qui achève la démonstration. ∎

II.3.5 Exercices et compléments

Exercice 1 (Régression non-linéaire). Étudier la variance asymptotique de l'estimateur des moindres carrés (II.30). On supposera que $Y_i - n_{\theta_*}(X_i)$ est indépendant de X_i.

Exercice 2 (Estimation de la médiane). Montrer que le théorème 15 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 Lebesgue sur un voisinage de θ_*. 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 $\widehat{\theta}_n$ l'estimateur de θ_* aux moindres carrés basé sur l'observation de $(X_k)_{0 \leqslant k \leqslant n}$. Proposer des conditions sur f pour avoir convergence presque sûre, normalité asymptotique (utiliser le théorème 60 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. 26. Montrer que $u = 1/p_{\theta_*}$ est optimal (on fera un calcul informel, oubliant les hypothèses techniques nécessaires).
Indication : Exprimer les matrices I et J en fonction de p_{θ_*}, $\partial_\theta p_{\theta_*}$ et de $v = u\partial_\theta p_{\theta_*} - \int u p_{\theta_*} \partial_\theta p_{\theta_*}$.

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

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

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).

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

$$\mathscr{L}(\theta, Y) = -\frac{1}{2}\sum_i \frac{(Y_i - M_i)^2}{V_i} - \log V_i, \quad \theta = (a,b,c)$$
$$M_i = aY_{i-1},$$
$$V_i = (b + c(Y_{i-1} - aY_{i-2}))^2.$$

Malheureusement, l'estimation va très mal se passer car la fonction $\mathscr{L}$ 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 [153]

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

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 qu'en définissant k_0 par

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

on a $k(\theta) \geqslant k_0(\theta)$ avec égalité seulement si $a = a_*$. Puis en utilisant que $\frac{a}{x} + \log(x) \geqslant 1 + \log(a)$, on montrera que $k_0(\theta)$ atteint son minimum en θ_*. En conclure que k est minimum en θ_*. 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 θ_* 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 ξ_k sont iid $\mathcal{N}(0,1)$. On sait que $f_* \in \Theta$ où

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

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

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

1. Montrer, en appliquant simplement un théorème d'analyse classique, que Θ est un ensemble compact pour la norme uniforme.
2. Montrer que presque sûrement $\|\widehat{f}_n - f_*\|_\infty \to 0$.

Exercice 7 (Vraisemblance non normalisée [98]). Score matching. On se place dans le cadre de l'exercice 4 p. 17, mais on propose une autre solution basée sur une fonction de contraste bien choisie. On dispose donc d'une famille paramétrique de densités $p_\theta(y) = e^{Q(y,\theta)}/Z(\theta)$, $y \in \mathbb{R}^d$, et l'on supposera dans la suite que les hypothèses techniques nécessaires au bon déroulement des intégrations et dérivations sont satisfaites. Par exemple $y \mapsto Q(y,\theta)$ peut être un réseau de neurones [154]. On note $p_*(y)$ la densité commune aux observations.

1. Montrer que pour tout θ la quantité

$$K(\theta, Y_1) = \Delta_y Q(Y_1, \theta) + \tfrac{1}{2}\|\nabla_y Q(Y_1, \theta)\|^2$$

est, à une constante indépendante de θ près, un estimateur non biaisé de

$$k(\theta) = \tfrac{1}{2}\int \|\nabla_y \log p_\theta(Y_1) - \nabla \log p_*(y)\|^2 p_*(y) dy. \tag{II.43}$$

Rq. : Le terme « score matching » couramment employé est impropre puisque le score est la dérivée par rapport au paramètre et non la rapport à la variable.

2. En déduire un algorithme d'estimation de θ ne faisant pas intervenir $Z(\theta)$. Pour justifier, on pourra se placer dans le cas où p_* appartient à la famille $(p_\theta)_{\theta\in\Theta}$, et donner des arguments informels.

Exercice 8 (Une application du critère d'Akaike-Takeuchi). Soit le modèle

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

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

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

1. Proposer un estimateur à minimum de contraste (Noter l'intérêt d'avoir retranché le terme $f(t_i)^2$).
2. Proposer un estimateur $\widehat{\sigma}^2$ de la variance commune aux e_{ij}, valide dès que $J \geqslant 2$ (commencer par le cas $J = 2$). Attention : on ne suppose pas que f appartienne à la famille g_θ.
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 $\widehat{\sigma}^2$.

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 θ qui minimise une certaine distance entre des statistiques observées (moments) et des statistiques prédites par θ. 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 χ^2

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

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

les probabilités empiriques. Les probabilités d'apparition des modalités sont paramétrées par un vecteur $\theta \in \Theta \subset \mathbb{R}^d$, $d \leqslant m-1$; une telle situation a été présentée au début du § II.2.5, le modèle de Hardy-Weinberg. On a donc m fonctions $p_1(\theta), \ldots p_m(\theta)$ dont la somme fait 1. L'estimateur au minimum de χ^2 est :

$$\widehat{\theta}_n = \operatorname*{argmin}_\theta n\sum_{j=1}^m \frac{(\widehat{p}_j - p_j(\theta))^2}{p_j(\theta)} = \operatorname*{argmin}_\theta \sum_{j=1}^m \frac{\widehat{p}_j^2}{p_j(\theta)}. \tag{II.44}$$

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 $\widehat{p}_j - p_j(\theta)$, 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 $p(\theta_*)$, $\widehat{\theta}_n$ est asymptotiquement gaussien et que la statistique

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

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

II.4.2 Définition et convergence

Un estimateur des moments généralisé est un estimateur de la forme

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}}\, H_n(\theta)^T S_n H_n(\theta), \tag{II.45}$$

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} H(\theta, Y_i) \tag{II.46}$$

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 :

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

où $\theta_\circ$ est une estimée initiale.

16 - 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 θ_* tel que $h(\theta_*) = 0$, alors $h(\widehat{\theta}_n)$ converge presque sûrement vers 0, et si θ_* est unique, $\widehat{\theta}_n$ converge presque sûrement vers θ_*.

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 $H_n(\widehat{\theta}_n)^T S_n H_n(\widehat{\theta}_n)$ aussi, et donc $H_n(\widehat{\theta}_n)$ tend vers zéro. Il suffit maintenant d'appliquer le théorème 9 et le théorème 10. ∎

17 - Théorème

On suppose que H satisfait l'hypothèse (FE) (p. 36). 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 $\widehat{\theta}_n$ converge en probabilité vers $\theta_* \in int(\Theta)$. Si

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

alors

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, J_*^{-1}J^T S I S J J_*^{-1}), \quad J_* = J^T S J \tag{II.47}$$

$$\sqrt{n}\, H_n(\widehat{\theta}_n) \xrightarrow{d} \mathcal{N}(0, QIQ^T), \quad Q = Id - JJ_*^{-1}J^T S. \tag{II.48}$$

Si I est inversible, un choix de S qui minimise la variance dans (II.47) 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

$$nH_n(\widehat{\theta}_n)^T S_n H_n(\widehat{\theta}_n) \xrightarrow{d} \chi^2_{Rang(I)-dim(\theta)}.$$

Démonstration. Rappelons que (FE) implique que $\nabla_\theta H_n(\widehat{\theta}_n) \to J$ p.s. (proposition 7). Il suffit alors d'appliquer le théorème 8 avec

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

Les hypothèses sont immédiatement vérifiées avec $\mathcal{D} = \mathcal{N}(0, J^T SISJ)$ et $J_* = J^T SJ$, et (II.47) s'ensuit. Pour (II.48) nous exploitons (II.6) et massivement le lemme de Slutsky :

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

ce qui prouve (II.48).

Pour l'optimalité de l'inverse de I nous renvoyons à l'exercice 1 p. 51. 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 $n\|RH_n(\widehat{\theta}_n)\|^2$; le vecteur $\sqrt{n}RH_n(\widehat{\theta}_n)$ converge en loi vers $\mathcal{N}(0, P)$, $P = RQIQ^T R$. On vérifie facilement que $QIQ^T = I - IK(K^T IK)^{-1}K^T I$, 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 χ^2 à $Trace(P) = Trace(IS) - dim(\theta)$ 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 11 donne bien la variance attendue (cf. aussi le § II.2.5).

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

Retour au minimum de χ^2. On trouve $I = \mathrm{Diag}(p(\theta_*)) - p(\theta_*)p(\theta_*)^T$ qui n'est pas inversible car $I\mathbf{1}_d = 0$. La matrice $S = \mathrm{Diag}(p(\theta_*))^{-1}$ satisfait bien $ISI = I$. L'hypothèse $J^T SJ > 0$ 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 θ_* est inconnu, on remplace θ_* par θ dans l'expression de S, ce qui conduit à (II.44). 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 17. On suppose l'existence d'une solution K à l'équation $IK = J$. Utiliser l'équation (II.26) pour démontrer que

$$(K^T IK)^{-1} \leqslant (J^T SJ)^{-1}J^T SISJ(J^T SJ)^{-1}.$$

Exercice 2 (Minimum de χ^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 $H(\theta, Y_i) = (1_{Y_i=k} - p_k(\theta))_{1\leqslant k\leqslant m}$. Calculer I et montrer (informellement) que le choix $S = Diag(p_k(\theta_*)^{-1}_{1\leqslant k\leqslant m})$ dans (II.45) conduit à une variance optimale.
   θ_* étant inconnu, on choisit de résoudre l'équation (II.44), ce qui fait l'objet du reste de l'exercice.
2. Quelle est la fonction H pour que (II.44) ait la forme (II.45) avec $S_n = Id$?
3. Quelle est, pour tout θ, 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 χ^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 θ_* satisfait une équation du type $g(\theta_*) = 0$, on pourra utiliser le minimum de contraste sous contraintes :

$$\widehat{\theta}_n^c = \arg \min_{g(\theta)=0} K_n(\theta). \tag{II.49}$$

On peut avoir

$$K_n(\theta) = \frac{1}{n} \sum K(\theta, Y_i), \tag{II.50}$$

mais on peut aussi définir $\widehat{\theta}_n^c$ par projection d'un premier estimateur $\widehat{\theta}_n$ en appliquant (II.49) avec

$$K_n(\theta) = (\widehat{\theta}_n - \theta)^T S (\widehat{\theta}_n - \theta) \tag{II.51}$$

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.50) et (II.51), 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.50), et toujours satisfaite dans le cas (II.51) 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. 57.

L'application des résultats de cette section aux tests de significativité relatifs à l'hypothèse nulle « $g(\theta_*) = 0$ » est présentée au § III.5.

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

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

$$\widehat{\theta}_n^c = \arg \min_{Rg(\theta)=r} \|\theta - \bar{Y}\|_F^2, \tag{II.52}$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \tag{II.53}$$

$$\|M\|_F^2 = \sum_{ij} M_{ij}^2 \qquad \text{(norme de Frobenius)}. \tag{II.54}$$

La solution de (II.52) s'obtient classiquement au travers de la décomposition en valeurs singulières de $\overline{Y}$, c'est la base de la théorie de l'analyse en composante principales. Cette écriture ressemble à (II.51), avec $\widehat{\theta}_n = \overline{Y}$, mais on a aussi, par une décomposition élémentaire,

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

qui correspond maintenant à (II.50).

II.5.2 Convergence

18 - Théorème

On suppose que les fonctions K_n convergent uniformément sur Θ vers une fonction $k(\theta)$ ayant un minimum unique θ_* sur Θ, que $g(\theta_*) = 0$ et que g est continue. Alors $\widehat{\theta}_n^c$ solution de (II.49) converge presque sûrement vers θ_*.

Remarque. Pour le cas (II.50), 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 Θ. ∎

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

$$\nabla K_n(\widehat{\theta}_n^c) + \lambda_n^T \nabla g(\widehat{\theta}_n^c) = 0$$
$$g(\widehat{\theta}_n^c) = 0$$

où λ_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 (θ, λ). On utilisera les hypothèses suivantes

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

(EC2) *Il existe un voisinage compact Θ_* de θ_* tel que presque sûrement la fonction $\theta \to K_n(\theta)$ est de classe C^2 sur Θ^*, et les fonctions $\theta \to \nabla^2 K_n(\theta)$ y convergent uniformément vers $\nabla^2 k$* [8].

Remarque. (EC2) peut être vérifiée à l'aide du théorème 6. Ces hypothèses font que le théorème des fonctions implicites permet de faire, au voisinage de θ_*, un changement de paramétrisation $\theta \mapsto \theta'$ tel que $(\theta'_i)_{i \leqslant m} = \nabla g(\theta)$; ceci ramène le problème au cas fondamental où les contraintes sont simplement 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.

19 - Théorème

On se place sous les hypothèses du théorème 18 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

$$\widehat{\theta}_n^c - \theta_* = P(\widehat{\theta}_n - \theta_*) + \frac{r''_n}{\sqrt{n}} \tag{II.55}$$

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

$$P = Id - J^{-1}G^T(GJ^{-1}G^T)^{-1}G \tag{II.56}$$

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

$$V_c = PJ^{-1}IJ^{-1}P^T. \tag{II.57}$$

Démonstration. Comme $\widehat{\theta}_n^c$ converge vers θ_*, alors $(\nabla K_n(\widehat{\theta}_n^c), \nabla g(\widehat{\theta}_n^c))$ converge vers $(0, \nabla g(\theta_*))$ (cf. la

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

proposition 7) 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 8 avec

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

Si $\widehat{\theta}_n^c$ n'est pas encore dans le voisinage de θ_* 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 θ au lieu de $\widehat{\theta}_n^c$ dans ∇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

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

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.56) 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. $V_cG = 0$), en raison des contraintes satisfaites. On a également $V_c \leqslant V$ 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

$$\boxed{\widehat{\theta}_n^c = \arg\min_{g(\theta)=0} \tfrac{1}{2}(\theta - \widehat{\theta}_n)^T V^{-1}(\theta - \widehat{\theta}_n)} \tag{II.58}$$

où V est la variance de $\sqrt{n}(\widehat{\theta}_n - \theta_*)$ (ou une estimée...). On voit facilement que $I = J = V^{-1}$, et (II.57) et (II.55) deviennent

$$V_c = V - VG^T(GVG^T)^{-1}GV \tag{II.59}$$

$$\widehat{\theta}_n^c - \theta_* = V_cV^{-1}(\widehat{\theta}_n - \theta*) + \frac{r_n''}{\sqrt{n}}. \tag{II.60}$$

Le choix proposé pour V est optimal, car un autre choix, disons W, pour la matrice de pondération de (II.58) conduit, en vertu de (II.57), à une variance de

$$\big(Id - WG^T(GWG^T)^{-1}G\big)V\big(Id - G^T(GWG^T)^{-1}GW\big) \tag{II.61}$$

qui est supérieure à V_c (appliquer l'exercice 3 p. 57 avec $A = WG^T(GWG^T)^{-1}$).

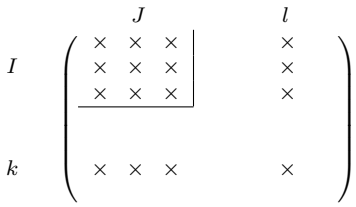
Noter que V ne dépend pas de g, ce qui va être exploité à la fin du § II.5.3 pour le calcul effectif de $\widehat{\theta}_n^c$.

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 $\mathrm{Rang}(\theta) = r$ se réécrit au voisinage de θ_* sous la forme $g(\theta) = 0$ où g est C^1 et ∇g de rang correspondant à son nombre de composantes. Pour ceci, notons que θ_* 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 θ_*, $\theta_{I,J}$ est, par continuité du déterminant, toujours de rang plein. Notons le $\mathscr{V}$. Sur $\mathscr{V}$, les matrices de rang r vont satisfaire bien entendu

$$\det((\theta_{ij})_{i\in I\cup\{k\}, j\in J\cup\{l\}}) = 0, \quad k \notin I,\ l \notin J \tag{II.62}$$



Réciproquement, si une matrice $\theta \in \mathscr{V}$ satisfait cet ensemble (II.62) de $(p-r)(q-r)$ conditions, alors le mineur θ_{IJ} 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.52) 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. [54] § 3.3 ; pour la convergence voir [115]), 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 : « $g(\theta_*)=0$ », le membre de droite de (II.63). 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.58) ; on va voir également un exemple en régression non-linéaire.

20 - Théorème

On se place sous les hypothèses du théorème 19 avec toujours $J=\nabla^2 k(\theta_*)$. Si $\sqrt{n}\,\nabla K_n(\theta_*)$ converge en loi vers $\mathcal{N}(0,J)$, alors

$$2n\big(K_n(\widehat{\theta}_n^c)-K_n(\widehat{\theta}_n)\big) \longrightarrow \chi^2_m \tag{II.63}$$

où m est la dimension de g, c.-à-d. la différence de dimension effective entre θ et θ^c.

Le cas $g(\theta)=\theta-\theta_*$ conduit à la région de confiance :

$$\mathcal{R}_\alpha = \Big\{\theta : 2\sum_{i=1}^{n} K_n(\theta) - K_n(\widehat{\theta}_n) \leqslant \chi^2_d(1-\alpha)\Big\}.$$

Si initialement $I \neq J$ et que l'on utilise (II.58), (II.63) devient

$$\boxed{n(\widehat{\theta}_n^c-\widehat{\theta}_n)^T J I^{-1} J (\widehat{\theta}_n^c-\widehat{\theta}_n) \longrightarrow \chi^2_m} \tag{II.64}$$

qui est utilisable en toute circonstance où I et J peuvent être raisonnablement estimées.

Démonstration. On peut appliquer le théorème 19. L'équation (II.55) donne

$$\begin{aligned}\widehat{\theta}_n^c-\widehat{\theta}_n &= -(Id-P)(\widehat{\theta}_n-\theta_*)+\frac{r_n''}{\sqrt{n}}\\ &= -J^{-1}G^T(GJ^{-1}G^T)^{-1}G(\widehat{\theta}_n-\theta_*)+\frac{r_n''}{\sqrt{n}}.\end{aligned}$$

Comme $\widehat{\theta}_n - \theta_* = n^{-1/2} J^{-1/2} X_n$ avec $X_n \to \mathcal{N}(0, Id)$, il vient,

$$\sqrt{n}(\widehat{\theta}_n^c - \widehat{\theta}_n) = -J^{-1} G^T (G J^{-1} G^T)^{-1} G J^{-1/2} X_n + r_n''$$

Par ailleurs, la formule de Taylor au voisinage de $\widehat{\theta}_n$ donne, pour un certain θ_n' du segment $[\widehat{\theta}_n - \widehat{\theta}_n^c]$

$$\begin{aligned} 2n\big(K_n(\widehat{\theta}_n^c) - K(\widehat{\theta}_n)\big) &= n(\widehat{\theta}_n - \widehat{\theta}_n^c)^T \nabla_\theta^2 K_n(\theta_n')(\widehat{\theta}_n - \widehat{\theta}_n^c) \\ &= \sqrt{n}(\widehat{\theta}_n - \widehat{\theta}_n^c)^T (J + r_n) \sqrt{n}(\widehat{\theta}_n - \widehat{\theta}_n^c) \end{aligned}$$

par application de la proposition 7. En remplaçant, on trouve que la loi limite est la même que celle de

$$\begin{aligned} X_n^T J^{-1/2} G^T (G J^{-1} G^T)^{-1} G J^{-1} G^T (G J^{-1} G^T)^{-1} G J^{-1/2} X_n \\ = X_n^T J^{-1/2} G^T (G J^{-1} G^T)^{-1} G J^{-1/2} X_n \end{aligned}$$

qui est le carré de la norme de PX_n, $P = (GJ^{-1}G^T)^{-1/2} G J^{-1/2}$. Comme $PP^T = Id_m$, la loi de PX_∞ 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 Θ 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 $\widehat{\theta}_n$ reste confinée dans un compact de Θ (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.31) p. 42 :

$$\begin{aligned} &K_n(\theta) = \tfrac{1}{n} \sum_{i=1}^n \tau \psi\big(Y_i - n_\theta(X_i)\big) \\ &\widehat{\theta}_R = \arg\min_\theta K_n(\theta) \\ &\tau = E\big[\psi''(e_1)\big] / E\big[\psi'(e_1)^2\big], \end{aligned} \tag{II.65}$$

la constante τ 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 τ sont satisfaites ainsi que les hypothèses du théorème 19 [9]. Alors dans ce cas on a bien $I = J$ et le théorème de Wilks s'applique. La constante τ 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.65). Donner des conditions sur la fonction $n_\theta(x)$ assurant que les hypothèses du théorème 19 seront satisfaites.

Exercice 2 (Une application du critère d'Akaike-Takeuchi [142]). On reprend l'exemple (II.65), sans constante τ ajoutée en facteur du contraste. Montrer que le critère de Takeuchi (page 44) devient

$$TIC(d) = \sum_{i=1}^n \psi(Y_i - n_{\widehat{\theta}^d}(X_i)) + d \frac{E[\psi'(e_1)^2]}{E[\psi''(e_1)]}.$$

9. A propos de $E\big[\psi''(e_1)\big]$. Si l'on utilise effectivement la fonction ψ donnée p. 42, 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 $E\big[\psi''(e_1)\big] = 2(p_e(-\alpha) + p_e(\alpha))$; on obtient le même résultat en calculant $\frac{d}{dt} E\big[\psi'(e_1 + t)\big]_{|t=0}$.

Exercice 3 (Démonstration de l'optimalité du choix de V dans la méthode en deux temps p.54). Soit A, G, S trois matrices, S étant définie positive. Montrer que

$$(Id - AG)S(Id - AG)^T \geqslant S - SG^T(GSG^T)^{-1}GS$$

si l'inverse existe (i. e. les lignes de G sont indépendantes), 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 :

$$H_n(\widehat{\theta}_n^c)^T + \lambda_n^T \nabla g(\widehat{\theta}_n^c) = 0$$
$$g(\widehat{\theta}_n^c) = 0.$$

On suppose que $(\widehat{\theta}_n^c, \lambda_n)$ 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 $\widehat{\theta}_n^c$ (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 $H_n(\widehat{\theta}_n^c)^T$ par $H_n(\widehat{\theta}_n^c)^T M^T$ 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.2.5). Utiliser l'exercice 3 ci-dessus.

II.6 Applications aux tests

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 [145]) mais on s'intéressera essentiellement à leur niveau asymptotique, dans le cas d'une hypothèse $g(\theta) = 0$ contre son contraire.

$$H_0 : g(\theta_*) = 0, \qquad H_1 : g(\theta_*) \neq 0$$

pour une certaine fonction g à valeurs dans $\mathbb{R}^q$. On notera $\widehat{\theta}^c$ l'estimateur sous la contrainte $g(\theta) = 0$.

II.6.1 Tests passant par un estimateur sous contrainte.

► Contraste quadratique. On considère l'estimateur en deux temps $\widehat{\theta}_c$ de la page 54. On a le test des scores (ou d'Hausman puisque le contraste est quadratique)

Rejeter H_0 si $\quad (\widehat{\theta}^c - \widehat{\theta})^T V^{-1} (\widehat{\theta}^c - \widehat{\theta}) \ \geqslant \ \chi^2_q(1-\alpha)$

où V est la variance de $\widehat{\theta} - \theta_*$ (ou une estimée). La validité de ce test est une conséquence de (II.60).

► Contraste général. On est ici dans le cadre du § II.5 avec les Y_i iid. Notons $H = \nabla K$, $\widehat{G} = \nabla g(\widehat{\theta}^c)$, $\widehat{J} = \sum_{i=1}^n \nabla H(\widehat{\theta}^c, Y_i)$ et la matrice d'information non-normalisée estimée $\widehat{I} = \sum_{i=1}^n H(\widehat{\theta}^c, Y_i) H(\widehat{\theta}^c, Y_i)^T$. On peut aussi les estimer en $\widehat{\theta}$. Le test de Wald devient :

Rejeter H_0 si $\quad g(\widehat{\theta})^T (\widehat{G}\widehat{J}^{-1}\widehat{I}\widehat{J}^{-T}\widehat{G}^T)^{-1} g(\widehat{\theta}) \ \geqslant \ \chi^2_q(1-\alpha)$

Dans le cas d'un minimum de contraste, le test du multiplicateur de Lagrange se déduit de la normalité de ce dernier car $\widehat{\lambda} = -\frac{1}{n}(\widehat{G}\widehat{G}^T)^{-1}\widehat{G}\sum_{i=1}^n \nabla K(\widehat{\theta}^c, Y_i)^T$ (cf. théorème 19) :

Rejeter H_0 si $\quad \widehat{\lambda}\widehat{G}\widehat{J}^{-1}\widehat{G}^T(\widehat{G}\widehat{J}^{-1}\widehat{I}\widehat{J}^{-1}\widehat{G}^T)^{-1}\widehat{G}\widehat{J}^{-1}\widehat{G}^T\widehat{\lambda} \ \geqslant \ \chi^2_q(1-\alpha)$

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 20) s'applique (cf. l'estimateur (II.65), ou encore (II.58)) ce qui conduit à

$$\boxed{\text{Rejeter } H_0 \text{ si} \quad 2\sum_i K(\widehat{\theta}^c) - K(\widehat{\theta}) \;\geqslant\; \chi^2_q(1-\alpha)} \qquad (\text{cas } I = J).$$

II.6.2 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 θ_0 une fois pour toutes mais où des tests réguliers (journaliers...) doivent être faits de manière simple et robuste[10]. 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 θ_0. *On va dans ce paragraphe normaliser les quantités, contrairement au paragraphe précédent.*

► SURVEILLANCE. Plaçons-nous dans la situation paramétrique[11], 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 $\widehat{\theta}^c = \theta_0$ 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.15)

$$\boxed{\begin{array}{l} H_n(\theta_0)^T I^{-1} H_n(\theta_0) \;\leqslant\; \chi^2_d(1-\alpha) \\ \quad H_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n H(\theta, Y_i) \\ \quad I = Var_{\theta_0}(H_n(\theta_0)) \end{array}} \qquad (\text{II.66})$$

► APPROCHE LOCALE. CAS D'UN SURPLUS DE FONCTIONS D'ESTIMATION. Pour étudier la puissance considérons l'alternative asymptotique locale H_δ : « $\theta = \theta_0 + \delta/\sqrt{n}$ ». Sous cette alternative on devine que l'on a, sous des hypothèses adéquates, la convergence en loi

$$H_n(\theta_0) = H_n(\theta_0) - H_n(\theta_0 + \delta/\sqrt{n}) + H_n(\theta_0 + \delta/\sqrt{n}) \xrightarrow{Loi} \mathcal{N}(-J\delta, I).$$

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 (II.66), qui doit maintenant être comparée à un $\chi^2_p(1-\alpha)$, est donc asymptotiquement sous H_δ 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

$$\boxed{H_n(\theta_0)^T I^{-1} J (J^T I^{-1} J)^{-1} J^T I^{-1} H_n(\theta_0) \;\leqslant\; \chi^2_d(1-\alpha) \quad (\text{cas } p > d)}$$

(c'est le test précédent appliqué à la fonction d'estimation fusionnée $H' = J^T I^{-1} H$, cf. le § II.2.5) est maintenant sous H_δ 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_δ est inchangé (il vaut $\delta J^T I^{-1} J\delta$) mais la 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 $g(\widehat{\theta}) \simeq G(\widehat{\theta} - \theta_0) \simeq -GJ^{-1}H_n(\theta_0)/\sqrt{n}$ (cf. la formule (II.6)). Sous H_0 la variance du terme ci-dessus est explicite et l'on obtient le test de Wald

10. Nous reprenons ici les idées exposées dans [32]. Voir les références pour des applications aux structures mécaniques soumises à des vibrations.

11. Ce n'est en fait pas nécessaire car on utilise essentiellement que I et J puissent être estimés.

$$H_n(\theta_0)^T J^{-T} G^T (G J^{-1} I J^{-T} G^T)^{-1} G J^{-1} H_n(\theta_0) \lesssim \chi^2_q(1-\alpha), \quad G = \nabla g(\theta_0)$$

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

$$H_n(\theta_0)^T I^{-1} J_1 (J_1^T I^{-1} J_1)^{-1} J_1^T I^{-1} H_n(\theta_0) \lesssim \chi^2_q(1-\alpha)$$
$$J_1 = J(J^T I^{-1} J)^{-1} G^T$$

L'espérance de la statistique sous H_δ est (asymptotiquement) $q + \delta^T G^T (G(J^T I^{-1} J)^{-1} G^T)^{-1} G\delta$.

II.7 Bornes de grandes déviations et estimées des moments

La convergence en loi de $\sqrt{n}(\widehat{\theta}_n - \theta_*)$ 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 $\widehat{\theta}$ est un estimateur à minimum de contraste, l'idée va être de majorer

$$\begin{aligned} E[f(\widehat{\theta})] &\leqslant E[f(\widehat{\theta}) \exp\{K(\theta_*, \omega) - K(\widehat{\theta}, \omega)\}] \\ &\leqslant E[\sup_\theta f(\theta) \exp\{K(\theta_*, \omega) - K(\theta, \omega)\}] \end{aligned}$$

où typiquement $f(\widehat{\theta}) = \|\widehat{\theta} - \theta_*\|^p$. Le sup est ensuite remplacé par une intégrale via l'inégalité de Sobolev (appendice D), et le calcul se termine après permutation de l'intégrale et de l'espérance.

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, σ et η qui apparaissent ensuite, et de mieux synthétiser l'information :

21 - Lemme

On suppose que $T \subset \mathbb{R}^d$ est la fermeture de son intérieur, lequel possède la propriété de cône (p. ex. convexe borné , cf. appendice D). 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 $t \in T$

$$E\big[\|\nabla_t L(t)\|^q e^{-L(t)} 1_{L(t) \geqslant 0}\big] \leqslant M\big(1 + \|t\|\big)^q e^{-\|t\|^2/2} \tag{II.67}$$
$$E\big[e^{-L(t)_+}\big] \leqslant M e^{-\|t\|^2/2} \tag{II.68}$$

où l'on a noté $L(t)$ pour $L(t, \omega)$. Soit une variable aléatoire $\widehat{t}(\omega)$ satisfaisant

$$L(\widehat{t}(\omega), \omega) \leqslant 0 \quad p.s.$$

alors il existe une constante $C_{q,T}$ ne dépendant que de q et de T telle que pour tout $1 \leqslant i \leqslant d$ et tout $\alpha \geqslant 0$

$$E\big[\exp(\alpha \widehat{t}_i(\omega))\big] \leqslant C_{q,T} M (1 + \alpha^d) e^{\frac{\alpha^2}{2}}. \tag{II.69}$$

La constante $C_{q,T}$ satisfait la propriété suivante : pour tout $t \in \mathbb{R}^d$ et $a \geqslant 1$, on a $C_{q,T+t} = C_{q,T}$ et $C_{q,aT} \leqslant C_{q,T}$.

Démonstration. Posons pour tout $\delta > 0$

$$r(t) = r(t,\omega) = e^{-\rho(L(t,\omega))}, \qquad \rho(x) = \frac{x^2 1_{x>0}}{\delta + x}$$
$$f(t) = \exp\{\alpha\,(t_i - \alpha)\}\,.$$

La fonction $\rho(x)$ a été choisie de sorte à être $\geqslant 0$, croissante, dérivable, nulle sur $\mathbb{R}_-$, à dérivée $\leqslant 1$ et telle que $\|\rho(x) - x_+\|_\infty \leqslant \delta$. En vertu de l'inégalité de Sobolev, éq. (D.2) p.125, et puisque $r(\widehat{t}) = 1$ on a

$$\begin{aligned}
E\big[f(\widehat{t})\big] &= E\big[f(\widehat{t})r(\widehat{t})\big] \\
&\leqslant E[\|rf\|_\infty] \\
&\leqslant C_q\varepsilon^{-d}E\int\Big(\varepsilon^q\|\nabla\log(r(z)f(z))\|^q + 1\Big)r(z)f(z)dz \\
&\leqslant 2^qC_q\varepsilon^{-d}E\int\Big(\varepsilon^q\|\nabla L(z,\omega)\|^q 1_{L(z,\omega)\geqslant 0} + \alpha^q\varepsilon^q + 1\Big)r(z)f(z)dz
\end{aligned}$$

d'où en faisant tendre δ vers 0 (et en notant e_i le i-ième vecteur de la base canonique) :

$$\begin{aligned}
E\big[f(\widehat{t})\big] &\leqslant 2^qC_q\varepsilon^{-d}E\int\Big(\varepsilon^q\|\nabla L(z,\omega)\|^q 1_{L(z,\omega)\geqslant 0} + 1 + \alpha^q\varepsilon^q\Big)e^{-L(z,\omega)_+}f(z)dz \\
&\leqslant 2^qC_q\varepsilon^{-d}\int\Big\{\varepsilon^q M(1+\|z\|)^q + M + M\varepsilon^q\alpha^q\Big\}e^{-\frac{\|z\|^2}{2}+\alpha(z_i-\alpha)}dz \\
&= 2^qC_q\varepsilon^{-d}e^{-\frac{\alpha^2}{2}}M\int\Big\{\varepsilon^q(1+\|z+\alpha e_i\|^q) + +\varepsilon^q\alpha^q\Big\}e^{-\|z\|^2/2}dz \\
&\leqslant C'_q\varepsilon^{-d}e^{-\frac{\alpha^2}{2}}M\big(\varepsilon^q(1+\alpha^q) + 1 + \varepsilon^q\alpha^q\big).
\end{aligned}$$

Il ne reste plus qu'à prendre $\varepsilon = 1/(1+\alpha)$ (qui est bien $\leqslant 1$).

La dernière affirmation du lemme vient de ce que $T + t_0$ et T sont isométriques et de que aT satisfait une meilleure propriété de cône que T. ∎

22 - Théorème

On suppose que $\Theta \subset \mathbb{R}^d$ est la fermeture de son intérieur, lequel possède la propriété de cône (p. ex. convexe borné, cf. appendice D). Θ 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 θ. On suppose que pour un $q > d$, certains $M, \sigma, \eta > 0$, un $\theta_* \in \Theta$, pour tout $\theta \in \Theta$

$$\begin{aligned}
&E\big[\|\nabla_\theta K_n(\theta)\|^q e^{-\eta(K_n(\theta)-K_n(\theta_*))}1_{K_n(\theta)\geqslant K_n(\theta_*)}\big] \\
&\qquad\qquad \leqslant M\big(\sqrt{n} + n\|\theta-\theta_*\|\big)^q e^{-n\|\theta-\theta_*\|^2/2\sigma^2} \qquad \text{(II.70)}\\
&E\big[e^{-\eta(K_n(\theta)-K_n(\theta_*))_+}\big] \leqslant Me^{-n\|\theta-\theta_*\|^2/2\sigma^2} \qquad \text{(II.71)}
\end{aligned}$$

où l'on a noté $K_n(\theta)$ pour $K_n(\theta,\omega)$. Soit $\widehat{\theta}_n$ qui satisfait

$$K_n(\widehat{\theta}_n,\omega) \leqslant K_n(\theta_*,\omega) \quad p.s.$$

alors il existe une constante C dépendant des données (q, M, σ, η et Θ), mais pas de θ_*, telle que pour tout $1 \leqslant i \leqslant d$ et tout $\alpha \geqslant 0$

$$P\big(\sqrt{n}|\widehat{\theta}_{ni} - \theta_{*i}| \geqslant \alpha\big) \leqslant C\big(1+\alpha^d\big)e^{-\frac{\alpha^2}{2\sigma^2}}. \qquad \text{(II.72)}$$

En particulier tout moment de $\sqrt{n}(\widehat{\theta}_n - \theta_*)$ est borné (quand n varie) et la suite $\sqrt{n}\|\widehat{\theta} - \theta_*\|/\sqrt{\log n}$ est bornée avec probabilité 1.

Démonstration. En appliquant le lemme 21 à la fonction

$$L(t,\omega) = \eta\Big(K_n\big(\frac{\sigma}{\sqrt{n}}t,\omega\big) - K_n(\theta_*,\omega)\Big)$$

sur $T = \sqrt{n}(\Theta - \theta_*)/\sigma$, et en posant $\widehat{t} = \sqrt{n}(\widehat{\theta}_n - \theta_*)/\sigma$, il vient

$$E\big[e^{\alpha\sqrt{n}(\widehat{\theta}_n-\theta_*)_i/\sigma}\big] \leqslant C_1(1+\alpha^d)e^{\frac{\alpha^2}{2}}.$$

où la constante C_1 dépend maintenant également de (M,σ,η). On en déduit

$$P\big(\sqrt{n}(\widehat{\theta}_n - \theta_*)_i \geqslant \alpha\big) \leqslant E\left[\exp\left\{\frac{\alpha\sqrt{n}}{\sigma^2}\left(\widehat{\theta}_i - \theta_{*i} - \alpha/\sqrt{n}\right)\right\}\right] \leqslant C_1(1+\alpha^d\sigma^{-d})e^{\frac{\alpha^2}{2\sigma^2}}.$$ ∎

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.29) sont $\mathcal{N}(0,\sigma_e^2)$, si $\theta \mapsto n_\theta(x)$ est de classe de classe $C^1(\Theta)$, et si

$$\forall x,\theta,\theta', \qquad C_1\|\theta-\theta'\| \leqslant |n_\theta(x) - n_{\theta'}(x)| \leqslant C_2\|\theta-\theta'\|$$

pour certains $0 < C_1 < C_2 < \infty$, alors les hypothèses plus fortes mais plus simples suivantes sont satisfaites pour l'estimateur (II.30)

$$E\big[\|\nabla_\theta K_n(\theta)\|^q\big] \leqslant M\big(\sqrt{n} + n\|\theta-\theta_*\|\big)^q, \quad \text{pour un } q > 2d$$
$$E\big[e^{-\eta(K_n(\theta)-K_n(\theta_*))}\big] \leqslant Me^{-\lambda n\|\theta-\theta_*\|^2}, \quad \eta = 1/\sigma_e^2,$$

les détails sont laissés en exercice. (II.70) se déduit alors de l'inégalité de Cauchy-Schwartz. Le théorème 22 s'applique donc dans ce cas.

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 θ_* 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.14)). 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 [13, 18, 11, 15], 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 $r(\theta, \theta') \geqslant 0$, nulle pour $\theta = \theta'$, typiquement $r(\theta, \theta') = |\theta - \theta'|^2$. Le risque d'un estimateur $\widehat{\theta}$ est la fonction suivante, perte moyenne de l'estimateur lorsque le vrai paramètre est θ :

$$R_\theta(\widehat{\theta}) = E_\theta[r(\theta, \widehat{\theta}(Y))].$$

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 $\widehat{\theta}$ est **admissible** s'il n'existe pas d'autre estimateur $\widehat{\theta}'$ strictement meilleur au sens où : $R_\theta(\widehat{\theta}') \leqslant R_\theta(\widehat{\theta})$ pour tout θ et $R_{\theta_0}(\widehat{\theta}') < R_{\theta_0}(\widehat{\theta})$ pour un certain θ_0. On verra que les estimateurs bayésiens sont généralement admissibles, ce qui fournit une grande classe d'exemples.

Par ailleurs, un estimateur $\widehat{\theta}$ est dit **minimax sur** Θ si et seulement si pour tout autre estimateur $\widehat{\theta}'$ on a

$$\sup_{\theta\in\Theta} R_\theta(\widehat{\theta}) \leqslant \sup_{\theta\in\Theta} R_\theta(\widehat{\theta}').$$

Pour éviter que les sup ne soient infinis, il faut généralement supposer Θ compact.

On verra dans l'exercice 10 p. 92 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

$$\forall\theta\in\Theta,\quad \forall\widehat{\theta}',\quad R_\theta(\widehat{\theta})\leqslant R_\theta(\widehat{\theta}')$$

où $\widehat{\theta}'$ 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 θ un minimum nul : il suffit de prendre $\widehat{\theta}'=\theta$. 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 $\widehat{\theta}'=\theta$).

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 π sur θ et à chercher un estimateur qui minimise le risque moyen $\int E_\theta[r(\theta,\widehat{\theta})]\pi(d\theta)$, donne une solution théoriquement simple au problème, voir § III.4.1. ; mais il reste à savoir quelle distribution poser sur θ. 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 $\widehat{\theta}$ est dit **r-non-biaisé** si et seulement si

$$\forall\theta,\theta'\in\Theta,\quad E_\theta[r(\theta,\widehat{\theta}(Y))]\leqslant E_\theta[r(\theta',\widehat{\theta}(Y))]. \tag{III.1}$$

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 θ_*.

Une statistique $S(Y)$ est dite **exhaustive** (ang. : **sufficient**) si la loi conditionnelle $P_\theta(Y|S(Y))$ ne dépend pas de θ[1] ; une fois que $S(Y)$ est connue, aucune autre information sur θ ne peut être inférée de Y, point qui sera précisé mathématiquement plus bas. Le problème d'estimation se réduit à estimer θ_* à partir d'une observation de S, ayant à sa disposition une famille lois (pour S) indexées par θ ; cette réduction diminue souvent considérablement le vecteur d'observation, car S aura typiquement la dimension de θ, 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** assure que S est une statistique exhaustive si et seulement s'il existe une factorisation de p_θ sous la forme

$$p_\theta(y)=q_\theta(S(y))p_0(y),$$

valide pour μ-presque tout y, où p_0 ne dépend pas de θ. En particulier, si S est exhaustive, l'estimateur du maximum de vraisemblance ne dépend que de S. On en trouvera une démonstration dans [13].

Si $\widehat{\theta}$ est un estimateur, alors $\breve{\theta}=E_{\theta_*}[\widehat{\theta}\,|S]$ est un meilleur estimateur de θ au sens de l'erreur quadratique (exercice 7 p. 66). C'est le **théorème de Rao-Blackwell** (Rao 1945, Blackwell 1947). En d'autres termes, le théorème de Rao-Blackwell assure que disposant d'une statistique suffisante S, on ne pourra extraire de l'échantillon davantage d'information utile à l'estimation de θ.

1. Pour tout ensemble mesurable A, il existe une fonction mesurable φ t.q. pour tout θ : $P_\theta(Y\in A|S(Y))=\varphi(S(Y))$ $P_\theta-$p.s.

Noter qu'en raison de l'exhaustivité, $\check{\theta}$ est effectivement calculable, c.-à-d. sans connaître θ_* (ce n'est malheureusement pas le cas de $E_{\theta_*}[\widehat{\theta}]$!) ; l'amélioration de $\widehat{\theta}$ 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. 66).

Une statistique exhaustive S est dite **minimale** si elle est fonction de toute autre statistique exhaustive [2]. Si une statistique exhaustive peut s'exprimer comme fonction mesurable des rapports $p_\theta(y)/p_{\theta_0}(y)$ pour θ appartenant à une famille dénombrable et θ_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 [3].

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_θ-presque sûrement pour tout θ ; 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. 66) 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 fausse 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. 67).

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 :

23 - Théorème (Lehmann-Scheffé, 1950)

Soit S une statistique exhaustive et complète et $g(\theta)$ une fonction de θ. Soit T un estimateur non-biaisé de variance finie de $g(\theta)$. Alors $T^* = E_\theta[T|S]$ ne dépend ni du choix de θ (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).

La démonstration est assez simple (exercice 8 p. 66). En particulier, pour toute fonction f, $f(S)$ est un estimateur UMVU de son espérance (appliquer le théorème à $T_f = f(S)$ et $g_f(\theta) = E_\theta[f(S)]$). 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. 67). 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. 91).
   (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. 66). L'intérêt des estimateurs biaisés augmente si θ est de grande dimension car on s'approche du cadre non paramétrique (p. ex. estimateurs à troncature pour le problème des *normal means*, estimateur lasso, estimateur ridge) ceci sort du cadre de ce cours.
   (c) Il peut ne pas exister d'estimateur non-biaisé (exercice 2 p. 66).
2. Il existe de nombreuses situations où il n'y a pas de statistique complète (exercices 6, 12 et 14).

2. Si T en est une autre, il existe φ telle que pour tout θ, p_θ-p.s., $S = \varphi(T)$.
3. On la produit théoriquement ainsi : on dit qu'une tribu $\mathscr{F}$ est exhaustive si $P_\theta(Y \in .|\mathscr{F})$ ne dépend pas de θ ; 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 [111].

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.

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 : S'intéresser à l'équation $E[\widehat{\theta}(X_1,\dots X_n)]=\frac{p}{1-p}$; 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 $\sum_{i\geqslant 0}|a_i|<\infty$ et si pour tout x d'un intervalle non vide de $]0,1[$ on a $\sum_{i\geqslant 0}a_i x^i=0$, alors les a_i sont tous nuls.*

Exercice 4 (Statistique exhaustive en régression). Montrer que dans le modèle (I.7) p. 9, $\widehat{\theta}$, éq. (I.10), est une statistique exhaustive.
Indication : Noter que le vecteur $Y-X\widehat{\theta}$ est orthogonal à toutes les colonnes de X.

Montrer que dans l'exemple (I.8) p. 9, $\sum Y_iX_i$ 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 λ. 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 $\widehat{\theta}=1_{X_1=0}$ est un estimateur non-biaisé de θ.
2. Calculer $P(X_1=0|S_n=k)$ pour tout $k\geqslant 0$.
3. Proposer un estimateur meilleur que $\widehat{\theta}$ 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_θ.

1. Vérifier que p a un maximum unique. *Indication* : $|\tanh x|\leqslant|x|$.
2. Montrer que X est une statistique exhaustive minimale.
   Indication : On pourra s'intéresser au maximum en θ 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 $\mathscr{F}$ une tribu, justifier l'inégalité $E[E[X|\mathscr{F}]^2]\leqslant E[X^2]$. 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 Θ 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 [4], 1961). Soit Y une observation de loi $\mathcal{N}(\mu,Id)$.

1. Montrer que l'estimateur au maximum de vraisemblance de μ est Y. En vertu de l'exercice 13 Y est une statistique exhaustive minimale et complète.

4. On trouvera des compléments intéressants au § 1 de [4] et au § 7.6 de [20], ainsi qu'une application au baseball et à la toxoplasmose dans [74]

2. On se place en dimension $d \geqslant 3$ et l'on considère l'estimateur de James et Stein :

$$\widehat{\mu}(Y) = \Big(1 - \frac{d-2}{\|Y\|^2}\Big)Y.$$

Montrer que son risque vaut

$$E\big[\|\widehat{\mu}(Y) - \mu\|^2\big] = d - (d-2)^2\ E\big[\|Y\|^{-2}\big].$$

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 μ 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. 83.

Exercice 11 (SURE-Shrinkage [67, 166]). On a vu, à l'exercice 10, que pour l'estimation de μ dans le modèle $Y \sim \mathcal{N}(\mu, \sigma^2 Id)$ (« normal means »), 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 σ, voire nuls, cas où 0 est plus proche de μ_i que Y_i. Soit l'estimateur pénalisé

$$\widehat{\mu} = \arg\min_{\mu} \frac{1}{2}\sum_i (Y_i - \widehat{\mu}_i)^2 + \lambda \sum_i |\widehat{\mu}_i|$$

où λ est un seuil à choisir. C'est un cas particulier de l'estimateur lasso (III.41). Montrer que

$$\widehat{\mu}_i(Y) = 1_{|Y_i|>\lambda}(Y_i - \lambda \operatorname{sign}(Y_i)).$$

Au vu de l'exercice 1 p. 16, on se propose de choisir λ qui minimise $\widehat{R}(\lambda)$ donné par (I.25). Montrer qu'ici

$$\widehat{R}(\lambda) = \sum_i Y_i^2 \wedge \lambda^2 + 2\sigma^2 \sum_i 1_{|Y_i|>\lambda} - n\sigma^2.$$

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 θ, 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_θ 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 θ 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 $|\mathrm{Re}(z)| + |h| < \varepsilon$, 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 $\sum$.

3. En déduire que f est analytique sur $|\mathrm{Re}(z)| < \varepsilon$ [5].

4. 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_θ ne dépend pas de θ. 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 θ, et S_0 aura la dimension de θ.

1. 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.
2. Montrer que si S est exhaustive minimale et s'il existe une fonction φ telle que $\varphi(S)$ soit une statistique libre non-constante, alors il n'existe pas de statistique exhaustive et complète.
3. En déduire qu'il n'existe pas de statistique exhaustive et complète dans les exemples de l'exercice 12.
4. (**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 θ).
Indication : Noter que pour toute fonction h on a $E_\theta[E_\theta[h(N)|S] - E_\theta[h(N)]] = 0$.
5. 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 [30]). 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)

1. $T \in L_2(Q)$, $T \in L_1(P_\theta)$ pour tout $\theta \in \Theta$, puis que T est non-biaisé
2. T est indépendant du choix de T_0 (vérifier que si T' en est un autre, alors $T - T' \perp H$)
3. 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 23 appartient donc pour tout θ_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_θ si $\theta \neq \theta_0$. L'exercice 11 p. 93 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 θ_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.

1. Vérifier que pour l'exemple 3 de l'exercice 12 avec $g(\theta) = \theta$ on trouve

$$T = \frac{n^{-1}}{2-\theta_0^2}\Big((\theta_0^2 - 2\theta_0 - 4)N_1 - (\theta_0^2 + 2\theta_0 - 4)N_2 + (2\theta_0^2 + \theta_0 - 2)N_3 - (2\theta_0^2 - \theta_0 - 2)N_4\Big).$$

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.

2. 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.
   (a) 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[6]).
   (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 $\mathscr{L}(\theta, Y)$. Il également la propriété importante d'être invariant par changement de variables sur θ. C'est un bon choix par défaut mais ses performances peuvent être assez décevantes lorsque l'échantillon est petit, particulièrement lorsque θ est vectoriel.

On se placera dans le cas d'un échantillons de v.a. iid $Y = (Y_1, ..Y_n)$ de loi P_{θ_*} et

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \mathscr{L}(\theta, Y_i). \tag{III.2}$$

Noter que $\mathscr{L}$ 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[7].

Le cas non indépendant devra être traité soit avec les résultats du chapitre II (th. 11, 13, 14, 12) ou ceux du IV.1.1.

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 $\mathscr{L}(\theta, y)$ infini (cf. la discussion précédant le théorème 14 p. 42) :

24 - Théorème

On suppose Θ compact, les Y_i iid de loi P_{θ_*}, et
   (a) La fonction $\mathscr{L}(\theta, y) = \log(p_\theta(y))$ satisfait (GNU) (la loi de Y est P_{θ_*})
   (b) L'application $\theta \longrightarrow P_\theta$ est injective

alors tout estimateur au maximum de vraisemblance (III.2) converge p.s. vers θ_*. On a la même conclusion si remplace (a) par les des hypothèses :
   (a') Pour tout $A \leqslant 0$, la fonction $(\theta, y) \mapsto \max(A, \log(p_\theta(y)))$ satisfait (GNU)
   (a") $E_{\theta_*}\left[|\log(p_{\theta_*}(Y))|\right] < \infty$.

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 avec $K(\theta, y) = -\log(p_\theta(y))$; le contraste moyen est ici $k(\theta) = D(P_{\theta_*} \| P_\theta) + k(\theta_*)$ où $D(.\|.)$ est la divergence de Kullback-Leibler

$$D(Q\|P) = \int \log(q(x)/p(x)) q(x) \mu(dx), \tag{III.3}$$

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 $\sum$.

7. La mesurabilité ne pose pas de problème car, en dehors de cas pathologiques, le maximum pourra être obtenu comme limite de maximums observés sur des grilles dyadiques finies de plus en plus fines.

p et q désignant les densités de ces mesures par rapport à une mesure commune μ (p. ex. $\mu = (P+Q)/2$). Il est classique que $D(Q\|P) \geqslant 0$, 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 13 et 14 avec $K(\theta, y) = -\log(p_\theta(y))$, il ne suffit de vérifier que l'unicité de l'extremum (point (b) du théorème 14) 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 : $\sqrt{n}(\widehat{\theta}_n - \theta_*) \to \mathcal{N}(0, I(\theta_*)^{-1})$ 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

$$I_{ij}(\theta) = E_\theta\Big[\frac{\partial \mathscr{L}(\theta, y)}{\partial \theta_i}\frac{\partial \mathscr{L}(\theta, y)}{\partial \theta_j}\Big] \tag{III.4}$$

en tout point θ intérieur à Θ tel que la fonction $\theta \mapsto \mathscr{L}(\theta, y)$ soit P_θ-p.s. différentiable, et que $I_{ii}(\theta) < +\infty$ pour tout i. Si l'on pense au maximum de vraisemblance, c'est la matrice I du théorème 11 p. 36.

Le calcul pratique de $I(\theta)$ fera intervenir deux point 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 27 plus bas).

25 - Proposition

Soit un ouvert $\mathscr{V} \subset \Theta$. On suppose que μ-p.s. la fonction $\theta \mapsto \mathscr{L}(\theta, y)$ y est deux fois dérivable (et donc $\theta \mapsto p_\theta(y)$ également) avec

$$\int \sup_{\theta \in \mathscr{V}} \|\nabla^2_\theta p_\theta(y)\| \mu(dy) < \infty. \tag{III.5}$$

Alors, pour tout $\theta_* \in \mathscr{V}$ tel que $I(\theta_*) < \infty$,

$$I(\theta_*) = -E_{\theta_*}[\nabla^2_\theta \mathscr{L}(\theta_*, Y)]. \tag{III.6}$$

Démonstration. La stratégie est de prendre l'espérance sous P_{θ_*} dans l'expression suivante

$$\nabla^2_\theta \mathscr{L}(\theta_*, y) = p_{\theta_*}(y)^{-1}\nabla^2_\theta p_{\theta_*}(y) - \nabla_\theta \mathscr{L}(\theta_*, y)\nabla_\theta \mathscr{L}(\theta_*, y)^T.$$

Il suffit de montrer que $\int \nabla^2_\theta p_{\theta_*}(y)\mu(dy) = 0$. Pour montrer cela, on va utiliser deux fois le théorème de dérivation sous le signe intégral afin de sortir le ∇^2_θ. L'inégalité de Cauchy-Schwartz conduit à

$$\int \|\nabla_\theta p_{\theta_*}(y)\|\mu(dy) \leqslant \Big(\int \frac{\|\nabla_\theta p_{\theta_*}(y)\|^2}{p_{\theta_*}(y)}\mu(dy)\Big)^{1/2}\Big(\int p_{\theta_*}(y)\mu(dy)\Big)^{1/2} = \mathrm{Trace}(I(\theta_*))^{1/2} < \infty$$

et donc, avec (III.5), par le théorème de dérivation sous le signe intégral :

$$\int \nabla_\theta^2 p_{\theta_*}(y)\mu(dy) = \nabla_\theta \int \nabla_\theta p_{\theta_*}(y)\mu(dy). \tag{III.7}$$

Comme pour ε petit on a

$$\sup_{\|\theta-\theta_*\|\leqslant\varepsilon} \|\nabla p_\theta(y)\| \leqslant \|\nabla p_{\theta_*}(y)\| + \varepsilon \sup_{\theta\in\mathscr{V}} \|\nabla^2 p_\theta(y)\|,$$

on peut appliquer de nouveau le théorème de dérivation sous le signe intégral et obtenir que le membre de droite de (III.7) est nul. ∎

La condition de R-régularité. Il est temps d'introduire cette condition technique [36] qui permet en particulier de permuter dérivation et intégrale dans les démonstrations ; elle utilise une définition plus générale de l'information de Fisher :

26 - Définition (R-Régularité. Information de Fisher)

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 μ-mesure pleine : la fonction $\theta \mapsto \nabla_\theta p_\theta(y)$ est continue au point θ_0.

(ii) Pour $\mu-$presque tout y : pour tous θ et h tels que le segment $[\theta, \theta+h]$ soit inclus dans Θ, alors $\int_0^1 |\langle h, \nabla_\theta p_{\theta+th}(y)\rangle| dt < \infty$ et :

$$p_{\theta+h}(y) - p_\theta(y) = \int_0^1 \langle h, \nabla_\theta p_{\theta+th}(y)\rangle dt. \tag{III.8}$$

(b) L'information de Fisher définie par

$$I(\theta) = \int \nabla_\theta p_\theta(y) \nabla_\theta p_\theta(y)^T p_\theta(y)^{-1} 1_{p_\theta(y)>0} \mu(dy) \tag{III.9}$$

est finie et continue sur Θ.

Notons qu'en conséquence de (a), pour tout $\theta_0 \in \Theta$, avec μ-mesure pleine, la fonction $\theta \mapsto p_\theta(y)$ est différentiable en θ_0. Si pour μ-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 $\nabla_\theta p_\theta(y) = \frac{1}{2} \text{sign}(y-\theta) e^{-|y-\theta|}$ et $I(\theta) = 1$; l'ensemble de mesure pleine du (i) est $\mathbb{R}\backslash\{\theta_0\}$. 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. 77.

Le lemme qui suit peut paraître technique mais sera d'un grand secours dans la suite :

27 - Lemme

On suppose l'expérience R-régulière. Soit f une fonction mesurable, et U un ouvert de $\mathbb{R}^d$ inclus dans Θ tels que

$$E_\theta[f(Y)^2] < c < \infty, \quad \theta \in U,$$

alors

$$\nabla_\theta E_\theta[f(Y)] = \int f(y) \nabla_\theta p_\theta(y) \mu(dy) \tag{III.10}$$

et cette quantité est une fonction continue de $\theta \in U$.

Ce lemme est une conséquence des résultats de l'appendice B (équation (B.9) p. 117) ; il est donc valide plus généralement dans le cas d'une expérience DMQU (cf. définition 46 p. 115), car une expérience R-régulière est DMQU en vertu des théorèmes 52 p. 119 et 54. Donnons-en ici 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 $t > 0$

$$E_{\theta+th}[f(Y)] = E_\theta[f(Y)] + \int f(y)\Big(\int_0^t \langle h, \nabla p_{\theta+sh}(y)\rangle ds\Big)\mu(dy).$$

On va appliquer le théorème de Fubini ; en effet l'inégalité de Cauchy-Schwartz,

$$\int_0^t \int |f(y)|\,\|\nabla p_{\theta+sh}(y)\|\mu(dy)ds \leqslant \int_0^t E_{\theta+sh}[f(Y)^2]^{1/2}\|I(s)\|^{1/2}ds < \infty$$

démontre l'applicabilité du théorème de Fubini, et donc

$$E_{\theta+th}[f(Y)] = E_\theta[f(Y)] + \int_0^t \Big(\int f(y)\langle h, \nabla p_{\theta+sh}(y)\rangle\mu(dy)\Big)ds$$

et il ne reste plus qu'à montrer la continuité de la fonction

$$\theta \mapsto \int f(y)\nabla p_\theta(y)\mu(dy).$$

Nous aurons besoin du lemme 48. Soit θ_n une suite convergente et les fonctions $w_n(y) = f(y)\nabla p_{\theta_n}(y)$. Décomposons w_n comme produit des fonctions

$$u_n(y) = f(y)\sqrt{p_{\theta_n}(y)}$$
$$v_n(y) = 1_{p_{\theta_n}(y)>0}\frac{\nabla p_{\theta_n}(y)}{\sqrt{p_{\theta_n}(y)}},$$

alors $u_n(y)$ est borné dans L_2 et $v_n(y)$ converge dans L_2 en raison point (ii) du lemme 48 et de l'hypothèse de continuité de $\theta \mapsto I(\theta)$. Il s'ensuit que $\int w_n(y)\mu(dy)$ converge bien vers la limite attendue en raison point (i) du lemme 48.

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. ∎

On en déduit en particulier que si l'expérience est R-régulière :

1. $E_\theta[\nabla_\theta \mathscr{L}(\theta, y)] = 0$ (faire $f = 1$ dans le lemme).
2. Donc l'information de Fisher est la variance du score $\nabla_\theta \mathscr{L}(\theta, Y)$.
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 11, mais un jeu d'hypothèses plus faible est obtenu en utilisant le théorème 15 et des résultats de l'appendice B :

28 - Théorème

Soit Θ une partie de $\mathbb{R}^d$ contenant un voisinage $\mathscr{V}$ de θ_*. On fait les hypothèses suivantes :

(i) L'expérience est R-régulière, et $I(\theta_*)$ est inversible.

(ii) L'estimateur au maximum de vraisemblance $\widehat{\theta}_n$ converge en probabilité vers θ_*.

(iii) De plus,

$$E_{\theta_*}\left[\sup_{\theta\in\mathscr{V}}\frac{\|\nabla_\theta p_\theta(y)\|^2}{p_\theta(y)^2}\right]<\infty. \tag{III.11}$$

Alors on a

$$\sqrt{n}\Big(\widehat{\theta}_n-\theta_*+\tfrac{1}{n}I(\theta_*)^{-1}\sum_{i=1}^n\frac{\nabla_\theta p_{\theta_*}}{p_{\theta_*}}(Y_i)\Big)\xrightarrow{P}0 \tag{III.12}$$

$$\sqrt{n}(\widehat{\theta}_n-\theta_*)\xrightarrow{Loi}\mathcal{N}(0,I(\theta_*)^{-1}). \tag{III.13}$$

Démonstration. Le résultat s'obtient en appliquant le théorème 15 à

$$K(\theta,Y)=-\mathscr{L}(\theta,Y)$$
$$\dot{K}(y)=\sup_\theta\|\nabla p_\theta(y)\|p_\theta(y)^{-1}.$$

L'équation (II.36) est une conséquence de (III.8) et du théorème 53 de l'appendice B avec $\psi(x)=\ln(x+\varepsilon)$, $\varepsilon>0$, en faisant tendre ε vers 0. L'équation (II.38), avec $J=-I(\theta_*)$, se déduit du lemme 50 et du théorème 54. ∎

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 29 plus bas).

Calcul pratique de la variance asymptotique. On préfère généralement l'estimateur sandwich [158] pour la variance de $\widehat{\theta}_n$, $\widehat{V}_n=\widehat{J}^{-1}\widehat{I}\widehat{J}^{-1}$, soit :

$$\widehat{V}_n=\Big(\sum\nabla^2\mathscr{L}(\widehat{\theta}_n,Y_i)\Big)^{-1}\Big(\sum\nabla\mathscr{L}(\widehat{\theta}_n,Y_i)\nabla\mathscr{L}(\widehat{\theta}_n,Y_i)^T\Big)\Big(\sum\nabla^2\mathscr{L}(\widehat{\theta}_n,Y_i)\Big)^{-1}. \tag{III.14}$$

qui ne présuppose pas la véracité du modèle paramétrique mais se base simplement sur le fait que $\mathscr{L}(\theta_*,Y_i)$ est une fonction de contraste. Voir l'exercice 4 p. 77 pour le cas de la régression linéaire.

Le cas non-compact. Si Θ 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 $\widehat{\theta}_n$ reste confinée dans un certain compact pouvant dépendre de la réalisation ω.

Remarque : Exhaustivité asymptotique. En reprenant la formule (II.34), donc avec un petit supplément d'hypothèses, on obtient que

$$p_{\theta_*}(Y_1,\dots Y_n)=e^{-\frac{n}{2}(\widehat{\theta}_n-\theta_*)^TI(\theta_*)(\widehat{\theta}_n-\theta_*)+\varepsilon_n}p_{\widehat{\theta}_n}(Y_1,\dots Y_n)$$

où ε_n converge en probabilité vers 0 quand n tend vers l'infini (et si les Y_i sont tirés selon p_{θ_*}) ; ceci peut s'interpréter comme de l'exhaustivité asymptotique de $\widehat{\theta}_n$. Plus spécifiquement, si $\varepsilon_n(\theta_*,Y)$ était identiquement nul, on aurait $p_\theta(y)=e^{-\frac{n}{2}(T(y)-\theta)^TI(\theta)(T(y)-\theta)}q(y)$, avec $y=(y_1,\dots y_n)$ et $T(y)=\widehat{\theta}(y)$, qui est simplement une famille exponentielle.

Le théorème de Wilks. Le théorème 20 permet de tester l'hypothèse H_0 : « $g(\theta_*)=0$ » (c.f. § III.5), en utilisant la statistique de rapport de vraisemblances de l'équation (III.15) : sous réserve de satisfaction

des hypothèses, si $g(\theta_*) = 0$, alors l'estimateur sous contrainte

$$\widehat{\theta}_n^c = \arg \min_{g(\theta)=0} \sum_{i=1}^n \mathscr{L}(\theta, Y_i)$$

satisfait

$$2\sum_{i=1}^n \mathscr{L}(\widehat{\theta}_n, Y_i) - \mathscr{L}(\widehat{\theta}_n^c, Y_i) \longrightarrow \chi^2_m \tag{III.15}$$

où m est la dimension de g, c.-à-d. la différence de dimension effective entre θ et θ^c.

Exemple : Test de « $\theta_* = \theta_0$ », région de confiance. On a en particulier

$$2\sum_{i=1}^n \mathscr{L}(\widehat{\theta}_n, Y_i) - \mathscr{L}(\theta_*, Y_i) \xrightarrow{P} \chi^2_d. \tag{III.16}$$

On pourra tester au niveau α l'hypothèse H_0 : « $\theta_* = \theta_0$ » (c.f. § III.5) par réfutation de cette dernière si $2\sum_{i=1}^n \mathscr{L}(\widehat{\theta}_n, Y_i) - \mathscr{L}(\widehat{\theta}_0, Y_i)$ est supérieur au quantile $\chi^2_d(1-\alpha)$. C'est le test du rapport de vraisemblance. L'ensemble

$$\mathcal{R}_\alpha = \left\{\theta : 2\sum_{i=1}^n \mathscr{L}(\widehat{\theta}_n, Y_i) - \mathscr{L}(\theta, Y_i) \leqslant \chi^2_d(1-\alpha)\right\}$$

est une région de confiance pour θ_* de niveau asymptotique α, qui s'ajoute aux régions (II.18) et (II.19).

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 σ (hypothèse H_0). On a classiquement

$$\widehat{\theta} = \frac{1}{n}\sum (Y_i - \overline{Y})(Y_i - \overline{Y})^T, \quad \overline{Y} = \frac{1}{n}\sum Y_i$$
$$\widehat{\sigma}^2 = \frac{1}{p} Tr(\widehat{\theta})$$
$$\widehat{\theta}^c = \widehat{\sigma}^2 Id.$$

Le membre de gauche de (III.15) vaut

$$S = np\ln(\widehat{\sigma}^2) - n\ln(\det(\widehat{\theta})$$

qui est asymptotiquement un $\chi^2_{p(p+1)/2-1}$ sous H_0. Le test de niveau 5% réfutera H_0 si S est supérieur au quantile à 0,95 de ce χ^2 , i. e. q t.q. $P(\chi^2_{p(p+1)/2-1} > q) = 0,95$.

Correction de Bartlett (1937). Un raffinement de (III.15) 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 [37]. 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.3.3 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 χ^2_d qui apparaît dans l'équation (III.16) n'est autre que le terme $(\widehat{\theta}_n - \theta_*)^T J(\widehat{\theta}_n - \theta_*)$ de (II.35). Le critère $TIC(d)$ devient ici le critère d'Akaike, qui est une estimée du constraste moyen en $\widehat{\theta}$, $k(\widehat{\theta})$, $k(\theta) = -2E_{\theta_*}[\mathscr{L}(\theta, Y)]$, estimée consistant à oublier des termes d'ordre inférieur et à remplacer ce χ^2_d par son espérance

$$\boxed{AIC(d) = -2\textstyle\sum_i \mathscr{L}(\widehat{\theta}_n^d, Y_i) + 2d}$$

Le raisonnement conduisant à cette approximation de $k(\widehat{\theta}_n^d)$ est valide pour $d \geqslant d_*$, car $I = J$ est réalisé dans TIC, les observations étant bien issues du modèle. Pour $d < d_*$, $AIC(d) - AIC(d_*)$ est d'ordre n, ce qui assure une marge confortable. Malheureusement, pour $d \geqslant d_*$, le terme négligé $2(\chi^2_d - d)$ est d'ordre $\sqrt{d}$, et ses fluctuations [8], vont rendre le minimiseur $\widehat{d}_n$ de AIC typiquement sensiblement plus grand que d_* : la remontée de $k(\widehat{\theta}^d)$ pour $d > d_*$ est trop douce et est noyée dans ces fluctuations (noter que maintenant on exploite le surajustement car c'est ce dernier qui fait que pour $d > d_*$, $k(\widehat{\theta}^d)$ augmente).

Ceci fait qu'en vue l'estimation de d_*, la compensation $2d$ ne suffit pas, et d'autres pénalisations ont été proposées, la plus connue étant $d \log n$, correspondant au critère BIC (cf. appendice G) dont on montre qu'elle garantit la consistance de $\widehat{d}_n$ lorsque n tend vers l'infini [9]. *Il faut distinguer le problème de prédiction (maximiser $l(\widehat{\theta}_n^d)$, $l(\theta) = E[\mathscr{L}(\theta, Y_1)]$) et celui de la sélection de variables (trouver d_* s'il existe).* Dans ces situations de modèles emboîtés, AIC est bon pour la prédiction et l'estimation tandis que BIC l'est pour la sélection [10]. On trouvera des compléments dans [16].

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 [11], même si l'objectif est la prédiction ; on trouvera des compléments intéressants dans [7], où le terme de pénalisation proposé pour la prédiction est d'ordre $2d + 2\log \nu_d$, où ν_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 G.

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 [12] 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.*

III.2.3 Bornes exponentielles. Convergence des moments

On s'intéresse ici à obtenir des théorèmes plus forts que la simple convergence en loi.

29 - Théorème

En plus de la compacité de $\Theta \subset \mathbb{R}^d$, de l'injectivité de $\theta \to P_\theta$, et de la R-régularité de l'expérience, on suppose que Θ est la fermeture de son intérieur, lequel possède la propriété de cône (p. 125) et que pour un $q > d$

$$\sup_\theta \int \left(\frac{\|\nabla_\theta p_\theta(y)\|}{p_\theta(y)} \right)^q p_\theta(y) \mu(dy) < \infty \tag{III.17}$$

(conséquence de la régularité si $d = 1$). Alors, si $I(\theta_*)$ est inversible, il existe C et σ tels que pour tous a et $n > 0$

$$P_{\theta_*}(\sqrt{n}\|\widehat{\theta}_n - \theta_*\| \geqslant a) \leqslant C e^{-a^2/2\sigma^2}. \tag{III.18}$$

Les constantes C et σ peuvent être choisies indépendantes de θ_* du moment que θ_* est restreint à un compact intérieur à Θ où I est inversible.
En particulier la suite $\sqrt{n}\|\widehat{\theta}_n - \theta_*\|/\sqrt{\log n}$ est bornée avec probabilité 1.
De plus sous les hypothèses du théorème 28, on a convergence de tout les moments de $\sqrt{n}(\widehat{\theta}_n - \theta_*)$ vers ceux de la gaussienne limite.

8. Voir [147] où ces fluctuations et leurs effets sont étudiées avec précision.
9. Il est montré dans [93] qu'une pénalisation d'ordre $d \log \log n$ suffit à obtenir la convergence forte de $\widehat{d}_n$ vers d_*.
10. Voir aussi [164] pour une discussion plus précise et un peu différente.
11. Leo Breiman qualifiait l'usage de AIC dans ce contexte de «quiet scandal in the statistical community» [42].
12. Sur CV et ses variantes, voir p. ex. [49, 43, 146] ; voir [104] pour une comparaison avec les méthodes de pénalisation.

Démonstration. On va utiliser le théorème 22. Il s'agit de vérifier les deux hypothèses, avec ici

$$K_n(\theta,\omega) = -\sum_{i=1}^{n} \mathscr{L}(\theta, Y_i).$$

Vérifions (II.71) pour $\eta = 1/2$, et sans prendre la partie positive, ce qui se réduit exactement à

$$\int \sqrt{p_\theta(y)}\sqrt{p_{\theta_*}(y)}\mu(dy) \leqslant e^{-\|\theta-\theta_*\|^2/2\sigma^2} \tag{III.19}$$

soit

$$1 - d_H(P_\theta, P_{\theta_*})^2/2 \leqslant e^{-\|\theta-\theta_*\|^2/2\sigma^2} \tag{III.20}$$

où d_H désigne la distance de Hellinger

$$d_H(P_\theta, P_{\theta'})^2 = \int (\sqrt{p_{\theta'}(y)} - \sqrt{p_\theta(y)}\,)^2 \mu(dy).$$

Il se trouve que la R-régularité implique que d_H satisfait

$$d_H(P_\theta, P_{\theta+h})^2 = \frac{1}{4} h^T [I(\theta) + \varepsilon_\theta(h)] h \tag{III.21}$$

avec pour tout compact K intérieur à Θ

$$\lim_{h\to 0} \sup_{\theta\in K} |\varepsilon_\theta(h)| = 0. \tag{III.22}$$

La démonstration de ce point est faite à l'appendice B.1 (équation (B.8) ; noter qu'une expérience R-régulière est DMQU en vertu des théorèmes 52 et 54).

Comme l'application $\theta \to d_H(P_\theta, P_{\theta_*})$ est continue, non nulle pour $\theta \neq \theta_*$, et que Θ est compact, (III.21) implique que pour un certain ε on a $d_H(P_\theta, P_{\theta_*})/2 \geqslant \varepsilon\|\theta - \theta_*\|^2$, d'où (III.20) pour un certain σ. L'uniformité en θ dans (III.22) fait que σ peut être rendu indépendant de $\theta \in K$.

Passons maintenant à (II.70). 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 $\eta = 1 - \frac{1}{2r}$ (l'équation (II.71) reste vraie pour cette valeur de η supérieure à $1/2$) :

$$\begin{aligned} E_{\theta_*}\left[\Big\|\sum \nabla\mathscr{L}(\theta, Y_i)\Big\|^{q_0}\Big(\frac{p_{\theta_*}(Y)}{p_\theta(Y)}\Big)^{-\eta}\right] &= E_\theta\left[\Big\|\sum \nabla\mathscr{L}(\theta, Y_i)\Big\|^{q_0}\Big(\frac{p_{\theta_*}(Y)}{p_\theta(Y)}\Big)^{1-\eta}\right] \\ &\leqslant E_\theta\left[\Big\|\sum \nabla\mathscr{L}(\theta, Y_i)\Big\|^{q}\right]^{1/p} E_\theta\left[\Big(\frac{p_{\theta_*}(Y)}{p_\theta(Y)}\Big)^{(1-\eta)r}\right]^{1/r} \\ &\leqslant C n^{q_0/2} e^{-\|\theta-\theta_*\|^2(1-\eta)/\sigma^2} \end{aligned}$$

la majoration du premier terme venant des théorèmes sur les sommes de variables aléatoires indépendantes, et celle du second venant de (III.19). Le théorème 22 s'applique donc avec q_0 (au lieu de q).

Concernant la dernière affirmation, noter que (III.18) 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 : $I(\theta) = \mathrm{Cov}_\theta(Y) = \nabla^2_\theta Z$.

Exemple de non-unicité. On observe Y_i, $i = 1, \ldots n$, 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 d'espérance m et de variance σ^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 σ tend vers 0. Θ n'est pas compact.

La tentative de compactifier Θ 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 σ-finie).

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 [113].

III.2.5 Exercices et compléments

Exercice 1 (Loi gamma). On considère des variables iid $T_1, \dots T_n$ de distribution gamma : $p(t) = \frac{1}{\Gamma(a)} b^{-a} t^{a-1} e^{-t/b}$. L'espérance 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 27 et l'inégalité $\|\nabla^2_\theta p_\theta(y)\| \leqslant p_\theta(y)(\|\nabla^2_\theta \mathscr{L}(\theta, y)\| + \|\nabla_\theta \mathscr{L}(\theta, y)\|^2)$.

En déduire, à l'aide du théorème 1, la variance asymptotique de $\widehat{\mu}_n$ en fonction de (μ, σ).

Exercice 2. Soit des variables $(X_i)_{1 \leqslant i \leqslant n}$ 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 (le plus simple est considérer que $Y_i = 0$ si $X_1 = 1$). 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 27 montrer que, si $I(\theta)$ est finie, $E_\theta[\nabla_\theta \mathscr{L}(\theta, y)] = 0$.

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 $\widehat{\theta}_n$ donné par (I.10). On pose $\widehat{u} = Y - X\widehat{\theta}_n$, vecteur des erreurs de prédiction. Montrer que l'estimateur sandwich (III.14) p. 73 vaut ici

$$\widehat{V}_n = (X^T X)^{-1} \Big(\sum_i \widehat{u}_i^2 X_i^T X_i \Big) (X^T X)^{-1}.$$

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 [13]). 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 27 ne s'applique pas si $f(x) = 1_{x=1}$.

Exercice 7 (Maximum de vraisemblance partielle). On reprend l'exercice 14 p. 29. Les paires (X_i, Y_i) sont désormais iid de densité $p_{\theta_*}(x, y)$. Exprimer la variance asymptotique de $\widehat{\theta}$ 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.

13. Exemple proposé par Arnaud Guyader.

***Exercice 8 (Vraisemblance, conditionnement, et monotonicité 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 μ est une probabilité. Soit φ une fonction mesurable et $Z = \varphi(Y)$.

1. Soit ν la loi image de μ par φ et la fonction

   $$q_\theta(Z) = E_\mu[p_\theta(Y)|Z].$$

   Montrer que si Y suit la loi P_θ, alors Z suit la loi $q_\theta(z)\nu(dz)$.
   Indication : Montrer que pour toute fonction borélienne bornée $\int f(z)q_\theta(z)\nu(dz) = E_\theta[f(Z)]$. On ramènera les deux termes en des intégrales en $\mu(dy)$ puis on exploitera la définition de q_θ.
2. Montrer que pour toute fonction mesurable bornée f on a

   $$E_\theta[f(Y)|Z] = q_\theta(Z)^{-1}E_\mu[f(Y)p_\theta(Y)|Z].$$

   Indication : Montrer que pour toute fonction borélienne bornée g, $E_\theta[g(Z)T_1] = E_\theta[g(Z)T_2]$, où T_1 et T_2 sont les deux membres. On notera que pour toute fonction h, $E_\theta[h(Y)] = E_\mu[h(Y)p_\theta(Y)]$.
3. Déduire des points précédents que si $\mathscr{L}_p(\theta, Y) = \log p_\theta(Y)$ et $\mathscr{L}_q(\theta, Z) = \log q_\theta(Z)$ on a

   $$\nabla_\theta\mathscr{L}_q(\theta, Z) = E_\theta[\nabla_\theta\mathscr{L}_p(\theta, Y)|Z].$$

4. En déduire que si $J(\theta)$ désigne l'information de Fisher liée au deuxième modèle on a $J(\theta) \leqslant I(\theta)$.

***Exercice 9 (Statistiques libres et information).** Soit N une statistique libre (exercice 14 p. 68).

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 $E_\theta[\nabla_\theta\mathscr{L}(\theta, Y)|N] = 0$.
2. Que penser de l'information de Fisher conditionnelle $-E_\theta[\nabla^2_\theta\mathscr{L}(\theta, Y)|N]$ pour estimer la variance ? Comparer dans le cas du premier exemple de l'exercice 12 p. 67.
3. Etudier également la question de la variance conditionnelle de $\widehat{\theta}$ dans le cas d'une suite Y_i iid $\mathcal{U}([\theta - 1, \theta + 1])$ avec $\widehat{\theta} = \frac{1}{2}(M + m), N = M - m$, $M = \max(Y_i), m = \min(Y_i)$.
4. Dans le cas où plusieurs statistiques libres existent (voir exercice 14 p. 68 point 1), le choix de N peut poser problème et Cox [57] 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. 67.

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 [14].

III.3.1 Borne de Cramér-Rao

30 - Théorème

On suppose l'expérience R-régulière avec une information de Fisher $I > 0$ sur Θ (cf. def. 26) ; alors pour tout estimateur non-biaisé $\widehat{g}$ d'une fonction C^1 vectorielle $g(\theta)$, tel que $E_\theta[|\widehat{g}|^2] < c < \infty$, $\theta \in \Theta$, la variance d'estimation est minorée comme suit :

$$E_\theta\Big[(\widehat{g} - g(\theta))(\widehat{g} - g(\theta))^T\Big] \geqslant \nabla g(\theta)I(\theta)^{-1}\nabla g(\theta)^T$$

où chaque ligne de ∇g contient le gradient de la coordonnée correspondante.

14. 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 [80].

Remarques. Cette borne implique que dans le cas de n échantillons indépendants, on a pour tout estimateur $\widehat{\theta}_n$ de θ :

$$nE_\theta\big[(\widehat{\theta}_n-\theta)(\widehat{\theta}_n-\theta)^T\big] \geqslant I(\theta)^{-1}$$

où l'information de Fischer est calculée sur la base d'un seul échantillon.

On observe alors que si l'estimateur $\widehat{\theta}_n$ est optimal au sens où le membre de gauche converge vers $I(\theta)^{-1}$ quand $n\to\infty$, alors $g(\widehat{\theta}_n)$ le sera également (cf. la delta-method p. 18).

Démonstration. En dérivant en θ l'identité

$$g(\theta)=\int \widehat{g}(y)\; p_\theta(y)\mu(dy),$$

il vient, en vertu du lemme 27, et du point 1 qui le suit :

$$\nabla g(\theta)=\int \widehat{g}(y)\;\nabla p_\theta(y)\mu(dy)=E_\theta\Big[\big(\widehat{g}(Y)-g(\theta)\big)\frac{\nabla p_\theta(Y)}{p_\theta(Y)}\Big]. \tag{III.23}$$

On peut alors directement conclure en utilisant l'inégalité de Cauchy-Schwarz matricielle

$$E\big[XZ^T\big]E\big[ZZ^T\big]^{-1}E\big[ZX^T\big]\leqslant E\big[XX^T\big] \tag{III.24}$$

avec $X=\widehat{g}(Y)-g(\theta)$ et $Z=\frac{\nabla p_\theta(Y)^T}{p_\theta(Y)}$ (vecteur colonne). L'équation (III.24) se démontre en notant que la différence des deux membres de (III.24) 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 :

$$\mathscr{L}(Y,\theta)=-\frac{1}{2\sigma^2}\sum_k\big(y_k-a\cos(\omega k+\varphi)\big)^2-n\log(\sigma).$$

Supposons pour simplifier qu'il s'agisse d'un problème de modulation de fréquence pour lequel seul ω est inconnu, c.-à-d. $\theta=\omega$. On a sous P_ω

$$\frac{d}{d\omega}\mathscr{L}(Y,\omega)=-\frac{1}{\sigma^2}\sum_k ka\sin(\omega k+\varphi)(y_k-a\cos(\omega k+\varphi))=-\frac{a}{\sigma^2}\sum_k k\sin(\omega k+\varphi)u_k.$$

Puis

$$E\big[(\partial_\omega\mathscr{L}(Y,\omega))^2\big]=\frac{a^2}{\sigma^2}\sum_k k^2\sin^2(k\omega+\varphi)=\frac{a^2}{2\sigma^2}\sum_k k^2\big(1-\cos(2k\omega+2\varphi)\big).$$

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 φ est uniforme sur $[0,2\pi]$:

$$E\big[(\partial_\omega\mathscr{L}(Y,\omega))^2\big]\simeq\frac{a^2}{2\sigma^2}\sum k^2\simeq\frac{a^2}{6\sigma^2}n^3.$$

On voit que ω n'intervient pas, contrairement au rapport signal sur bruit a/σ, et que la dépendance en n est n^3. Les observations ne sont pas iid.

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 satisfaite.

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 θ sur un ouvert arbitrairement petit, c'est l'inégalité de van Trees [15]. On va se donner une densité de probabilité π sur Θ et l'on considérera la mesure de probabilité ainsi construite sur l'espace produit :

$$E[f(Y,\theta)] = \iint f(y,\theta)p_\theta(y)\mu(dy)\pi(\theta)d\theta = \iint E_\theta\big[f(Y,\theta)\big]\pi(\theta)d\theta$$

(noter que E n'a pas l'indice θ). Le risque sera alors calculé sur la base de cette espérance $E[r(\theta,\widehat{\theta}(Y))] = \int R_\theta(\widehat{\theta})\pi(d\theta)$ (cf. équation (III.25)) ; 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 $R_\theta(\widehat{\theta})$ sur le support (choisi petit) de π.*

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. 84.

31 - Théorème (Inégalité de van Trees)

On suppose l'expérience R-régulière sur $\Theta \subset \mathbb{R}^d$, d'information de Fisher $I(\theta)$ (cf. déf. 26). Soit π une densité C^1 sur $\mathbb{R}^d$ à support compact inclus dans Θ. Soit g une fonction vectorielle C^1 définie sur Θ. On pose

$$\bar{I} = \int I(\theta)\pi(\theta)d\theta$$
$$I_\pi = \int \frac{\nabla\pi(\theta)\nabla\pi(\theta)^T}{\pi(\theta)}d\theta$$
$$G = \int \nabla g(\theta)\pi(\theta)d\theta$$

(chaque ligne de ∇g contient le gradient de la coordonnée correspondante). Soit $y \mapsto \widehat{g}(y)$ un estimateur de $g(\theta)$. On suppose que les quantités ci-dessus sont finies et que

$$E\Big[\|\widehat{g}(Y)\|^2 + \|g(\theta)\|^2\Big] < \infty.$$

Alors

$$E\Big[\big(\widehat{g}(Y) - g(\theta)\big)\big(\widehat{g}(Y) - g(\theta)\big)^T\Big] \geqslant G\big(\bar{I} + I_\pi\big)^{-1}G^T. \tag{III.25}$$

Remarques. Dans le cas de n observations, $\bar{I}$ doit simplement être multipliée par n et le reste est inchangé ; le terme I_π a alors une contribution très faible. Si $g(\theta) = \theta$ alors $G = Id$.

Démonstration. Tout va se déduire de l'identité suivante où ∇ désigne le gradient en θ (vecteur ligne) :

$$\iint \big(\widehat{g}(y) - g(\theta)\big)\Big(\nabla p_\theta(y)\pi(\theta) + p_\theta(y)\nabla\pi(\theta)\Big)\mu(dy)d\theta = \int \nabla g(\theta)\pi(\theta)d\theta. \tag{III.26}$$

Cette identité est immédiate sous des hypothèses plus fortes de régularité par une simple intégration par parties dans l'intégrale en θ 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

15. Pour une discussion des extensions possibles consulter [85].

du lemme 27

$$\iint \widehat{g}(y)\nabla p_\theta(y)\pi(\theta)\mu(dy)d\theta = \int \Big(\nabla \int \widehat{g}(y)p_\theta(x)\mu(dy)\Big)\pi(\theta)d\theta$$
$$= -\int \Big(\int \widehat{g}(y)p_\theta(y)\mu(dy)\Big)\nabla\pi(\theta)d\theta$$

qui est l'opposé de l'autre terme mettant en jeu $\widehat{g}$ dans (III.26) ; cette partie de l'intégrale est donc nulle. Concernant le terme avec $g(\theta)$, il conduit bien au membre de droite de (III.26) puisque $\int \nabla p_\theta(y)\mu(dy) = 0$ et $\int p_\theta(y)\mu(dy) = 1$. (III.26) est bien démontré.

On va maintenant appliquer l'inégalité de Cauchy-Schwarz matricielle (III.24) sous la mesure $\nu(dy, d\theta) = p_\theta(y)\pi(\theta)\mu(dy)d\theta$ avec

$$X = \widehat{g}(y) - g(\theta), \qquad Z = \frac{\nabla p_\theta(y)}{p_\theta(y)} + \frac{\nabla\pi(\theta)}{\pi(\theta)}.$$

On a donc

$$\nu\big(XX^T\big) \geqslant \nu\big(XZ^T\big)\nu\big(ZZ^T\big)^{-1}\nu\big(ZX^T\big). \tag{III.27}$$

Comme sur l'ensemble A_θ de régularité de p_θ (cf. Définition 26) $\nabla p_\theta(y) = 0$ 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.26) vaut $\nu(XZ^T)$. Calculons $\nu(ZZ^T)$, on obtient quatre termes en développant :

$$\nu(ZZ^T) = \int \Big(\frac{\nabla p_\theta(y)\nabla p_\theta(y)^T}{p_\theta(y)}\pi(\theta) + p_\theta(y)\frac{\nabla\pi(\theta)\nabla\pi(\theta)^T}{\pi(\theta)}$$
$$- \nabla\pi(\theta)\nabla p_\theta(y))^T - \nabla p_\theta(y)\nabla\pi(\theta)^T\Big)\mu(dy)d\theta$$
$$= \bar{I} + I_\pi - 0 - 0$$

car les deux derniers termes sont nuls en vertu du lemme 27 ; il ne reste plus qu'à remplacer dans (III.27). ∎

Borne minimax asymptotique. On a donc dans le cas scalaire, pour n observations indépendantes

$$\int E_\theta\big[n(\widehat{\theta}_n - \theta)^2\big]\pi(\theta)d\theta \geqslant \Big(\int I(\theta)\pi(\theta)d\theta + \tfrac{1}{n}I_\pi\Big)^{-1}. \tag{III.28}$$

En raison du facteur $\frac{1}{n}$ et du caractère arbitraire de π, 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 π a son support dans un petit ouvert $U \subset \Theta$ on obtient en majorant trivialement le membre de gauche de (III.28)

$$\sup_{\theta\in U} E_\theta\big[n(\widehat{\theta}_n - \theta)^2\big] \geqslant \Big(\int I(\theta)\pi(\theta)d\theta + \tfrac{1}{n}I_\pi\Big)^{-1}.$$

En prenant la liminf sur n puis en faisant tendre π vers une masse de Dirac au point de plus mauvaise information de Fisher de U, il vient la borne minimax asymptotique

$$\varliminf_n \sup_{\theta\in U} E_\theta\big[n(\widehat{\theta}_n - \theta)^2\big] \geqslant \sup_{\theta\in U} I(\theta)^{-1}. \tag{III.29}$$

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

$$E_\theta\big[(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T\big] = I(\theta)^{-1} \quad \text{pour tout} \quad \theta \in \Theta.$$

Un théorème classique (exercice 3 p. 83) 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 $\widehat{\theta}_n = \widehat{\theta}_n(Y_1, ..Y_n)$ est asymptotiquement efficace si

$$\rho_n(\theta) = nE_\theta\big[(\widehat{\theta}_n - \theta)(\widehat{\theta}_n - \theta)^T\big] \longrightarrow I(\theta)^{-1} \quad \text{pour tout} \quad \theta \in \Theta. \tag{III.30}$$

Une variante est de demander la convergence en loi

$$\sqrt{n}(\widehat{\theta}_n - \theta) \longrightarrow \mathcal{N}(0, I(\theta)^{-1}) \quad \text{pour tout} \quad \theta \in \Theta. \tag{III.31}$$

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.30) 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.30) ou (III.31) 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. 83 ; le défaut de cet estimateur est que la convergence n'est plus uniforme en θ : si θ_* 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.30), il faut définir l'efficacité de $\widehat{\theta}_n$ comme la convergence *uniforme* de $\rho_n(.)$ vers $I(.)$. En effet, une conséquence de (III.29) est que si l'expérience est régulière, alors une convergence uniforme ne peut se faire que vers une limite partout supérieure à $I(.)$. On peut montrer que sous des hypothèses générales plus fortes que la R-régularité, l'estimateur au maximum de vraisemblance est efficace (pour des cas généraux, cf. th. 12, ou bien [11] § III.1 th.1.1).

Ces idées ont été développées par Hajek et Le Cam dans les années 1960-1970 [16]. 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 obtenu lorsque l'on se restreint à la classe des estimateurs réguliers au sens donné dans l'énoncé :

32 - Théorème (Théorème de convolution de Hajek. 1970. Cas R-régulier)

On suppose l'expérience R-régulière. On suppose également que $\widehat{\theta}_n$ est une suite d'estimateurs régulière au point θ_* au sens où pour tout $h \in \mathbb{R}^d$

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta_* - \frac{h}{\sqrt{n}}\Big) \xrightarrow{Loi} X, \quad \text{sous } P_{\theta_* + h/\sqrt{n}} \tag{III.32}$$

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

$$\big(\sqrt{n} I(\theta_*)(\widehat{\theta}_n - \theta_*) - S_n(\theta_*), S_n(\theta_*)\big) \xrightarrow{Loi} (Y, Z)$$

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

$$\sqrt{n} I(\theta_*)(\widehat{\theta}_n - \theta_*) - S_n(\theta_*) \xrightarrow{Loi} 0. \tag{III.33}$$

16. On en trouvera une description approfondie dans les chapitres 7 et 8 de [18], ou le chapitre 2 de [36], ou le chapitre I.10 de [11]

Bien différencier les expressions «régularité de l'expérience» (R-régularité) et «régularité de la suite d'estimateurs» (éq. III.32).

La force de ce théorème est qu'il est vrai dans le cadre général de l'hypothèse LAN (p. 109) de régularité de l'expérience, i. e. sans se restreindre au cas des suites d'observations indépendantes (th. 41 p.110).

Pour l'existence d'un estimateur satisfaisant (III.33), voir le th. 44, dû à Le Cam. On montre que sous les hypothèses du th. 28, l'estimateur au maximum de vraisemblance satisfait (III.32) (th. 42).

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.32). Noter que sous les hypothèses du th. 41, deux estimateurs $\widehat{\theta}_n$, et $\widehat{\theta}'_n$ réguliers au sens (III.32) et efficaces, sont équivalents au sens où $\sqrt{n}(\widehat{\theta}_n - \widehat{\theta}'_n) \xrightarrow{P} 0$.

Des résultats du même type se trouvent montrés plus directement au chapitre 6.2 de [8].

Un bon bilan des aspects théoriques de ces questions se trouve au chapitre 2 de [36].

Développements récents. Il apparaît qu'en vue de l'estimation d'un paramètre θ 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 [124], 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[17]. Il est démontré que forcément, ces estimateurs ne sont pas efficaces au sens minimax local [163, 116], même si leur variance asymptotique peut être satisfaisante[18] comme l'est celle de l'estimateur de Hodge (sur cet effet trompeur, voir [116] 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. 75.

III.3.4 Exercices et compléments

Exercice 1. Soit la famille paramétrique $\mathcal{N}(\theta, \sigma^2)$, σ 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 $\widehat{\theta} = g(Y)$ on a $E\big[(\widehat{\theta} - \theta)^2\big] \geqslant \sigma^2$. *Indication* : On pourra choisir pour π 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 $|\widehat{\theta}_n| \leqslant n^{-1/4}$ et sinon $\widehat{\theta}_n$, a des performances asymptotiques meilleures si $\theta_* = 0$ (au sens où $I(\theta)^{-1}$ dans (III.31) est inférieur) 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[19].

Exercice 4 (Borne de Barankin [30]). Soit $\theta_1, \dots \theta_n \in \Theta$ et $c_1, \dots c_n \in \mathbb{R}$. Utiliser la relation $Var(T) \geqslant Cov(T, V)^2 / Var(V)$, avec $V = p_\theta(X)^{-1} \sum_{i=1}^n c_i p_{\theta_i}(X)$ pour montrer que si T est un estima-

17. 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 [48]. Tout le problème ici réside dans les tout petits coefficients.

18. C'est le cas de SCAD et de *Adaptive lasso*. Concernant ce dernier voir [165].

19. Voir [13], ou [12] Th.2.5.12, pour des compléments.

teur non-biaisé de $g(\theta)$ alors

$$Var_\theta(T) \geqslant \frac{|\sum c_i(g(\theta_i) - g(\theta))|^2}{Var_\theta(V)} \tag{III.34}$$

(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

$$Var_\theta(T) \geqslant \sup_h \frac{(g(\theta+h) - g(\theta))^2}{\int \frac{p^2_{\theta+h}(x)}{p_\theta(x)}\mu(dx) - 1}.$$

En déduire la borne de Cramér-Rao dans le cas monodimensionnel.

Barankin montre que le sup du membre de droite de (III.34) sur tous les c_i et les θ_i correspond à la variance de l'estimateur de l'exercice 15 p. 68 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 θ.

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 $|w(x)| \leqslant 1/2$. Soit un estimateur $\widehat{g}$ de $g(\theta) = p_\theta(0)$, basé sur n échantillons indépendants tirés sous p_θ. 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

   $$\sup_{\theta\in\Theta} E_\theta\big[(\widehat{g}(Y) - p_\theta(0))^2\big] \geqslant w(0)^2\Big(2n\|w\|_2^2 + I_\pi\Big)^{-1}, \qquad \|w\|_2^2 = \int_{-1/2}^{-1/2} w(x)^2 dx.$$

   Indication : On minorera le sup par l'espérance sous π.

2. Montrer que si w prend la forme $w(x) = \frac{1}{M}\varphi(Mx)$, où $M > 1$ et φ satisfait les mêmes conditions que w précédemment, alors la borne devient

   $$\sup_{\theta\in\Theta} E_\theta\big[(\widehat{g}(Y) - p_\theta(0))^2\big] \geqslant \varphi(0)^2\Big(2n\|\varphi\|_2^2/M + M^2 I_\pi\Big)^{-1}.$$

   Un calcul de variations montre qu'un bon choix pour minimiser I_π est de prendre $\pi(\theta) = 2\sin^2(\pi\theta)1_{0\leqslant\theta\leqslant 1}$. Vérifier que dans ce cas $I_\pi = 4$.

3. Soit l'ensemble de densités sur $[-1/2, 1/2]$

   $$\mathscr{E} = \Big\{p : |p(x) - p(y)| \leqslant |x - y|, p \geqslant 0, \int p(x)dx = 1\Big\}.$$

   En remarquant que $\mathscr{E}$ 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 $\widehat{g}(Y)$ de la densité en 0, on a

   $$\sup_{p\in\mathscr{E}} E_p\big[(\widehat{g}(Y) - p(0))^2\big] \geqslant \tfrac{1}{6}n^{-2/3}.$$

   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 θ 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 :

PROBLEM.

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 $P(\theta_1 < \theta < \theta_2 | Y)$?

Ceci suppose l'introduction d'une distribution pour θ. On se donne donc une probabilité a priori $\pi(d\theta)$ sur θ, et la paire (θ, Y) est probabilisée avec la loi

$$p_\theta(y)\mu(dy)\pi(d\theta).$$

Tout se passe comme si θ était d'abord tiré avec la loi π, puis y tiré conditionnellement à θ avec la loi de densité $p_\theta(y) = p(y|\theta)$. On verra que, dans les faits, la distribution π représente davantage un apport d'information qu'une réelle distribution du paramètre.

On en déduit la loi de θ sachant Y (probabilité a posteriori) :

$$p(d\theta|Y) = \frac{p_\theta(Y)\pi(d\theta)}{\int p_\tau(Y)\pi(d\tau)} \tag{III.35}$$

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 μ, 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 θ) d'un vecteur gaussien centré (X, Y) conditionnellement à Y comme étant $\mathcal{N}(R_{XY}R_{YY}^{-1}Y, R_{XY}R_{YY}^{-1}R_{YX})$, 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[20], colinéarité en régression, etc.). Le rôle de π comme apport d'information supplémentaire, même vague, est ici de stabiliser l'estimation.
   L'estimateur ridge en régression linéaire (exercice 4 p. 91) est un exemple classique.

Elle permet également de donner une interprétation parlante aux estimateurs pénalisés (p. ex. 4 p. 91).

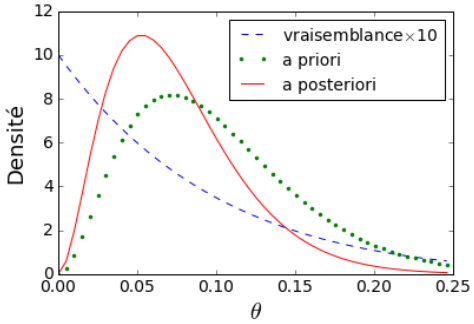
Mentionnons finalement qu'il arrive, dans les cas non-réguliers, qu'un estimateur bayésien soit asymptotiquement strictement meilleur que celui du maximum de vraisemblance (exercice 2 p. 90).

Exemple. On s'intéresse au taux de mortalité θ 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 θ ?

Notons d'abord que la présence de répétitions (plusieurs hôpitaux) est un argument pour considérer θ 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 $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$. La distribution Bêta$(3, 27)$ a pour espérance 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, par la formule de Bayes, la distribution a posteriori Bêta(3,37) pour θ :

$$p(\theta|y = 10) \propto p(y = 10|\theta)\frac{p(\theta)}{p(y = 10)} \propto (1-\theta)^{10}\Big(\theta^2(1-\theta)^{26}\Big) = \theta^2(1-\theta)^{36}.$$

20. P. ex. la prédiction des votes Carter/Ford (présidentielles 1976) à l'aide de 6 variables socio-économiques et géographiques pour chaque état dans [23] § 5.2.



Vraisemblance : $(1-\theta)^{10}$
A priori : $\frac{1}{B(3,27)}\,\theta^2(1-\theta)^{26}$
A posteriori : $\frac{1}{B(3,37)}\,\theta^2(1-\theta)^{36}$.

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 π rend son interprétation sujette à discussions homériques et récurrentes [21] ; car bien entendu π sera toujours soumise à un certain arbitraire. Tout dépend donc de l'importance de cette dépendance et du crédit que l'on prête à π.

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 π. L'intervalle de crédibilité est plus simple à exprimer car θ 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 fausse ($\theta \notin \mathcal{R}$ 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 posteriori.

Exemple [53]. 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. 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 espérance $\alpha\beta t_i$; 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 :

$$\pi(\alpha, \beta) \propto \{\alpha(1+\beta)^2\}^{-1}.$$

C'est la mesure uniforme sur $\mathbb{R}\times[0,1]$ pour $(\log\mu, 1/(1+\beta))$ [22]. La loi de (α, β) sachant les observations s'obtient par la formule de Bayes

$$p(\alpha, \beta|Y) \propto \frac{1}{\alpha(1+\beta)^2}\prod_{i=1}^{n}\frac{\beta^{Y_i}\Gamma(Y_i+\alpha)}{(1+t_i\beta)^{\alpha+Y_i}\Gamma(\alpha)}\,. \tag{III.36}$$

Bien que $\pi(\alpha, \beta)$ ne soit pas d'intégrale finie, la mesure conditionnelle ci-dessus est bien finie [23]. Les auteurs peuvent ainsi obtenir par simulation la distribution a posteriori de μ dans différents sous-groupes liés aux types d'infraction (*traffic de drogue*, *possession de drogue*, *cambriolage*, *vol avec agression*), à la peine (*emprisonnement* ou *pas*) et à la période. Ils observent en particulier que les trafiquants de drogue ayant subit une peine de prison ont un taux d'arrestation pour des délits non liés à la drogue significativement inférieur aux autres.

21. Cf. [79] et aussi la discussion de Larry Wasserman, qui suit l'article.
22. Le fait que cette loi ne soit pas une probabilité (distribution impropre) est un défaut couramment accepté en théorie bayésienne (cf. l'estimateur de Pitman plus bas).
23. Nous ne détaillons pas cet exercice : Remarquer que si $Y_i > 0$, $\Gamma(\alpha+Y_i) = (\alpha+Y_i-1)...(\alpha+1)\alpha\Gamma(\alpha)$, ce qui permet alors de majorer (III.36) 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 α puis en β.

L'estimateur bayésien. La méthode bayésienne consiste à minimiser en $\widehat{\theta}$ le risque calculé sur la base de la probabilité totale :

$$R(\widehat{\theta}) = \int R_\theta(\widehat{\theta})\pi(d\theta) = \iint r(\theta, \widehat{\theta}(y))p_\theta(y)\pi(d\theta)\mu(dy)$$

puis à définir les estimateurs bayésiens par la propriété :

> *$\widehat{\theta}_b$ est bayésien si pour tout estimateur $\widehat{\theta}$ on a $R(\widehat{\theta}_b) \leqslant R(\widehat{\theta})$.*

Ceci donne

$$\widehat{\theta}_b(y) = \arg\min_{\theta'} \int r(\theta, \theta')p_\theta(y)\pi(d\theta) = \arg\min_{\theta'} E\big[r(\theta, \theta')|Y = y\big] \quad \mu - p.p. \tag{III.37}$$

(au moins lorsque cette fonction est mesurable). Deux observations sont en théorie très importantes :

1. [Unicité implique admissibilité] *Si $\mu(\widehat{\theta}_b(y) \neq \widehat{\theta}'_b(y)) = 0$ pour tout autre estimateur π-bayésien $\widehat{\theta}'_b$, alors $\widehat{\theta}_b$ est admissible.*
   En effet, si $\widehat{\theta}_b$ n'est pas admissible, alors il existe un autre estimateur $\widehat{\theta}$ strictement meilleur ; $\widehat{\theta}$ est forcément bayésien et comme pour un certain θ on a $E_\theta[r(\theta, \widehat{\theta})] < E_\theta[r(\theta, \widehat{\theta}_b)]$, nécessairement $P_\theta(\widehat{\theta}(y) \neq \widehat{\theta}_b(y)) > 0$.
2. [Risque constant implique minimax] *S'il existe c tel que $\pi(R_\theta(\widehat{\theta}_b) = c) = 1$ et $\sup_\theta R_\theta(\widehat{\theta}_b) \leqslant c$, alors $\widehat{\theta}_b$ est minimax* (cf. exercice 10 p. 92 ; résultat dû à Hodge et Lehmann, 1951).

Ceci donne donc une méthode pour fabriquer des estimateurs admissibles minimax : *Trouver une distribution π qui satisfasse ces deux conditions.* Ceci n'est pas forcément simple ; il faut noter qu'une telle distribution π est **la plus défavorable** au sens où elle maximise le risque bayésien (exercice 10 p. 92) ; 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, π se concentre autour de $\frac{1}{2}$ ce qui conduit à un estimateur dont le risque est fortement détérioré ailleurs (vis-à-vis du maximum de vraisemblance) : le risque obtenu est constant égal à $\frac{1}{4(1+\sqrt{n})^2}$ (cf. la discussion au § III.1.1, et le § III.3.3).

Dans le cas du risque quadratique, on peut être plus explicite :

33 - Théorème (Moyenne a posteriori)

On suppose que $r(\theta, \theta') = \|\theta - \theta'\|^2$ et que $\int \|\theta\|^2\pi(d\theta) < \infty$, alors

$$\widehat{\theta}_b = E[\theta|Y] = \frac{\int \theta p_\theta(Y)\pi(d\theta)}{\int p_\theta(Y)\pi(d\theta)}. \tag{III.38}$$

De plus, si $Z(y) = \int p_\theta(y)\pi(d\theta)$ est μ-p.s. non-nul, $\widehat{\theta}_b$ est admissible.

Démonstration. Pour tout autre estimateur $\widehat{\theta}'$, on a

$$\begin{aligned} R(\widehat{\theta}') &= E\big[\|\theta - \widehat{\theta}'(Y)\|^2\big] \\ &= E\big[\|\theta - E[\theta|Y] + \widehat{\theta}_b(Y) - \widehat{\theta}'(Y)\|^2\big] \\ &= E\big[\|\theta - E[\theta|Y]\|^2\big] + E\big[\|\widehat{\theta}_b(Y) - \widehat{\theta}'(Y)\|^2\big] \quad \text{(Pythagore)} \\ &= R(\widehat{\theta}_b) + \int \|\widehat{\theta}_b(y) - \widehat{\theta}'(y)\|^2 Z(y)\mu(dy). \end{aligned}$$

Donc $\widehat{\theta}_b$ est bien bayésien. Si $\widehat{\theta}'$ est toujours meilleur que $\widehat{\theta}_b$ (i. e. $\forall\theta,\ R_\theta(\widehat{\theta}') \leqslant R_\theta(\widehat{\theta}_b)$), alors le deuxième terme est nécessairement nul, et l'hypothèse faite sur Z implique que $\mu\big(\{\widehat{\theta}_b(y) \neq \widehat{\theta}'(y)\}\big) = 0$. Donc $R_\theta(\widehat{\theta}') = R_\theta(\widehat{\theta}_b)$ pour tout θ ce qui empêche $\widehat{\theta}'$ d'être strictement meilleur que $\widehat{\theta}_b$, d'où l'admissibilité. ∎

Sous des hypothèses générales, dans une asymptotique où le nombre d'échantillons tend vers l'infini mais π *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. [11] Th. 8.1 et Th. 8.2, ou [41] Th. 36.9). Ceci vient essentiellement de ce que dans le numérateur de (III.35), 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_{θ_*}. Densité a posteriori des observations. La loi $\pi(d\theta)$ induit, par l'application $\theta \mapsto p_\theta(.)$, une mesure sur l'espace des densités de probabilité par rapport à μ. Sur cet espace on peut considérer le risque (divergence de Kullback-Leibler)

$$r(\widehat{p}, p_\theta) = D(p_\theta \| \widehat{p}) = \int \log\Big(\frac{p_\theta(y)}{\widehat{p}(y)}\Big) p_\theta(y)\mu(dy).$$

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.37) sans oublier la contrainte $\int \widehat{p}(y)\mu(dy) = 1$, que sous des hypothèses raisonnables l'estimateur

$$\widehat{p}(x) = \int p_\theta(x) p(d\theta | y = Y) = \frac{\int p_\theta(x) p_\theta(Y)\pi(d\theta)}{\int p_\theta(Y)\pi(d\theta)} \tag{III.39}$$

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_{θ_*} 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 π_k uniforme sur $[-k, k]$ on obtient l'estimateur bayésien $\widehat{\theta}_b^k$ pour le risque quadratique

$$\widehat{\theta}_p^k(Y) = \frac{\int_{-k}^k \theta p_0(Y - \theta) d\theta}{\int_{-k}^k p_0(Y - \theta) d\theta}, \qquad p_0(Y - \theta) = \prod_{i=1}^n p_0(Y_i - \theta).$$

L'estimateur de Pitman s'obtient en faisant tendre k vers l'infini

$$\widehat{\theta}_p(Y) = \frac{\int \theta p_0(Y - \theta) d\theta}{\int p_0(Y - \theta) d\theta}. \tag{III.40}$$

Il faut bien entendu des hypothèses adéquates d'intégrabilité. On a donc au final une distribution π impropre (la mesure de Lebesgue). On montre que cet estimateur (de risque constant) est minimax sous certaines hypothèses (exercice 10 p. 92 pour le cas où p_0 est supposé à support compact, et [12] Th.5.3.6 pour le cas général).

Notons que si π est une mesure finie, $\widehat{\theta}_b$ est forcément biaisé ; $\widehat{\theta}_p$ en revanche peut être non-biaisé (exercice 5 p. 91).

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

$$p_\theta(y) = e^{\langle T(\theta), y\rangle - Z(\theta)}$$

alors le choix

$$\pi(d\theta) = \pi_{\alpha,\beta}(\theta)d\theta = c_{\alpha,\beta} e^{\langle T(\theta), \alpha\rangle - \beta Z(\theta)} d\theta$$

où α et β sont des paramètres à choisir et $c_{\alpha,\beta}$ est la constante de normalisation, conduit à la loi a posteriori (loi de θ 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 β **observations fictives** de valeur α/β.

Si la loi de Y_i est normale, gamma, Poisson ou binomiale, on obtient respectivement pour θ une loi normale, gamma, gamma ou bêta.

On trouvera une application à l'inference variationnelle à l'exercice 13 p. 94.

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 [68]. En effet, si l'on suppose que θ est une fonction mesurable de la suite iid $(Y_i)_{1\geqslant 1}$ [24], on a pour toute fonction borélienne bornée f

$$E\big[f(\theta)|Y_1,\dots Y_n\big] \longrightarrow E\big[f(\theta)|Y_1,\dots\big] = f(\theta) \quad p.s.$$

Moyennant une hypothèse d'intégrabilité de θ, en particulier si Θ est borné, ceci règle le cas du risque quadratique en prenant $f(\theta) = \theta$ [25].

Bilan. La question du choix de π 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. Utilisé en inference variationnelle (exercice 13 p. 94).
- 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. 92. 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 θ, 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 [26]. 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 θ 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 π_β, et estimer le paramètre β 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 6 p. 91 pour un exemple où β 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 θ 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 θ est un vecteur ayant beaucoup de composantes nulles, Johnstone et Silverman [102] choisissent pour chaque composante le même mélange d'une masse de Dirac en 0 et d'une loi exponentielle :

    $$\pi_{a,w}(d\theta) = \psi_{a,w}(d\theta_1)\dots\psi_{a,w}(d\theta_d), \quad \psi_{a,w}(dx) = (1-w)\delta_0(dx) + \tfrac{1}{2}wae^{-a|x|}dx.$$

    Ici $\beta = (a,w) \in \mathbb{R}_+ \times [0,1]$. En dépit du caractère très simpliste de la famille ψ_β, les résultats sont très bons.
  - ▶ Bayes empirique non-paramétrique : Il s'agit de calculer π au maximum de vraisemblance *sous la contrainte* que les θ_j sont iid sous π (i. e. $\pi(d\theta) = \psi(d\theta_1)\dots\psi(d\theta_d)$) [27]. Ici $\beta = \psi$. L'estimée $\widehat{\psi}$ sera typiquement une somme de masses de Dirac [101], solution d'un problème assez lourd à mettre en œuvre [28].
    Voir aussi les exercices 7 et 8 où (III.38) est estimé directement sans passer par π !
    Si θ est de dimension 1, la contrainte ne joue plus et c'est le maximum de vraisemblance.

24. 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_θ sont toutes différentes ; alors θ est l'argument du maximum en τ de $\int \log p_\tau(y)p_\theta(dy) = \lim n^{-1}\sum_{i=1}^n \log p_\tau(Y_i)$; si pour tout θ 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.

25. On trouvera des compléments dans [84].

26. Il est à noter que cette mesure est asymptotiquement la plus défavorable (cf. plus haut p. 87) pour le risque de Kullback [52]. 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. [5] § 2.9.

27. Une idée de Kiefer et Wolfowitz (1956), voir l'exercice 12 p. 93, mais qu'ils n'utilisent pas pour estimer θ ensuite.

28. Une mise en œuvre pour des modèles spécifiques est réalisée dans le programme R REBayes [89].

Nous renvoyons à [5] pour des traitements d'exemples et des approfondissements.

Des méthodes d'esprit bayésien ont été proposées pour traiter des cas où l'on sort du cadre paramétrique strict au sens ou la dimension de l'espace des paramètres n'est plus contrôlée, pouvant même être supérieure au nombre d'individus, le résultat de Barron et Leung présenté à l'exercice 14 en est un excellent exemple ; elles ont débouché dans les années 2000 sur l'**agrégation à poids exponentiels** (exponentially weighted aggregate (EWA)) [26], sujet de l'exercice 15. Le but originel de ces méthodes est de combiner un grand nombre d'estimateurs. L'approche n'est plus paramétrique, et donc plutôt discriminative que générative (cf. p. 6). Leur efficacité théorique peut être étonnante, et si ces estimateurs sont généralement incalculables exactement, des approximations obtenues par méthode de Monte-Carlo peuvent être très satisfaisantes.

III.4.2 Le maximum a posteriori (MAP).

L'estimation de θ 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 [29], il vient :

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} \pi(\theta) p_\theta(Y).$$

Cet estimateur correspond également à l'estimateur bayésien limite $\varepsilon \to \infty$ quand $r(\theta, \theta') = \varphi(\|\theta - \theta'\|)$ où φ 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 θ prend des valeurs discrètes : ce choix minimise la probabilité d'erreur $P(\widehat{\theta} \neq \theta|Y)$ (a fortiori, cet estimateur minimise $P(\widehat{\theta} \neq \theta)$).

L'estimateur lasso pour le modèle linéaire $Y \sim \mathcal{N}(X\beta, \sigma^2 Id)$

$$\widehat{\beta}_L = \arg\min_{\beta} \tfrac{1}{2}\|Y - X\beta\|^2 + \lambda \sum_j |\beta_j| \tag{III.41}$$

est un estimateur MAP pour une distribution exponentielle des coefficients. L'estimateur ridge est à la fois MAP et bayésien (exercice 4 p. 91).

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 pénalisée. 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 à posteriori, le MAP étant considéré comme une information trop succincte.

L'exercice 2 propose un exemple où l'estimateur bayésien est asymptotiquement strictement meilleur que le maximum de vraisemblance mais pas le MAP. L'exercice 4 propose un exemple où l'estimateur bayésien est égal au MAP, et où le maximum de vraisemblance peut ne même pas être défini (cas $p > n$).

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 θ 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 (Bayésien supérieur au maximum de vraisemblance et au MAP). On considère l'exemple de l'exercice 5 p. 7 dont on pourra admettre la conclusion. Soit l'estimateur bayésien pour la mesure $\pi(dx) = x^{-2}dx$, cette distribution impropre étant 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. Vérifier que ce n'est pas le cas du maximum a posteriori.

Une étude plus poussée montre que le résultat reste vrai du moment que $\pi(dx) = f(x)dx$ avec f continue bornée > 0.

Exercice 3 (Formule de Tweedie [73]). Soit $Y \sim \mathcal{N}(\theta, 1)$ et une certaine loi a priori π pour θ. On désigne par $f(y)$ la loi inconditionnelle de Y (mélange de gaussiennes). On supposera π à support

29. On a bien entendu $p(y) = \int p(\theta, y)d\theta = \int p_\theta(y)\pi(\theta)d\theta$.

compact ce qui rend f très régulière et rapidement décroissante à l'infini. Exprimer l'estimateur bayésien $\widehat{\theta}(Y)$ en fonction de Y, $f(Y)$ et $f'(Y)$ seulement (formule de Tweedie). Montrer que, en toute généralité, $R(\widehat{\theta}) = E[\theta^2] - E[\widehat{\theta}^2]$ (utiliser $R(\widehat{\theta}) = E\big[\|\theta - E[\theta|Y]\|^2\big]$). En déduire que le risque vaut

$$R(\widehat{\theta}) = 1 - \int \frac{f'(y)^2}{f(y)} dy.$$

Exercice 4 (L'estimateur ridge est bayésien). Soit le modèle linéaire gaussien (I.7), soit

$$Y = X\beta_* + u, \quad Y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n\times p}, \quad u \sim \mathcal{N}(0, \sigma_*^2 Id_n). \tag{III.42}$$

L'estimateur au maximum de vraisemblance $\widehat{\beta}_{MV} = (X^TX)^{-1}X^TY$ pose des problèmes si le nombre de variables p est grand car la matrice X^TX risque d'être mal conditionnée ; il arrive même que $p > n$. On a proposé l'estimateur pénalisé, appelé estimateur ridge :

$$\widehat{\beta}_R = \arg\min_{\beta} \tfrac{1}{2}\|Y - X\beta\|^2 + \lambda\|\beta\|^2. \tag{III.43}$$

où $\lambda > 0$ est un paramètre à choisir (possiblement par validation croisée).

1. Vérifier que
$$\widehat{\beta}_R = \big(X^TX + \lambda Id_p\big)^{-1} X^TY. \tag{III.44}$$
2. Soit la loi a priori $\beta_* \sim \mathcal{N}(0, \frac{\sigma_*^2}{\lambda} Id_p)$. Il est clair que $\widehat{\beta}_R$ est l'estimateur MAP. Vérifier que c'est également l'estimateur bayésien $\widehat{\beta}_R = E[\beta_*|Y]$.
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 $X = Id$ (cas des « normal means »). Vérifier que dans ce cas, la validation croisée pour le choix de λ ne peut fonctionner. On décide de choisir λ par minimisation du risque estimé par SURE, $\widehat{R} = \widehat{R}(\lambda)$ de l'équation (I.25) p. 17. Montrer que l'on retrouve alors quasiment l'estimateur de James-Stein de l'exercice 10 p. 66.

Exercice 5 (Un estimateur bayésien est toujours biaisé). Pour le $\widehat{\theta}$ du § III.4.1 avec risque quadratique, montrer qu'on a

$$R(\widehat{\theta}) = -\int \theta^T b(\theta)\pi(d\theta)$$

où $b(\theta) = E[\widehat{\theta}|\theta] - \theta$ est le biais. On pourra utiliser (et justifier) que $R(\widehat{\theta}) = E[\theta^T(\theta - \widehat{\theta})]$.

Attention, ceci ne s'applique que si la mesure π 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 $n = 1$.

Nous approfondirons à l'exercice 11 plus bas.

Exercice 6 (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 $\theta \sim \mathcal{N}(0, \tau^2 I_d)$. σ est connu.

1. Exprimer (θ, Y) comme fonction linéaire d'une autre paire $(U, V) \sim \mathcal{N}(0, I_{2d})$. En déduire que (θ, Y) est un vecteur gaussien dont on donnera la matrice de covariance.
2. Calculer l'estimateur bayésien $\widehat{\theta}_\tau$ de θ 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 θ.
4. Quelle est la loi de Y (i. e. inconditionnelle à θ) ?
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 θ pour le cas où τ est inconnu (σ est toujours connu). Vérifier qu'il s'agit de l'estimateur de James-Stein de l'exercice 10 p. 66 (on pourra supposer $\sigma = 1$).

Exercice 7 (Estimateur de Robbins-Turing [47, 71]). Cet estimateur a été un des instruments de la méthode d'estimation de Turing et Good dans le décryptage du code Enigma [29].

On observe $Y_i \sim \mathcal{P}(\theta_i)$, $i = 1, \dots n$, p. ex. Y_i est le nombre d'occurences du i-ième mot du dictionnaire dans les œuvres de Victor Hugo. Soit une distribution π pour les paramètres θ_i indépendants ; le paramètre est donc $\theta = (\theta_i)_{1 \leqslant i \leqslant n}$. Vérifier que $E\big[\theta_i|Y\big] = E\big[\theta_i|Y_i\big]$, et que (III.38) conduit à

$$E\big[\theta_i|Y_i\big] = (Y_i + 1)\frac{P(Y_i + 1)}{P(Y_i)}$$

où P est la distribution de Y_i résultant du modèle bayésien. Cette identité suggère l'estimateur :

$$\widehat{\theta}_i = (Y_i + 1)\frac{\widehat{P}(Y_i + 1)}{\widehat{P}(Y_i)}, \quad \widehat{P}(k) = \mathrm{Card}\big(\{j : Y_j = k\}\big)/n.$$

Une méthode de lissage a été mise au point pour diminuer la variabilité du quotient [82].

Noter que conditionnellement aux θ_i et à la somme N des Y_i, la suite des Y_i suit une distribution multinomiale de paramètres $p_1, \dots p_n$, et N, $p_i = \theta_i / \sum \theta_j$; ce sont ces paramètres p_i (fréquence théorique de chaque mot) qui sont en pratique réellement recherchés, le modèle poissonien n'étant qu'un artifice classique pour travailler sur des variables indépendantes.

On pose $\widehat{p}_i = \widehat{\theta}_i / \sum \widehat{\theta}_j$. Vérifier que $\sum \widehat{\theta}_i = N$, et $\sum_{Y_i = 0} \widehat{\theta}_i = n\widehat{P}(1)$. Sans même connaître le dictionnaire, la proportion de mots non encore observés est donc estimée à la proportion de mots observés une fois.

Exercice 8 (Dans le même style). $\mathcal{E}(\lambda)$ désigne la loi exponentielle d'espérance $1/\lambda$. On observe une suite indépendante $Y_i \sim \mathcal{E}(\theta_i)$, $1 \leqslant i \leqslant n$, chaque θ_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 π.
2. Calculer l'estimateur bayésien $E[\theta_i|Y_i]$. Il dépend de π.
3. L'exprimer comme fonction de p et Y_i, mais sans faire apparaître π. 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 [46]).
5. (Un petit supplément) Reprendre le point 3 pour le calcul de $E[\theta_i^{-1}|Y_i]$. Quel est l'avantage ?

Exercice 9. On considère l'exemple de l'exercice 4 p. 7 avec $p > n$. On note $Z = (X, Y)$. Vérifier que :

$$\mathscr{L}(Z, \theta) = -\sum_{i=1}^{n} \frac{(X_i - \mu_i)^2}{2\sigma^2} - \sigma \sum_{j=1}^{p} Y_j + (p - n)\log\sigma \tag{III.45}$$

$$I(\theta) = \sigma^{-2}\begin{pmatrix} Id & 0 \\ 0 & 2n + p \end{pmatrix}. \tag{III.46}$$

1. Calculer l'estimateur bayésien de $\theta = (\mu_1, ...\mu_n, \sigma)$ si π 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 σ pour la mesure de Jeffreys $\pi(d\theta) = \det(I(\theta))^{1/2} d\theta$, et montrer que l'on retrouve le problème original.

Exercice 10 (Estimateurs bayésiens et estimateurs minimax). Soit $\pi(d\theta)$ une distribution sur Θ et $\widehat{\theta}_b$ un estimateur bayésien associé. On se propose de justifier la construction proposée au § III.4.1, et de démontrer le point 2 de la page 87.

1. Soit $\widehat{\theta}$ tel que pour tout θ' : $R_{\theta'}(\widehat{\theta}) \leqslant \int R_\theta(\widehat{\theta}_b)\pi(d\theta)$

   (i) Montrer que $\widehat{\theta}$ est minimax, que $R_\theta(\widehat{\theta})$ est π-presque sûrement égal au risque minimax, partout inférieur à ce dernier, et que $\widehat{\theta}$ est π-bayésien.
   Indication : Montrer directement que pour tout estimateur $\widehat{\theta}'$ on a $\sup_\theta R_\theta(\widehat{\theta}) \leqslant \sup_\theta R_\theta(\widehat{\theta}')$.

(ii) Vérifier que ceci démontre le point 2 de la page 87 en considérant $\widehat{\theta} = \widehat{\theta}_b$ (immédiat).

(iii) Vérifier que tout estimateur π-bayésien $\widehat{\theta}_b$ satisfaisant les hypothèses du point 2 satisfait $R_{\theta'}(\widehat{\theta}) \leqslant \int R_\theta(\widehat{\theta}_b)\pi(d\theta)$ pour tout $\widehat{\theta}$ et tout θ' (immédiat en utilisant le point 2).

(iv) En déduire que si $\widehat{\theta}_b$ est un estimateur π-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 π est la distribution *la plus défavorable*.

On va voir dans la suite comment l'égalisation du risque se fait par un choix de π qui donne un poids élevé aux points où l'estimation est plus difficile ($\widehat{\theta}_b$ est biaisé vers ces points). π 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

$$\widehat{\theta} = \frac{\overline{Y} + a}{1 + b}, \quad \overline{Y} = \frac{1}{n}\sum Y_i.$$

(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 $\frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}\ t^{u-1}(1-t)^{v-1}$. Son espérance 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 33 pour montrer l'admissibilité. Calculer l'information de Fisher et vérifier que lorsque n tend vers l'infini, la distribution π 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 $R_\theta(\widehat{\theta}) \leqslant \sup_k \int R_\theta(\widehat{\theta}^k_b)\pi_k(d\theta)$ pour une suite infinie d'estimateurs bayésiens $\widehat{\theta}^k_b$.

Application : On considère l'estimateur de Pitman avec les notations de la page 88. On suppose que p_0 est à support inclus dans un certain intervalle $[-A, A]$.

(a) Montrer que pour tout θ et tout k, $R_\theta(\widehat{\theta}^k_p) \leqslant 4A^2$ et vaut $R_\theta(\widehat{\theta}_p)$ si $k \geqslant |\theta| + 2A$.

(b) En déduire que $R_\theta(\widehat{\theta}_p) = \lim_k \int R_\theta(\widehat{\theta}^k_b)\pi_k(d\theta)$. Il s'ensuit que $\widehat{\theta}_p$ est minimax. Vérifier qu'il est de risque constant.

Exercice 11 (Estimateur de Barankin. Estimateur bayésien non-biaisé). Soit l'estimateur de l'exercice 15 p. 68, 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

$$\int \big(\widehat{\theta}(x) - \theta_q\big)^2 q(x)\mu(dx), \qquad \theta_q = \int \widehat{\theta}(x)q(x)\mu(dx)$$

sous la contrainte

$$\theta = \int \widehat{\theta}(x)p_\theta(x)\mu(dx), \quad \theta \in \Theta. \tag{III.47}$$

On choisit q de la forme $q(x) = \int p_\theta(x)\pi(d\theta)$ pour une certaine mesure π. Démontrer que $\widehat{\theta}(x)$ est l'estimateur π-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.47). *Indication* : Montrer que tous les estimateurs $\widehat{\theta}$ satisfaisant (III.47) ont le même θ_q, puis utiliser $\widehat{\theta}(x) - \theta = (\widehat{\theta}(x) - \theta_q) + (\theta_q - \theta)$.

***Exercice 12 (NPMLE : Bayes empirique non-paramétrique [30]).** On se donne le modèle suivant pour une suite d'observations indépendantes à valeurs dans $\mathbb{R}^d$:

$$Y_k \sim f(y; \beta, \alpha_k)dy, \quad k = 1 \dots n$$

30. Complètement inspiré de Kiefer et Wolfowitz [105]. On trouvera des compléments intéressants par exemple dans [90], [63] et [87], et une mise en œuvre dans le programme REBayes [89].

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

$$Y_{ij} = \beta + \alpha_i e_{ij}, \quad 1 \leqslant j \leqslant n_0, \quad 1 \leqslant i \leqslant n \tag{III.48}$$

où les e_{ij} sont iid $\mathcal{N}(0,1)$, $Y_{i\cdot}$ peut être une série de n_0 mesures faites par l'opérateur i ; ou encore

$$Y_{ij} = \alpha_i + \beta e_{ij}, \quad 1 \leqslant j \leqslant n_0, \quad 1 \leqslant i \leqslant n. \tag{III.49}$$

$Y_{i\cdot}$ peut être une série de n_0 mesures faites sur le sujet i. 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 [105] proposent de postuler que les α_k forment une suite iid de distribution μ inconnue (*indépendance et homogénéïté* des paramètres), puis d'estimer $\theta = (\beta, \mu)$ au maximum de vraisemblance. On a donc le contraste

$$K(y,\theta) = -\log \int f(y;\beta,\alpha)\mu(d\alpha), \qquad k(\theta) = -E\Big[\log \int f(Y_1;\beta,\alpha)\mu(d\alpha)\Big].$$

On se donne pour Θ 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 ([39] Th. 6.8), ce qui signifie

$$\Big(d_P(\mu,\mu_n) \longrightarrow 0\Big) \Longleftrightarrow \Big(\forall f \text{ continue bornée sur } A,\ \mu_n(f) \longrightarrow \mu(f)\Big).$$

La compacité de A fait que $\mathcal{M}$ est compact pour cette métrique ([39] Th. 5.1). Soit la distance sur Θ

$$d(\theta,\theta') = |\beta - \beta'| + d_P(\mu,\mu').$$

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 (μ_*,β_*) unique tel que la loi de Y_1 soit $\int f(y,\alpha,\beta_*)\mu_*(d\alpha)$.

Montrer les points suivants :

1. (Θ, d) est compact.
2. Sous (H1), (H2) et (H3), les hypothèses (a,b,c) du théorème 14 sont satisfaites.
   Indication : Si μ_n et ν_n sont deux suites de mesures convergent faiblement vers μ et ν, alors $\mu_n \times \nu_n$ converge faiblement vers $\mu \times \nu$ ([39] Th. 2.8).

Montrer que :

1. L'exemple (III.48) satisfait les hypothèses si A est de la forme $[a,b]$ avec $a > 0$.
2. L'exemple (III.49) satisfait les hypothèses si $n_0 > 1$, et $B = [a,b]$, $a > 0$. Expliquer pourquoi $n_0 > 1$ est nécessaire.

Signalons au passage que l'estimée $\widehat{\mu}_n$ peut être recherchée parmi les combinaisons linéaires de $n+1$ masses de Dirac [31].

Exercice 13 (Inférence variationnelle). Soit une paire de v.a. (U,X) où sont explicites les densités de U, notée π, et $p(X|U)$ de X sachant U. Seul X est observé et l'on cherche à exprimer $p(U|X)$; c'est un problème purement numérique pour lequel l'inférence variationnelle propose une solution *approchée*. Il faudra pour cela que p et π admettent les formes (III.54) et (III.55).

Ce problème apparaît si $U = \theta$ est le paramètre d'un modèle bayésien et que l'on recherche une forme explicite de $p(\theta|X)$, ou lorsque U est une variable latente d'un certain modèle (p. ex. la variable Z de l'algorithme EM, exercice 13 p. 29. Pour un exemple typique voir [62] § 5.2).

Comme $p(U|X) = p(X|U)\pi(U)/p(X)$, seul le calcul de $p(X)$ pose problème. Comme $p(U|X) =$

31. Pour toute mesure μ, notons $h(\mu) = (\int f(Y_i,\alpha,\widehat{\beta})\mu(d\alpha))_{1\leqslant i\leqslant n}$. L'ensemble des vecteurs $H = \{h(\mu) : \mu \in \mathcal{M}\} \subset \mathbb{R}^n$ est l'enveloppe convexe de $H_D = \{h(\mu) : \mu \text{ masse de Dirac}\}$. En vertu du théorème de Carathéodory, $h(\widehat{\mu})$, comme tout point de H, est combinaison de $n+1$ points de H_D, points qui fournissent la solution désirée.

$\int p(X|U=u)\pi(u)du$ ceci demande un calcul d'intégrale numériquement prohibitif si la dimension de u n'est pas petite.

On note d'abord que la densité $\psi(.) = P(.|X)$ est solution de

$$\min_{\psi} -\int \ln p(u|X)\psi(u)du + \int \ln(\psi(u))\psi(u)du \tag{III.50}$$

(divergence de Kullback-Leibler entre $p(.|X)$ et $\psi(.)$) où le minimum est pris sur les mesures de probabilité. C'est donc aussi la solution de

$$\min_{\psi} -\int \ln p(u,X)\psi(u)du + \int \ln(\psi(u))\psi(u)du \tag{III.51}$$

puisque les expressions ne diffèrent que de $\ln p(X)$. On se propose de minimiser sur un ensemble plus petit sur lequel le minimum sera facilement calculable ; l'ensemble choisi est l'ensemble des densités rendant les composantes de u indépendantes [32] : $\psi(u) = \prod_j \psi_j(u_j)$, l'objectif étant de remplacer un problème en dimension $n = \dim(n)$ par n problèmes en dimension 1 si les calculs se découplent bien. De plus le résultat obtenu permet une interprétation plus simple dans la suite car les u_i étant indépendants, on peut calculer leurs moments facilement.

1. On note $u^{(i)} = (u_1, \dots u_{i-1}, u_{i+1}, \dots u_n)$, $\psi^{(i)}(u^{(i)}) = \prod_{k\neq i} \psi_k(u_k)$, etc. Montrer que pour chaque j, ψ_j est solution de

$$\min_{\psi_j} -\int \left(\int \ln p(u,X)\psi^{(j)}(u^{(j)})du^{(j)} \right)\psi_j(u_j)du_j + \int \ln(\psi_j(u_j))\psi_j(u_j)du_j \tag{III.52}$$

2. En déduire, en utilisant ce qui a été observé plus haut concernant les solutions de (III.50) et (III.51) que

$$\psi_j(u_j) \propto \exp\left(\int \ln p(u,X)\psi^{(j)}(u^{(j)})du^{(j)} \right) \tag{III.53}$$

   Ce qui fait un jeu de n équations pour les fonctions $(\psi_1, \dots \psi_n)$. Ce système fonctionnel se ramène à un système vectoriel si ces fonctions sont paramétrées. On voit dans la suite qu'une paramétrisation adéquate rend les intégrales facilement calculables, ce qui permettra une résolution effective.

3. Supposons qu'existent des fonctions vectorielles l_j telles que $p(x|u)$ soit de la forme

$$\ln p(x|u) = \sum_j t_j(x)^T l_j(u_j) + \sum_{j<k} l_j(u_j)^T m_{jk}(x) l_k(u_k) + \rho(x). \tag{III.54}$$

   où chaque $t_j(x)$ (resp. $m_{jk}(x)$) est une fonction vectorielle (resp. matricielle) de dimensions adéquates. On postule alors une forme analogue conjuguée pour π :

$$\ln \pi(u) = \sum_j \tau_j^T l_j(u_j) + \sum_{j<k} l_j(u_j)^T \mu_{jk} l_k(u_k) + \sum_j \beta_j(u_j). \tag{III.55}$$

   Montrer que, en posant $m_{kj} = m_{jk}^T$ pour $k < j$, on a

$$\ln \psi_j(u_j) = l_j(u_j)^T \left(\tau_j + t_j(X) + \sum_{k\neq j} (\mu_{jk} + m_{jk}(X)) \int l_k(u_k)\psi_k(u_k)du_k \right) + \beta_j(u_j) + cst.$$

   Chaque ψ_j appartient donc à la famille exponentielle associée à l_j avec densité e^{β_j}, i.e. $\psi_j(v) \propto \exp(\bar{\tau}_j^T l_j(v) + \beta_j(v))$, et si l'intégrale ci-dessus a une forme explicite (ce qui est le cas en général pour une famille exponentielle classique), la formule ci-dessus fournit une équation de point fixe pour les $\bar{\tau}_j$ (ψ apparaît dans les deux membres), qu'il ne reste plus qu'à résoudre.

32. D'autres propositions existent [83].

4. Appliquer au modèle hiérarchique suivant

$$\begin{aligned}\tau &\sim \text{Gamma}(k_0, b_0) \qquad i.e. \quad f(\tau) = \frac{\tau^{k_0-1} e^{-\frac{\tau}{b_0}}}{\Gamma(k_0) b_0^{k_0}},\\ \mu|\tau &\sim \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1})\\ X &\sim \mathcal{N}(\mu, \tau^{-1} Id_n)\end{aligned}$$

avec $u = (\tau, \mu)$, en identifiant les fonctions l_j, β_j, t_j, m_{jk}, ρ, et les coefficients τ_j et μ_{jk}.

On trouvera des compléments dans, par exemple [152], ou la page Wikipedia *Variational Bayesian methods*.

***Exercice 14 (Agrégation de modèles linéaires avec inégalité oracle [117]).** 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 OLS $\widehat{\mu}_t(y) = P_t y$, $t = 1, \dots T$, basé chacun sur un modèle de régression de matrice de design X_t (cf. p. 9 ; pour chacun de ces modèles, P_t sera donc la matrice de projection sur l'espace image de X_t, $P_t = X_t(X_t^T X_t)^{-1} X_t^T$ si X_t 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_t de rang plein d'une certaine matrice de design X ayant possiblement plus de colonnes (variables) que de lignes (individus).

Noter qu'on ne cherche pas ici, contrairement à l'exercice suivant, un prédicteur utilisable sur de nouvelles données, mais simplement un estimateur de μ basé sur les données présentes, comme le suggère la mesure de performance (III.59). Le lecteur verra qu'on peut cependant, en posant $\beta_t = (X_t^T X_t)^{-1} X_t^T$, proposer (sans garantie) le prédicteur $\sum w_t \langle x'_t, \beta_t \rangle$, où les x'_t sont les variables explicatives associées à chaque modèle pour le nouvel échantillon, et les w_t sont les poids donnés plus bas.

Il s'agit d'éviter le surajustement qui se produirait si l'on prenait simplement le modèle de moindre MSE. Plutôt que d'essayer de choisir le meilleur estimateur par une procédure de sélection à la Mallows (cf. exercice 1 p. 16), Barron et Leung proposent de réaliser une combinaison linéaire bien choisie $\widehat{\mu}$ de ces derniers.

L'estimateur se présente ainsi

$$\widehat{\mu} = \sum_t w_t \widehat{\mu}_t = \sum_t w_t P_t y \tag{III.56}$$

$$w_t = \frac{e^{-\widehat{r}_t/4\sigma^2} \pi_t}{\sum_{t'} e^{-\widehat{r}_{t'}/4\sigma^2} \pi_{t'}} \tag{III.57}$$

$$\widehat{r}_t = \|y - \widehat{\mu}_t\|^2 + 2\sigma^2 Tr(P_t) - n\sigma^2. \tag{III.58}$$

Les π_t sont des poids à choisir librement, dont la somme fait 1. On notera que $\widehat{r}_t$ est un estimateur sans biais de $E[\|\mu - \widehat{\mu}_t\|^2]$ (cf. exercice 1 p. 16) ; on aura également besoin de l'équation (I.25) appliquée à $\widehat{\mu}$:

$$\|\mu - \widehat{\mu}\|^2 = \|y - \widehat{\mu}\|^2 + 2\sigma^2 Tr(\nabla_y \widehat{\mu}) - n\sigma^2 + \xi$$

où ξ est une variable centrée.

1. Montrer que $2\sigma^2 \nabla_y w_t = w_t(\widehat{\mu}_t - \widehat{\mu})$, puis en déduire

$$2\sigma^2 Tr(\nabla_y \widehat{\mu}) + \|\widehat{\mu} - y\|^2 = \sum w_t \widehat{r}_t + n\sigma^2.$$

2. Justifier les inégalités

$$\sum_t w_t \widehat{r}_t \leqslant 4\sigma^2 \log \sum_t w_t e^{\widehat{r}_t/4\sigma^2} \leqslant \widehat{r}_{t_0} - 4\sigma^2 \log \pi_{t_0}, \quad 1 \leqslant t_0 \leqslant T.$$

3. Déduire des points précédents que pour tout t_0

$$E\big[\|\mu - \widehat{\mu}\|^2\big] \leqslant E\big[\|\mu - \widehat{\mu}_{t_0}\|^2\big] - 4\sigma^2 \log \pi_{t_0}. \tag{III.59}$$

L'estimateur fusionné fait donc aussi bien que le meilleur de tous, à un terme additionnel près, relativement modeste, même si le nombre de modèles comparés T dépasse largement le nombre d'individus n (penser au cas $\pi_t = 1/T$).

Leung et Barron montrent que dans le cas où l'on considère tous les sous-modèles associés à une certaine matrice de design X, il est favorable de choisir les π_t de sorte que, en notant N_t le nombre de modèles t' de même nombre de degrés de liberté que t (i. e. $Rang(P_{t'}) = Rang(P_t)$), alors $\pi_t N_t$ soit à peu près constant ([117] § V). 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 [138].

L'estimation de σ est un point délicat mais assez important, on trouve des pistes dans [137].

***Exercice 15 (Agrégation à poids exponentiels (EWA)).** Soit Y_i une suite iid, Θ une partie convexe d'un espace vectoriel, et π une mesure de probabilité sur Θ. Soit l'estimateur randomisé $\bar\theta_n$ basé sur des observations indépendantes du contraste K :

$$\widehat\theta_j = \frac{\int \theta e^{-S_j(\theta)}\pi(d\theta)}{\int e^{-S_j(\theta)}\pi(d\theta)}, \quad S_j(\theta) = K(\theta, Y_1) + \cdots + K(\theta, Y_j), \quad S_0(\theta) = 0 \tag{III.60}$$

$$\bar\theta_n \sim \mathcal{U}(\{\widehat\theta_0, \dots \widehat\theta_n\}). \tag{III.61}$$

La mesure π sera typiquement discrète, possiblement uniforme, sur une famille de candidats présélectionnés (estimateurs sur un autre ensemble de données, etc.).

La suite $(Y_0, Y_1, \dots)$ est iid et le risque de $\bar\theta_n$ est $E[k(\bar\theta_n)]$ avec $k(\theta) = E\big[K(\theta, Y_i)\big]$, soit $E[k(\bar\theta_n)] = E\big[K(\bar\theta_n, Y_0)\big]$.

On préférerait obtenir des résultats sur $\widehat\theta_n$ que sur l'estimateur randomisé $\bar\theta_n$, mais ceci reste très difficile dans un cadre général. On suppose que

(H) Pour tout θ_0, la fonction $\theta \mapsto E[e^{-K(\theta,Y_0)+K(\theta_0,Y_0)}]$ est concave[33].

C'est le cas si p.s. $\theta \mapsto e^{-K(\theta,Y_0)}$ est concave, ce qui n'est malheureusement pas réalisé si $K(\theta, y) = (\theta - y)^2$ à moins que y ne soit borné, ainsi que Θ ; nous y reviendrons. Noter que (H) interdit que $\Theta = \mathbb{R}$ (fonction concave $\geqslant 0$), sauf dans la situation triviale où K ne dépend pas de θ.

1. Le risque de $\bar\theta_n$.
   (a) Soient les mesures $\pi_j(d\theta) = \frac{e^{-S_j(\theta)}\pi(d\theta)}{\int e^{-S_j(\theta')}\pi(d\theta')}$. En utilisant (H) et l'inégalité de Jensen, montrer que pour tout j

$$E\Big[\log \int e^{-K(\theta,Y_0)+K(\widehat\theta_j,Y_0)}\pi_j(d\theta)\Big] \leqslant 0. \tag{III.62}$$

   (b) Déduire de (III.62) une borne pour $E[k(\widehat\theta_j]$, puis montrer que

$$E\big[k(\bar\theta_n)\big] \leqslant \frac{-1}{n+1} E\Big[\log \int e^{-S_{n+1}(\theta)}\pi(d\theta)\Big]. \tag{III.63}$$

   (c) Montrer l'inégalité générale pour un processus $X(\theta, \omega)$

$$E\left[\log \int e^{X(\theta,\omega)}\pi(d\theta)\right] \geqslant \log \int e^{E[X(\theta,\omega)]}\pi(d\theta).$$

   Indication : Remarquer que c'est élémentaire si $E[X(\theta, \omega)] = 0$ pour tout θ ; se ramener à cette situation en posant $\pi'(d\theta) = e^{E[X(\theta,\omega)]}\pi(d\theta)/\int e^{E[X(\theta,\omega)]}\pi(d\theta)$.

33. (H) est ici une version simplifiée de l'hypothèse plus faible due à [103] (th. 4.2) reprise par [25] (éq. (5.1)) qui est, pour tout θ_0, l'existence d'un "sur-gradient" en $\theta = \theta_0$: $E[e^{-K(\theta,Y_0)+K(\theta_0,Y_0)}] \leqslant 1 + l_{\theta_0}(\theta - \theta_0)$, pour une certaine forme linéaire l_{θ_0}. C'est la formule que l'on obtient, en supposant K différentiable, si l'on applique la formule clé (III.62) à une mesure π_j qui donne un poids ε à θ et $1-\varepsilon$ à θ_0, avec $\widehat\theta_j = \varepsilon\theta + (1-\varepsilon)\theta_0$, et en faisant $\varepsilon \to 0$. Elle semble donc difficilement améliorable dans le cadre de cette approche.
Cette hypothèse implique $e^{-k(\theta)+k(\theta_0)} \leqslant 1 + l_{\theta_0}(\theta - \theta_0) \leqslant e^{l_{\theta_0}(\theta-\theta_0)}$ et donc la convexité de k.
Il est expliqué dans [103], bornes inférieures à l'appui, pourquoi, dans ce contexte, un estimateur basé sur la sélection d'un candidat donnera généralement un résultat asymptotiquement moins bon en termes de risque moyen.

En déduire que pour tout $\theta_\circ \in \Theta$

$$E\big[k(\bar\theta_n)\big] \leqslant k(\theta_\circ) - \frac{1}{n+1}\log \int e^{-(n+1)(k(\theta)-k(\theta_\circ))}\pi(d\theta). \tag{III.64}$$

Il se trouve que le deuxième terme peut être rendu raisonnablement petit dans des circonstances de sélection de modèle où π est choisi comme une somme de mesures à densité sur des espaces euclidiens de dimension différente, grâce à la contribution de l'intégrale au voisinage de $\theta_\circ$ [61]. La suite de l'exercice concerne le cas où π est purement atomique.

(d) En déduire que si π est une mesure sur une partie finie $\{\theta_1, \dots \theta_T\} \subset \Theta$, en notant $\pi_t = \pi(\theta_t)$,

$$E\big[k(\bar\theta_n)\big] \leqslant k(\theta_t) - \frac{\ln \pi_t}{n+1}, \qquad 1 \leqslant t \leqslant T \tag{III.65}$$

(inégalité qui peut se déduire directement de (III.63)). Noter que π_t peut même être remplacé dans (III.65) par $\pi(\{\theta \in \Theta : k(\theta) \leqslant k(\theta_t)\})$.

L'estimateur $\bar\theta_n$ a donc un risque quasiment aussi bon que le meilleur des candidats θ_t, même si T est, disons, une puissance de n (penser p. ex. au cas $\pi(\theta_t) = \frac{1}{T}$, $1 \leqslant t \leqslant T$). Le facteur $1/(n+1)$ est ici petit car le nombre de paramètres étant incontrôlé, une vitesse $1/n$ pour le risque avec T si grand n'est pas à l'ordre du jour, comme pour l'exercice 14.

2. L'ESTIMATEUR MOYENNÉ. Montrer en exploitant la convexité de k (cf. note 33) que l'estimateur $\frac{1}{n+1}\sum_0^n \widehat\theta_j$ a un risque inférieur à $\bar\theta_n$. Il est malheureusement sensiblement plus coûteux à calculer ; mais toute moyenne partielle, issue de tirages aléatoires uniformes dans $\{\widehat\theta_1, \dots \widehat\theta_n\}$, aura aussi un risque inférieur.

3. APPLICATION : APPRENTISSAGE. Soit un problème de régression de réponse Y et de variable explicative X. La paire (X, Y) a une loi jointe inconnue et l'on en dispose d'un échantillon iid $((X_1, Y_1), \dots (X_n, Y_n))$. On possède plusieurs candidats prédicteurs $x \mapsto f_t(x)$, $1 \leqslant t \leqslant T$, par exemple un catalogue d'estimations basées sur un échantillon pilote antérieur (de distribution possiblement différente). Il s'agit d'estimer la meilleure combinaison de ces derniers, le problème étant qu'ils sont très nombreux, comme à l'exercice 14, p. ex., $T \gg n$, et il faut donc éviter le surajustement potentiel. L'évaluation d'un prédicteur f se fait via une fonction de perte $r(y, f(x))$, p. ex. $r(y, u) = (y-u)^2$. Soit $\eta > 0$, on pose

$$\Theta = \Big\{ f = \sum_t w_t f_t : w \in \mathbb{R}_+^T, \sum w_t = 1 \Big\}, \quad K(f, (x, y)) = \eta r(y, f(x)),$$
$$k(f) = \eta R(f), \quad R(f) = E[r(Y_1, f(X_1))]$$

(Θ est de dimension finie. Pour rester dans le cadre habituel $\Theta \subset \mathbb{R}^T$, on aurait pu travailler sur les coordonnées en prenant pour Θ l'ensemble des vecteurs de poids w possibles, les candidats θ_t devenant les vecteurs de la base canonique, mais cela ne fait qu'alourdir les notations).

Noter que la concavité d'une fonction $f \mapsto \varphi(f(x))$ sur Θ, se ramène à la concavité de φ sur un convexe contenant les images des $f \in \Theta$.

(a) En classification on a $Y_0 \in \{-1, 1\}$. On considère $r(y, u) = \ln(1 + e^{-uy})$, ou $r(y, u) = e^{-uy}$ (cf. [61] § 7 pour d'autres possibilités). Vérifier que, si les fonctions f_t sont uniformément bornées, on peut trouver η pour que (H) soit satisfait.

(b) On suppose ici que $Y_i \sim \mathcal{N}(f_*(X_i), \sigma^2)$, et $r(y, u) = (y-u)^2$. On pose $M_f = \sup_t \|f_t(X_1) - f_*(X_1)\|_\infty$. Montrer que si $2\eta(M_f^2 + \sigma^2) \leqslant 1$ alors (H) est satisfait.

Indication : On commencera par vérifier que pour tous b et $a > 0$, la fonction $t \mapsto e^{-a(t+b)^2}$ est concave sur $\{t : 2a(t+b)^2 \leqslant 1\}$ (il suffit de le vérifier pour $b = 0$).

(c) Vérifier que l'estimateur (III.60) conduit à

$$w_{tk} = \pi_t \exp\Big(-\eta \sum_{j=1}^k r(Y_j, f_t(X_j)) \Big), \quad \widehat f_k = \frac{\sum_t w_{tk} f_t}{\sum w_{sk}}$$

avec la garantie que l'estimateur $\bar f_n$, choisi uniformément parmi les $\widehat f_k$, satisfasse pour tout t

$$E[R(\bar f_n)] \leqslant R(f_t) - \frac{\log \pi_t}{\eta(n+1)}.$$

4. Application : Estimation de densité ([31] § 3). On possède plusieurs candidats $p_1, \dots p_T$ pour une densité $p_*(x)$. On cherche à estimer p_* à l'aide d'un échantillon $Y = (Y_1, \dots Y_n)$ de n variables iid de loi p_*, comme combinaison linéaire des p_t. On pose

$$\Theta = \Big\{ p = \sum_t w_t p_t : w \in \mathbb{R}_+^T : \sum w_t = 1 \Big\},$$

$$K(p, y) = -\ln p(y), \qquad k(p) = -\int \ln(p(x)) p_*(x) \mu(dx).$$

L'hypothèse de concavité de $p \mapsto e^{-K(p,y)}$ est bien entendu satisfaite. Un minimum de k minimise la divergence de Kullback-Leibler entre p et p_*.

(a) Vérifier que l'estimateur (III.60) conduit à

$$\widehat{p}_j(x) = \frac{\sum_t p_t(x) p_t(Y^j) \pi_t}{\sum_t p_t(Y^j) \pi_t}, \quad p_t(Y^j) = \prod_{i=1}^{j} p_t(Y_i),$$

que l'on pourra comparer à (III.39), et que pour tout $t \leqslant T$ (cf. (III.3))

$$E[D(p_* \| \bar{p}_n)] \leqslant D(p_* \| p_t) - \frac{\log \pi_t}{n+1}.$$

(b) Montrer que $\widehat{p}_n$ est π-bayésien pour le risque de Kullback $D(p_* \| \widehat{p}_n)$.
Indication : Utiliser (III.37) ; on montrera que $\sum_t \{D(p_t \| p') - D(p_t \| \widehat{p}_n)\} p_t(Y^n) \pi_t \leqslant 0$ pour toute mesure de probabilité p' (faire apparaître une divergence de Kullback-Leibler).

5. L'approche déterministe (mélange d'experts [109]). On suppose ici que la suite Y_i est déterministe et que pour tout y, la fonction $\theta \mapsto e^{-K(\theta,y)}$ est concave. Il y a T experts dont chacun propose à chaque instant i une valeur du paramètre θ_{ti}, typiquement basée sur les observations passées. Cette valeur est censée minimiser $K(\theta, Y_{i+1})$. L'équation (III.60) est remplacée par

$$\widehat{\theta}_j = \frac{\int \theta_{tj} e^{-S_j(t)} \pi(dt)}{\int e^{-S_j(t)} \pi(dt)}, \quad S_j(t) = K(\theta_{t0}, Y_1) + \cdots + K(\theta_{t,j-1}, Y_j), \quad S_0(t) = 0 \tag{III.66}$$

où π est une mesure sur $\{1, \dots T\}$.

(a) Montrer, en utilisant la concavité, que

$$K(\widehat{\theta}_i, Y_{i+1}) \leqslant -\ln \int e^{-S_{i+1}(t)} \pi(dt) + \ln \int e^{-S_i(t)} \pi(dt)$$

(b) En déduire que pour tout t

$$\sum_{i=0}^{n} K(\widehat{\theta}_i, Y_{i+1}) \leqslant S_{n+1}(t) - \ln \pi(t).$$

Le terme de gauche s'interprète comme l'erreur cumulée au cours de l'algorithme qui apprend au fur et à mesure. Noter que si les Y_i sont iid et $\theta_{ti} = \theta_t$ on a $\sum_i E[K(\widehat{\theta}_i, Y_{i+1})] = (n+1) E[k(\bar{\theta})]$, et l'on retrouve (III.65), mais sous l'hypothèse plus forte de concavité de $\theta \mapsto e^{-K(\theta,y)}$.

III.5 Les tests classiques et leur seuil asymptotique

Cette section fait écho au § II.6, mais nous sommes ici dans le cadre paramétrique. Nous considérons à nouveau les tests du type

$$H_0 : g(\theta_*) = 0, \qquad H_1 : g(\theta_*) \neq 0$$

pour une certaine fonction g à valeurs dans $\mathbb{R}^q$. On notera $\widehat{\theta}^c$ l'estimateur sous la contrainte $g(\theta) = 0$.

Notons la vraisemblance $\mathscr{L}(\theta) = \mathscr{L}(\theta, Y_1, ..Y_n)$ et omettons l'indice n qui reste sous-entendu. Soit $\widehat{\theta}^c = \widehat{\theta}^c_n$ l'estimateur au maximum de vraisemblance de θ sous la contrainte $g(\theta) = 0$. On montre sans problème à l'aide du théorème 19 p. 53, comme on l'a déjà fait pour démontrer le théorème de Wilks (p. 74), que l'on a sous H_0 la convergence en loi de S vers un χ^2_q pour chacun des quatre choix suivants de la statistique S :

$$S = \begin{cases} 2(\mathscr{L}(\widehat{\theta}) - \mathscr{L}(\widehat{\theta}^c)) & \text{(statistique du rapport de vraisemblance, cf. (II.63))} \\ \nabla\mathscr{L}(\widehat{\theta}^c)\widehat{I}^{-1}\nabla\mathscr{L}(\widehat{\theta}^c) & \text{(statistique des scores, ou de Rao)} \\ (\widehat{\theta} - \widehat{\theta}^c)\widehat{I}(\widehat{\theta} - \widehat{\theta}^c) & \text{(statistique d'Hausman, cf. (II.64))} \\ g(\widehat{\theta})^T(\widehat{G}\widehat{I}^{-1}\widehat{G}^T)^{-1}g(\widehat{\theta}) & \text{(statistique de Wald)} \end{cases}$$

où $\widehat{I}$ est l'estimée (ou la vraie valeur) de la matrice d'information non-normalisée (i. e. correspondant à l'échantillon complet), par exemple $\widehat{I} = -\nabla^2\mathscr{L}(\widehat{\theta}^c)$ et $\widehat{G}$ l'estimée de $\nabla g(\theta_*)$, par exemple $\widehat{G} = \nabla g(\widehat{\theta}_c)$. Ces quatre statistiques sont donc **asymptotiquement libres**. D'où les tests :

Rejeter H_0 si $\quad S \geqslant \chi^2_q(1-\alpha)$

où $\chi^2_q(.)$ désigne la fonction quantile du χ^2_q. Ces statistiques sont asymptotiquement égales sous H_0. Un exemple est le test de sphéricité de la page 74.

L'intérêt ici 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 aussi de montrer que S est asymptotiquement libre, ce qui explique son intérêt et valide (asymptotiquement !) la méthode par simulation, et par bootstrap [33, 92].

IV

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 [8], dont deux théorèmes sont rappelés en appendice F.

IV.1 Le maximum de vraisemblance en situation générale

IV.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 $\widehat{\theta}_n$ l'estimateur au maximum de vraisemblance basé sur l'échantillon $(Y_0, Y_1, \dots Y_n)$. On supposera que $\widehat{\theta}_n$ converge presque sûrement vers le vrai paramètre θ_* (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 [8] § 6.2 pour plus de détails). On notera

$$\begin{aligned} L_n(\theta) &= \log p_\theta(Y_0, \dots Y_n) - \log p_\theta(Y_0) \\ L'_n(\theta) &= \nabla_\theta L_n(\theta) \qquad \text{(vecteur colonne)} \\ L''_n(\theta) &= \nabla^2_\theta L_n(\theta) \qquad \text{(matrice)} \end{aligned}$$

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 $p^n_\theta(y) = p_\theta(Y_n = y | Y_0, \dots Y_{n-1})$ est R-régulière au sens de la définition p. 71.

La propriété essentielle est que $L'_n(\theta_*)$ est une martingale sous P_{θ_*} : 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

$$\begin{aligned} E_{\theta_*}\left[L'_n(\theta_*) - L'_{n-1}(\theta_*) | Y^{n-1}\right] &= \int \nabla_\theta \log p_{\theta_*}(y_n | Y^{n-1}) p_{\theta_*}(y_n | Y^{n-1}) \ \mu(dy_n) \\ &= \int \nabla_\theta p_{\theta_*}(y_n | Y^{n-1}) \ \mu(dy_n) \\ &= 0. \end{aligned}$$

On a utilisé à la fin le lemme 27. Les quantités suivantes joueront un rôle central :

$$U_n = L'_n(\theta_*) - L'_{n-1}(\theta_*) = \nabla_\theta \log p_{\theta_*}(Y_n|Y_0, \dots Y_{n-1}) \quad \text{(vecteur colonne)}$$
$$I_n = \sum_{i=1}^{n} E\big[U_i U_i^T | \mathscr{F}_{i-1}\big], \quad \mathscr{F}_i = \sigma(Y_0, ...Y_i).$$

Il faut bien voir que dans le cas iid, on a $I_n = nI$. Considérons par exemple le modèle ARCH

$$\sigma_n^2 = \theta_0 + \sum_{i=1}^{q} \theta_i Y_{n-i}^2 \tag{IV.1}$$
$$Y_n = \sigma_n \varepsilon_n \tag{IV.2}$$

où ε_n est une suite i.i.d $\mathcal{N}(0,1)$ et les paramètres sont $\geqslant 0$ avec $\sum_i \theta_i < 1$; on vérifie simplement que

$$\log p_\theta(y_n|y_0, \dots y_{n-1}) = -\frac{y_n^2}{2\sigma_n^2} - \frac{1}{2}\log(\sigma_n^2), \quad \sigma_n^2 = \theta_0 + \sum_{i=1}^{q} \theta_i y_{n-i}^2$$
$$U_n = \frac{Y_n^2 - \sigma_n^2}{2\sigma_n^4} \nabla_\theta \sigma_n^2 = \frac{\varepsilon_n^2 - 1}{2\sigma_n^2} \big(1, Y_{n-1}^2, \dots Y_{n-q}^2\big)$$
$$I_n = \sum_{i=1}^{n} \frac{1}{2\sigma_i^4} \big(1, Y_{i-1}^2, \dots Y_{i-q}^2\big) \big(1, Y_{i-1}^2, \dots Y_{i-q}^2\big)^T.$$

La vérification des hypothèses de la proposition qui vient pour ce modèle sont laissées en exercice 2 p. 104.

On a alors sous ces notations :

34 - Proposition (Normalité asymptotique des scores)

On suppose que pour tout n avec probabilité 1 sous P_{θ_*} 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. 71. S'il existe une matrice p.s. inversible I telle que

$$n^{-1} I_n \xrightarrow{P} I \tag{IV.3}$$
$$n^{-1} \sum_{i=1}^{n} E\big[\|U_i\|^2 1_{\|U_{ni}\| > \varepsilon\sqrt{n}} | \mathscr{F}_{i-1}\big] \xrightarrow{P} 0, \quad \text{pour tout } \varepsilon > 0. \tag{IV.4}$$

Alors la suite $n^{-1/2} L'_n(\theta_*)$ converge en loi, conditionnellement à I, vers $\mathcal{N}(0, I)$ [1].

Remarque. La dernière condition est impliquée par $n^{-1-\eta} \sum_{i=1}^{n} E\big[\|U_i\|^{2+2\eta}\big] \to 0$ pour un $\eta > 0$, qui est utilisée dans la majorité des exemples.

Démonstration. On a $n^{-1/2} L'_n(\theta_*) = \sum X_{ni}$, $X_{ni} = U_i/\sqrt{n}$. Les hypothèses du théorème 61 p. 130 sont satisfaites puisque

$$\sum_{i=1}^{n} E\big[X_{ni} X_{ni}^T | \mathscr{F}_{n,i-1}\big] = n^{-1} I_n \xrightarrow{P} I.$$

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 φ continue bornée, on a la convergence

$$E[\varphi(I, n^{-1/2} L'_n(\theta_*))] \to \int E[\varphi(I, y) e^{-\frac{1}{2} y^T I^{-1} y} det(I)^{-1/2}] (2\pi)^{d/2} dy.$$

Pour avoir la normalité asymptotique de $\widehat{\theta}_n$, il faut faire comme dans la démonstration du théorème 8 p. 33 :

$$0 = L'_n(\widehat{\theta}_n) = L'_n(\theta_*) + L''_n(\theta'_n)(\widehat{\theta}_n - \theta_*)$$

avec $\theta'_n \in [\widehat{\theta}_n, \theta_*]$ (dans le cas vectoriel, θ'_n change d'une ligne à l'autre de la matrice L''_n) ; d'où

$$n^{1/2}(\widehat{\theta}_n - \theta_*) = -\Big\{nL''_n(\theta'_n)^{-1}I\Big\}\Big\{n^{-1/2}I^{-1}L'_n(\theta_*)\Big\}.$$

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 $E[L''_n - L''_{n-1} + U_nU_n^T|\mathscr{F}_{n-1}] = 0$. Il est alors naturel de supposer que $n^{-1}L''_n(\theta_n)$ tend vers $-I$, ce qui est l'analogue de la quatrième hypothèse du théorème 8. D'où le théorème :

35 - Théorème

En plus des hypothèses de la proposition 34, on suppose que les fonctions $L_n(\theta)$ sont, avec probabilité 1, de classe C^2 sur un voisinage fixe de θ_* et que $n^{-1}L''_n(\theta_n)$ converge en probabilité vers $-I$ pour toute suite de variables aléatoires θ_n convergeant presque sûrement vers θ_*.
Alors tout estimateur $\widehat{\theta}_n$ tel que $L'_n(\widehat{\theta}_n) = 0$ et tel que $\widehat{\theta}_n$ converge vers θ_* satisfait

$$n^{1/2}(\widehat{\theta}_n - \theta_*) \longrightarrow \mathcal{N}(0, I^{-1}) \quad \text{en loi conditionnellement à } I.$$

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 (IV.1,IV.2), on obtient

$$L''_n(\theta) = \sum_{i=1}^{n} \frac{1}{2\sigma_i(\theta)^6}(\sigma_i^2(\theta) - 2Y_i^2)M_{i-1}$$
$$M_{i-1} = \big(1, Y_{i-1}^2, \dots Y_{i-q}^2\big)\big(1, Y_{i-1}^2, \dots Y_{i-q}^2\big)^T.$$

Comme $Y_i = \sigma_i(\theta_*)\varepsilon_i$, il vient, en notant $\sigma_{i*} = \sigma_i(\theta_*)$

$$\begin{aligned} L''_n(\theta) + I_n &= (L''_n(\theta) - L''_n(\theta_*)) + (L''_n(\theta_*) + I_n) \\ &= \sum_{i=1}^{n}\left(\frac{1}{2\sigma_i(\theta)^4} - \frac{1}{2\sigma_{i*}^4} - \left(\frac{1}{\sigma_i(\theta)^6} - \frac{1}{\sigma_{i*}^6}\right)\sigma_{i*}^2\varepsilon_i^2\right)M_{i-1} + \sum_{i=1}^{n}\frac{1-\varepsilon_i^2}{\sigma_{i*}^4}M_{i-1}. \end{aligned} \tag{IV.5}$$

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 θ assez proche de θ_*, et que le deuxième terme est une martingale (exercice 2 p. 104).

IV.1.2 Exemple : Chaînes de Markov stationnaires

Supposons que $Y_n, n \geqslant 0$ 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

$$L_n(\theta) = \sum_{i=1}^{n}\log\big(q_\theta(Y_i|Y_{i-1})\big) = \sum_{i=1}^{n} l(\theta, Z_i), \quad Z_i = (Y_i, Y_{i-1}).$$

Après division par n le théorème 13 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 θ_* comme unique minimum ; il vaut

$$\begin{aligned} k(\theta) &= -E\big[\log(q_\theta(Y_i|Y_{i-1}))\big] = -\int\int \log(q_\theta(y|x))\, q_{\theta_*}(y|x)\pi_{\theta_*}(x)\mu(dy)\mu(dx) \\ &= \int D(Q^x_{\theta_*}|Q^x_\theta)\ \pi_{\theta_*}(x)\ \mu(dx) + k(\theta_*), \qquad Q^x_\theta(dy) = q_\theta(y|x)\mu(dy). \end{aligned}$$

Le premier terme est $\geqslant 0$ et ne s'annule que si $\pi_{\theta_*}(x)\mu(dx)$-p.s., $Q^x_{\theta_*} = Q^x_\theta$. Il faut donc que pour tout θ, il existe un ensemble A de π_{θ_*}-probabilité non nulle tel que pour tout $x \in A$ les probabilités de transition $Q^x_{\theta_*}$ et Q^x_θ diffèrent (identifiabilité). En d'autres termes, il suffit que pour tout $\theta \neq \theta_*$ on ait $\int |q_\theta(y|x)) - q_{\theta_*}(y|x)|\pi_{\theta_*}(x)\mu(dy)\mu(dx) > 0$.

La condition de régularité imposée à $p^n_\theta(y)$ (proposition 34) est satisfaite si pour μ-presque tout y_0 la densité $q_\theta(y|y_0)$ satisfait l'hypothèse de régularité habituelle.

On obtient ensuite

$$\begin{aligned}
&U_n = \nabla_\theta \log(q_{\theta_*}(Y_n|Y_{n-1})),\\
&E[U_n U_n^T|\mathscr{F}_{n-1}] = \int \nabla_\theta \log(q_{\theta_*}(y|Y_{n-1}))\nabla_\theta \log(q_{\theta_*}(y|Y_{n-1}))^T q_{\theta_*}(y|Y_{n-1})\mu(dy)\\
&L''_n(\theta) - L''_{n-1}(\theta) = \nabla^2 \log(q_\theta(Y_n|Y_{n-1})) = \nabla^2_\theta l(\theta, Z_n).
\end{aligned}$$

La stationnarité et l'ergodicité implique que I_n/n converge p.s. vers $E[I_1]$ dès que $E_{\theta_*}\left[\|U_1\|^2\right] < \infty$.

L'hypothèse (IV.4) est satisfaite si par exemple $E_{\theta_*}\left[\|U_1\|^3\right] < \infty$.

Il ne reste plus qu'à vérifier l'hypothèse du théorème 35. En vertu de la proposition 7, si $\nabla^2 l(\theta, Z_n)$ satisfait (GNU) au voisinage de θ_*, la suite $L''_n(\theta_n)/n$ converge vers $E[\nabla^2_\theta l(\theta_*, Z)]$, quantité qui vaut $-E[I_1]$ si pour μ-presque tout x il existe un voisinage $\mathscr{V}$ de θ_*, tel que $\int \sup_{\theta\in\mathscr{V}} \|\nabla^2_\theta q_\theta(y|x)\|\mu(dy) < \infty$ (cf. la proposition 25 p. 70 appliquée ici à la loi conditionnelle à Y_{n-1}).

Alors $\sqrt{n}(\widehat{\theta}_n - \theta_*)$ a pour variance asymptotique $E[U_n U_n^T]^{-1}$ si cette quantité est finie. Rappelons les hypothèses :

1. $\widehat{\theta}_n$ reste dans un compact de Θ et θ_* est intérieur à Θ.
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 $\mathscr{V}$ de θ_* où presque sûrement $\nabla^2_\theta l(\theta, Z_i)$ existe et satisfait (GNU).
5. Pour μ-presque tout y_0 la densité $q_\theta(y|y_0)$ satisfait l'hypothèse de régularité (cf. p. 71) (noter que la condition (a) est déjà satisfaite).
6. Pour μ-presque tout y_0, $\int \sup_{\theta\in\mathscr{V}} \|\nabla^2_\theta q_\theta(y|y_0)\|\mu(dy) < \infty$.
7. $E_{\theta_*}\left[\|\nabla_\theta l(\theta_*, Z_i)\|^3\right] < \infty$.

IV.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 (IV.1,IV.2) que l'on suppose stationnaire ergodique [2]. On suppose que $\theta^*_i > 0$ pour tout $i > 0$.

1. Montrer que pour une certaine constante $C(\theta_*)$ on a $\|U_i\| \leqslant C(\theta_*)|\varepsilon_i^2 - 1|$.
2. Vérifier les hypothèses de la proposition 34 (utiliser la remarque qui la suit).
3. Vérifier l'hypothèse additionnelle du théorème 35 (exploiter (IV.5)).

Exercice 3 (Processus de branchement [3]). 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 φ, donc $Y_n = \sigma_n\varepsilon_n = \varepsilon_n\psi(\varepsilon_n^2, \varepsilon_{n-1}^2, \dots)$ pour une certaine fonction ψ, ce qui implique la stationnarité et l'ergodicité car la suite ε_n est iid.

3. On trouvera des compléments dans [8] p.178.

1. Z_n représente l'effectif d'une génération au temps n. Interpréter les p_j.
2. Justifier les relations

$$E\big[Z_{n+1}|Z_n\big] = mZ_n$$
$$E\big[(Z_{n+1} - mZ_n)^2|Z_n\big] = \sigma^2 Z_n.$$

   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 $\widehat{m} = Z_{n+1}/Z_n$. Proposer par un calcul heuristique, un ordre de grandeur de la variance d'estimation.

Exercice 4. On considère le processus $(Y_i)_{i\geqslant 0}$ 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 (IV.3, IV.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 θ_n convergeant vers θ_*. 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 $\widehat{\theta}_n$ converge vers θ_* ; démontrer la normalité asymptotique de l'estimateur.

IV.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 ε_i sont iid $\mathcal{N}(0, \sigma^2)$. On s'intéresse ici à la situation $|\theta_*| > 1$, cas étudié par White en 1958 [159]. On obtient l'estimateur au maximum de vraisemblance :

$$\widehat{\theta}_n = \frac{\sum_{i=1}^n Y_i Y_{i-1}}{\sum_{i=1}^n Y_{i-1}^2} = \theta_* + \frac{\sum_{i=1}^n \varepsilon_i Y_{i-1}}{\sum_{i=1}^n Y_{i-1}^2}.$$

Notons que

$$Y_i = \theta_*^i Z + r_i$$

avec

$$Z = Y_0 + \theta_*^{-1}\varepsilon_1 + \theta_*^{-2}\varepsilon_2 + \cdots = Y_0 + T$$
$$r_i = -\theta_*^{-1}\varepsilon_{i+1} - \theta_*^{-2}\varepsilon_{i+2} + \ldots$$

et la suite r_i est stationnaire gaussienne. On peut vérifier alors simplement que

$$\theta_*^n(\widehat{\theta}_n - \theta_*) = \frac{Z\sum_{i=1}^n \theta_*^{-n+i-1}\varepsilon_i + u_n}{Z^2\sum_{i=1}^n \theta_*^{-2n+2i-2} + v_n}$$

où u_n et v_n convergent dans L_1 vers 0. Comme $Z = Y_0 + T$, ceci se réécrit

$$\theta_*^n(\widehat{\theta}_n - \theta_*) = \Big(\sum_{i=1}^n \theta_*^{-2n+2i-2} + v_n\Big)^{-1}\Big(\frac{T'_n}{Y_0+T} + Z^{-1}u_n\Big), \quad T'_n = \sum_{i=1}^n \theta_*^{-n+i-1}\varepsilon_i.$$

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 $\mathcal{N}\big(0, \sigma^2(\theta_*^2-1)^{-1}Id\big)$, on a finalement

$$\theta_*^n(\widehat{\theta}_n - \theta_*) \longrightarrow (\theta_*^2-1)\frac{X'}{X + \sigma^{-1}\sqrt{\theta_*^2-1}\,Y_0}\ , \qquad (X, X') \sim \mathcal{N}\big(0, Id\big).$$

Si $Y_0 = 0$, le rapport suit une distribution de Cauchy.

Exercice. Quelle est la limite et la distribution asymptotique de l'estimateur :

$$\widehat{\theta}_n = \frac{\sum_{i=1}^n Y_i Y_{i-1}}{\sum_{i=0}^n Y_i^2}\ ?$$

IV.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 σ_*^2. La suite Y_i n'est pas iid.

36 - Définition

L'estimateur de β^* aux moindres carrés ordinaires (OLS, ordinary least squares) est

$$\widehat{\beta} = \arg\min_\beta \sum_i (Y_i - X_i\beta)^2.$$

C'est l'estimateur de β^* au maximum de vraisemblance sous l'hypothèse de normalité de u. On a le résultat élémentaire

37 - Proposition

$\widehat{\beta}$ est un estimateur sans biais avec, en notant $R = \sum_i X_i^T X_i$:

$$\widehat{\beta} = R^{-1}\sum X_i^T Y_i = \beta^* + R^{-1}\sum X_i^T u_i$$
$$Var(\widehat{\beta}) = \sigma_*^2 R^{-1}.$$

On s'intéresse maintenant à l'asymptotique quand le nombre d'observations n tend vers l'infini. Notons

$$\widehat{\beta}_n = R_n^{-1}\sum_{i=1}^n X_i^T Y_i.$$

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. [110]). Noter que dans la suite, X_i devra être bien entendu $\mathscr{F}_{i-1}$-mesurable. On a le théorème suivant :

38 - Théorème

Soit une suite de tribus croissante $\mathscr{F}_n$ telle que X_i soit $\mathscr{F}_{i-1}$-mesurable et telle que les u_i forment une suite d'accroissements de martingale avec :

$$E\big[u_i|\mathscr{F}_{i-1}\big] = 0,$$
$$\sup_i E\big[u_i^2|\mathscr{F}_{i-1}\big] < \infty, \quad p.s.$$

Si presque sûrement $\sup_n \|R_n\| \, \|R_n^{-1}\| < \infty$ et $R_n^{-1} \to 0$, alors $\widehat{\beta}_n \to \beta^*$ presque sûrement.
S'il existe une matrice Q déterministe telle que p.s. $\frac{1}{n} R_n \to Q > 0$, $E\big[u_i^2|\mathscr{F}_{i-1}\big] \to \sigma_*^2$ p.s., et si

$$\sup_i E\big[|u_i|^\eta|\mathscr{F}_{i-1}\big] < \infty, \quad p.s.$$

pour un $\eta > 2$ alors

$$\sqrt{n}(\widehat{\beta}_n - \beta) \longrightarrow \mathcal{N}(0, \sigma_*^2 Q^{-1}) \quad \text{en loi.}$$

Démonstration. Commençons par la convergence presque sûre. On a besoin du lemme suivant dont la démonstration est faite à l'appendice C :

39 - Lemme (de Kronecker)

Soit R_n une suite croissante de matrices symétriques $(R_n - R_{n-1} \geqslant 0)$ telle que R_n^{-1} tend vers 0, et $\sum M_n$ une série convergente de matrices (non nécessairement absolument convergente), alors la suite

$$\|R_n\|^{-1} \sum_{i=1}^{n} R_i M_i$$

tend vers zéro.

On a

$$\widehat{\beta}_n - \beta^* = R_n^{-1} \sum_{i=1}^{n} X_i^T u_i = \big(R_n^{-1}\|R_n\|\big)\Big(\|R_n\|^{-1} \sum_{i=1}^{n} R_i\big(R_i^{-1} X_i^T u_i\big)\Big).$$

Comme $\|R_n^{-1}\|.\|R_n\|$ 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 59

$$\sum_{i \geqslant l} \|R_i^{-1} X_i^T\|^2 E\big[u_i^2|\mathscr{F}_{i-1}\big] < \infty$$

ce qui est réalisé si

$$\mathrm{Tr} \sum_{i \geqslant l} R_i^{-1} X_i^T X_i R_i^{-1} < \infty. \tag{IV.6}$$

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

$$A(A^{-1} - B^{-1})A \leqslant B - A.$$

Ceci vient simplement de ce que la différence des deux membres vaut

$$B - A - A + AB^{-1}A = (B - A)B^{-1}(B - A) \geqslant 0.$$

On a donc

$$\sum_{i=l}^{n} R_i^{-1} X_i^T X_i R_i^{-1} \leqslant \sum_{i=l}^{n} R_{i-1}^{-1} - R_i^{-1} = R_{l-1}^{-1} - R_n^{-1} \leqslant R_{l-1}^{-1}$$

ce qui implique bien (IV.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 60 avec $X'_{ni} = X_i u_i / \sqrt{n}$. Les conditions (F.1) et (F.2) sont satisfaites si

$$\frac{1}{n} \sum_{i=1}^{n} E\big[\|X_i u_i\|^2 1_{\|X_i u_i\| > \varepsilon\sqrt{n}} | \mathscr{F}_{i-1}\big] \xrightarrow{P} 0, \qquad \text{pour tout } \varepsilon > 0$$

$$\frac{1}{n} \sum_{i=1}^{n} X_i^T X_i E\big[u_i^2 | \mathscr{F}_{i-1}\big] \xrightarrow{P} \sigma_*^2 Q.$$

La seconde relation est élémentaire. Pour la première,

$$\begin{aligned}
\frac{1}{n} \sum_{i \leqslant n} E\big[\|X_i u_i\|^2 1_{\|X_i u_i\| > \varepsilon\sqrt{n}} | \mathscr{F}_{i-1}\big] &\leqslant \frac{1}{n} \sum_{i \leqslant n} E\big[\|X_i u_i\|^\eta | \mathscr{F}_{i-1}\big] (\varepsilon\sqrt{n})^{2-\eta} \\
&\leqslant \frac{1}{n} \sum_{i \leqslant n} \|X_i\|^\eta E\big[|u_i|^\eta | \mathscr{F}_{i-1}\big] (\varepsilon\sqrt{n})^{2-\eta} \\
&\leqslant \Big(\frac{1}{n} \sum_{i \leqslant n} \|X_i\|^2\Big) \, (\sup_{i \leqslant n} \|X_i\| / \varepsilon\sqrt{n})^{\eta-2} \, \sup_{i \leqslant n} E\big[|u_i|^\eta | \mathscr{F}_{i-1}\big] \\
&\leqslant C(\omega) \, \sup_{i \leqslant n} \big(\|X_i\| / \varepsilon\sqrt{n}\big)^{\eta-2}.
\end{aligned}$$

Il ne suffit plus que de montrer que $\sup_{i \leqslant n} \|X_i\| / \sqrt{n}$ tend vers zéro. Mais la convergence de $n^{-1} R_n$ implique que $\|X_i\| / \sqrt{i}$ tend vers zéro car

$$\frac{\|X_i\|^2}{i} = \operatorname{Tr}\left(\frac{R_i}{i} - \frac{R_{i-1}}{i}\right) \to 0$$

et donc

$$\sup_{i \leqslant n} \|X_i\| / \sqrt{n} \leqslant \sup_{i \leqslant n_0} \|X_i\| / \sqrt{n} + \sup_{n_0 \leqslant i \leqslant n} \|X_i\| / \sqrt{i}.$$

Le deuxième terme peut être rendu arbitrairement petit par un choix adéquat de n_0 et le premier tend vers zéro. ∎

A

Hypothèse LAN. Théorèmes de Hajek et Le Cam

L'objet de ce chapitre est la démonstration du théorème de convolution de Hajek, théorème 41, ainsi que le théorème d'existence d'un estimateur efficace de Le Cam, théorème 44. Ces deux résultats ont été présentés page 82. Ils permettent d'aborder le cas d'une suite générale d'expériences statistiques (cf. I.3.1).

L'hypothèse essentielle est l'hypothèse LAN (local asymptotic normality) ; dans le cas indépendant, i.e. d'expériences statistiques de la forme $P_\theta^{\otimes n}$, elle est impliquée par la R-régularité (théorème 55). Énonçons cette hypothèse :

(LAN) *P_θ^n est une suite d'expériences statistiques. Il existe pour tout $\theta \in \Theta$ une matrice $I(\theta)$ déterministe et une suite de v.a. $S_n(\theta)$ convergeant en loi sous P_θ^n vers $\mathcal{N}(0, I(\theta))$ telles que*

$$\log\left(\frac{P^n_{\theta+h/\sqrt{n}}(Y)}{P^n_\theta(Y)}\right) = h^T S_n(\theta) - \frac{1}{2}h^T I(\theta)h + \varepsilon_n(\theta, h) \tag{A.1}$$

avec pour tout $M < \infty$ et tout $\alpha > 0$

$$\lim_{n\longrightarrow\infty} \sup_{|h|\leqslant M} P^n_\theta(|\varepsilon_n(\theta, h)| > \alpha) = 0.$$

Dans le cas d'une expérience de la forme $P_\theta^{\otimes n}$, le score $S_n(\theta)$ sera essentiellement

$$S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\nabla p_\theta}{p_\theta}(Y_i).$$

Cette hypothèse exprime que le score est asymptotiquement localement une statistique suffisante. Cette approche peut être vue comme une façon de mesurer l'écart entre les lois P_θ sur des processus $(Y_1, Y_2, \dots)$, bien que les mesures correspondantes soient étrangères (en raison du nombre infini de variables). La propriété LAN est satisfaite pour des chaînes de Markov générales [97] et même pour des processus à source markovienne cachée [38]. Le théorème 5 peut être un outil pour vérifier cette hypothèse. Il est simple de vérifier qu'elle est satisfaite pour le modèle AR(1) gaussien avec variance connue.

A.1 Théorème convolution de Hajek

Il faut commencer par un lemme classique :

40 - Lemme (Le Cam, 1960)

Soit $(\Omega, \mathscr{B})$ un espace d'évènements et $\mathscr{B}_n$ une suite croissante de sous-tribus de $\mathscr{B}$. Soit $\{P_n\}$ et $\{Q_n\}$ deux suites de lois définies sur $\mathscr{B}_n$. Soit Λ_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 $\mathscr{B}_n$-mesurables. Si

$$(T_n, \Lambda_n) \xrightarrow{d} (T, \Lambda) \quad \text{sous} \quad P_n$$

avec $E[\Lambda] = 1$, alors

$$T_n \xrightarrow{d} T' \quad \text{sous} \quad Q_n$$

où la loi de T' est donnée par

$$E[f(T')] = E[f(T)\Lambda] \tag{A.2}$$

pour toute f borélienne bornée.

Démonstration. Pour simplifier, on note $P_n[.]$, $Q_n[.]$, $R_n[.]$, les espérances sous les lois correspondantes. Montrons avant tout que $R_n[\Omega]$ tend vers 0 :

$$\begin{aligned} R_n[\Omega] &= Q_n[\Omega] - P_n[\Lambda_n] \leqslant 1 - P_n[\Lambda_n 1_{\Lambda_n < c}] \\ \overline{\lim_n} R_n[\Omega] &\leqslant 1 - E[\Lambda 1_{\Lambda < c}] \end{aligned}$$

qui est arbitrairement petit. Soit maintenant f une fonction positive continue bornée, et c un point de continuité de la distribution de Λ :

$$\begin{aligned} Q_n[f(T_n)] &= P_n[f(T_n)\Lambda_n] + R_n[f(T_n)] \\ &= P_n[f(T_n)\Lambda_n 1_{\Lambda_n < c}] + P_n[f(T_n)\Lambda_n 1_{\Lambda_n \geqslant c}] + O(R_n[\Omega]) \\ &= E[f(T)\Lambda 1_{\Lambda < c}] + o_c(1) + O(1) P_n[\Lambda_n 1_{\Lambda_n \geqslant c}] + o(1) \\ &= E[f(T)\Lambda 1_{\Lambda < c}] + O(1)(1 - P_n[\Lambda_n 1_{\Lambda_n < c}]) + o_c(1) \\ &= E[f(T)\Lambda 1_{\Lambda < c}] + O(1) E[\Lambda 1_{\Lambda \geqslant c}] + o_c(1) \\ &= E[f(T)\Lambda] + O(1) E[\Lambda 1_{\Lambda \geqslant c}] + o_c(1) \end{aligned}$$

qui est arbitrairement proche de $E[f(T)\Lambda]$ pour n grand pour un choix adéquat de c. ∎

Passons à la démonstration du théorème de convolution de Hajek.

41 - Théorème (Théorème de convolution de Hajek)

Le théorème 32 reste vrai si l'on remplace l'hypothèse de R-régularité par l'hypothèse LAN, le score S_n étant donné par (A.1).

Démonstration. On va appliquer le lemme 40 avec

$$\begin{aligned} T_n &= \sqrt{n}(\widehat{\theta}_n - \theta) \\ P_n &= P_\theta \\ Q_n &= P_{\theta + h/\sqrt{n}} \\ \Lambda_n &= \frac{p_{\theta + h/\sqrt{n}}(Y)}{p_\theta(Y)} \end{aligned}$$

(on pose $\Lambda_n = 0$ si le dénominateur est nul). L'hypothèse LAN et (III.32) donnent

$$\ln \Lambda_n \xrightarrow{Loi} \mathcal{N}\big(-\tfrac{1}{2}h^T I(\theta)h, h^T I(\theta)h\big), \quad \text{sous } P_\theta$$
$$T_n \xrightarrow{Loi} X, \quad \text{sous } P_\theta.$$

Par le théorème de Prohorov sur la *tightness* des distributions, de toute suite d'entiers, on peut extraire une sous-suite telle que la paire (T_{n_k}, Λ_{n_k}) converge en loi (i.e. loi conjointe) vers une certaine limite (T, Λ) sous P_θ. On peut alors appliquer le lemme 40 à une telle sous-suite. On sait par hypothèse (III.32) que $T' \sim X + h$. Pour identifier complètement la limite notons que l'équation (A.2) avec $f(x) = e^{ia^T x}$ devient

$$E\big[e^{ia^T(X+h)}\big] = E\big[e^{ia^T X + h^T Z - h^T I(\theta)h/2}\big]$$

où $Z \sim \mathcal{N}(0, I(\theta))$, X suit sa loi, et la loi liée est encore inconnue. En posant $Y = I(\theta)X - Z$ (conformément aux notations du théorème 32) et en choisissant a de la forme $a = I(\theta)b$:

$$e^{ia^T h} E\big[e^{ib^T(Y+Z)}\big] = E\big[e^{ib^T(Y+Z) + h^T Z - h^T I(\theta)h/2}\big]$$

soit

$$\begin{aligned} E\big[e^{ib^T Y + (ib+h)^T Z}\big] &= e^{ia^T h + h^T I(\theta)h/2} E\big[e^{ib^T(Y+Z)}\big] \\ &= e^{(h+ib)^T I(\theta)(h+ib)/2} e^{b^T I(\theta)b/2} E\big[e^{ib^T(Y+Z)}\big]. \end{aligned}$$

Comme $Z \sim \mathcal{N}(0, I(\theta))$, les deux membres sont des fonctions analytiques de h, et donc cette identité s'applique à $h = ic - ib$ pour c arbitraire, ce qui prouve bien l'indépendance de Y et Z car le membre de droite est le produit d'une fonction de c par une fonction de b. La distribution obtenue étant indépendante de la sous-suite, on a bien convergence. ∎

On a une sorte de réciproque :

42 - THÉORÈME

On suppose que l'hypothèse LAN est satisfaite. Soit un estimateur tel que pour tout $\theta \in \Theta$

$$\sqrt{n}(\widehat{\theta}_n - \theta) - I(\theta)^{-1} S_n(\theta) \xrightarrow{Loi} 0, \quad \text{sous } P_\theta, \tag{A.3}$$

par exemple l'estimateur au maximum de vraisemblance sous les hypothèses du théorème 28. Alors l'hypothèse (III.32) du théorème de convolution de Hajek est satisfaite.

Démonstration. On va appliquer le théorème 40 avec

$$T_n = \sqrt{n}(\widehat{\theta}_n - \theta) - h$$

et pour Λ_n l'exponentielle du membre de droite de (A.1). L'hypothèse LAN étant satisfaite (dans le théorème 28, la R-régularité est satisfaite, et donc LAN). On trouve grâce à (A.3)

$$(T, \Lambda) \sim (I(\theta)^{-1} S - h, \exp(h^T S - \frac{1}{2} h^T I(\theta) h)$$

avec

$$S \sim \mathcal{N}(0, I(\theta)),$$

puis

$$X \sim \mathcal{N}(0, I(\theta)^{-1})$$

ce qui achève la démonstration. ∎

A.2 Existence d'un estimateur efficace

Nous reprenons la construction particulièrement astucieuse de Le Cam [114] pp. 68 et suivantes. On commence par une propriété de régularité des scores, qu'on utilisera par la suite. Elle exprime que I est en un certain sens la dérivée de S_n et rappelle la formule (III.6), mais il faut pour cela choisir une certaine version des scores.

Posons, en référence à l'hypothèse LAN (p. 109)

$$L_{\theta,h} = \log\left(\frac{P^n_{\theta+h/\sqrt{n}}(Y)}{P^n_\theta(Y)}\right) = h^T S_n(\theta) - \frac{1}{2}h^T I(\theta)h + \varepsilon_n(\theta,h).$$

43 - Lemme

On se place sous l'hypothèse LAN (p. 109) et l'on suppose que l'application $\theta \mapsto I(\theta)$ est continue. Soit le score $S'_n(\theta)$ dont la i-ème coordonnée vaut

$$S'_n(\theta)_i = L_{\theta,e_i} + \frac{1}{2}I(\theta)_{ii} \tag{A.4}$$

où e_i est le i-ème vecteur de la base canonique. Alors l'hypothèse LAN est encore satisfaite avec $S_n(\theta)$ remplacé par $S'_n(\theta)$ et l'on a pour tout $\alpha > 0$:

$$\sup_{|h|\leqslant M} P_\theta\Big(|S'_n(\theta+h/\sqrt{n}) - S'_n(\theta) + I(\theta)h| > \alpha\Big) \longrightarrow 0. \tag{A.5}$$

Remarque. Une inspection de la démonstration montre que le choix des e_i est arbitraire, on aurait pu prendre n'importe quelle base, même dépendante continûment de θ, $I(\theta)_{ii}$ devenant $e_i^T I(\theta)e_i$.

Démonstration. Pour vérifier que l'hypothèse LAN est encore vérifiée, notons simplement que (A.1) avec $h = e_i$ se réécrit

$$S_n(\theta)_i = S'_n(\theta)_i - \varepsilon_n(\theta, e_i)$$

donc en reportant cette identité dans (A.1), on voit que LAN est encore vérifiée avec $S'_n(\theta)$ et

$$\varepsilon'_n(\theta,h) = \varepsilon_n(\theta,h) - \sum_i h_i \varepsilon_n(\theta,e_i).$$

Pour (A.5), noter que

$$\begin{aligned}
S'_n(\theta+\tfrac{h}{\sqrt{n}})_i - S'_n(\theta)_i &= L_{\theta+h/\sqrt{n},e_i} - L_{\theta,e_i} + \tfrac{1}{2}I(\theta+\tfrac{h}{\sqrt{n}})_{ii} - \tfrac{1}{2}I(\theta)_{ii} \\
&= L_{\theta,e_i+h} - L_{\theta,h} - L_{\theta,e_i} + \tfrac{1}{2}I(\theta+\tfrac{h}{\sqrt{n}})_{ii} - \tfrac{1}{2}I(\theta)_{ii} \\
&= \tfrac{1}{2}(e_i+h)^T I(\theta)(e_i+h) - \tfrac{1}{2}h^T I(\theta)h - \tfrac{1}{2}e_i^T I(\theta)e_i \\
&\quad + \varepsilon_n(\theta, e_i+h) - \varepsilon_n(\theta,e_i) - \varepsilon_n(\theta,h) \\
&\quad + \tfrac{1}{2}I(\theta+\tfrac{h}{\sqrt{n}})_{ii} - \tfrac{1}{2}I(\theta)_{ii} \\
&= \big(I(\theta)h\big)_i + \varepsilon_n(\theta,e_i+h) - \varepsilon_n(\theta,e_i) - \varepsilon_n(\theta,h) + \tfrac{1}{2}I(\theta+\tfrac{h}{\sqrt{n}})_{ii} - \tfrac{1}{2}I(\theta)_{ii}
\end{aligned}$$

ce qui démontre le résultat. ∎

44 - Théorème (Existence d'un estimateur efficace. Le Cam 1969)

On se place sous l'hypothèse LAN et l'on suppose que l'application $\theta \mapsto I(\theta)$ est continue. On suppose que l'on dispose d'un estimateur θ_n qui converge à vitesse $n^{-1/2}$ vers θ_* au sens où

$$\lim_{A\to\infty} \sup_n P\big(\sqrt{n}|\theta_n - \theta_*| > A\big) = 0. \tag{A.6}$$

Alors l'estimateur

$$\widehat{\theta}_n = [\theta_n] + n^{-1/2} I([\theta_n])^{-1} S'_n([\theta_n]) \tag{A.7}$$

où S' est donné par (A.4) et $[.]$ désigne l'arrondi coordonnée par coordonnée à $1/\sqrt{n}$ près, satisfait

$$n^{1/2}(\widehat{\theta}_n - \theta_*) - I(\theta_*)^{-1} S_n(\theta_*) \xrightarrow{P} 0.$$

Démonstration. Posons $I = I(\theta_*)$, $I_n = I([\theta_n])$, et remarquons que

$$\begin{aligned}
&n^{1/2}(\widehat{\theta}_n - \theta_*) - I^{-1} S_n(\theta_*) \\
&\quad = n^{1/2}([\theta_n] - \theta_*) + I_n^{-1} S'_n([\theta_n]) - I^{-1} S_n(\theta_*) \\
&\quad = n^{1/2}(Id - I_n^{-1} I)([\theta_n] - \theta_*) \\
&\qquad + I_n^{-1}(S'_n([\theta_n]) - S'_n(\theta_*) + n^{1/2} I([\theta_n] - \theta_*)) \\
&\qquad + I_n^{-1}(S'_n(\theta_*) - S_n(\theta_*)) + (I_n^{-1} - I^{-1}) S_n(\theta_*).
\end{aligned}$$

En raison de (A.6) et par continuité de $\theta \to I(\theta)$ le premier terme tend en probabilité vers 0. La convergence en loi de $S_n(\theta_*)$ et de $S'_n(\theta_*)$ ainsi que celle de I_n vers I en probabilité implique que le dernier tend également vers 0 en probabilité.

Il ne reste plus qu'à montrer la convergence en loi du deuxième vers 0, en exploitant le lemme 43. On ne peut pas utiliser directement (A.5) 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 θ_*, valeurs que nous notons δ_i ; elles sont en nombre inférieur à $(2A)^d$. On a donc

$$\begin{aligned}
&P\big(|n^{1/2}([\theta_n] - \theta_*) + J S'_n([\theta_n]) - J S'_n(\theta_*)| \geqslant \alpha\big) \\
&\quad \leqslant P\big(n^{1/2}|[\theta_n] - \theta_*| > A\big) + \sum_i P\big(|n^{1/2} I(\delta_i - \theta_*) + S'_n(\delta_i) - S'_n(\theta_*)| > \alpha\big) \\
&\quad \leqslant P\big(n^{1/2}|[\theta_n] - \theta_*| > A\big) + (2A)^d \sup_{|h|\leqslant A} P\big(|Ih + S'_n(\theta_* + h/\sqrt{n}) - S'_n(\theta_*)| > \alpha\big).
\end{aligned}$$

Un choix de A assez grand permet de rendre le premier terme arbitrairement petit et le second peut alors être réduit en augmentant n. ∎

Un petit supplément : Régularité de S_n sous l'hypothèse ULAN. Introduisons l'hypothèse

(ULAN) *L'hypothèse* LAN *est satisfaite et renforcée par : pour tout $M < \infty$, tout $\alpha > 0$ et tout compact K intérieur à Θ*

$$\lim_{n\longrightarrow\infty} \sup_{\theta\in K} \sup_{|h|\leqslant M} P_\theta(|\varepsilon_n(\theta, h)| > \alpha) = 0. \tag{A.8}$$

On a voir que cette hypothèse, qui est généralement satisfaite lorsque LAN l'est, implique que (A.5) est satisfait avec $S' = S$. On peut donc dans ce cas prendre $S' = S$ dans (A.7).

45 - Théorème (Régularité des scores)

On se place sous l'hypothèse ULAN (p. 113). Alors (A.5) est vrai avec S_n au lieu de S'_n.
Si $I(\theta)$ est borné sur un compact K intérieur à Θ, cette convergence y est uniforme en θ.

Remarque. L'hypothèse LAN est certainement insuffisante pour ce résultat car elle ne caractérise pas suffisamment $S_n(\theta)$. En effet, si l'on se donne un point θ_0 et que l'on définit $S'_n(\theta) = S_n(\theta) + n^{-1/4}\|\theta - \theta_0\|^{-1}1_{\theta\neq\theta_0}\theta_0$, alors LAN est toujours vérifié si l'on remplace S_n par S'_n mais (A.5) sera mis en défaut pour $\theta = \theta_0$.

Démonstration. Reprenons les notations de l'équation (A.1). Notons $L_{h,k} = \mathscr{L}_n(\theta+(h+k)/\sqrt{n}) - \mathscr{L}_n(\theta + h/\sqrt{n})$, $S_h = S_n(\theta + h/\sqrt{n})$, $\varepsilon_{h,k} = \varepsilon_n(\theta + h/\sqrt{n}, k)$, alors

$$\begin{aligned}
L_{0,h} &= h^T S_0 - \frac{1}{2}h^T I(\theta)h + \varepsilon_{0,h} \\
L_{h,k} &= k^T S_h - \frac{1}{2}k^T I(\theta)k + \varepsilon_{h,k} - \frac{1}{2}k^T(I(\theta + h/\sqrt{n}) - I(\theta))k \\
L_{0,h+k} &= (h+k)^T S_0 - \frac{1}{2}(h+k)^T I(\theta)(h+k) + \varepsilon_{0,h+k}
\end{aligned}$$

et en retranchant à la troisième équation la somme des deux premières, les termes $L_{\cdot,\cdot}$ s'éliminent et il vient

$$k^T(S_h - S_0) + k^T I(\theta)h = \varepsilon_{0,h+k} - \varepsilon_{h,k} - \varepsilon_{0,h} + \frac{1}{2}k^T(I(\theta + h/\sqrt{n}) - I(\theta))k.$$

Des quatre restes, tous sauf $\varepsilon_{h,k}$ tendent clairement vers 0 en probabilité sous la loi P_θ, la convergence étant uniforme en θ, h, k. Pour $\varepsilon_{h,k}$, on écrit avec les notations du lemme 40

$$\begin{aligned}
P_\theta\big(|\varepsilon_{h,k}| > \alpha\big) &= P_\theta\big(|\varepsilon_{h,k}| > \alpha, \Lambda_n \leqslant \eta\big) + P_\theta\big(|\varepsilon_{h,k}| > \alpha, \Lambda_n \geqslant \eta\big) \\
&\leqslant P_\theta\big(\Lambda_n \leqslant \eta\big) + \eta^{-1}E_\theta\big(1_{|\varepsilon_{h,k}|>\alpha}\Lambda_n\big) \\
&\leqslant P_\theta\big(\log(\Lambda_n) \leqslant A\big) + \eta^{-1}P_{\theta+h/\sqrt{n}}\big(|\varepsilon_{h,k}| > \alpha\big), \quad A = -\log(\eta) \\
&\leqslant P_\theta\big(|\varepsilon_{0,h}| > A/2\big) + P_\theta\big(|h^T S_n(\theta) - \tfrac{1}{2}h^T I(\theta)h| > A/2\big) + \eta^{-1}P_{\theta+h/\sqrt{n}}\big(|\varepsilon_{h,k}| > \alpha\big).
\end{aligned}$$

Tous ces termes peuvent être rendus arbitrairement proches de zéro en choisissant η petit (A grand) puis n grand, ceci uniformément en θ et h grâce à (A.8) et à la convergence en loi de $S_n(\theta)$, la convergence étant uniforme en θ pour les premier et troisième terme. Si $I(\theta)$ est borné, la convergence est également uniforme pour le deuxième en raison de l'inégalité de Chebyshev. ∎

B

Compléments sur les expériences régulières

On va montrer que la R-régularité (cf. p. 71) implique l'hypothèse plus générale de différentiabilité en moyenne quadratique uniforme (DMQ). Cette dernière hypothèse est mathématiquement plus adaptée à l'étude de certains aspects des expériences de la forme $P_\theta^{\otimes n}$; elle conduira à (B.9) qui généralise (III.10) et à l'estimée des distances de Hellinger (B.8) qui a été utilisée dans la démonstration du théorème 29. L'hypothèse DMQ 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 $L_2(\mu)$, la variante 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 $\nabla p_\theta / 2\sqrt{p_\theta}$ 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_θ. Un autre jeu d'hypothèses est adapté aux suites d'expériences statistiques plus générales, c'est l'hypothèse LAN (p. 109), et sa version uniforme ULAN (p. 113). On va montrer les implications suivantes (théorèmes 52, 54, et 55) :

$$\begin{array}{ccccccccc} \text{R-régularité} & \Longrightarrow & \text{Régularité} & \Longrightarrow & \text{DMQU} & \Longrightarrow & \text{DMQ} & & \text{(Cas i.i.d)} \\ & & & & \Downarrow & & \Downarrow & & \\ & & & & \text{ULAN} & \Longrightarrow & \text{LAN} & & \text{(Cas général)} \end{array}$$

Une référence pour ce chapitre est [36] (chapitre 2, ou appendice A.5) ou encore [18].

B.1 Différentiabilité en moyenne quadratique

46 - Définition (Différentiabilité en moyenne quadratique)

L'expérience est dite différentiable en moyenne quadratique uniformément sur K compact intérieur à Θ (DMQU), s'il existe une fonction R_θ telle que pour tout $\theta \in K$:

$$\sqrt{p_{\theta+h}}(y) = \sqrt{p_\theta}(y) + \tfrac{1}{2} h^T R_\theta(y) + \varphi_{\theta,h}(y) \tag{B.1}$$

avec

$$\sup_{\theta \in K} \int \varphi_{\theta,h}(y)^2 \mu(dy) = o(|h|^2). \tag{B.2}$$

Si K est un singleton, $K = \{\theta\}$, on dit simplement que l'expérience est différentiable en moyenne quadratique (DMQ) en θ. L'expérience est différentiable en moyenne quadratique (DMQ) sur Θ si elle l'est en tout point de Θ.

Si $\sqrt{p_\theta}$ est différentiable, R_θ est simplement $\nabla_\theta p_\theta/\sqrt{p_\theta}$. Dans la suite on posera

$$\nabla_\theta p_\theta = R_\theta \sqrt{p_\theta} \tag{B.3}$$

$$S_\theta = \frac{R_\theta}{\sqrt{p_\theta}} 1_{p_\theta>0} = \frac{\nabla_\theta p_\theta}{p_\theta} 1_{p_\theta>0} \quad \text{(le score)} \tag{B.4}$$

$$I(\theta) = \int R_\theta(y) R_\theta(y)^T \mu(dy) = E_\theta[S_\theta(Y) S_\theta(Y)^T]. \tag{B.5}$$

47 - Lemme

Si l'expérience est DMQ en tout point du segment $[\theta, \theta+h]$ de $\mathbb{R}^d$ et si $I(.)$ y est borné, alors, pour presque tout y,

$$\sqrt{p_{\theta+h}}(y) = \sqrt{p_\theta}(y) + \tfrac{1}{2} \int_0^1 h^T R_{\theta+th}(y) dt \tag{B.6}$$

(cet ensemble de y de μ-mesure pleine dépend de h). De plus

$$E_\theta\big[S_\theta(y)\big] = 0. \tag{B.7}$$

Démonstration. L'équation (B.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

$$\int \psi(y)(\sqrt{p_{\theta+h}(y)} - \sqrt{p_\theta(y)})\mu(dy) = \frac{1}{2} \iint_0^1 \psi(y) h^T R_{\theta+th}(y)\mu(dy)$$

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 (B.6) est orthogonale à ψ, donc à toute fonction de $L_2(\mu)$. Pour le dernier point

$$\begin{aligned}\tfrac{1}{2} E_\theta\big[h^T S_\theta(y)\big] &= \int \Big(\sqrt{p_{\theta+h}}(y) - \sqrt{p_\theta}(y) - \varphi_{\theta,h}(y)\Big)\sqrt{p_\theta}(y)\mu(dy) \\ &= -\frac{1}{2}\int \Big(\sqrt{p_{\theta+h}}(y) - \sqrt{p_\theta}(y)\Big)^2 - \int \varphi_{\theta,h}(y)\sqrt{p_\theta}(y)\mu(dy).\end{aligned}$$

Donc

$$\big|\tfrac{1}{2} E_\theta\big[h^T S_\theta(y)\big]\big| = O(|h|^2) + o(|h|) = o(|h|).$$

En remplaçant h par εh et en faisant $\varepsilon \to 0$ on obtient le résultat. ∎

Nous passons maintenant à un résultat général qui sera utilisé plusieurs fois dans la suite.

48 - Lemme

Soit un espace Ω 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 μ-presque partout, $\sup_n \|u_n\|_2^2 < \infty$ et si $\|v_n - v\|_2 \to 0$, alors $\mu(u_n v_n)$ converge vers $\mu(uv)$.

(ii) Si $w_n 1_{|w|>0}$ converge μ-presque partout vers w et si $\overline{\lim}_n \|w_n\|_2^2 \leqslant \|w\|_2^2 < \infty$, alors $\|w_n - w\|_2 \to 0$.

Démonstration. Commençons par le premier point. On pose $\widetilde{u}_n = u_n - u$ et $\widetilde{v}_n = v_n - v$; noter qu'en vertu du lemme de Fatou $\|u\|_2 < \infty$; on a pour tout $A > 0$

$$\begin{aligned}|\mu(u_n v_n) - \mu(uv)| &= |\mu(u_n \widetilde{v}_n) + \mu(\widetilde{u}_n v)| \\ &\leqslant \|u_n\|_2 \|\widetilde{v}_n\|_2 + |\mu(\widetilde{u}_n v 1_{|\widetilde{u}_n| \leqslant A|v|})| + |\mu(\widetilde{u}_n v 1_{|\widetilde{u}_n| > A|v|})| \\ &\leqslant \|u_n\|_2 \|\widetilde{v}_n\|_2 + |\mu(\widetilde{u}_n v 1_{|\widetilde{u}_n| \leqslant A|v|})| + A^{-1}\mu(\widetilde{u}_n^2).\end{aligned}$$

Le théorème de convergence dominée s'applique au deuxième terme car $|\widetilde{u}_n v| 1_{|\widetilde{u}_n| \leqslant A|v|} \leqslant Av^2$, et par conséquent

$$\overline{\lim_n} |\mu(u_n v_n) - \mu(uv)| \leqslant A^{-1} \sup_n \mu(\widetilde{u}_n^2)$$

qui peut être rendu arbitrairement petit en prenant A grand.

Concernant le deuxième point, notons que

$$\|w_n - w\|_2^2 = 2\mu\big(w(w - w_n 1_{|w|>0})\big) + \mu(w_n^2) - \mu(w^2).$$

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 $\leqslant 0$. ∎

49 - Théorème

Soit K une partie compacte de Θ. On suppose l'expérience DMQU sur K (donc DMQ en θ si $K = \{\theta\}$). Alors

$$\sup_{\theta \in K} I(\theta) < \infty.$$

De plus on a l'estimée des distances de Hellinger

$$d_H(P_\theta, P_{\theta+h})^2 = \int \Big(\sqrt{p_{\theta+h}(y)} - \sqrt{p_\theta(y)}\Big)^2 \mu(dy) = \frac{1}{4} h^T I(\theta) h + o(|h|^2) \tag{B.8}$$

le $o(|h|^2)$ étant uniforme en $\theta \in K$.

Pour tout $\theta \in K$ et toute fonction ψ 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

$$\frac{d}{dt} E_{\theta+th}[\psi(Y)]_{|t=0} = E_\theta[\psi(Y) h^T S_\theta(Y)]. \tag{B.9}$$

Remarque. En considérant le cas $K = \{\theta\}$, on voit que le théorème couvre également le cas DMQ.

Démonstration. L'équation (B.1) se réécrit

$$\tfrac{1}{2} h^T R_\theta = \sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \varphi_{\theta,h}. \tag{B.10}$$

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 (B.8) est alors une conséquence immédiate de (B.1).

Passons à (B.9). En remplaçant dans (B.10) h par εh et en multipliant par $2\sqrt{p_\theta}$ il vient

$$\begin{aligned} \varepsilon h^T S_\theta p_\theta &= 2\Big(\sqrt{p_{\theta+\varepsilon h}}\sqrt{p_\theta} - p_\theta - \varphi_{\theta,\varepsilon h}\sqrt{p_\theta}\Big) \\ &= -(\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta})^2 - p_\theta + p_{\theta+\varepsilon h} - 2\varphi_{\theta,\varepsilon h}\sqrt{p_\theta} \end{aligned}$$

et en multipliant par $\psi(y)$ puis en intégrant sous μ il vient

$$E_{\theta+\varepsilon h}[\psi(Y)] - E_\theta[\psi(Y)] - \varepsilon E_\theta[\psi(Y) h^T S_\theta(Y)] = \int \Big((\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta})^2 + 2\varphi_{\theta,\varepsilon h}\sqrt{p_\theta}\Big)\psi d\mu.$$

Il suffit donc de montrer que

$$\int (\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta})^2 \psi d\mu = o(\varepsilon) \tag{B.11}$$

car la majoration du dernier terme est immédiate avec l'inégalité de Cauchy-Schwarz. Mais

$$\varepsilon^{-1}(\sqrt{p_{\theta+\varepsilon h}}-\sqrt{p_\theta})^2\psi \leqslant |R_\theta|\,|\sqrt{p_{\theta+\varepsilon h}}-\sqrt{p_\theta}|\psi+\varepsilon^{-1}|\varphi_{\theta,\varepsilon h}|\,|\sqrt{p_{\theta+\varepsilon h}}-\sqrt{p_\theta}|\psi = u_\varepsilon+v_\varepsilon.$$

Comme pour toute suite ε_n tendant vers 0 on a en vertu de (B.6)

$$\sqrt{p_{\theta+\varepsilon_n h}}(y)-\sqrt{p_\theta}(y)=\tfrac{1}{2}\int_0^{\varepsilon_n} h^T R_{\theta+th}(y)dt$$

(l'ensemble de μ-mesure pleine peut être choisi commun à tous les ε_n), u_{ε_n} tend presque sûrement vers 0 ; de plus, comme $u_{\varepsilon_n}^2 \leqslant 2(p_{\theta+\varepsilon_n h}+p_\theta)\psi^2+2|R_\theta|^2$, cette suite est bornée dans $L_2(\mu)$; le premier point du lemme 48, 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 (B.2). ∎

Le résultat qui suit intervient dans la démonstration du th. 28.

50 - Lemme

On suppose que l'expérience est DMQ en θ_* avec $I(\theta_*)<\infty$, et que pour un voisinage $\mathscr{V}$ de θ_* on a

$$\forall\theta\in\mathscr{V},\quad |\ln p_\theta(y)-\ln p_{\theta_*}(y)|\leqslant|\theta-\theta_*|L(y) \tag{B.12}$$

avec

$$E_{\theta_*}\left[L(y)^2\right]<\infty.$$

Alors, la fonction $\theta\mapsto l(\theta)=E_{\theta_*}\left[\ln p_\theta(Y)\right]$ admet le développement de Taylor en θ_* à l'ordre deux

$$l(\theta_*+h)-l(\theta_*)=-\frac{1}{2}h^T I(\theta_*)h+o(|h|^2).$$

Démonstration. On a pour $x>-1$

$$\begin{aligned}&\ln(1+x)-x+2(\sqrt{1+x}-1)^2=f(x)r(x)\\&f(x)=|x|\ln(1+x)^2 1_{x\leqslant 1}+\left(\sqrt{1+x}-1\right)^2 1_{x>1}\\&|r(x)|\leqslant C\end{aligned}$$

en effet ceci est valide au voisinage de -1, 0, et $+\infty$. Avec $x=p_{\theta_*+h}(y)/p_{\theta_*}(y)-1$, et en notant que $\ln(1+x)\leqslant|h|L(y)$, il vient

$$\begin{aligned}&\ln(p_{\theta_*+h}(y))-\ln(p_{\theta_*}(y))+p_{\theta_*+h}(y)/p_{\theta_*}(y)-1+2(\sqrt{p_{\theta_*+h}(y)}-\sqrt{p_{\theta_*}(y)})^2 p_{\theta_*}(y)^{-1}\\&\qquad=\left(|x||h|^2L(y)^2 1_{x\leqslant 1}+\left(\sqrt{p_{\theta_*+h}(y)/p_{\theta_*}(y)}-1\right)^2 1_{x>1}\right)r(x)\end{aligned}$$

et en prenant l'espérance sous p_{θ_*},

$$\begin{aligned}&E_{\theta_*}\left[\ln(p_{\theta_*+h}(y))-\ln(p_{\theta_*}(y))\right]+2\int\left(\sqrt{p_{\theta_*+h}(y)}-\sqrt{p_{\theta_*}(y)}\right)^2\mu(dy)=\bar r(h)\\&|\bar r(h)|\leqslant C\int|p_{\theta_*+h}(y)/p_{\theta_*}(y)-1||h|^2L(y)^2 1_{p_{\theta_*+h}(y)\leqslant 2p_{\theta_*}(y)}p_{\theta_*}(y)\mu(dy)\\&\qquad\qquad+C\int\left(\sqrt{p_{\theta_*+h}(y)}-\sqrt{p_{\theta_*}(y)}\right)^2 1_{p_{\theta_*+h}(y)>2p_{\theta_*}(y)}\mu(dy).\end{aligned}$$

En raison de (B.8), il ne suffit plus que de montrer que $|h|^{-2}\bar{r}(h)$ tend vers 0 quand $h \to 0$. Pour le deuxième terme nous utilsons (B.1) :

$$|h|^{-2}|\bar{r}(h)| \leqslant C \int \big|p_{\theta_*+h}(y)/p_{\theta_*}(y) - 1\big| L(y)^2 1_{p_{\theta_*+h}(y) \leqslant 2p_{\theta_*}(y)} p_{\theta_*}(y)\mu(dy)$$
$$+ C \int |R_\theta(y)|^2 1_{p_{\theta_*+h}(y) > 2p_{\theta_*}(y)} + C|h|^{-2} \int \varphi_{\theta,h}(y)^2 \mu(dy).$$

Les deux premières intégrales tendent vers 0 par application du théorème de convergence dominée et la troisième en raison de (B.2). ∎

B.2 Liens entre les jeux d'hypothèses

51 - 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 Θ dans $L_2(\mu)$.

52 - Théorème (Propriété DMQU pour l'expérience régulière)

Une expérience régulière est DMQU sur tout compact intérieur à Θ.

Démonstration. La fonction $\varphi_{\theta,h}$ définie par (B.1) satisfait presque sûrement en raison de (B.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

$$\int \varphi_{\theta,h}(y)^2\mu(dy) \leqslant \frac{|h|^2}{4}\int_0^1 \int |R_{\theta+th}(y) - R_\theta(y)|^2 \mu(dy)dt \leqslant \frac{|h|^2}{4} \sup_{|\theta-\theta'|\leqslant|h|} \int |R_{\theta'}(y) - R_\theta(y)|^2\mu(dy).$$

D'où le résultat. ∎

53 - Lemme

Soit f une fonction intégrable sur $[0,1]$ et $F(t) = \int_0^t f(u)du$. Alors pour toute fonction ψ de régularité C^1 sur l'intervalle $F([0,1])$

$$\psi(F(1)) - \psi(F(0)) = \int_0^1 \psi'(F(t))f(t)dt.$$

Démonstration. Pour toute fonction ψ de régularité C^2, on a pour tout $n > 0$

$$\begin{aligned}
\psi(F(1)) - \psi(F(0)) &= \sum_{k=1}^n \psi\big(F(\tfrac{k}{n})\big) - \psi\big(F(\tfrac{k-1}{n})\big) \\
&= \sum_{k=1}^n \psi'(c_k) \int_{(k-1)/n}^{k/n} f(t)dt, \qquad \text{pour des } c_k \in [F(\tfrac{k}{n}), F(\tfrac{k-1}{n})] \\
&= \sum_{k=1}^n \int_{(k-1)/n}^{k/n} \psi'(F(t))f(t)dt + O(\|\psi''\|_\infty) \int_{(k-1)/n}^{k/n} |F(t) - c_k|\,|f(t)|dt \\
&= \int_0^1 \psi'(F(t))f(t)dt + O(\|\psi''\|_\infty) \sup_{|t-s|<\frac{1}{n}} |F(t) - F(s)| \int_0^1 |f(t)|dt
\end{aligned}$$

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 ψ de régularité C^1 par approximation de ψ par une suite ψ_n de régularité C^2 telle que $\|\psi_n - \psi\|_\infty + \|\psi_n' - \psi'\|_\infty \to 0$. ∎

54 - Théorème (Régularité de l'expérience R-régulière)

Soit un modèle R-régulier au sens de la définition 26 p. 71, où $I(\theta)$ est donné par (B.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 48 implique la convergence de $R_{\theta'}$ vers R_θ dans $L_2(\mu)$ lorsque θ' tend vers θ.

On définit $\varphi_{\theta,h}$ par (B.1) et il ne reste qu'à démontrer (B.2). Posons $F(t) = p_{\theta+th}(y)$ et $f(t) = \langle h, \nabla_\theta p_{\theta+th}(y)\rangle$. En vertu du lemme 53 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 $I(.)$ soit borné implique que μ-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 ε 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 52. ∎

55 - Théorème (Hypothèse LAN sous DMQ)

Sous l'hypothèse DMQ sur Θ et d'indépendance des observations, l'hypothèse LAN (p. 109) est satisfaite, avec

$$S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_\theta}{\sqrt{p_\theta}}(Y_i) 1_{p_\theta(Y_i) > 0}.$$

Si la différentiabilité est uniforme sur K (i.e. DMQU, en particulier sous l'hypothèse de régularité de l'expérience, et à plus forte raison de R-régularité) ULAN (p. 113) est satisfait sur K.

Démonstration. On veut montrer que

$$\varepsilon_n(\theta, h) = \sum_{i=1}^n \log p_{\theta_n}(Y_i) - \log p_\theta(Y_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{h^T R_\theta}{\sqrt{p_\theta}}(Y_i) + \frac{1}{2} h^T I(\theta) h, \quad \theta_n = \theta + h/\sqrt{n}$$

satisfait pour tout $M < \infty$ et tout $\alpha > 0$

$$\lim_{n \longrightarrow \infty} \sup_{|h| \leqslant M} P_\theta(|\varepsilon_n(\theta, h)| > \alpha) = 0.$$

Soit

$$A_n = \big\{y : |\tfrac{h^T}{\sqrt{n}} R_\theta(y)| + 2|\varphi_{\theta,h/\sqrt{n}}(y)| \leqslant \sqrt{p_\theta(y)}\big\}$$
$$\Omega_n = \big\{\omega : Y_i \in A_n, 1 \leqslant i \leqslant n\big\}.$$

Montrons que $P(\Omega_n) \to 1$ uniformément en $|h| \leqslant M$:

$$\begin{aligned}
P(\Omega\backslash\Omega_n) &\leqslant \sum_{i=1}^n P\big(Y_i \notin A_n\big) \\
&\leqslant nP\Big(|\tfrac{h^T}{\sqrt{n}} R_\theta(Y_1)| > \tfrac{1}{2}\sqrt{p_\theta(Y_1)}\Big) + nP\Big(|\varphi_{\theta,h/\sqrt{n}}(Y_1)| > \tfrac{1}{2}\sqrt{p_\theta(Y_1)}\Big) \\
&\leqslant n\int \frac{(2\frac{h^T}{\sqrt{n}} R_\theta(y))^2}{p_\theta(y)} 1_{|h^T R_\theta(y)| > \frac{1}{2}\sqrt{np_\theta(y)}} p_\theta(y)\mu(dy) + n\int \frac{4\varphi_{\theta,h/\sqrt{n}}(y)^2}{p_\theta(y)} p_\theta(y)\mu(dy) \\
&= 4\int (h^T R_\theta(y))^2 1_{|h^T R_\theta(y)| > \frac{1}{2}\sqrt{np_\theta(y)}} 1_{p_\theta(y)>0}\mu(dy) + 4n\int \varphi_{\theta,h/\sqrt{n}}(y)^2\mu(dy).
\end{aligned}$$

Le premier terme tend vers 0 par application du théorème de convergence dominée et le second en raison de (B.2). Comme $P(\Omega_n) \to 1$, il suffit de montrer que $\varepsilon_n(\theta,h)1_{\Omega_n}$ converge en probabilité vers 0 uniformément en $|h| \leqslant M$; tout va être basé sur une décomposition étrangement complexe de $\varepsilon_n(\theta,h)$. Définissons q_n par

$$\sqrt{q_n}(y) = \sqrt{p_\theta}(y) + \tfrac{1}{2\sqrt{n}} h^T R_\theta(y).$$

La variable $q_n(Y_i)$ est bien définie sur Ω_n (le membre de droite est $\geqslant 0$), et $\varepsilon_n(\theta,h)$ y est la somme des quatre termes suivants :

$$\begin{aligned}
\varepsilon_n'(\theta,h) &= 2\sum_{i=1}^n \log\frac{\sqrt{p_{\theta_n}}}{\sqrt{q_n}}(Y_i) - \frac{\varphi_{\theta,h/\sqrt{n}}(Y_i)}{\sqrt{p_\theta}} \\
\varepsilon_n''(\theta,h) &= 2\sum_{i=1}^n \frac{\varphi_{\theta,h/\sqrt{n}}(Y_i)}{\sqrt{p_\theta}} \\
\varepsilon_n'''(\theta,h) &= 2\sum_{i=1}^n \log\frac{\sqrt{q_n}}{\sqrt{p_\theta}}(Y_i) - \frac{1}{2}\frac{h^T R_\theta(y)}{\sqrt{n}\sqrt{p_\theta}}(Y_i) + \frac{1}{8n}\Big(\frac{h^T R_\theta(y)}{\sqrt{p_\theta}}(Y_i)\Big)^2 \\
\varepsilon_n''''(\theta,h) &= \frac{1}{2}h^T I(\theta)h - \frac{1}{4n}\sum_{i=1}^n \Big(\frac{h^T R_\theta(y)}{\sqrt{p_\theta}}(Y_i)\Big)^2.
\end{aligned}$$

Commençons par $\varepsilon_n'(\theta,h)$: sur Ω_n, $|\frac{\varphi_{\theta,h/\sqrt{n}}(y)}{\sqrt{p_\theta}}(Y_i)| \leqslant 1/2$ et $\frac{1}{2} \leqslant \frac{\sqrt{q_n}}{\sqrt{p_\theta}}(Y_i) \leqslant 2$, ce qui joue un rôle important dans les calculs qui suivent (nous omettons l'argument y) :

$$\begin{aligned}
\Big|\log\frac{\sqrt{p_{\theta_n}}}{\sqrt{q_n}} - \frac{\varphi_{\theta,h/\sqrt{n}}}{\sqrt{p_\theta}}\Big| &= \Big|\log\Big(1 + \frac{\varphi_{\theta,h/\sqrt{n}}}{\sqrt{q_n}}\Big) - \frac{\varphi_{\theta,h/\sqrt{n}}}{\sqrt{p_\theta}}\Big| \qquad \text{(cf. (B.1))} \\
&\leqslant \Big|\frac{\varphi_{\theta,h/\sqrt{n}}}{\sqrt{q_n}} - \frac{\varphi_{\theta,h/\sqrt{n}}}{\sqrt{p_\theta}}\Big| + C\frac{\varphi^2_{\theta,h/\sqrt{n}}}{q_n} \\
&= \frac{\varphi_{\theta,h/\sqrt{n}}}{\sqrt{q_n p_\theta}}\frac{1}{2\sqrt{n}}|h^T R_\theta| + C\frac{\varphi^2_{\theta,h/\sqrt{n}}}{q_n} \\
&\leqslant \frac{1}{\sqrt{n}}\frac{\varphi_{\theta,h/\sqrt{n}}}{p_\theta}|h^T R_\theta| + 4C\frac{\varphi^2_{\theta,h/\sqrt{n}}}{p_\theta}.
\end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwartz et (B.2) on obtient que premier terme a une espérance $o(|h|^2/n)$, et le deuxième satisfait le même borne ; $\varepsilon_n'(\theta,h)$ converge donc vers 0 en probabilité.

La variance de $\varepsilon_n''(\theta,h)$ tend vers 0 à cause de (B.2) et son espérance vaut, en raison de (B.7),

$$\begin{aligned}
E[\varepsilon_n''(\theta,h)] &= nE\left[\frac{2\varphi_{\theta,h/\sqrt{n}}(Y_1)+h^TR_\theta(Y_1)}{\sqrt{p_\theta}(Y_1)}\right] \\
&= 2n\int\Big(\sqrt{p_{\theta_n}}(y)-\sqrt{p_\theta}(y)\Big)\sqrt{p_\theta}(y)\mu(dy) \\
&= -n\int\Big(\sqrt{p_{\theta_n}}(y)-\sqrt{p_\theta}(y)\Big)^2\mu(dy) \\
&= -\tfrac{1}{4}h^TI(\theta)h+no(|h^2|/n)
\end{aligned}$$

donc

$$\varepsilon_n''(\theta,h)\xrightarrow{P}-\tfrac{1}{4}h^TI(\theta)h.$$

Pour traiter ε''', commençons par remarquer qu'en vertu de la formule de Taylor pour la fonction logarithme au voisinage de 1, on a

$$1_{-1/2\leqslant x}\Big|\log(1+x)-x+\tfrac{1}{2}x^2\Big|\leqslant x^2r(x)$$

où $r(x)\to 0$ si $x\to 0$, et $\|r\|_\infty<\infty$. Comme $\frac{\sqrt{q_n}}{\sqrt{p_\theta}}=1-\frac{1}{2}\frac{h^TR_\theta(y)}{\sqrt{n}\sqrt{p_\theta}}$, on obtient que

$$\begin{aligned}
E[|\varepsilon_n'''(\theta,h)|1_{\Omega_n}] &\leqslant 2n\int\Big(\sqrt{q_n(y)}-\sqrt{p_\theta(y)}\Big)^2r\Big(\frac{\sqrt{q_n}}{\sqrt{p_\theta}}(y)-1\Big)1_{A_n}(y)\mu(dy) \\
&\leqslant \frac{1}{2}\int h^TR(\theta)(y)R_\theta(y)^Thr\Big(\frac{\sqrt{q_n}}{\sqrt{p_\theta}}(y)-1\Big)1_{A_n}(y)\mu(dy)
\end{aligned}$$

qui tend vers 0 en vertu du théorème de convergence dominée. Finalement, comme ε'''' converge presque sûrement vers $\frac{1}{4}h^TI(\theta)h$, le résultat est démontré. ∎

C

Démonstration du lemme de Kronecker matriciel

Il s'agit du lemme 39. Tout repose sur une sommation d'Abel. 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

$$\begin{aligned}
\|R_n\|^{-1}\Big\|\sum_{i=n_0}^{n-1}(R_{i+1} - R_i)S_i\Big\| &\leqslant \|R_n\|^{-1}\sum_{i=n_0}^{n-1}\|R_{i+1} - R_i\|\ \|S_i\| \\
&\leqslant \varepsilon(n_0)\|R_n\|^{-1}\sum_{i=n_0}^{n-1}\|R_{i+1} - R_i\|, \quad \varepsilon(n_0) = \sup_{i\geqslant n_0}\|S_i\| \\
&\leqslant \varepsilon(n_0)\|R_n\|^{-1}\sum_{i=n_0}^{n-1}\mathrm{Tr}(R_{i+1} - R_i) \\
&\leqslant \varepsilon(n_0)\ \|R_n\|^{-1}\mathrm{Tr}(R_n) \\
&\leqslant \varepsilon(n_0)d.
\end{aligned}$$

On a donc montré que

$$\overline{\lim_n}\,\|R_n\|^{-1}\Big\|\sum_{i=1}^{n} R_i M_i\Big\| \leqslant \varepsilon(n_0)d$$

pour tout n_0. Par conséquent la limite est nulle. ∎

D

Une inégalité de Sobolev

Ω est ici un ouvert de $\mathbb{R}^d$. On dira que Ω possède la **propriété de cône** s'il existe un cône tronqué C d'intérieur non-vide et de sommet 0 tel pour tout point $x \in \Omega$, il existe une rotation R telle que $x + RC \subset \Omega$; ceci signifie qu'il n'y a pas de pointe à la frontière de Ω, par exemple un demi-espace. **Tout ouvert borné convexe possède la propriété de cône** ([88] th. 1.2.2.2 et cor. 1.2.2.3).

Pour toute fonction g différentiable sur Ω et tout $q > d$, on a l'inégalité de Sobolev (lemme 5.15 de [21], ou théorème 4.12 de [22] case A avec $m = 1$ et $j = 0$) :

$$\|g\|_\infty^q \leqslant C_q \int_\Omega \Big(\|\nabla g(z)\|^q + |g(z)|^q \Big) dz \tag{D.1}$$

où C_q ne dépend que de q et de C. Pour $\varepsilon \leqslant 1$ on peut appliquer la même inégalité à $g(\varepsilon z)$ sur Ω/ε (qui vérifie la condition de cône avec même C) et l'on obtient

$$\begin{aligned} \|g\|_\infty^q &\leqslant C_q \int_{\Omega/\varepsilon} \Big(\varepsilon^q \|\nabla g(\varepsilon z)\|^q + |g(\varepsilon z)|^q \Big) dz \\ &= C_q \varepsilon^{-d} \int_\Omega \Big(\varepsilon^q \|\nabla g(z)\|^q + |g(z)|^q \Big) dz. \end{aligned}$$

Il pourra être 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

$$\|h\|_\infty \leqslant C_q \varepsilon^{-d} \int_\Omega \Big(\varepsilon^q \|\nabla \log h(z)\|^q + 1 \Big) |h(z)| dz, \qquad \varepsilon \leqslant 1. \tag{D.2}$$

E

Une borne sur les processus empiriques

Nous présentons ici des résultats un peu techniques utiles à la démonstration du théorème 15. Commençons par une inégalité très importante due à Arkadi Nemirovski (2000) :

56 - Théorème Inégalité de Nemirovski

Soit $X_i \in \mathbb{R}^T$, $T \geqslant 3$, une suite de variables indépendantes vectorielles. En notant $S(t) = \sum_{i=1}^n X_i(t)$, on a

$$E\Big[\sup_t \big|S(t) - E[S(t)]\big|^2\Big] \leqslant 8\ln(2T) E\Big[\sup_t \sum_{i=1}^n X_i(t)^2\Big]. \tag{E.1}$$

Dans la version originale de Nemirovski, $\sup_t |\sum_i X_i^2|$ est remplacé par $\sum_i \sup_t |X_i^2|$, c'est une inégalité beaucoup plus simple à montrer (cf. p. ex. [69] p.141 où l'on voit que cette inégalité, plus faible, reste vraie pour des accroissements de martingale), qui nous suffira pour la démonstration du théorème qui suit. Nous référons à [48] §14.10.1 pour une démonstration de (E.1) (ou [1] Th. 11.2 et Ex. 11.8), et à [1] §11.2 pour une discussion plus approfondie.

57 - Théorème

Soit $\theta \mapsto L_i(\theta)$ une suite processus indépendants à valeurs réelles définis sur une boule $\{\theta : \|\theta\| \leqslant r\} \subset \mathbb{R}^d$. On suppose qu'il existe des variables indépendantes $\dot{L}_i$ telles que pour tout θ, avec probabilité 1,

$$\big|L_i(\theta) - L_i(\theta')\big| \leqslant \dot{L}_i \|\theta - \theta'\|, \text{ pour tout } \theta' \tag{E.2}$$

$$E\big[L_i(\theta)\big] = 0 \tag{E.3}$$

$$L_i(0) = 0 \tag{E.4}$$

et que les constantes suivantes sont finies

$$C(L) = \sup_i E\big[\sup_\theta L_i(\theta)^2\big]^{\frac{1}{2}} \tag{E.5}$$

$$C(\dot{L}) = \max_i E\big[\dot{L}_i^2\big]^{\frac{1}{2}}. \tag{E.6}$$

Alors,

$$E\Big[\sup_{\|\theta\|<r} n^{-1}\Big|\sum_{i=1}^n L_i(\theta)\Big|^2\Big]^{\frac{1}{2}} \leqslant C_0 d C(L)\left(1 + \log\Big(\frac{C(\dot{L})r}{C(L)}\Big)\right)^{\frac{1}{2}} \leqslant C_0 d C(\dot{L}) r \tag{E.7}$$

où C_0 est une constante universelle.

Remarque. La première inégalité sera utilisée dans un contexte particulier où $r \to 0$, 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 $\Theta_x = \{\theta : \|\theta\| \leqslant r\} \cap \big((2x)\mathbb{Z}^d\big)$. On note

$$S_n(\theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} L_i(\theta), \qquad \varphi(x) = E\Big[\sup_{\theta\in\Theta_x} |S_n(\theta)|^2\Big]^{\frac{1}{2}}.$$

En vertu du théorème de Lebesgue, la limite de $\varphi(x)$ quand x tend vers 0 est bien $\varphi(0)$, la racine carrée du le membre de gauche de (E.1). On a pour $x \leqslant y$, en notant $\widehat{\theta}$ le point de Θ_y le plus proche de θ

$$\sup_{\theta\in\Theta_x} |S_n(\theta)| \leqslant \sup_{\theta\in\Theta_x} |S_n(\widehat{\theta})| + \sup_{\theta\in\Theta_x} |S_n(\widehat{\theta}) - S_n(\theta)| = \sup_{\theta\in\Theta_y} |S_n(\theta)| + \sup_{\theta\in\Theta_x} |S_n(\widehat{\theta}) - S_n(\theta)|.$$

Pour traiter le dernier terme, on va appliquer l'inégalité de Nemirovski avec $X_i = n^{-\frac{1}{2}}(L_i(\theta) - L_i(\widehat{\theta}))_{\theta\in\Theta_x}$, et $T = \Theta_x$. Comme $\|\widehat{\theta} - \theta\|^2 \leqslant dy^2$, on a $\sup_\theta X_i(\theta)^2 \leqslant n^{-1} dy^2 \dot{L}_i^2$ (en vertu de (E.2) avec $\theta' = \widehat{\theta}$), et par ailleurs $|\Theta_x| \leqslant (r/x)^d$. Donc

$$E\Big[\sup_{\theta\in\Theta_x} |S_n(\widehat{\theta}) - S_n(\theta)|^2\Big] \leqslant 8\ln(2(r/x)^d) n^{-1} dy^2 E\Big[\sum \dot{L}_i^2\Big].$$

En rassemblant ces équations on obtient

$$\varphi(x) \leqslant \varphi(y) + \Big(8d\ln(2(r/x)^d) C(\dot{L})^2 y^2\Big)^{\frac{1}{2}}.$$

Soit $n_0 \geqslant 1$. En faisant $x = 2^{-n-1}r$ et $y = 2^{-n}r$ et en sommant sur $n \geqslant n_0$, il vient

$$\begin{aligned}
\varphi(0) &\leqslant \varphi(2^{-n_0}r) + \sum_{n=n_0}^{\infty} \Big(8d\ln(2^{(n+1)d+1}) C(\dot{L})^2 r^2 2^{-2n}\Big)^{\frac{1}{2}}. \\
&\leqslant \varphi(2^{-n_0}r) + C_0 d C(\dot{L}) r \sum_{n=n_0}^{\infty} n^{\frac{1}{2}} 2^{-n} \\
&\leqslant \varphi(2^{-n_0}r) + C_0 d C(\dot{L}) r n_0^{\frac{1}{2}} 2^{-n_0}
\end{aligned}$$

où la constante C_0 a changé. 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 $E[\|X_i\|_\infty^2] \leqslant n^{-1}C(L)^2$, d'où

$$\varphi(2^{-n_0}r)^2 \leqslant 8\ln(2^{dn_0+1}) C(L)^2 \leqslant C_0 d n_0 C(L)^2.$$

Donc, si C_0 a été choisi $\geqslant 1$,

$$\varphi(0) \leqslant C_0 n_0^{\frac{1}{2}} d\Big(C(L) + C(\dot{L}) r 2^{-n_0}\Big).$$

On obtient alors la première inégalité en prenant $n_0 = \big\lceil \log_2\big(C(\dot{L})r/C(L)\big)\big\rceil$. La deuxième inégalité vient de ce que $C(L) \leqslant C(\dot{L})r$ et $x \mapsto (1 + \ln x)/x$ est décroissante pour $x \geqslant 1$. ∎

58 - Lemme

Soit U une variable aléatoire positive telle pour tout $x > 0$

$$E\big[U 1_{U\leqslant x}\big] \leqslant C_1\sqrt{x} + C_2,$$

alors pour tout $x > 0$

$$P(U > x) \leqslant \frac{4C_1}{\sqrt{x}} + \frac{2C_2}{x}.$$

Démonstration. Posons $x_n = 4^n x$, alors

$$P(x_n \leqslant U \leqslant x_{n+1}) \leqslant x_n^{-1} E\big[U 1_{U\leqslant x_{n+1}}\big] \leqslant C_1 2^{1-n} x^{-\frac{1}{2}} + C_2 x^{-1} 4^{-n}$$

et il ne reste plus qu'à sommer sur $n \geqslant 0$. ∎

F

Théorèmes-limite pour les martingales

Nous énonçons un théorème classique, puis le théorème 2.17 et le corollaire 3.1 de [8].

59 - Théorème

Soit $S_n = \sum_{i=1}^n X_i$ une $\mathscr{F}_n$-martingale, alors, sur l'ensemble $\{\omega : \sum E[X_i^2|\mathscr{F}_{i-1}] < \infty\}$, S_n converge p.s.

Un tableau de martingales est une famille de variables et de tribus $\{S_{ni}, \mathscr{F}_{ni}, 1 \leqslant i \leqslant n\}$ telles que

$$E\big[S_{n,i+1}|\mathscr{F}_{ni}\big] = S_{ni}.$$

C'est-à-dire que pour tout n fixé, la suite S_{ni}, $1 \leqslant i \leqslant n$, 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, $X_{ni} = \frac{1}{\sqrt{n}} Y_i$.

60 - Théorème

Soit $\{S_{ni}, \mathscr{F}_{ni}, 1 \leqslant i \leqslant n, 1 \leqslant n\}$ un tableau de martingales vectorielles :

$$E\big[S_{n,i+1}|\mathscr{F}_{ni}\big] = S_{ni}.$$

Supposons que les variables $X_{ni} = S_{ni} - S_{n,i-1}$ satisfont les conditions suivantes quand $n \to \infty$

$$\sum_i E\big[\|X_{ni}\|^2 1_{\|X_{ni}\|>\varepsilon}|\mathscr{F}_{n,i-1}\big] \xrightarrow{P} 0, \quad \text{pour tout } \varepsilon > 0 \tag{F.1}$$

$$\sum_{i=1}^n E\big[X_{ni} X_{ni}^T|\mathscr{F}_{n,i-1}\big] \xrightarrow{P} V. \tag{F.2}$$

pour une certaine matrice V. Alors

$$S_{nn} \xrightarrow{d} \mathcal{N}(0, V). \tag{F.3}$$

Remarque. La condition (F.1) est satisfaite si pour un $\eta > 0$, $\sum_i E[\|X_{ni}\|^{2+\eta}] \to 0$.

61 - Théorème

On se place sous les hypothèses du théorème 60 sauf que l'on autorise V à être une variable aléatoire. On suppose en outre que $\mathscr{F}_{ni} \subset \mathscr{F}_{n+1,i}$. Dans ce cas on a la même conclusion, conditionnellement à V, au sens où pour tout $u \in \mathbb{R}^d$ et toute fonction φ continue bornée

$$E\big[\varphi(V)e^{i\langle u,S_{nn}\rangle}\big] \to E\big[\varphi(V)e^{-\langle Vu,u\rangle/2}\big]. \tag{F.4}$$

G

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 [94] 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 [55]. 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, $y \in \mathbb{R}^q$, la formulation exacte est que le coût de codage de y avec précision ε sur chaque coordonnée est $-\log(\varepsilon^q p(y))$ (i. e. on s'est ramené au cas discret en attribuant la probabilité $\varepsilon^q p(y_i)$ à chaque valeur ε-discrétisée $y_i \in \varepsilon\mathbb{Z}^q$, à un terme d'ordre supérieur près). En raison du caractère additif constant du terme $-\log(\varepsilon^q)$, 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 θ bien choisi et y codé sur la base de p_θ. 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 Θ augmente, le coût de codage de y diminuera probablement tandis que celui de θ 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 (G.5) ou (G.6) plus bas. Rappelons que dans le cas continu le terme additionnel $-q\log(\varepsilon)$ a été omis car il ne dépend pas de θ.

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 » ([156] § 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 θ, 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 [140] 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 $p_{\widehat{\theta}}(y)$, où $\widehat{\theta}(Y)$ est l'estimateur au maximum de vraisemblance ; ce coût de codage doit aussi prendre en compte le codage des composantes de $\widehat{\theta}$ 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 δ_k avec laquelle on transmet $\widehat{\theta}$, est bien décrite dans [94], donnons-en les grands traits. On se place dans le cadre d'observations iid, ce qui permet de mieux mettre en valeur l'essence du calcul.

On suppose qu'il y a une suite de modèles Θ_t, $1 \leqslant t \leqslant T$. Soit $\pm\delta_k/2$ la précision (à déterminer) du codage de chaque coordonnée $\widehat{\theta}_k$; ce codage sera basé sur une probabilité a priori π_t sur Θ_t d'entropie finie [1] ; le $\widehat{\theta}$ codé sera noté $\check{\theta}$. On obtient le coût de codage

$$-\sum_{i=1}^{n} \log \varepsilon^q p_{\check{\theta}}(Y_i) - \log\Big((\prod \delta_k) \pi_t(\widehat{\theta}) \Big) \tag{G.1}$$

où q est la dimension de Y_i, les logarithmes sont négatifs (on raisonne avec ε très petit et on verra que δ_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 δ_k comme suit (elle sera de l'ordre de la précision sur $\widehat{\theta}$, c'est-à-dire $1/\sqrt{n}$). Si l'on pose $\check{\delta} = \check{\theta} - \widehat{\theta}$, l'équation plus haut s'approche en utilisant la formule de Taylor à l'ordre 2, et en ne gardant que les termes dépendants de δ

$$-\sum_{i=1}^{n} \log p_{\check{\theta}}(Y_i) - \sum_k \log \delta_k \simeq -\sum_{i=1}^{n} \log p_{\widehat{\theta}}(Y_i) + \tfrac{1}{2} \check{\delta}^T H_n(\widehat{\theta}) \check{\delta} - \sum_k \log \delta_k \tag{G.2}$$

$$H_n(\theta) = -\nabla^2 \sum_{i=1}^{n} \log p_\theta(Y_i) \tag{G.3}$$

(le terme du premier ordre est nul car $\widehat{\theta}$ 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 en δ, chaque coordonnée de θ doit être discrétisée à un pas différent, et, en approchant $\check{\delta}_k$ par δ_k, on trouve $\delta_k^{-2} \simeq c H_n(\theta)_{kk}$ (c est une constante d'ordre 1 liée à la distribution effective des $\check{\delta}_k \in [-\delta_k, \delta_k]$) et l'on obtient une première approximation de (G.1) :

$$-\sum_{i=1}^{n} \log p_{\widehat{\theta}(Y)}(Y_i) + \frac{d_t}{2} \log n + \frac{1}{2} \log\det(\tfrac{1}{n} H_n(\widehat{\theta})) - \log \pi_t(\widehat{\theta}) \tag{G.4}$$

où d_t désigne la dimension de Θ_t. 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 $\widehat{\theta}$ dans une base de vecteurs propres de $H(\widehat{\theta})$, et l'on retrouve (G.4). *Si l'on considère n grand,* en remarquant que le terme $\log \pi_t(\widehat{\theta})$ est a priori d'ordre $O(d_t)$ si π_t est une mesure produit, le critère (G.4) devient essentiellement

$$BIC(t) = -\log p_{\widehat{\theta}(Y)}(Y) + \frac{d_t}{2} \log n \tag{G.5}$$

où $\widehat{\theta}(Y)$ est l'estimateur au maximum de vraisemblance sous le modèle t (Bayesian Information Criterion

1. On va voir que le choix de π_t n'intervient asymptotiquement pas. Dans [140] le codage est la partie entière $\widehat{\theta}/\delta_k$, écrite en base 2, ce qui fait essentiellement un coût de $\log(|\widehat{\theta}|/\delta_k)_+$ bits, qui ne change pas l'asymptotique.

de Gideon Schwarz) [2]. La référence [149] recommande de conserver la forme plus complète

$$SIC(t) = -\log p_{\widehat{\theta}(Y)}(Y) + \frac{1}{2}\log\det(H_n(\widehat{\theta})), \tag{G.6}$$

la dénomination SIC provenant de [94] éq. (17), où son intérêt est fortement remis en cause par des simulations (cf. § 4.1.2 et 4.1.3 de l'article).

Considérons le cas de la régression linéaire gaussienne $Y_i = X_i\beta + u_i, u_i \sim \mathcal{N}(0,\sigma^2), \beta \in \mathbb{R}^p$, cf. p. 9; ou encore vectoriellement $Y = X\beta + u$. Il s'agit ici de sélectionner une matrice de design X parmi plusieurs possibilités offertes. Le paramètre est $\theta = \beta$, et l'on trouve pour chaque $X \in \mathbb{R}^{n\times p_X}$, supposée de rang plein, après retrait de termes constants

$$\begin{aligned}
&\widehat{\beta} = (X^TX)^{-1}X^TY\\
&H(\widehat{\theta}) = \sigma^{-2}X^TX\\
&BIC(X) = \frac{1}{2\sigma^2}\|Y - X\widehat{\beta}\|^2 + \frac{p_X}{2}\log n\\
&SIC(X) = \frac{1}{2\sigma^2}\|Y - X\widehat{\beta}\|^2 + \frac{1}{2}\log\det\Big(\frac{X^TX}{\sigma^2}\Big).
\end{aligned}$$

Le codage en deux temps consiste à choisir X, puis transmettre une paire $(\widehat{\beta},\widehat{u})$ ainsi que l'indice de X dans la famille, à charge au récepteur de calculer $X\widehat{\beta}+\widehat{u}$.

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 $\widehat{\theta}(y)$ 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 [3].

Ce problème de choix de codage a conduit à une spirale mathématico-philosophique émaillée de discussions byzantines et de calculs sophistiqués alimentant de nombreux articles visant à gagner en conceptualisation. Rien de bien clair ne semble s'en dégager en ce qui concerne l'estimation. Finalement, la théorie initiale de Wallace et Boulton (MML, codage en deux temps) reste une sage option.

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 postulant une probabilité a priori sur les paramètres, $p(\theta) = \sum_t \alpha_t\pi_t(\theta)$ où $\alpha_t > 0$ et π_t est une probabilité sur Θ_t [144]; dans cette approche, $\widehat{t}$ sera obtenu par maximisation de la probabilité a posteriori $\alpha_t\int p_\theta(Y)\pi_t(\theta)d\theta$. Il montre que, lorsque les modèles ont tous une dimension différente, cette méthode revient asymptotiquement à maximiser $BIC(t)$, indépendamment du choix des π_t et des α_t (sous certaines conditions raisonnables...). Ici aussi, c'est la forme (G.6) qui apparaît naturellement. Dans ce calcul, $\exp(-BIC)$ **apparaît comme une version approchée de la probabilité a postériori de** t **sachant** Y.

Christophe Giraud a souligné que si de nombreux modèles ont même dimension, le calcul de Schwarz doit être modifié ([7] p. 49), et de la même façon le troisième terme de (G.4) doit tenir compte de ce qu'un terme de codage de t doit être pris en compte. Ceci ajoute à BIC un terme $-\log(\psi_t)$ où ψ est une certaine probabilité sur $\{1,\dots T\}$. Si ν_d est le nombre de modèles de dimension d, un choix raisonnable

2. Wallace et Freeman dans [157] obtiennent (G.4) par un calcul différent (§ 5.1, formule p.245) : Ils cherchent la stratégie $y \mapsto \check{\theta}(y)$ qui minimise le coût de codage *moyen* sous l'hypothèse que les données ont pour densité $\int p_\theta(y)\pi(d\theta)$, où π est une loi sur l'ensemble de tous les paramètres; le développement n'est pas fait en $\widehat{\theta}$ 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 δ est centré et que les calculs y sont faits en espérance.

Dans un cadre où l'on connaît π, c'est bien entendu plus logique que le codage en deux temps avec $\widehat{\theta}(y)$.

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.

3. P. ex. il propose dans [139] de coder Y sur la base de la mesure $p_{\widehat{\theta}(y)}(y)/\int p_{\widehat{\theta}(z)}(z)dz$, où ici $\widehat{\theta}$ 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.

est $\psi_t \propto \nu_{d_t}^{-1}$ (donner le même poids à chaque dimension), ce qui conduit à

$$BIC(t) = -\sum_{i=1}^{n} \log p_{\widehat{\theta}(Y)}(Y_i) + \frac{d_t}{2}\log n + \log \nu_{d_t}.$$

Le codage est donc en trois temps : t puis θ puis Y. 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 $d \ll p_0$, $\log \nu_d \simeq d \log(p_0/d)$. On aurait pu prendre p. ex. $\psi_d \propto d^{-2}\nu_d^{-1}$ pour pénaliser les grandes dimension mais cela n'aurait que marginalement changé le dernier terme.

Exercice (Le nombre d'individus doit dépasser la dimension du plus grand modèle). On reprend le modèle de régression mais désormais σ est inconnu, et l'on choisit $\theta = (\beta, \tau = \sigma^{-2})$. Montrer qu'on trouve, $d_X = p_X + 1$ et

$$\widehat{\beta} = (X^T X)^{-1} X^T y, \qquad \widehat{\tau} = \widehat{\sigma}^{-2} = n\|Y - X\widehat{\beta}\|^{-2}$$
$$H(\widehat{\theta}) = \begin{pmatrix} \widehat{\sigma}^{-2} X^T X & 0 \\ 0 & \frac{n}{2}\widehat{\sigma}^4 \end{pmatrix}$$
$$BIC(X) = \frac{1}{2} n \log \widehat{\sigma}^2 + \frac{p_X + 1}{2} \log n$$
$$SIC(X) = \frac{1}{2}(n - p_X + 2) \log \widehat{\sigma}^2 + \frac{1}{2} \log \det(X^T X).$$

On suppose que l'on veut choisir parmi certains sous-ensembles de régresseurs extraits d'une famille donnée de cardinal $p_0 > n$, conduisant à une matrice X de rang plein ; X est donc obtenue par extraction de colonnes d'une grande matrice X_0. On suppose également qu'une de ces matrices est $n \times n$, donc inversible. Montrer qu'alors BIC ou SIC choisira toujours un modèle à n variables ($\widehat{\sigma} = 0$).

D'un point de vue codage, cela semble absurde de coder n valeurs de θ, plutôt que de coder les Y_i directement. Ce paradoxe vient de ce que le développement limité (G.2) est injustifié si $n = p$ à cause de la singularité du logarithme en $\widehat{\sigma} = 0$ qui donne un avantage infini à ne pas avoir à transmettre u.

Exercice. On reprend l'exercice 4 p. 7. Montrer que l'estimateur suivant inspiré de SIC

$$\widehat{\theta} = \arg\min_{\theta} \; -\log p_\theta(Z) + \tfrac{1}{2} \log \det(I(\theta))$$

se comporte correctement sur cet exemple, grâce à la pénalisation. On pourra utiliser les formules explicitées à l'exercice 9 p. 92.

Bibliographie

[1] S. Boucheron, G. Lugosi et P. Massart. *Concentration inequalities.* A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013.

[2] A. DasGupta. *Asymptotic theory of statistics and probability.* Springer, 2008.

[3] P. J. Diggle. *Statistical analysis of spatial and spatio-temporal point patterns.* CRC Press, 2014.

[4] B. Efron. *Large-scale inference.* Cambridge Univ. Press, 2010.

[5] A. Gelman et al. *Bayesian data analysis.* CRC Press, 2014.

[6] E. Giné et R. Nickl. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge Univ. Press, 2016.

[7] C. Giraud. *Introduction to high-dimensional statistics.* T. 139. CRC Press, 2015.

[8] P. Hall et C. C. Heyde. *Martingale limit theory and its application.* Academic Press, 1980.

[9] T. Hastie, R. Tibshirani et J. Friedman. *The elements of statistical learning.* 2e éd. Springer, 2009.

[10] P. J. Huber et E. M. Ronchetti. *Robust statistics.* Wiley, 2009.

[11] I. A. Ibragimov et R. Z. Hasminskiĭ. *Statistical estimation.* Springer, 1981.

[12] E. L. Lehmann et G. Casella. *Theory of point estimation.* Springer, 1998.

[13] A. Monfort. *Cours de statistique mathématique.* Economica, 1997.

[14] A. Pakes et D. Pollard. « Simulation and the asymptotics of optimization estimators ». In : *Econometrica* 57.5 (1989), p. 1027-1057.

[15] J. Pfanzagl. *Parametric statistical theory.* Walter de Gruyter, 1994.

[16] C. R. Rao et Y. Wu. « On model selection ». In : *Model selection.* T. 38. Inst. Math. Statist., 2001, p. 1-64.

[17] C. P. Robert. *The Bayesian choice.* 2e éd. Springer-Verlag, 2001.

[18] A. W. van der Vaart. *Asymptotic statistics.* Cambridge Univ. Press, 1998.

[19] A. W. van der Vaart et J. A. Wellner. *Weak convergence and empirical processes.* Springer, 1996.

[20] L. Wasserman. *All of nonparametric statistics.* Springer, 2006.

Références

[21] R. A. Adams. *Sobolev spaces*. Academic Press, 1975.

[22] R. A. Adams et J. J. F. Fournier. *Sobolev spaces*. Elsevier/Academic Press, 2003.

[23] 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.

[24] Y. Atchadé. *Markov Chain Monte Carlo confidence intervals*. Rapp. tech. 2013.

[25] J.-Y. Audibert. « Fast learning rates in statistical inference through aggregation ». In : *Ann. Statist.* 37.4 (2009), p. 1591-1646.

[26] J.-Y. Audibert. « PAC-Bayesian statistical learning theory ». Thèse de doct. Univ. Paris 6 et Paris 7, 2004.

[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] F. Balabdaoui et Y. Kulagina. « The Enigma behind the Good–Turing formula ». In : Snapshots of modern mathematics from Oberwolfach (2021). https ://publications.mfo.de/handle/mfo/3875.

[30] E. Barankin. « Locally best unbiased estimates ». In : *Ann. Math. Statistics* 20 (1949), p. 477-501.

[31] A. R. Barron. « Are Bayes Rules Consistent in Information ? » In : *Open Problems in Communication and Computation*. Sous la dir. de T. M. Cover et B. Gopinath. Springer, 1987, p. 85-91.

[32] 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.

[33] R. Beran. « Prepivoting test statistics : a bootstrap view of asymptotic refinements ». In : *J. Amer. Statist. Assoc.* 83.403 (1988), p. 687-697.

[34] J. Berkson. « Minimum chi-square, not maximum likelihood ! » In : *Ann. Statist.* 8.3 (1980).

[35] J. Besag. « Spatial interaction and the statistical analysis of lattice systems ». In : *J. Roy. Statist. Soc. Ser. B* 36 (1974), p. 192-236.

[36] P. J. Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, 1993.

[37] 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.

[38] 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.

[39] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 1999.

[40] 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.

[41] A. Borovkov. *Statistique Mathématique*. Mir, 1987.

[42] 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.

[43] 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.

[44] P. Brémaud. *Point processes and queues*. Springer, 1981.

[45] 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.

[46] 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.

[47] 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.

[48] P. Bühlmann et S. van de Geer. *Statistics for high-dimensional data*. Springer, 2011.

[49] 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.

[50] S. X. Chen et I. Van Keilegom. « A review on empirical likelihood methods for regression ». In : *TEST* 18.3 (2009), p. 415-447.

[51] S. N. Chiu et al. *Stochastic geometry and its applications*. Third. John Wiley & Sons, 2013.

[52] 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.

[53] 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`, p. 1260-1270.

[54] 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.

[55] T. M. Cover et J. A. Thomas. *Elements of information theory*. Second. Wiley, 2006.

[56] D. Cox. « Partial likelihood ». In : *Biometrika* 62.2 (1975), p. 269-276.

[57] D. Cox. « The choice between alternative ancillary statistics ». In : *J. Roy. Statist. Soc. Ser. B* 33 (1971), p. 251-255.

[58] D. Cox et N. Reid. « A note on pseudolikelihood constructed from marginal densities ». In : *Biometrika* 91.3 (2004), p. 729-737.

[59] N. Cressie. *Statistics for spatial data*. Wiley, 1991.

[60] S. Dachian et Y. Kutoyants. « Hypotheses testing : Poisson versus self-exciting ». In : *Scand. J. Statist.* 33.2 (2006), p. 391-408.

[61] A. S. Dalalyan et A. B. Tsybakov. « Mirror averaging with sparsity priors ». In : *Bernoulli* 18.3 (2012), p. 914-944.

[62] J.-J. Daudin, F. Picard et S. Robin. « A mixture model for random graph ». In : *Statistics and Computing* 18 (juin 2008), p. 173-183.

[63] L. Dicker et S. Zhao. *Nonparametric empirical Bayes and maximum likelihood estimation for high-dimensional data analysis*. Rapp. tech. 2014.

[64] O. Diekmann, H. Heesterbeek et T. Britton. *Mathematical tools for understanding infectious disease dynamics*. Princeton University Press, 2013.

[65] 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.

[66] 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.

[67] 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.

[68] J. Doob. « Application of the theory of martingales ». In : *Le Calcul des Probabilités et ses Applications*. CNRS, 1949, p. 23-27.

[69] L. Dümbgen et al. « Nemirovski's inequalities revisited ». In : *Amer. Math. Monthly* 117.2 (2010), p. 138-160.

[70] N. Dunford et J. Schwartz. *Linear operators. Part I*. Wiley, 1988.

[71] B. Efron. « Robbins, empirical Bayes and microarrays ». In : *Ann. Statist.* 31.2 (2003), p. 366-378.

[72] 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.

[73] B. Efron. « Tweedie's Formula and Selection Bias ». In : *Journal of the American Statistical Association* 106.496 (2011), p. 1602-1614.

[74] 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.

[75] H.-r. Fang et Y. Saad. « Two classes of multisecant methods for nonlinear acceleration ». In : *Numer. Linear Algebra Appl.* 16.3 (2009), p. 197-221.

[76] 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.

[77] T. Fleming et D. Harrington. *Counting processes and survival analysis*. Wiley, 1991.

[78] C. Francq et J.-M. Zakoïan. « Estimating linear representations of nonlinear processes ». In : *J. Statist. Plann. Inference* 68.1 (1998), p. 145-165.

[79] D. A. S. Fraser. « Is Bayes posterior just quick and dirty confidence ? » In : *Statist. Sci.* 26.3 (2011), p. 299-316.

[80] M. Fréchet. « Sur l'extension de certaines évaluations statistiques au cas de petits échantillons ». In : *Rev. Inst. Intern. Stat.* 11 (1943), p. 182-205.

[81] C. Gaetan et X. Guyon. *Modélisation et statistique spatiales*. T. 63. Springer-Verlag, 2008.

[82] W. A. Gale et G. Sampson. « Good-Turing smoothing without tears ». In : *Journal of Quantitative Linguistics* 2 (1995). `https://www.grsampson.net/AGtf1.html`.

[83] S. Gershman, M. Hoffman et D. Blei. *Nonparametric variational inference*. 2012. arXiv : 1206.4665 [cs.LG].

[84] 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`. 1996.

[85] 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.

[86] V. Godambe et B. Kale. « Estimating functions : an overview ». In : t. 7. Oxford Univ. Press, 1991, p. 3-20.

[87] E. Greenshtein et Y. Ritov. *Generalized maximum likelihood estimation of the mean of parameters of mixtures, with applications to sampling*. arXiv : 2107.09296. 2021.

[88] Grisvard. *Elliptic Problems in Nonsmooth Domains*. Monographs and studies in mathematics 24. Pitman Advanced Pub. Program, 1985.

[89] J. Gu et R. Koenker. « REBayes : Empirical Bayes Mixture Methods in R ». In : *Journal of Statistical Software* 82.8 (2075).

[90] J. Gu et R. Koenker. « Unobserved heterogeneity in inconme dynamics : an empirical Bayes perspective ». In : *Journal of Business & Economic Statistics* (2015).

[91] P. Hall. « Cross-validation in density estimation ». In : *Biometrika* 69.2 (1982), p. 383-390.

[92] P. Hall et S. Wilson. « Two guidelines for bootstrap hypothesis testing ». In : *Biometrics* 47.2 (1991), p. 757-762.

[93] E. Hannan. « The estimation of the order of an ARMA process ». In : *Ann. Statist.* 8.5 (1980), p. 1071-1081.

[94] 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.

[95] A. Hawkes et D. Oakes. « A cluster process representation of a self-exciting process ». In : *J. Appl. Probability* 11 (1974), p. 493-503.

[96] 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.

[97] 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.

[98] A. Hyvärinen. « Estimation of non-normalized statistical models by score matching ». In : *J. Mach. Learn. Res.* 6 (2005), p. 695-709.

[99] T. Inglot. « Inequalities for quantiles of the chi-square distribution ». In : *Probab. Math. Statist.* 30.2 (2010), p. 339-351.

[100] C. Jackson et al. « Multistate Markov models for disease progression with classification error ». In : *The Statistician* 52.2 (2003), p. 193-209.

[101] W. Jiang et C.-H. Zhang. « General maximum likelihood empirical Bayes estimation of normal means ». In : *Ann. Statist.* 37.4 (2009), p. 1647-1684.

[102] 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.

[103] A. Juditsky, P. Rigollet et A. B. Tsybakov. « Learning by mirror averaging ». In : *Ann. Statist.* 36.5 (2008), p. 2183-2206.

[104] M. Kearns et al. « An experimental and theoretical comparison of model selection methods ». In : *Machine Learning* 27 (1997).

[105] 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.

[106] N. Kiefer, T. Vogelsang et H. Bunzel. « Simple robust testing of regression hypotheses ». In : *Econometrica* 68.3 (2000), p. 695-714.

[107] J. F. C. Kingman. *Poisson processes*. The Clarendon Press, 1993.

[108] Y. Kitamura. *Empirical Likelihood Methods in Econometrics : Theory and Practice*. Rapp. tech. `cowles.yale.edu/sites/default/files/files/pub/d15/d1569.pdf`. 2006.

[109] J. Kivinen et M. K. Warmuth. « Averaging expert predictions ». In : *Computational learning theory (Nordkirchen, 1999)*. T. 1572. Lecture Notes in Comput. Sci. Springer, Berlin, 1999, p. 153-167.

[110] 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.

[111] D. Landers et L. Rogge. « Minimal sufficient σ-fields and minimal sufficient statistics. Two counterexamples ». In : *Ann. Math. Statist.* 43 (1972), p. 2045-2049.

[112] R. Latała et D. Matlak. « Royen's Proof of the Gaussian Correlation Inequality ». In : *Geometric Aspects of Functional Analysis* (2017). arXiv : 1512.08776, 265–275.

[113] L. M. Le Cam. « Maximum Likelihood : An Introduction ». In : *International Statistical Review / Revue Internationale de Statistique* 58.2 (1990), p. 153-171.

[114] 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.

[115] K. Lee et D. Stöger. *Randomly Initialized Alternating Least Squares : Fast Convergence for Matrix Sensing*. arXiv : 2204.11516. 2022.

[116] 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.

[117] G. Leung et A. R. Barron. « Information theory and mixing least-squares regressions ». In : *IEEE Trans. Inform. Theory* 52.8 (2006).

[118] M. N. M. van Lieshout. *Markov point processes and their applications*. Imperial College, 2000.

[119] 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.

[120] C. Mallows. « Some Comments on C_p ». In : *Technometrics* 15.4 (1973), p. 661-675.

[121] 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.

[122] G. McLachlan et D. Peel. *Finite mixture models*. Wiley, 2000.

[123] G. J. McLachlan et T. Krishnan. *The EM algorithm and extensions*. Wiley, 2008.

[124] N. Meinshausen et P. Bühlmann. « High-dimensional graphs and variable selection with the lasso ». In : *Ann. Statist.* 34.3 (2006), p. 1436-1462.

[125] S. Meyn et R. L. Tweedie. *Markov chains and stochastic stability*. Second. Cambridge University Press, 2009.

[126] G. Mohler et al. « Self-exciting point process modeling of crime ». In : *J. Amer. Statist. Assoc.* 106.493 (2011), p. 100-108.

[127] R. M. Neal. « Connectionist learning of belief networks ». In : *Econometrica* 56.71 (1992), p. 71-113.

[128] W. Newey. *Introduction à la théorie des bornes d'efficacité semi-paramétriques*. Annales d'économie et de statistique, 1989.

[129] J. Neyman et E. Scott. « Statistical approach to problems of cosmology ». In : *J. Roy. Statist. Soc. Ser. B* 20 (1958), p. 1-43.

[130] 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.

[131] Y. Ogata et D. Vere-Jones. « Inference for earthquake models : a self-correcting model ». In : *Stochastic Process. Appl.* 17.2 (1984), p. 337-347.

[132] 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.

[133] 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.

[134] J. W. Pratt. « Concavity of the log likelihood ». In : *J. Amer. Statist. Assoc.* 76.373 (1981), p. 103-106.

[135] 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.

[136] 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.

[137] S. Reid, R. Tibshirani et J. Friedman. « A study of error variance estimation in Lasso regression ». In : *Statist. Sinica* 26.1 (2016), p. 35-67.

[138] P. Rigollet et A. Tsybakov. « Exponential screening and optimal rates of sparse estimation ». In : *Ann. Statist.* 39.2 (2011), p. 731-771.

[139] J. Rissanen. « Fisher information and stochastic complexity ». In : *IEEE Trans. Inform. Theory* 42.1 (1996), p. 40-47.

[140] J. Rissanen. « Modeling by Shortest Data Description ». In : *Automatica* 14 (1978), p. 465-471.

[141] R. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[142] E. Ronchetti. « Robust model selection in regression ». In : *Statist. Probab. Lett.* 3.1 (1985), p. 21-23.

[143] W. Rudin. *Functional analysis*. McGraw-Hill, 1991.

[144] G. Schwarz. « Estimating the dimension of a model ». In : *Ann. Statist.* 6.2 (1978), p. 461-464.

[145] R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley, 1980.

[146] J. Shao. « Linear model selection by cross-validation ». In : *J. Amer. Statist. Assoc.* 88.422 (1993), p. 486-494.

[147] R. Shibata. « Selection of the order of an autoregressive model by Akaike's information criterion ». In : *Biometrika* 63.1 (1976), p. 117-126.

[148] 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.

[149] 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.

[150] C. Stone. « Adaptive maximum likelihood estimators of a location parameter ». In : *Ann. Statist.* 3 (1975), p. 267-284.

[151] D. Strauss et M. Ikeda. « Pseudolikelihood estimation for social networks ». In : *J. Amer. Statist. Assoc.* 85.409 (1990), p. 204-212.

[152] M.-N. Tran, T.-N. Nguyen et V.-H. Dao. *A practical tutorial on Variational Bayes*. 2021. arXiv : 2103.01327.

[153] L. Truquet. *A new smoothed QMLE for AR processes with LARCH errors*. Rapp. tech. hal-00284776. 2008.

[154] P. Vincent. « A connection between score matching and denoising autoencoders ». In : *Neural Comput.* 23.7 (2011), p. 1661-1674.

[155] A. Wald. « Note on the consistency of the maximum likelihood estimate ». In : *Ann. Math. Statistics* 20 (1949), p. 595-601.

[156] C. S. Wallace et D. M. Boulton. « An information measure for classification ». In : *Computer Journal* 11.2 (1968), p. 185-194.

[157] C. Wallace et P. Freeman. « Estimation and inference by compact coding ». In : *J. Roy. Statist. Soc. Ser. B* 49.3 (1987), p. 240-265.

[158] H. White. « Maximum likelihood estimation of misspecified models ». In : *Econometrica* 50.1 (1982), p. 1-25.

[159] J. S. White. « The limiting distribution of the serial correlation coefficient in the explosive case ». In : 29 (1958), p. 1188-1197.

[160] R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Elsevier, 2012.

[161] R. Wolpert et K. Ickstadt. « Poisson/gamma random field models for spatial statistics ». In : *Biometrika* 85.2 (1998), p. 251-267.

[162] 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.

[163] Y. Yang. « Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation ». In : *Biometrika* 92.4 (2005), p. 937-950.

[164] Y. Zhang et Y. Yang. « Cross-validation for selecting a model selection procedure ». In : *J. Econometrics* 187.1 (2015), p. 95-112.

[165] H. Zou. « The adaptive lasso and its oracle properties ». In : *J. Amer. Statist. Assoc.* 101.476 (2006), p. 1418-1429.

[166] 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), 63
agrégation
    de modèles linéaires, 96
    à poids exponentiels, 97
AIC, 74, 75
Akaike, 49
algorithme EM, 29
ancillary, *voir* libre (statistique)
ARCH, 102
ARMA, 10, 34

Barankin (borne), 84
Barankin (estimateur), 68, 84, 93
Barron-Leung (estimateur), 96
Bartlett (correction), 74
Basu (théorème), 68
bayésienne (méthode), 84
Bernoullis corrélés, 26
biais, 18, 64
BIC, 75, 131
branchement (processus), 104

Chapman-Robbins (borne), 84
complète (statistique), 65
conditionnellement binônial (modèle), 14
conjuguées (familles), 88
consistance, 18
contraintes (estimation), 52
contraste, *voir* minimum de contraste
convergence d'estimateurs
    en probabilité, 24
    normalité asymptotique, *voir* normalité asymptotique
    ps, 23, 42, 50, 52, 69
Cox (processus), 16
C_p-Mallows, 16, 44
Cramér-Rao, 78, 83
critère de sélection, 44, 75, 131, 134

delta-method, 18
discriminative vs generative (training), 6, 20

efficacité, 81
EM (algorithme), 29
empirical Bayes, 89
entropie, 23
équation d'estimation, 21, 34
estimateur
    à minimum de contraste, *voir* M-estimateurs
    Bayes empirique, 89, 93
    bayésien, 84, 90
    de Barankin, 68, 84, 93
    de James-Stein, 66, 91
    de Pitman, 88
    de Robbins-Turing, 92
    des moments, 22, 49
    M-estimateur, 19, 41
    OLS, 9, 16, 106
    par agrégation (AWE), 97
    par agrégation (Barron-Leung), 96
    par pseudo-vraisemblance, 20, 26, 29, 48
    ridge, 16, 85, 90, 91
    robuste, 20, 42
    sandwich, 73, 77
    sous contraintes, 52
exhaustivité, 64
expérience R-régulière, 71
expérience régulière, 119

familles exponentielles, 7, 65, 83
fonction d'estimation, 21, 34
fusion, 38, 40

generative vs discriminative (training), 6, 20
Gibbs (processus), 15
Godambe (matrice d'information), 36

Hajek (théorème de convolution), 82, 110
Hawkes (processus), 12
Hellinger (distance), 76
Hodges (estimateur), 82
Huber (estimateur), 19, 42
hypothèses
    EC, 53
    FE, 36
    GNU, 31
    GNU', 35

LAN, 109, 115, 120
ME, 45
ULAN, 113, 115

impropre (distribution), 86, 88
inférence variationnelle, 94
information
matrice, 70, 71
monotonicité, 78
inégalité de Cauchy-Schwarz matricielle, 79

jackknife, 30
James-Stein (estimateur), 66, 91

Kronecker (lemme), 107
Kullback-Leibler (divergence), 69

LAN, 109, 115, 120
LARCH, 48
lasso, 17, 67, 90
Le Cam (estimateur efficace), 83, 113
Le Cam (paradoxe), 6
Lehmann-Scheffé, 64, 65
libre (statistique), 68, 78
log-concave, 28
loi des grands nombres uniforme, 25, 32

M-estimateur, 19, 41
Mallows, 16, 44
MAP, 90
martingales, 129
maximum d'entropie, 23
maximum de vraisemblance, 69, 101
avec normalisation inconnue, 17
conditionnelle, 20, 27, 29, 77
dans les familles exponentielles, 8
du maximum de vraisemblance, 26
empirique, 41
et exhaustivité, 64
mis en défaut, 7, 77, 85, 90
normalité asymptotique, 72
par algorithme EM, 29
partielle, 20, 27–29, 77
MDL, 131
mélange de lois, 8
mélange d'experts, 99
minimale (statistique), 68
minimax (estimateur), 63, 92
minimum de χ^2, 49
minimum de contraste, 26, 27, 41
MML, 131
moindres carrés conditionnels, 9
moments (méthode), 22, 49

Neyman-Fisher, 64
Neyman-Scott, 15, 16
non-biaisé, 18, 64
non-paramétrique, 6
normal means, 67, 91
normalité asymptotique, 36, 38, 45, 50, 53, 73, 103, 113

OLS (estimateur), 9, 16, 106
oracle, 96, 97
overfitting, 44

paramétrique, 6
pénalisation, 16, 44, 67, 75, 131, 134
Pitman (estimateur), 88
Poisson (processus), 11, 15–17, 28
ponctuel (processus), 15, 16, 28, 34
Prekopa-Leindler (théorème), 28
pseudo-vraisemblance, 20, 26–29, 48, 77

quasi-vraisemblance, 48

R-régulière (expérience), 71, 77
Rao-Blackwell, 64, 66
Rao-Blackwellisation, 65, 66
région de confiance, 36, 74
régression, 9
catégorielle ordonnée, 27
linéaire, 9, 16, 96, 106
logistique, 9
non-linéaire, 41, 43, 56
régulière (expérience), 71, 77
ridge (estimateur), 16, 85, 90, 91
risque, 63
Robbins (estimateur), 92
robuste (estimateur), 20, 42

sandwich (estimateur), 73, 77
score matching, 48
semi-paramétrique, 6
sensibilité (matrice), 36, 45, 50
SIC, 133, 134
Sobolev (inégalité), 25, 38, 60, 125
stabilisation de variance, 30
Stein (estimateur SURE), 16, 67, 91
Stein (estimateur), 66, 91
Strauss (processus), 15, 17, 22
sufficient, *voir* exhaustivité
surajustement, 44, 96
SURE, 16, 67, 91

Takeuchi (critère), 44, 49, 56, 74
test
d'Hausman, 57, 99
de Wald, 57, 99
des scores, 57, 99
du multiplicateur de Lagrange, 57
du rapport de vraisemblance, 74, 99
théorème-limite central, *voir* normalité asymptotique

TIC, 44, 49, 56, 74
totale (statistique), 65
Turing, 92
Tweedie (formule), 90

ULAN, 113, 115
UMR, 64
UMVU, 65

validation croisée, 27, 75
    mise en échec, 91
    vs SURE, 91
variationnelle (inférence), 94
vraisemblance
    et conditionnement, 78
    non normalisée, 17, 48
    partielle, 20, 27–29, 77

Wilks, 55, 73

Z-estimateur, 21, 34