OLIVIER RIOUL

MODÈLES PROBABILISTES ET STATISTIQUES

ENSTA STA 221

2021

Table des matières

Liste des notations

$\mathcal{B}(p)$	loi de Bernoulli de paramètre $p \in [0,1]$
$\mathcal{B}(n,p)$	loi binomiale de longueur n et de paramètre $p \in [0,1]$
$\mathrm{B}(\alpha,\beta)$	loi bêta de paramètres $\alpha > 0$ et $\beta > 0$.
$B(\alpha,\beta)$	fonction bêta $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
$\mathbb{B}(\widehat{\theta})$	biais
$D(p\|\|q)$	divergence de Kullback-Leibler
$\mathbb{E}(\cdot)$	espérance
$\mathcal{E}(\lambda)$	loi exponentielle de paramètre $\lambda > 0$ (loi gamma $\Gamma(1,\lambda)$)
$\Gamma(\alpha)$	fonction Gamma
$\Gamma(\alpha,\beta)$	loi gamma de paramètres $\alpha > 0$ et $\beta > 0$.
$\Gamma^{-1}(\alpha,\beta)$	loi inverse gamma de paramètres $\alpha > 0$ et $\beta > 0$.
J_θ	information de Fisher
$\mathcal{N}(\mu,\sigma^2)$	loi normale (gaussienne) de moyenne μ et variance σ^2
$\mathcal{N}(\boldsymbol{\mu},\mathbf{C})$	loi normale de moyenne $\boldsymbol{\mu}$ et de matrice de covariance $\mathbf{C}$
$\mathcal{P}(\lambda)$	loi de Poisson de paramètre $\lambda > 0$
$\mathbb{P}(\cdot)$	(loi de) probabilité
$R(\widehat{\theta})$	risque
$S_\theta(X)$	score
θ	paramètre
$\widehat{\theta}$	estimateur
$\mathcal{U}[a,b]$	loi uniforme sur l'intervalle $[a,b]$
$\mathbb{V}(\widehat{\theta})$	variance

$\lvert \cdot \rvert$	norme euclidienne sur $\mathbb{R}^n$
$\lVert \cdot \rVert$	norme euclidienne sur $\mathbb{R}^N$
$X_N \xrightarrow{\mathcal{L}} X$	convergence en loi (en distribution)
$X_N \xrightarrow{\mathbb{P}} X$	convergence en probabilité
$X_N \to X$ *p.s.*	convergence presque sûre

Liste des exercices

Table des figures

Introduction

Ce cours intitulé « Modèles probabilistes & statistiques » comprend, comme on peut s'y attendre, deux parties :

1. **Modèles statistiques** : essentiellement le problème de l'*estimation* statistique
2. **Modèles probabilistes** : essentiellement les *chaînes de Markov*.

Il est parfois délicat de distinguer clairement ce qui est *probabiliste* de ce qui est *statistique* :

« *probabiliste* » se réfère à l'étude théorique des *lois* de probabilité, fondée sur la théorie de la mesure, mais pas seulement [1] : elle développe les notions d'indépendance, de conditionnellement, de convergence…

« *statistique* » se réfère à l'étude pratique de données réelles (à l'aide des probabilités) : elle développe ses propres méthodes : estimation, confiance, tests…

Un autre terme, plus recherché, est peut être moins bien compris :

« *stochastique* » se réfère initialement à l'art de conjecturer [2] ; au 19e siècle on l'utilise pour l'application de la théorie de probabilités au monde réel (donc assez proche du sens de « statistique »). Mais plus récemment, ce terme a changé de sens sous l'influence de l'école russe [3] : il est devenu simplement un vague synonyme d'« aléatoire ». (Son intérêt est sans doute son caractère international car on utilise le même mot dans tous les langues.) Ainsi « processus stochastique » signifie simplement « processus aléatoire », etc.

1. Selon une remarque attribuée à Mark Kac : « La théorie des probabilités est une théorie de la mesure *avec une âme* »

2. Jacob Bernoulli, *Ars Conjectandi sive Stochastice*, 1713.

3. Khinchine et Kolmogorov dans les années 1930 : sous Staline, toute application de la science au monde réel qui n'était pas strictement déterministe était suspect.

Plan du cours

I. Estimation statistique

La marge de droite est laissée à l'étudiant pour prendre des notes.

- Généralités sur les modèles probabilistes
  - estimation, intervalles de confiance, tests d'hypothèse
  - fréquentiste vs. bayesien.
- MSE [4]=erreur quadratique moyenne, variance, biais
- estimateur MVU [5] (non biaisé de variance minimale)
- score, information de Fisher, borne de Cramér-Rao, estimateur (asymptotiquement) efficace
- Estimateurs :
  - ML [6]=maximum de vraisemblance, propriétés asymptotiques, d'invariance
  - BLUE [7]=meilleur linéaire non biaisé
  - LSE [8]=aux moindres carrés
- Modème bayesien, a priori et a posteriori
- Risque quadratique, estimateur MMSE [9]=erreur quadratique moyenne minimale
- Risque 0-1 (succès-échec), estimateur MAP [10] =maximum a posteriori et estimateur ML
- Estimateurs LMMSE=linéaires MMSE, filtrage de Wiener

4. *mean-squared error*
5. *minimum variance unbiased*
6. *maximum likelihood*
7. *best linear unbiased estimator*
8. *least squares estimator*
9. *minimum mean-squared error*
10. *maximum a posteriori probability*

II. Chaînes de Markov

- Modèles markoviens
  - Généralités
  - Homogénéité
  - Finitude

- Machines à états, matrice de transition, graphes de transition
- États transitoires, récurrents
- Irréductibilité, loi stationnaire (invariante), théorème de Perron-Frobenius
- Périodicité, convergence dans le cas irréductible apériodique
- Temps de retour, temps de passages, nombre de visites
- Propriété de Markov forte, théorème ergodique

III. . . et enfin, si le temps le permet, une application qui combine les deux parties **I** et **II** :

- Échantillonnage de Gibbs [11]

11. Geman/Geman, 1984.

. . . le tout saupoudré de nombreux **exercices** au fur et à mesure du cours. Chaque exercice s'énonce sur une page recto ; le verso contient quelques **indications** pour sa résolution [12].

12. Non, non, pas de corrigé rédigé tout fait ! Les corrections ne se feront le cas échéant qu'en cours, au tableau.

Première partie

Estimation statistique

1
Modèle statistique

Qu'est ce qu'un modèle statistique ? C'est d'abord :

• *des « données »* (des « observations », des « mesures ») que l'on modélise comme une **suite de** N **variables aléatoires** (réelles, voire complexes ou même vectorielles) :

$$X = (X_1, X_2, \dots, X_N)$$

regroupés dans un vecteur aléatoire X. Souvent (mais pas toujours) X est un « N-échantillon », constitué de N variables aléatoires i.i.d. [1]

On distinguera les v.a. $X = (X_1, X_2, \dots, X_N)$ de leurs réalisations $x = (x_1, x_2, \dots, x_N)$.

1. indépendantes et identiquement distribuées.

C'est ensuite :

• *une loi* $\mathbb{P}_\theta$ que suivent les données X, où θ est un **paramètre** inconnu (ou regroupe plusieurs paramètres inconnus), à valeurs dans un certain ensemble Θ. On a donc une famille (souvent infinie) de lois de probabilité

$$\Big\{\mathbb{P}_\theta \ ; \ \theta \in \Theta\Big\}$$

et les données observées dépendent des paramètres inconnus θ par l'intermédiaire de leur loi : on peut écrire

$$X \sim \mathbb{P}_\theta$$

Dans le cas i.i.d., on écrira que chaque $X_i \sim \mathbb{P}_\theta$ (loi indépendante de i puisque les lois des X_i sont identiques). On peut alors se permettre de laisser tomber l'indice i et de noter X n'importe quel X_i : attention alors à ne pas confondre ce X et le vecteur $\underline{X} = (X_1, X_2, \dots, X_N)$!

Paramétrique ou non paramétrique ?

Si $\theta \in \mathbb{R}$, il n'y a qu'un seul paramètre (cas *scalaire*). Si $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n$, il y a n paramètres en tout (cas *vectoriel*). Dans tous ces cas on dit que le modèle est **paramétrique** (car il n'y a qu'un nombre fini de paramètres inconnus).

Si par contre, θ est une fonction continue, par exemple, c'est plus compliqué car θ ne peut pas être décrit simplement par la donnée d'un nombre fini de réels. On dit alors que le modèle est « **non-paramétrique** ».

Un modèle statistique, c'est enfin :

• *une problématique* (ce qu'on cherche à faire) : typiquement, on peut vouloir :

– **estimer** θ à partir des données, à l'aire d'un estimateur $\widehat{\theta}(X)$ — qui doit être « proche » de θ selon un certain critère moyen (on minimise un « risque »). On peut vouloir à la place estimer $\alpha = g(\theta)$, une fonction donnée de θ.

– **trouver un intervalle** (ou une région) **de confiance** sur θ à partir des données, disons un intervalle $I(X)$ où la probabilité que θ tombe dans l'intervalle est suffisamment grande : $\mathbb{P}\{\theta \in I(X)\} \geqslant 1 - \varepsilon$ (confiance de « niveau $1 - \varepsilon$ »).

– **formuler un test** à partir des données de sorte à comparer une hypothèse nulle $H_0 : \theta \in \Theta_0$ contre une autre hypothèse $H_1 : \theta \in \Theta_1$ à l'aide d'un test $T(X) \in \{0, 1\}$, tel que $\mathbb{P}_\theta(T(X) = 1) \leqslant \varepsilon$ (niveau du test).

RÉSUMÉ: un modèle statistique, c'est :

- des données $X = (X_1, X_2, \dots, X_N)$
- qui suivent une loi $\mathbb{P}_\theta$
- avec une problématique sur θ (par exemple l'estimer).

Le modèle est :

- *paramétrique* si le nombre de paramètres est fini
- *non-paramétrique* sinon.

Ne pas confondre :

$$\begin{cases} N & \text{nombre d'observations} \\ n & \text{nombre de paramètres} \end{cases}$$

Exemple: un sondage

Cet exemple montre que plusieurs problématiques (estimation, intervalle de confiance, test d'hypothèses) peuvent se poser naturellement pour une même observation $X \sim \mathbb{P}_\theta$.

Un vote a lieu entre deux candidats A et B. Le sondage fournit un N-échantillon $X = (X_1, X_2, \dots, X_N)$ (i.i.d.) où

$$X_i = \begin{cases} 0 & \text{si } i \text{ vote pour } A, \\ 1 & \text{si } i \text{ vote pour } B \end{cases}$$

La loi $\mathbb{P}_\theta$ est celle de Bernoulli $\mathcal{B}(\theta)$ de paramètre $\theta \in [0,1]$. Ainsi $\mathbb{P}_\theta(X_i = 1) = \theta$, $\mathbb{P}_\theta(X_i = 1) = 1 - \theta$.

1. Le modèle est-il paramétrique ?
2. Que signifient les quantités $N_B = \sum_{i=1}^N X_i$ et $N_A = N - N_B$?
3. Quelle est la loi conjointe de $X = (X_1, X_2, \dots, X_N)$?
4. Donner un **estimateur** $\widehat{\theta}(X)$ de θ que vous trouvez le plus naturel.

On cherche à déterminer si B est vainqueur.

5. Quelle serait la forme d'un **intervalle de confiance** qui permettrait de conclure avec une certaine fiabilité ?
6. J'affirme que B a gagné si $\frac{N_B}{N} > \frac{1}{2}$. Formuler le **test d'hypothèses** correspondant.

Dans la suite, on ne s'intéressera qu'au problème de l'estimation.

Exemple: un sondage

Indications:

1. Ici $\theta \in [0,1]$ est l'unique paramètre.
2. $\sum_{i=1}^{N} X_i$ compte le nombre de votes $X_i = 1$.
3. En utilisant l'indépendance des X_i, l'écrire en fonction de θ, N_A et N_B.
4. L'estimateur « de la moyenne »...
5. J'aurais confiance en le fait que B gagne si la probabilité que le sondage le donne nettement vainqueur ($\frac{N_B}{N} > \frac{1}{2} + \varepsilon$) est assez grande.
   (Ici $\varepsilon > 0$ dépend des données !)
6. On peut prendre par exemple comme hypothèse nulle : $\theta < 1/2$.

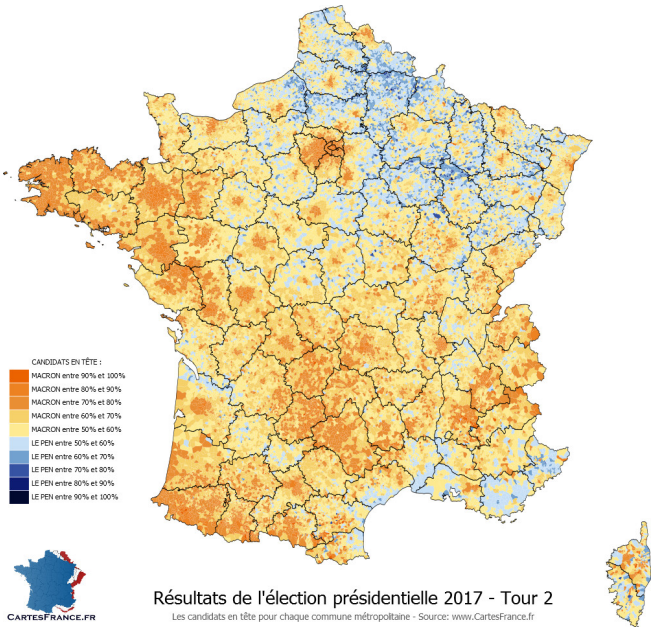


FIGURE 1.1: Sondage par communes lors du deuxième tour d'une élection présidentielle

Exemple: le problème de la régression

On observe un N-échantillon de variables réelles « explicatives » X_i et « expliquées » Y_i selon le modèle

$$Y_i = \theta(X_i) + W_i \qquad (i = 1, \dots, N)$$

où $y = \theta(x)$ est une fonction inconnue et W_i sont des échantillons i.i.d. de bruit faible dont la loi est donnée.

Attention, l'observation est ici (X, Y), vue comme le vecteur de composantes i.i.d. $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

1. Dessiner dans le plan (x, y) un nuage de points correspondant aux observations. Le modèle est-il paramétrique ?

On suppose maintenant que la fonction $\theta(x)$ est linéaire[2] : $\theta(x) = ax + b$.

2. Et maintenant, le modèle est-il paramétrique ?

On note $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ et de même $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

3. Donner un estimateur de b « naturel » en fonction de $\overline{X}$, $\overline{Y}$ et d'un estimateur de a.
4. (Plus difficile) Donner un estimateur de a qui vous semble pas trop mal.

2. Selon la tradition française on devrait dire « fonction affine »

Dans la suite, on ne s'intéressera qu'au problème de l'estimation paramétrique.

Exemple: le problème de la régression

Indications:

1. On peut par exemple imposer que θ soit une fonction régulière (continue...).

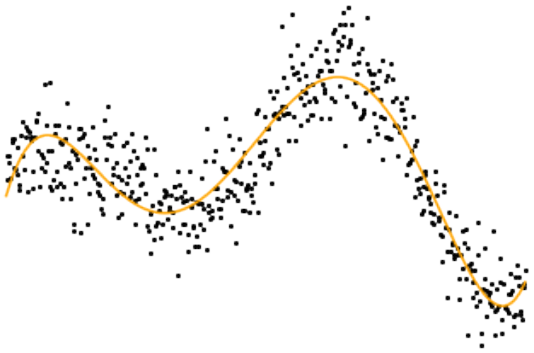


FIGURE 1.2: Régression non linéaire

2. Ici θ est uniquement déterminé par la pente a et l'interception b.

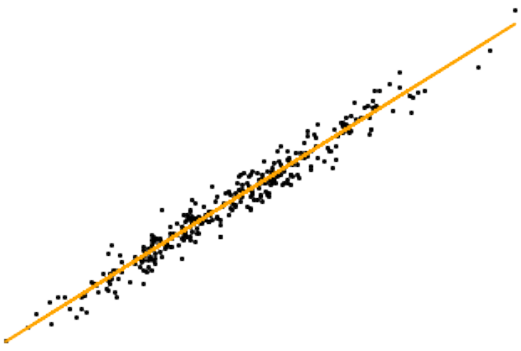


FIGURE 1.3: Régression linéaire

3. On supposera que les W_i sont de moyenne nulle dans le modèle $Y_i = aX_i + b + W_i$.
4. Voir le cours de probabilités (droite de régression linéaire aux moindres carrés).

Exemple: signal dans du bruit

On considère le modèle linéaire suivant :

$$X = \theta\, S + W$$

où $S = (S_1, \ldots, S_N)$ est un signal connu, et $W = (W_1, \ldots, W_N)$ est un bruit gaussien centré. Ici $\theta \in \mathbb{R}$ représente l'*amplitude* du signal et X dépend donc linéairement de θ.

Noter que le modèle peut s'écrire

$$X_i = \theta\, S_i + W_i \qquad (i = 1, \ldots, N)$$

ou encore en notation scalaire[3] : $X = \theta S + W$ en laissant tomber l'indice i.

1. Donner la loi $\mathbb{P}_\theta$ lorsque le bruit W est *blanc* de variance σ^2.
2. Donner la loi $\mathbb{P}_\theta$ dans le cas général où W est corrélé : $W \sim \mathcal{N}(0, \mathbf{C})$ de matrice de covariance $\mathbf{C}$ (qu'on supposera inversible).
3. Interpréter le cas particulier du modèle $X = \theta + W$ (comment interpréter cette équation ?)
4. Donner alors un estimateur d'amplitude.

3. L'équation $X = \theta\, S + W$ peut ainsi se lire indifféremment en vectoriel ou en scalaire.

Exemple: signal dans du bruit

Indications:

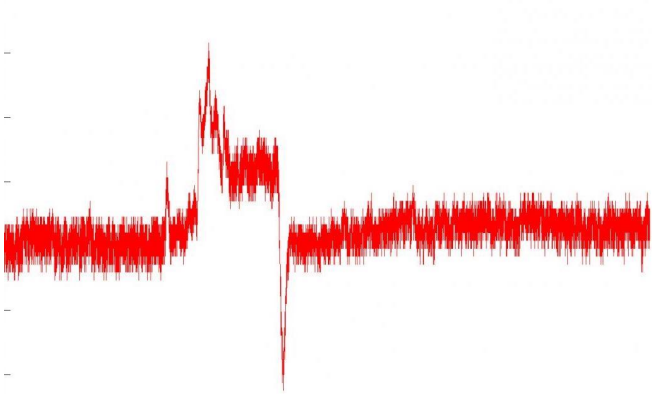


FIGURE 1.4: Signal dans du bruit gaussien

1. Dire qu'un bruit gaussien W est *blanc* revient à dire que les W_i sont i.i.d.$\sim \mathcal{N}(0, \sigma^2)$. On peut écrire la densité $p_\theta(x)$ en fonction de la norme euclidienne[4] $\|x - \theta S\|$.
2. On rappelle (voir cours de probabilités) que $W \sim \mathcal{N}(0, \mathbf{C})$ a pour densité

$$p(w) = \frac{e^{-\frac{1}{2} w^t \mathbf{C}^{-1} w}}{\sqrt{(2\pi)^N |\mathbf{C}|}}.$$

   où w est vu comme un vecteur colonne et $|\mathbf{C}|$ est le déterminant de la matrice $\mathbf{C}$.
3. Ici $X = \theta + W$ doit être interprété en notation scalaire.
4. Un estimateur $\widehat{\theta}(X)$ de la moyenne (le bruit étant de moyenne nulle).

4. on notera indifféremment $\|x\|$ ou $|x|$ quand x est un vecteur

2
Conventions sur le modèle statistique

On fait un certain nombre d'hypothèses toujours sous-entendues dans ce qui va suivre. Ainsi, pour simplifier, on supposera désormais que le modèle statistique est :

• *« identifiable » :* cela signifie que $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ implique $\theta = \theta'$. Autrement dit, la loi paramétrée par θ caractérise $\theta \in \Theta$ de manière unique. Sans ça, il n'y a clairement aucun espoir.

C'est pratiquement toujours vérifié, à moins de faire un modèle stupide. Par exemple, un modèle $\mathbb{P}_\theta$ de variance θ^2 paramétrée par $\theta \in \mathbb{R}$ n'est pas identifiable, puisque $\pm\theta$ correspondent à la même variance. Mais c'est ridicule de vouloir estimer θ positif ou négatif : il suffit de poser que θ est positif ($\Theta = \mathbb{R}_+$) pour rendre le modèle identifiable.

• *« canonique » :* cela signifie que $\mathbb{P}_\theta$ est défini directement sur l'espace des observations. En effet on pourrait décrire $X \in \mathbb{R}^N$ de deux manières :

1. comme une fonction de $(\Omega, \mathfrak{A}, \mathbb{P})$ vers l'espace des valeurs $\mathbb{R}^N$, c'est la définition théorique qui impose de préciser l'univers Ω
2. directement par une distribution de probabilité $\mathbb{P}_\theta$ sur $\big(\mathbb{R}^N, \mathfrak{B}(\mathbb{R}^N)\big)$ où $\mathfrak{B}(\mathbb{R}^N)$ désigne la tribu des boréliens de $\mathbb{R}^N$.

C'est la deuxième option qu'on choisit : le modèle canonique décrit $\mathbb{P}_\theta$ directement sur $\big(\mathbb{R}^N, \mathfrak{B}(\mathbb{R}^N)\big)$, on s'affranchit complètement de la considération du mystérieux univers Ω !

- « *dominé* » par une « mesure dominante » μ (σ-finie). Cela signifie que :

$$\forall\, \theta \in \Theta, \quad \mathbb{P}_\theta \ll \mu$$

où le symbole « $\ll$ » signifie « *est absolument continue par rapport à* ». Autrement dit $\mu(A) = 0 \implies \mathbb{P}_\theta(A) = 0$ pour tout borélien $A \in \mathfrak{B}(\mathbb{R}^N)$.

Bien entendu μ ne dépend pas elle-même de θ ![1] Elle doit dominer toute loi $\mathbb{P}_\theta$ pour toute valeur du paramètre θ.

1. Ce serait trop facile : $\mathbb{P}_\theta \ll \mathbb{P}_\theta$.

D'après le théorème de Radon-Nikodym (voir cours de probabilités) dire que $\mathbb{P}_\theta \ll \mu$ revient à dire que $\mathbb{P}_\theta$ admet une *densité* p_θ par rapport à μ :

$$\mathbb{P}_\theta\{X \in A\} = \int_A p_\theta(x)\, \mathrm{d}\mu(x)$$

Autrement dit, pour spécifier le modèle, il suffit de spécifiée la densité de probabilité $p_\theta(x)$ sur $\mathbb{R}^N$ par rapport à la mesure dominante choisie μ. On écrira simplement

$$X \sim p_\theta$$

Exemples de domination

1. μ = mesure de Lebesgue : $p_\theta(x)$ est alors une densité de probabilité au sens usuel (p.d.f.)[2].
   Par exemple pour un seul ($N = 1$) échantillon gaussien $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)$ paramétré par sa moyenne,

$$p_\theta(x) = \frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

   La mesure de Lebesgue est souvent notée $\mu = \mathrm{d}x$.
2. μ = mesure de comptage (sur un certain ensemble discret $\mathcal{X}$) : $p_\theta(x)$ est alors une distribution de probabilité discrète (p.m.f.)[3] définie pour $x \in \mathcal{X}$, de sorte que

$$p_\theta(x) = \mathbb{P}_\theta(X = x)$$

   La mesure de comptage sur $\mathcal{X}$ est $\mu = \sum_{x \in \mathcal{X}} \delta_x$ où δ_x est la mesure (masse) de Dirac en x.
3. On peut imaginer des cas plus généraux, qui ne sont ni discrets, ni continus, mais pas exemple un mélange des deux[4].

2. *probability density function*

3. *probability mass function*

4. voir exercice sur les modèles dominés

RÉSUMÉ: un modèle statistique (identifiable, canonique, dominé) pour l'estimation :

- des données $X = (X_1, X_2, \ldots, X_N) \in \mathbb{R}^N$
- qui suivent une *densité* $p_\theta(x)$ sur $\mathbb{R}^N$ par rapport à une mesure μ :

$$\begin{cases} \text{cas discret :} & \text{mesure de comptage} \\ \text{cas continu :} & \text{mesure de Lebesgue} \end{cases}$$

- la densité p_θ caractérise θ
- on veut estimer θ à l'aide d'un estimateur $\widehat{\theta}(X)$

Exercice: exemples de modèles dominés

Il est important de bien comprendre la notion fondamentale de modèle définie par une densité (par rapport à une mesure dominante μ), qu'on soit dans le cas discret ou dans le cas continu.

Identifier une mesure dominante μ et préciser la densité correspondante $p_\theta(x)$ dans les cas suivants :

On prend d'abord pour simplifier $N = 1$. On donnera une *formule simple* de $p_\theta(x)$ comme une fonction de x.

1. $\mathbb{P}_\theta = \mathcal{B}(\theta)$ (Bernoulli de paramètre θ)
2. $\mathbb{P}_\theta = \mathcal{B}(N, \theta)$ (binomiale de longueur N et paramètre θ)

   Cet exemple correspond à la loi de $\sum_i X_i$ pour un N-échantillon i.i.d. X de composantes $\sim \mathcal{B}(\theta)$.
3. $X \sim \mathbb{P}_\theta = \mathcal{U}[-\theta, \theta]$ (uniforme sur l'intervalle $[-\theta, \theta]$)
4. $X = U^+$ où $U \sim \mathcal{U}[-\theta, \theta]$
5. Un *contre-exemple* : $X = \theta \in \mathbb{R}$. Qu'en pensez-vous ?

Dans le cas général :

6. Si $\mathbb{P}_\theta \ll \mu$ et $\mathbb{P}'_\theta \ll \mu'$, la loi $\lambda\mathbb{P}_\theta + (1-\lambda)\mathbb{P}'_\theta$ (où $0 < \lambda < 1$) est-elle dominée et si oui par quelle mesure ?
7. Dans le cas d'un N-échantillon i.i.d. X, préciser sa densité en fonction de la densité p_θ de chaque observation X_i.
   (a) Application au cas $\mathbb{P}_\theta = \mathcal{B}(\theta)$ (exemple du sondage)
   (b) Application au cas $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)$ (exemple d'une amplitude dans du bruit)

Exercice: exemples de modèles dominés

Indications:

1. Penser à la formule $\theta^x(1-\theta)^{1-x}$. Il y a plusieurs choix pour μ : $\delta_0 + \delta_1$, $\sum_{n\in\mathbb{N}} \delta_n$, $\sum_{n\in\mathbb{Z}} \delta_n$, ...
2. Penser au coefficient binomial[5] $\binom{N}{x}$. On peut choisir μ indépendant de N.

5. Prononcer en anglais « *N choose x* » ou en français « x parmi N »

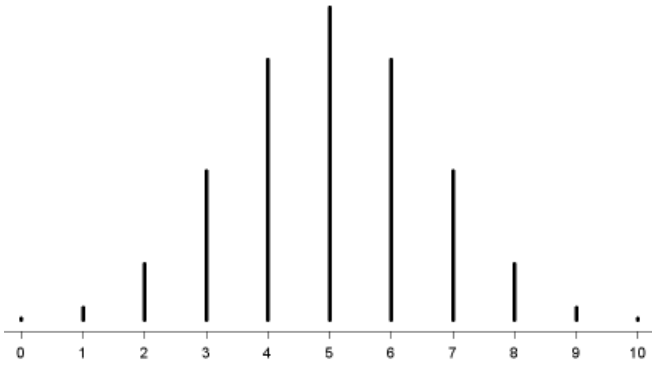


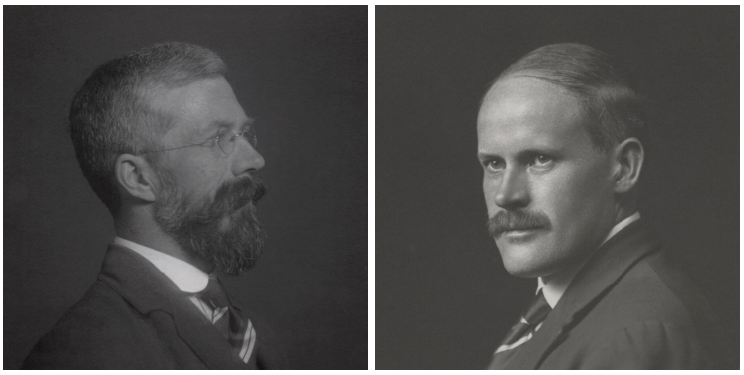
FIGURE 2.1: Une densité binomiale d'ordre 10

3. Une densité usuelle (p.d.f.) constante.
4. On rappelle de $U^+ = \max(U, 0)$. C'est un cas *mixte discret-continu* : un mélange d'un Dirac en 0 et d'une densité sur $\mathbb{R}_+$.
5. $\mathbb{P}_\theta = \delta_\theta$ (Dirac) et $\delta_\theta \ll \mu$ implique $\mu(\{\theta\}) > 0$ pour tout $\theta \in \mathbb{R}$. Mais (cf. cours de probabilités) une mesure ne peut avoir qu'un nombre dénombrable de masses.

   C'est un cas « dégénéré » : $X = \theta$, il n'y a rien à faire !
6. $\mu + \mu'$. En mélangeant des lois on peut toujours trouver une mesure dominante.
7. $p_\theta(\underline{x}) = \prod_{i=1}^N p_\theta(x_i)$ pour $\underline{x} = (x_1, x_2, \ldots, x_N)$

3
Fréquentiste vs. bayesien : deux approches opposées

Le développement moderne des statistiques a été initié dans les années 1920-1930 par Fisher, qui a rapidement imposé une vision dite *classique* ou « *fréquentiste* ». À peu près à la même époque, Jeffreys défend une approche dite « *bayesienne* » dont certains concepts remontent au 18e siècle (Bayes et Laplace).



(a) Ronald Aylmer Fisher en 1931 (b) Harold Jeffreys en 1926

FIGURE 3.1: Les fondateurs des écoles fréquentiste et bayesienne.

Les deux écoles fréquentiste et bayesienne se sont rudement critiquées, chacune reprochant à l'autre d'être mathématiquement infondée. Encore aujourd'hui, le débat fait rage, et de nombreux arguments ont été développés pour démontrer la supériorité de l'une par rapport à l'autre.

Les deux approches sont effectivement très différentes :

- *L'approche fréquentiste* (ou classique) :

  θ est *inconnu*, mais *déterministe*.

C'est l'approche vue jusqu'ici où $\mathbb{P}_\theta$ est paramétrée par le **paramètre** θ (qui est **fixé**). Ici

$$p_\theta(x) \text{ est souvent notée } p(x;\theta)$$

Dans cette approche le nombre N d'observations est typiquement très grand (d'où le terme « fréquentiste » : on détermine la loi à partir des données).

- *L'approche bayesienne* (utilise l'inférence bayesienne) :

  θ est « *connu a priori* », mais *aléatoire*.

Plus précisément on connait (ou on choisit) sa loi *a priori* $p(\theta)$, de sorte que θ est considérée en fait comme une **variable aléatoire**. Dans ce cas

$$p_\theta(x) \text{ est une densité conditionnelle } p(x|\theta)$$

et naturellement la loi conjointe de X et θ est

$$p(x,\theta) = p(\theta)p(x|\theta)$$

On peut alors trouver la loi *a posteriori* $p(\theta|x)$ de θ (une fois l'observation faite) en appliquant la formule de Bayes [1]. C'est ce qu'on appelle l'*inférence bayesienne*.

1. [Voir plus loin p. 111]

Dans cette approche le nombre N d'observations peut être relativement faible.

On se limite dans un premier temps à l'estimation classique (fréquentiste) avant de passer à l'estimation bayesienne à partir du chapitre 12

RÉSUMÉ: Deux approches :
L'approche fréquentiste où

$$\theta \text{ est } \begin{cases} \text{déterministe (paramètre fixé)} \\ \text{inconnu} \end{cases}$$

L'approche bayesienne où

$$\theta \text{ est } \begin{cases} \text{aléatoire} \\ \text{connu (par sa loi a priori)} \end{cases}$$

Ne pas confondre :

$$\begin{cases} p(x,\theta) & \text{densité conjointe (cas bayesien)} \\ p(x;\theta) & \text{densité paramétrée (cas fréquentiste)} \\ p(x|\theta) & \text{densité conditionnelle (cas bayesien)} \end{cases}$$

4
Estimation paramétrique et risque

Un **estimateur** $\widehat{\theta}$ de $\theta \in \mathbb{R}^n$ est n'importe quelle fonction des données

$$\widehat{\theta}(X) = \widehat{\theta}(X_1, X_2, \ldots, X_N)$$

à valeurs dans $\mathbb{R}^n$ qui va servir à estimer θ. On l'appelle parfois estimateur **ponctuel** car (étant donnée une observation des données) il ne fournit qu'une valeur dans $\mathbb{R}^n$ (un point), et non une plage de valeurs comme pour un intervalle ou une région de confiance.

N'importe quelle fonction de X (de $\mathbb{R}^N$ dans $\mathbb{R}^n$) définit un estimateur ! Par exemple, on pourrait prendre $\widehat{\theta}(X) = 0$ ou $=$ constante, même si un tel estimateur risque fort de ne pas être très bon...

Cependant, $\widehat{\theta}(X)$ doit être seulement une fonction des données X, elle ne doit pas dépendre elle-même de θ qui est inconnu et qu'on cherche précisément à estimer ![1]

1. Ce serait tricher : il suffirait de prendre comme estimateur la fonction constante $\widehat{\theta}(X) = \theta$

Puisqu'un estimateur peut être quelconque, il faut un critère pour savoir s'il est bon ou mauvais :

Risque

Le risque $R(\widehat{\theta})$ est une fonction de coût moyen entre $\widehat{\theta} = \widehat{\theta}(X)$ et θ, qu'on cherche à rendre le plus petit possible. « Moyen » signifie qu'on prend l'**espérance** sur la loi de $X \sim \mathbb{P}_\theta$ du coût choisi. Ainsi $R(\theta)$ est une quantité moyenne, déterministe, qui ne dépend que de θ.

L'espérance est notée $\mathbb{E}$ ou parfois $\mathbb{E}_\theta$ pour indiquer qu'elle dépend elle-même du paramètre θ (puisque qu'elle porte sur la loi $\mathbb{P}_\theta$ de X).

Erreur quadratique moyenne (MSE)

Le critère le plus souvent utilisé est le **risque quadratique**, aussi appelé **erreur quadratique moyenne** ou MSE[2] :

2. *mean-squared error*

$$\boxed{R_\theta(\widehat{\theta}) = \mathrm{MSE}_\theta = \mathbb{E}\{|\widehat{\theta} - \theta|^2\}}$$

La norme $|\cdot|$ (aussi notée $\|\cdot\|$) est ici la norme euclidienne sur $\mathbb{R}^n$. Sous forme développée cela s'écrit :

$$\mathrm{MSE} = \mathbb{E}_\theta\{|\widehat{\theta}(X) - \theta|^2\} = \int_{\mathbb{R}^N} |\widehat{\theta}(x) - \theta|^2 p_\theta(x)\,\mathrm{d}\mu(x)$$

Dans le cas habituel d'un seul paramètre ($n = 1$) on a simplement $\mathrm{MSE} = \mathbb{E}\{(\widehat{\theta} - \theta)^2\}$. Il existe aussi d'autres fonctions de coût, par exemple pour d'autres normes (autre qu'euclidienne) sur $\mathbb{R}^n$.

Estimation vs. détection

Lorsque θ prend des valeurs discrètes (typiquement un ensemble fini de valeurs) on peut utiliser le critère de **probabilité d'erreur** :

$$R_\theta(\widehat{\theta}) = \mathbb{P}_e(\widehat{\theta}) = \mathbb{P}\{\widehat{\theta} \neq \theta\}$$

qui est bien une fonction de coût moyen : $\mathbb{P}_e(\widehat{\theta}) = \mathbb{P}_\theta\{\widehat{\theta}(X) \neq \theta\} = \mathbb{E}_\theta\{1_{\widehat{\theta}(X)\neq\theta}\}$ où le coût $1_{\widehat{\theta}\neq\theta}$ vaut 1 si $\widehat{\theta}(X) \neq \theta$ et 0 sinon. Dans ce cas très particulier on parle **détection** plutôt que d'estimation.

Estimation optimale ?

Le choix du risque R_θ sert à deux choses.

- Tout d'abord, si on construit un estimateur $\widehat{\theta}(X)$, on peut calculer le risque correspondant pour mesurer si cet estimateur est bon ou mauvais. On peut ainsi comparer deux estimateurs $\widehat{\theta}_1$ et $\widehat{\theta}_2$: si $R_\theta(\widehat{\theta}_1) \leqslant R_\theta(\widehat{\theta}_2)$ (pour **tout** $\theta \in \Theta$) alors on peut dire que $\widehat{\theta}_1$ est **meilleur** que $\widehat{\theta}_2$.

  Mais il est possible que $R_\theta(\widehat{\theta}_1) < R_\theta(\widehat{\theta}_2)$ pour certaines valeurs de θ et que $R_\theta(\widehat{\theta}_1) > R_\theta(\widehat{\theta}_2)$ pour d'autres. Dans ce cas on ne peut rien dire…

• Le risque R_θ permet aussi, en théorie, de trouver l'estimateur *optimal*, c'est-à-dire celui qui minimise ce risque. Par exemple, pour le MSE, on doit trouver la fonction $\widehat{\theta}(\cdot)$ qui minimise l'intégrale

$$\widehat{\theta}^* = \arg\min_{\widehat{\theta}} \int_{\mathbb{R}^N} |\widehat{\theta}(x) - \theta|^2 p_\theta(x)\, \mathrm{d}\mu(x).$$

Même si on arrive à résoudre ce problème de minimisation, il est néanmoins tout à fait possible qu'une telle fonction optimale $\widehat{\theta}^*(\cdot)$ dépende de θ !

Or, comme on l'a vu, c'est totalement interdit car $\widehat{\theta}(X)$ doit être une fonction seulement des données. Par conséquent, **il n'existe pas toujours d'estimateur optimal** [3].

3. Voir exercice ci-dessous pour un exemple

Sauf mention contraire on se limite dans la suite au risque quadratique (MSE).

RÉSUMÉ:

- Un estimateur est *optimal* s'il minimise le risque quadratique (erreur quadratique moyenne MSE) :

$$\widehat{\theta}^* = \arg\min_{\widehat{\theta}} \underbrace{\mathbb{E}_\theta\{(\widehat{\theta}(X) - \theta)^2\}}_{\mathrm{MSE}_\theta(\widehat{\theta})}$$

- Un estimateur $\widehat{\theta}_1$ est **meilleur** (strictement) que $\widehat{\theta}_2$ si

$$(\forall \theta \in \Theta) \qquad \mathrm{MSE}_\theta(\widehat{\theta}_1) \leqslant \mathrm{MSE}_\theta(\widehat{\theta}_2)$$

(avec une inégalité stricte pour au moins une valeur de θ).

- Un estimateur sera donc optimal s'il est meilleur que tous les autres, mais un tel estimateur n'existe pas toujours.

Exercice: estimateurs (in)admissibles

On dit qu'un estimateur est inadmissible s'il est toujours plus mauvais qu'un autre estimateur (ici pour le MSE). (Plus mauvais strictement, pour au moins une valeur de θ). Il est admissible dans le cas contraire. La notion d'admissibilité, comme on va le voir, est assez rigide.

On veut estimer la moyenne d'un N-échantillon $X = (X_1, X_2, \ldots, X_N)$ (de carré intégrable, i.i.d.) à l'aide de l'estimateur de la moyenne

$$\overline{X}_N = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

1. Montrer que $\overline{X}_1 = X_1$ est inadmissible.

On précise le modèle : $\mathcal{N}(\theta, \sigma^2)$ où $\theta \in [0, 1]$.

2. Montrer que $\overline{X}_N$ est aussi inadmissible.
3. Montrer cependant que l'estimateur constant $= \frac{1}{2}$ est *admissible* (bien qu'il soit certainement très mauvais).

Exercice: estimateurs (in)admissibles

Indications:

1. On est dans le cas où les moyennes sont égales et il suffit de comparer les variances de $\overline{X}_1$ et $\overline{X}_N$.
2. À cause de l'hypothèse $0 \leqslant \theta \leqslant 1$ on peut construire un estimateur contraint au même intervalle :

$$\widehat{\theta}_N(X) = \begin{cases} 0 & \text{si } \overline{X}_N \leqslant 0 \\ \overline{X}_N & \text{si } 0 \leqslant \overline{X}_N \leqslant 1 \\ 1 & \text{si } \overline{X}_N \geqslant 1 \end{cases}$$

   Comparer alors $|\widehat{\theta}_N - \theta|$ et $|\overline{X}_N - \theta|$.
3. Si on avait un estimateur $\widehat{\theta}$ meilleur que $\frac{1}{2}$, on aurait $\mathbb{E}\{(\widehat{\theta} - \theta)^2\} \leqslant \mathbb{E}\{(1/2 - \theta)^2\}$ pour tout $\theta \in [0, 1]$ et donc en particulier pour $\theta = 1/2$!

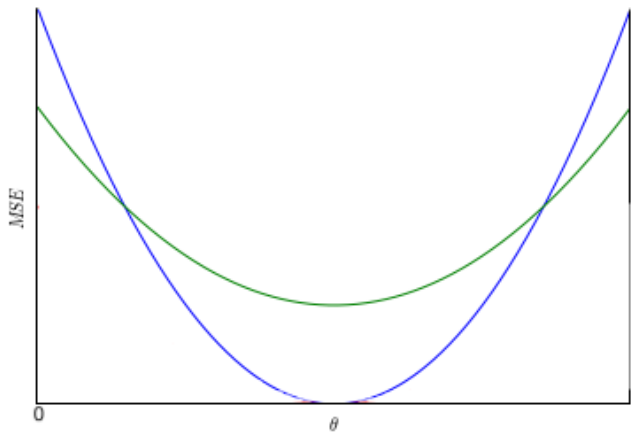


FIGURE 4.1: Exemples de MSE(θ) d'estimateurs admissibles

5
Compromis biais-variance

Biais

Un *biais* statistique est une erreur systématique qui se produit lors de l'estimation de θ par $\widehat{\theta}$. Comme c'est une erreur constante quelles que soient les valeurs observées X, on la retrouve en moyennant l'estimateur sur X. Le biais $\mathbb{B}(\widehat{\theta})$ de $\widehat{\theta}$ est donc

$$\mathbb{B} = \mathbb{B}_\theta(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta = \mathbb{E}_\theta\big(\widehat{\theta}(X) - \theta\big)$$

et il dépend en général de θ. S'il y a un biais (non nul), on dit que l'estimateur est **biaisé**.

On aimerait intuitivement ne pas avoir cette erreur systématique et donc garantir un biais nul : $\mathbb{B} = 0$ (pour tout $\theta \in \Theta$). Dans ce cas, l'estimateur est **non biaisé** ou **sans biais**. Cela revient à dire que sa moyenne égale le paramètre à estimer : $\mathbb{E}(\widehat{\theta}) = \theta$.

On verra que certains estimateurs non biaisés peuvent quand même être meilleurs ! [1]

1. voir exercice p. 53

Variance

Supposons pour simplifier que $\theta \in \mathbb{R}$ (cas scalaire, un seul paramètre). La *variance* d'un estimateur $\widehat{\theta}$ est simplement la variance de la variable aléatoire réelle $\widehat{\theta}(X)$:

$$\mathbb{V} = \mathbb{V}_\theta(\widehat{\theta}) = \mathrm{Var}\{\widehat{\theta}(X)\} = \mathbb{E}_\theta\{\big(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\big)^2\}$$

Dans le cas général où $\theta \in \mathbb{R}^n$ on peut considérer la *matrice de covariance* du vecteur aléatoire $\widehat{\theta}$ (une matrice carrée $n \times n$) et définir la variance comme la trace de cette matrice.

On se limite ici pour simplifier au cas scalaire.

MSE = biais^2 *+ variance*

Toujours dans le cas scalaire, on peut décomposer le MSE :

$$\begin{aligned}\mathrm{MSE} &= \mathbb{E}\{(\widehat{\theta}-\theta)^2\}\\ &= \mathbb{E}\{(\widehat{\theta}-\mathbb{E}(\widehat{\theta})+\mathbb{B})^2\}\\ &= \mathbb{E}\{\big(\widehat{\theta}-\mathbb{E}(\widehat{\theta})\big)^2\}+\mathbb{E}\{\mathbb{B}^2\}+0\\ &= \mathbb{V}+\mathbb{B}^2\end{aligned}$$

où on a développé le carré et tenu compte du fait que la biais est constant (non aléatoire). Dans le développement le terme croisé s'annule car $\mathbb{E}\{(\widehat{\theta}-\mathbb{E}(\widehat{\theta}))\,\mathbb{B}\} = \big(\mathbb{E}(\widehat{\theta})-\mathbb{E}(\widehat{\theta})\big)\cdot\mathbb{B} = 0$.

Le risque quadratique (qu'on veut minimiser) se décompose donc en

$$\boxed{\mathrm{MSE} = \mathbb{B}^2+\mathbb{V}}$$

Par conséquent, pour minimiser le risque, on cherche à la fois à minimiser (voire annuler) le biais et à minimiser la variance. Ce sont deux minimisations concurrentes qui peuvent amener un **compromis biais-variance**. Par exemple, tolérer un petit biais peut permette de réduire davantage la variance de sorte à globalement réduire le MSE.

RÉSUMÉ:

- Le biais et la variance d'un estimateur $\widehat{\theta}$ sont

$$\mathbb{B}_\theta(\widehat{\theta}) = \mathbb{E}_\theta\big(\widehat{\theta}(X)\big) - \theta$$
$$\mathbb{V}_\theta(\widehat{\theta}) = \mathbb{E}_\theta\{\big(\widehat{\theta}(X) - \mathbb{E}(\widehat{\theta})\big)^2\}$$

- Le MSE se décompose en carré du biais + variance :

$$\text{MSE} = \mathbb{B}^2 + \mathbb{V}$$

- Un bon estimateur aura donc à la fois un biais faible (voire nul) et une variance faible.

- Un estimateur est *non biaisé* si son biais est nul pour tout θ, c'est à dire que son espérance égale le paramètre :

$$\mathbb{E}(\widehat{\theta}) = \theta.$$

Un tel estimateur sera d'autant meilleur que sa variance est faible.

- Un estimateur MVU (*minimum variance unbiased estimator*) est un estimateur non biaisé de *variance minimale*. [cf. chapitre suivant]

Exercice: avec ou sans biais

Cet exercice montre un estimateur biaisé n'est pas forcément optimal, et même, qu'il n'existe pas toujours !

Soit le modèle $X \sim \mathcal{N}(\mu, \theta)$ (i.i.d.) : on estime la variance à moyenne μ connue à l'aide d'un N-échantillon gaussien. Un estimateur naturel est la variance empirique :

$$\widehat{\theta}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

1. Est-il non biaisé ? quel est sa variance ? son MSE ?
2. Même question pour l'estimateur

$$\widehat{\theta}'(X) = \frac{1}{N+2}\sum_{i=1}^{N}(X_i - \mu)^2$$

Conclure.

Soit maintenant la modèle i.i.d. de Bernoulli $X \sim \mathcal{B}(\frac{1}{\theta})$: on estime l'inverse du paramètre d'une loi de Bernoulli.

3. Justifier que $S = \sum_{i=1}^{N} X_i$ est une statistique suffisante [2] et donc qu'on peut se limiter à des estimateurs du type $\widehat{\theta}(S)$.
4. Résoudre l'équation $\mathbb{E}(\widehat{\theta}(S)) = \theta$ (pour tout θ) et conclure.

2. voir exercice p. 39

Exercice: avec ou sans biais

Indications:

1. Noter que $\mathbb{E}\{(X_i - \mu)^2\} = \theta$ est la variance des échantillons. Ainsi MSE=$\mathbb{V}$. On rappelle que si $Z \sim \mathcal{N}(0, \sigma^2)$ alors le moment d'ordre 4 vaut $\mathbb{E}(Z^4) = 3\sigma^4$.
2. L'espérance de $\widehat{\theta}$ est multipliée par $\frac{N}{N+2}$ donc le biais change ($\mathbb{B}^2$ augmente). La variance est multipliée par $\left(\frac{N}{N+2}\right)^2$ dont $\mathbb{V}$ diminue. Tous calculs faits le MSE diminue aussi.
3. Voir l'exercice p. 39, S suit la loi binomiale $\mathcal{B}(N, \frac{1}{\theta})$
4. $\mathbb{E}(\widehat{\theta}(S))$ est un polynôme en $\frac{1}{\theta}$, qui ne peut être égal à θ pour tout θ !

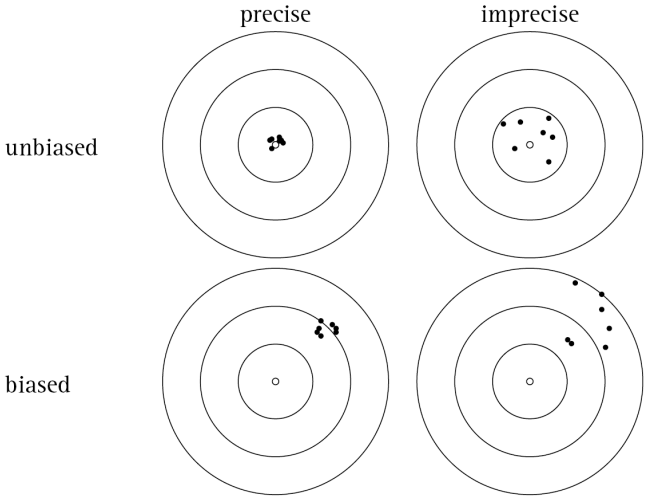


FIGURE 5.1: Biais et variance lors d'un tir sur cible

6
Estimation non biaisée optimale

Estimateur MVU

Un estimateur MVU, comme son nom l'indique [1] est un estimateur *non biaisé* de *variance minimale*. C'est donc un estimateur optimal pour le MSE mais sous la contrainte de biais nul :

$$\widehat{\theta}_{\text{MVU}} = \arg\min_{\widehat{\theta}} \left\{ \mathbb{E}\{|\widehat{\theta} - \theta|^2\} \;\middle|\; \mathbb{E}(\widehat{\theta}) = \theta \right\}$$

Comme on l'a vu, il peut ne pas exister, et un autre estimateur biaisé peut être meilleur.

Cependant, c'est le type d'estimateur pour lequel on a tout de même des bons résultats théoriques : comme on va le voir dans ce qui suit, sous certaines conditions (plus ou moins restrictives) il existe des méthodes permettant d'en trouver.

Voici deux théorèmes qui, combinés, permettent de construire un estimateur MVU à l'aide d'une statistique suffisante [2] :

Théorème de Rao-Blackwell (1947). [3] *Si S est une statistique suffisante, alors l'espérance conditionnelle $\widehat{\theta}^* = \mathbb{E}(\widehat{\theta}|S)$ est un estimateur de même biais mais meilleur que que $\widehat{\theta}$.*

Théorème de Lehmann-Scheffé (1955). [4] *Si, de plus, S est une statistique « complète », c'est à dire que pour tout fonction déterministe f,*

$$(\forall \theta \in \Theta) \quad \mathbb{E}_\theta\big(f(S)\big) = 0 \implies f(S) = 0 \; p.s.$$

alors $\widehat{\theta}^* = \mathbb{E}(\widehat{\theta}|S)$ est un estimateur MVU. En particulier c'est le cas de S.

Malheureusement, trouver une statistique suffisante complète est souvent hors de portée...

1. *minimum variance unbiased estimator, aussi appelé UMVUE : uniformly minimum variance unbiased estimator*

2. voir exercice p. 39

3. prouvé en exercice p. 57

4. prouvé en exercice p. 59

7
Score et information de Fisher

Sous certaines hypothèses de **régularité**, on peut déterminer les notions utiles de score et d'information de Fisher.

Modèle statistique régulier

• L'hypothèse principale de régularité est que la densité p_θ est **dérivable en** θ. Cela n'impose nullement que $p_\theta(x)$ doive être une densité continue. Par exemple, la distribution de Bernoulli $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ ($x \in \{0,1\}$) est bien dérivable en θ.

Comme on dérive par rapport au paramètre θ et non par rapport à la variable x, on note la dérivée comme une dérivée *partielle* : $\frac{\partial}{\partial\theta} p_\theta(x)$. Dans le cas vectoriel ($\theta \in \mathbb{R}^n$) il s'agit d'un gradient $\nabla_\theta p_\theta$.

• Pour éviter les problèmes de dérivation d'intégrales du type $\int f(x) p_\theta(x)\,\mathrm{d}\mu(x)$ (liés aux bornes ou régions d'intégration), on suppose que le support des lois $\mathbb{P}_\theta$ est fixe, indépendant de θ. Par exemple ce pourrait être tout l'espace $\mathbb{R}^N$.

• Enfin on suppose réunies les conditions d'intégration sous le signe $\int$ dans les intégrales qu'on va considérer dans les calculs :

$$\frac{\partial}{\partial\theta}\int f(x) p_\theta(x)\,\mathrm{d}\mu(x) = \int f(x)\frac{\partial}{\partial\theta} p_\theta(x)\,\mathrm{d}\mu(x)$$

La vérification rigoureuse de tout cela n'est pas toujours facile…

On suppose dans la suite le modèle régulier.
Pour simplifier, on se place dans le cas scalaire $\theta \in \mathbb{R}$.

Score

Le **score** (ou « informant ») sur θ est défini par

$$\boxed{S_\theta(X) = \frac{\partial}{\partial\theta} \log p_\theta(X) = \frac{\frac{\partial}{\partial\theta} p_\theta(X)}{p_\theta(X)}}$$

où log désigne de logarithme naturel (ln). L'idée de passer par le **logarithme** de p_θ simplifie souvent les calculs (on retrouvera $\log p_\theta(X)$ sous le nom de *log-vraisemblance* [1]).

1. Voir p. 81

Naturellement $S_\theta(X)$ dépend de θ en général, ce n'est donc pas un estimateur. Néanmoins il est **centré** : $\mathbb{E}\{S_\theta(X)\} = 0$ car

$$\begin{aligned}
\mathbb{E}\Big\{ \frac{\frac{\partial}{\partial\theta} p_\theta(X)}{p_\theta(X)} \Big\} &= \int \frac{\frac{\partial}{\partial\theta} p_\theta}{p_\theta} \cdot p_\theta \, \mathrm{d}\mu \\
&= \int \frac{\partial}{\partial\theta} p_\theta \, \mathrm{d}\mu \\
&= \frac{\partial}{\partial\theta} \int p_\theta \, \mathrm{d}\mu \\
&= \frac{\partial}{\partial\theta} 1 \\
&= 0
\end{aligned}$$

Information de Fisher

L'**information de Fisher** est la variance du score :

$$\boxed{J_\theta = \mathbb{V}\big(S_\theta(X)\big) = \mathbb{E}\{S_\theta(X)^2\}} \geqslant 0$$

(où on a tenu compte du fait que $\mathbb{E}\{S_\theta(X)\} = 0$ dans la dernière formule).

Il y a un lien intéressant avec la dérivée du score (donc la dérivée seconde de $\log p_\theta(X)$) :

$$\frac{\partial}{\partial\theta} S_\theta(X) = \frac{\partial}{\partial\theta} \Big(\frac{\frac{\partial}{\partial\theta} p_\theta(X)}{p_\theta(X)} \Big) = \frac{\frac{\partial^2}{\partial\theta^2} p_\theta(X)}{p_\theta(X)} - \Big(\frac{\frac{\partial}{\partial\theta} p_\theta(X)}{p_\theta(X)} \Big)^2$$

En prenant l'espérance, on voit que $\mathbb{E}\Big\{ \frac{\frac{\partial^2}{\partial\theta^2} p_\theta(X)}{p_\theta(X)} \Big\} = 0$ par un calcul similaire à celui fait ci-dessus avec une dérivée seconde à la place de la dérivée.

On obtient par conséquent

$$\boxed{J_\theta = -\mathbb{E}\Big\{\frac{\partial}{\partial\theta} S_\theta(X)\Big\} = -\mathbb{E}\big\{\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\big\}}$$

Il suffit de dériver encore le score et de prendre l'espérance pour trouver l'information de Fisher (à un signe près). C'est souvent plus facile que de calculer directement la variance du score.

Comment s'interprète l'information de Fisher ? Elle mesure, en quelque sorte, l'information [2] apportée par X sur θ.

La formule ci-dessus permet de l'interpréter comme la courbure moyenne du log-vraisemblance $\log p_\theta(X)$ en θ. En effet, cette courbure, donnée par la dérivée seconde $\frac{\partial^2}{\partial\theta^2}\log p_\theta$ et en moyenne négative puisque $\mathbb{E}\big\{\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\big\} = -J_\theta \leqslant 0$. Ceci correspond à un maximum de vraisemblance [3] en θ d'autant plus « piqué » ou « précis » (grande courbure ou variation de pente) que J_θ est grand. Autrement dit, plus J_θ est grand, plus en moyenne $p_\theta(X)$ apportera une connaissance précise de θ.

2. Le lien avec la théorie de l'information de Shannon est approfondi dans mon livre *Théorie de l'information et du codage*, Hermes Science : London, 2007. Voir aussi l'exercice p. 95

3. voir la notion de maximum de vraisemblance p. 81

Cas d'un modèle i.i.d.

De façon générale, comme $X = (X_1, X_2, \ldots, X_N)$, le score et l'information de Fisher dépendent aussi du nombre N d'échantillons et on pourrait les noter respectivement $S_{\theta,N}(X)$ et $J_{\theta,N}$.

Dans le cas d'un modèle $X = (X_1, X_2, \ldots, X_N)$ **i.i.d.**, la densité se factorise en produit $p_\theta(x) = \prod_{i=1}^N p_\theta(x_i)$. En prenant le logarithme et en dérivant on obtient une somme : $S_{\theta,N}(X) = \sum_{i=1}^N S_\theta(X_i)$ où chaque $S_\theta(\cdot)$ dans la somme est identique (puisque les X_i sont identiquement distribués) et correspond au cas d'un seul échantillon. On le note pour cette raison $S_{\theta,1}(X_i)$. On a donc $S_{\theta,N}(X) = \sum_{i=1}^N S_{\theta,1}(X_i)$ où les termes $S_{\theta,1}(X_i)$ sont encore i.i.d., de même variance notée $J_{\theta,1}$. En prenant la variance de la somme on obtient le somme des N variances individuelles égales :

$$\boxed{J_\theta = N \cdot J_{\theta,1}}$$

L'information de Fisher est ainsi proportionnelle au nombre de données : naturellement, plus il y en a, plus cela apporte de l'information sur θ.

En pratique cette relation est très utile pour le calcul : dans le cas i.i.d., *il suffit de traiter d'abord le cas $N = 1$ puis de multiplier par N.*

Cas général de plusieurs paramètres $\theta \in \mathbb{R}^n$

Dans le cas général de n paramètres, la dérivée selon θ doit être remplacé par le gradient :

$$S_\theta(X) = \nabla_\theta \log p_\theta(X)$$

Le score est un vecteur (disons un vecteur colonne) de taille n. C'est un vecteur aléatoire centré.

On définit alors la **matrice d'information de Fisher** par

$$\mathbf{J}_\theta = \mathbb{E}\{S_\theta(X) S_\theta(X)^t\} = -\mathbb{E}\{\nabla_\theta \nabla_\theta^t \log p_\theta(X)\}$$

qui est la matrice de covariance du score, une matrice carrée de taille $n \times n$. C'est également l'opposé du Hessien moyen du log-vraisemblance.

RÉSUMÉ: pour un paramètre $\theta \in \mathbb{R}$:

- *vraisemblance :* la fonction $\theta \mapsto p_\theta(X)$
[cf. chapitre 9]

- *log-vraisemblance :* la fonction $\theta \mapsto \log p_\theta(X)$

- *score :* la dérivée $S_\theta(X) = \dfrac{\partial}{\partial\theta} \log p_\theta(X)$
(de moyenne nulle)

- *information de Fisher :* variance du score :
$J_\theta = -\mathbb{E}\Big\{ \dfrac{\partial^2}{\partial\theta^2} \log p_\theta(X) \Big\} \geqslant 0$

- *Cas d'un modèle i.i.d. :*

$$S_{\theta,N}(X) = \sum_{i=1}^{N} S_{\theta,1}(X_i)$$
$$J_\theta = N \cdot J_{\theta,1}$$

8
Borne de Cramér-Rao

Les notions de score et d'information de Fisher permettent d'obtenir une borne importante sur le MSE. Soit $\widehat{\theta}(X)$ un estimateur **non biaisé**. On le compare au score $S_\theta(X)$ avec l'*inégalité de Cauchy-Schwarz* [1] :

$$\mathrm{Cov}(\widehat{\theta}(X), S_\theta(X))^2 \leqslant \mathbb{V}(\widehat{\theta}(X)) \cdot \mathbb{V}(S_\theta(X))$$

1. cf. cours de probabilités : $|\mathrm{Cov}(X,Y)| \leqslant \sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)}$ avec égalité si et seulement si X et Y sont linéairement dépendantes, du type $aX + bY = c$ p.s.

Ici $\mathbb{V}(S_\theta(X)) = J_\theta$ (information de Fisher), $\mathbb{V}(\widehat{\theta}) =$MSE puisque $\widehat{\theta}$ est non biaisé. De plus, comme $S_\theta(X)$ est centré, la covariance de $\widehat{\theta}(X)$ et $S_\theta(X)$ égale l'espérance de leur produit :

$$\begin{aligned}
\mathrm{Cov}(\widehat{\theta}(X), S_\theta(X)) &= \mathbb{E}\{\widehat{\theta}(X) S_\theta(X)\} \\
&= \int \widehat{\theta} \cdot \frac{\frac{\partial}{\partial\theta} p_\theta}{p_\theta} \cdot p_\theta \,\mathrm{d}\mu \\
&= \int \widehat{\theta} \cdot \frac{\partial}{\partial\theta} p_\theta \,\mathrm{d}\mu \\
&= \frac{\partial}{\partial\theta} \int \widehat{\theta} \cdot p_\theta \,\mathrm{d}\mu \\
&= \frac{\partial}{\partial\theta} \mathbb{E}(\widehat{\theta}) \\
&= \frac{\partial}{\partial\theta} \theta \\
&= 1
\end{aligned}$$

En reportant dans l'inégalité de Cauchy-Schwarz, on obtient la **borne de Cramér-Rao** (CRB [2] due aux statisticiens français **Fréchet et Darmois** :

2. *Cramér-Rao bound*

$$\boxed{\mathbb{V}(\widehat{\theta}) \geqslant \frac{1}{J_\theta}}$$

Plus l'information de Fisher sur θ est grande, plus la CRB est petite et donc plus l'estimation peut être précise (variance faible).

Estimateur efficace

Un estimateur est dit **efficace** s'il est non biaisé et *atteint la CRB* : $\mathrm{MSE} = \mathbb{V}(\widehat{\theta}) = \frac{1}{J_\theta}$.

Comme cette borne s'applique à tous les estimateurs non biaisés, un estimateur efficace a une variance minimale : **un estimateur efficace est MVU**. L'inverse n'est pas forcément vrai, car la borne de Cramér-Rao n'est pas nécessairement atteinte, comme on va le voir maintenant.

Dire qu'un estimateur est efficace revient à dire que l'inégalité de Cauchy-Schwarz ci-dessus est une *égalité*, c'est-à-dire que $\widehat{\theta}(X)$ et $S_\theta(X)$ sont linéairement dépendantes : $\widehat{\theta}(X) = \alpha S_\theta(X) + \beta$ où $\alpha > 0$ puisque la covariance $\mathrm{Cov}(\widehat{\theta}(X), S_\theta(X)) = 1$ est positive. En prenant les moyennes il vient $\theta = \alpha \cdot 0 + \beta$ d'où $\beta = \theta$. En prenant les variances il vient $1/J_\theta = \alpha^2 J_\theta$ d'où $\alpha = 1/J_\theta$. On obtient l'expression :

$$\widehat{\theta}(X) = \frac{S_\theta(X)}{J_\theta} + \theta$$

qui ne fournit l'estimateur efficace **que** si cette expression **ne dépend pas de** θ ! Dans le cas contraire, l'estimateur efficace n'existe pas et la borne de Cramér-Rao n'est pas atteinte.

Cas général de plusieurs paramètres $\theta \in \mathbb{R}^n$

L'inégalité de Cramér-Rao se généralise pour la *matrice de covariance* $\mathrm{Cov}(\widehat{\theta})$ de l'estimateur non biaisé [3]

$$\mathrm{Cov}(\widehat{\theta}) \geqslant \mathbf{J}_\theta^{-1}$$

C'est une inégalité entre deux matrices symétriques qui signifie que la différence est positive [4] : $\mathbf{A} \geqslant \mathbf{B} \iff \mathbf{A} - \mathbf{B} \geqslant 0$. La borne de Cramér-Rao (CRB) est ici l'inverse de la matrice d'information de Fisher $\mathbf{J}_\theta$.

Un estimateur efficace (qui atteint la CRB) est nécessairement de la forme $\widehat{\theta}(X) = \mathbf{J}_\theta^{-1} S_\theta(X) + \theta$ si cette expression ne dépend pas de θ.

3. voir exercice p. 73

4. une matrice symétrique est positive si ses valeurs propres sont positives

RÉSUMÉ:

- *Borne de Fréchet-Darmois-Cramér-Rao :*

$$\mathbb{B}(\widehat{\theta}) = 0 \implies \mathbb{V}(\widehat{\theta}) \geqslant \frac{1}{J_\theta}$$

- *Estimateur efficace :* tel que

$$\mathbb{B}(\widehat{\theta}) = 0$$
$$\mathbb{V}(\widehat{\theta}) = \frac{1}{J_\theta}$$

- Estimateur efficace $\implies$ MVU
  (réciproque fausse)

Exercice: estimateurs efficaces

Certains modèles i.i.d. simples permettent de calculer facilement l'estimateur efficace (MVU). D'autres non…

Pour chacun des modèles i.i.d. suivants, déterminer score, information de Fisher, borne de Cramér-Rao, et estimateur efficace (s'il existe).

1. $\mathcal{B}(\theta)$ (Bernoulli)
2. $\mathcal{P}(\theta)$ (Poisson)
3. $\mathcal{N}(\theta, \sigma^2)$
4. $\mathcal{N}(\mu, \theta)$
5. $\mathcal{E}(1/\theta)$ (exponentielle de moyenne θ)
6. $\mathcal{E}(\theta)$
7. $\mathcal{U}[0, \theta]$ (uniforme)

Exercice: estimateurs efficaces

Indications:
Appliquer les formules du cours (cas i.i.d.) en commençant par $N = 1$.

1. Noter que $\mathcal{B}(\theta)$ a pour variance $\theta(1-\theta)$. Vérifier a posteriori la variance de $\overline{X}_N$.
2. Noter que $\mathcal{P}(\theta)$ a pour variance θ. Vérifier a posteriori la variance de $\overline{X}_N$.
3. Vérifier a posteriori la variance de $\overline{X}_N$.

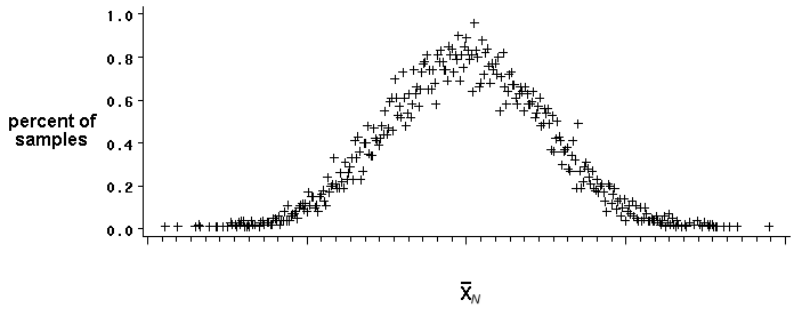


FIGURE 8.1: Estimation efficace par la moyenne

4. Noter que le moment d'ordre 4 de $\mathcal{N}(0, \sigma^2)$ est $3\sigma^4$. Vérifier a posteriori la variance de la variance empirique $\frac{1}{N}\sum_i (X_i - \mu)^2$.
5. Noter que $\mathcal{E}(1/\theta)$ a pour variance θ^2. Vérifier a posteriori la variance de $\overline{X}_N$.
6. Noter que $\mathcal{E}(\theta)$ a pour variance $1/\theta^2$ et que $-\theta^2\overline{X}_N + 2\theta$ dépend évidemment de θ.
7. Le modèle est régulier ? Pourquoi ?

Exercice: bornes de Cramér-Rao généralisées

La borne dite de Cramér-Rao a été obtenue par Fréchet en 1943 dans le cas scalaire i.i.d., étendu ensuite par Darmois en 1945 au cas vectoriel non nécessairement i.i.d.; indépendamment par Rao en 1945 (suite au cas d'égalité traité par Aitken en 1941) et par Cramér, généralisé aux estimateurs biaisés, en 1946.

Extension aux estimateurs biaisés :

On considère modèle régulier et un estimateur $\widehat{\theta}$ de biais $\mathbb{B}_\theta$.

1. Recalculer la covariance du score et de l'estimateur en justifiant que $\mathbb{B}_\theta$ est dérivable en θ.
2. En déduire la borne de Cramér-Rao généralisée :

$$\mathrm{MSE} \geqslant \frac{\left(1 + \frac{\partial}{\partial\theta}\mathbb{B}_\theta\right)^2}{J_\theta} + \mathbb{B}_\theta^2$$

3. Cette borne peut-elle être inférieure à celle du cas non biaisé $\frac{1}{J_\theta}$? Comme exemple, pour un modèle i.i.d. $\mathcal{N}(\mu, \theta)$, considérer l'estimateur $\widehat{\theta}^*(X) = \frac{1}{N+2}\sum_{i=1}^{N}(X_i - \mu)^2$.

Extension cas vectoriel $\theta \in \mathbb{R}^n$:

On revient au cas d'un estimateur non biaisé $\widehat{\theta}$.

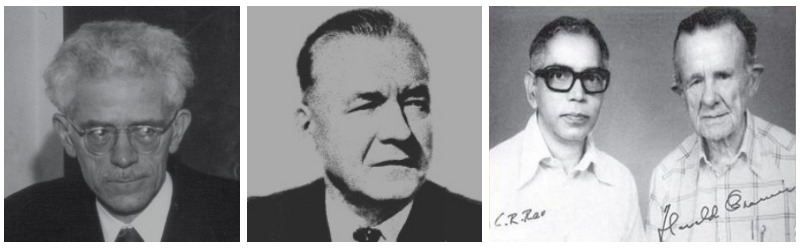
4. Montrer que $\mathbb{E}\big(S_\theta(X)\widehat{\theta}(X)^t\big) = \mathbf{I}_n$ (matrice identité de taille $n \times n$).
5. En développant la matrice de covariance de $\mathbf{J}_\theta^{-1}S_\theta(X) - \widehat{\theta}(X)$, démontrer la borne matricielle de Cramér-Rao : $\mathrm{Cov}(\widehat{\theta}) - \mathbf{J}_\theta^{-1} \geqslant 0$ et discuter le cas d'égalité.

Exercice: bornes de Cramér-Rao généralisées

Indications:

1. Calcul similaire au cas non biaisé, en tenant compte du fait que $\mathbb{E}(\widehat{\theta}) = \theta + B_\theta$.
2. Calcul similaire au cas non biaisé. Utiliser l'expression biais-variance du MSE. Noter que la borne dépend de l'estimateur ! (via son biais).
3. Oui car la dérivée du biais vaut $-\frac{2}{N+2} < 0$ dans l'exemple énoncé (dont on a vu qu'il est non biaisé mais meilleur que l'estimateur efficace [5]), on trouve $\text{MSE} = \text{CRB} = \frac{2\theta^2}{N+2} < \frac{1}{J_\theta} = \frac{2\theta^2}{N}$.
4. Calcul similaire au cas scalaire en notant que $\nabla_\theta \theta^t = \mathbf{I}_n$. Noter qu'il s'agit de la matrice de covariance du score et de l'estimateur (car le score est de moyenne nulle).
5. Dans le développement, la covariance du premier terme $\mathbf{J}_\theta^{-1} S_\theta(X)$ est $\mathbf{J}_\theta^{-1}\mathbf{J}_\theta\mathbf{J}_\theta^{-1} = \mathbf{J}_\theta^{-1}$, et le terme d'inter-covariance se simplifie grâce au résultat de la question précédente. Enfin, on rappelle qu'une matrice de covariance est toujours positive et ne s'annule que lorsque le vecteur aléatoire est constant.

5. voir p. 53 et p. 71



(a) Maurice Fréchet en 1942 (b) Georges Darmois vers 1945 (c) Calyampudi Radhakrishna Rao et Harald Cramér

FIGURE 8.2: Fréchet, Darmois, Cramér et Rao

Exercice: non existence d'estimateurs optimaux

L'existence d'un estimateur efficace ou même d'un estimateur MVU n'est pas garantie !

On considère le modèle statistique $X = (X_1, X_2)$ où

$$X_1 \sim \mathcal{N}(\theta, 1)$$
$$X_2 \sim \begin{cases} \mathcal{N}(\theta, 1) & \text{si } \theta > 0 \\ \mathcal{N}(\theta, 2) & \text{sinon} \end{cases}$$

1. Calculer la borne de Cramér-Rao.
2. Existe-t-il un estimateur efficace ?

On considère les deux estimateurs suivants :

$$\widehat{\theta}_1 = \frac{X_1 + X_2}{2}$$
$$\widehat{\theta}_2 = \frac{2X_1 + X_2}{3}$$

3. Calculer leurs biais et variances.
4. Existe-t-il un estimateur MVU ?

Exercice: non existence d'estimateurs optimaux

Indications:

1. Discuter suivant le signe de θ.
2. L'expression $\frac{S_\theta(X)}{J_\theta} + \theta$ dépend-elle de θ ?
3. Pour la variance, discuter suivant le signe de θ.

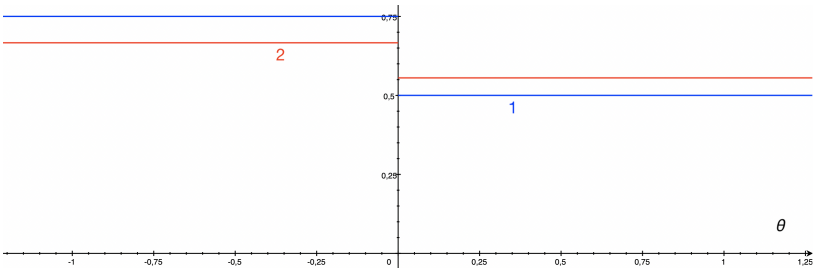


FIGURE 8.3: MSE de deux estimateurs suivant la valeur du paramètre

4. Dans ce cas, un estimateur MVU atteindrait la CRB (pourquoi ?).

Exercice: reparamétrisation et efficacité asymptotique

*Un changement de paramètre $\alpha = g(\theta)$ modifie complètement la borne de Cramér Rao et le problème d'estimation. Même si $\widehat{\theta}$ est efficace, l'estimateur "**plug-in**" $g(\widehat{\theta})$ peut ne plus l'être ou seulement asymptotiquement.*

Soit $g(\cdot)$ un difféomorphisme et $\alpha = g(\theta)$.

1. Comment le score, l'information de Fisher et la borne de Cramér-Rao sont-ils modifiés par ce changement de paramètre ?

On considère le modèle i.i.d. $\mathcal{N}(\theta, \sigma^2)$ où $\widehat{\theta} = \overline{X}_N$ est efficace.

2. Que dire de $\widehat{\alpha} = g(\overline{X}_N)$ si $g(\theta) = a\theta + b$ est une fonction linéaire ?

On désire maintenant estimer $\alpha = g(\theta) = \theta^2$.

3. Calculer le biais et la variance de $\widehat{\alpha} = g(\overline{X}_N)$.
4. Conclure lorsque $N \to +\infty$.

Exercice: reparamétrisation et efficacité asymptotique

Indications:

1. La loi de l'observation n'ayant pas changé ($p_\alpha = p_\theta$), appliquer la règle $\frac{\partial}{\partial \alpha} = \frac{\partial \theta}{\partial \alpha}\frac{\partial}{\partial \theta}$.
2. Le biais reste nul, la variance est multipliée par b^2, et $a\widehat{\theta} + b$ reste efficace.
3. Le biais devient non nul. Pour la variance appliquer la formule du moment d'ordre 4 d'une v.a. gaussienne $X \sim \mathcal{N}(\mu, \sigma^2) : \mathbb{E}(X^4) = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4$.
4. Quel est le MSE_α limite ? On dit que $\widehat{\alpha} = \widehat{\theta}^2$ est *asymptotiquement efficace*.

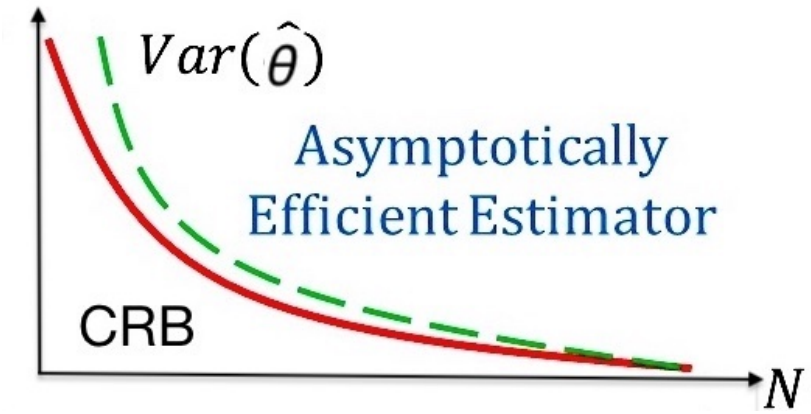


FIGURE 8.4: Estimation asymptotiquement efficace

Exercice: pseudo-inverse et déconvolution

Un exemple classique d'un modèle vectoriel est

$$X = \mathbf{H}\,\theta + W$$

où on représente les vecteurs sous forme colonne : X et W de longueur N, θ de longueur n, et $\mathbf{H}$ une matrice (déterministe) $N \times n$. La matrice $\mathbf{H}$ peut représenter un « filtrage » linéaire d'un signal θ dans du bruit additif W, on cherche dans ce cas à résoudre un problème de « déconvolution » : retrouver le signal θ à partir de l'observation X filtrée et bruitée.

On considère le modèle ci-dessus où $\mathbf{H}$ de rang plein, $N > n$, et W est un bruit gaussien **blanc**, c'est-à-dire un vecteur gaussien i.i.d. $\mathcal{N}(0, \sigma^2)$.

1. Calculer la CRB.
2. Montrer que l'estimateur efficace est

   $$\widehat{\theta} = (\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}^t X$$

   où $(\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}^t$ est la *pseudo-inverse* de $\mathbf{H}$.
3. Vérifier a posteriori que cet estimateur est bien MVU.

On suppose maintenant que $W \sim \mathcal{N}(0, \mathbf{C})$ où la matrice de covariance $\mathbf{C}$ est inversible (le modèle n'est plus i.i.d.).

4. *Blanchiment* : Montrer qu'on peut factoriser $\mathbf{C}^{-1} = \mathbf{D}^t\mathbf{D}$ et se ramener à un modèle équivalent avec bruit $\mathbf{D}W$ blanc. En déduire l'estimateur efficace

   $$\widehat{\theta} = (\mathbf{H}^t\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{C}^{-1}X$$

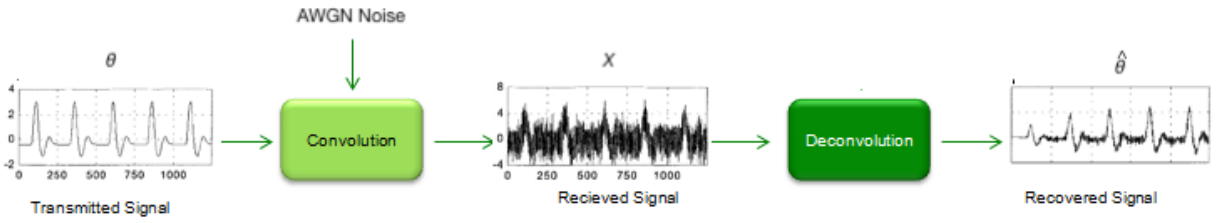
5. *Calcul direct* : Calculer directement l'estimateur efficace et vérifier qu'il est bien MVU.

Exercice: pseudo-inverse et déconvolution

Indications:

1. Pour alléger les calculs, procéder directement en matriciel avec les formules
$$\begin{cases} \nabla_\theta(\theta^t \mathbf{A}\theta) = 2\mathbf{A}\theta \\ \nabla_\theta(\theta^t \mathbf{A}) = \mathbf{A} \end{cases}$$
2. Montrer que $\mathbf{J}_\theta^{-1} S_\theta(X) + \theta = (\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}^t X$.
3. Calculer son bias et sa covariance et comparer avec la CRB.

FIGURE 8.5: Principe de la déconvolution dans du bruit additif



4. Par le théorème spectral, $\mathbf{C} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^t$ où $\mathbf{\Lambda}$ est diagonale et $\mathbf{P}$ est orthogonale. Poser alors $\mathbf{D} = \mathbf{\Lambda}^{-1/2}\mathbf{P}^t$. Le modèle équivalent (statistique suffisante) est $X' = \mathbf{D}X$. Montrer qu'il suffit alors de remplacer X par X' et de même $\mathbf{H}$ par $\mathbf{H}' = \mathbf{D}\mathbf{H}$ dans l'expression de l'estimateur.
5. Similaire au cas ci-dessus $\mathbf{C} = \mathbf{I}$, avec un score $S_\theta(X) = -\frac{1}{2}\nabla_\theta\big((X-\mathbf{H}\theta)^t\mathbf{C}^{-1}(X-\mathbf{H}\theta)\big) = \mathbf{H}^t\mathbf{C}^{-1}(X-\mathbf{H}\theta)$ et $J_\theta = \mathbf{H}^t\mathbf{C}^{-1}\mathbf{H}$.

9
Maximum de vraisemblance

Hormis quelques cas simples, un estimateur optimal est difficile à trouver (s'il existe). À l'inverse, le principe de *maximum de vraisemblance* (ML [1]) donne un estimateur simple à trouver dans de nombreux cas, mais sans aucune garantie d'optimalité.

1. *maximum likelihood*

Vraisemblance

Étant donné une observation $X = x$, on peut dire qu'une valeur de θ d'autant plus « vraisemblable » que la densité p_θ évaluée en x est élevé. On appelle donc **vraisemblance** la quantité [2] $p_\theta(X)$, *vue comme une fonction de* θ (et non de X !).

2. en anglais *likelihood*, noté parfois $\ell_X(\theta) = p_\theta(X)$

C'est un léger abus de notation car θ désigne normalement le « vrai » paramètre de la loi des données $X \sim p_\theta$, qui est fixé. La fonction de vraisemblance devrait donc être notée comme une fonction d'une autre variable, par exemple : $\theta' \mapsto p_{\theta'}(X)$.

Passer par le logarithme simplifie souvent les calculs, et on appelle **log-vraisemblance** (LL [3]) la fonction $\log p_\theta(X)$.

3. *log-likelihood*, noté parfois $\ell\ell_X(\theta) = \log p_\theta(X)$

Estimateur ML

Si la fonction de vraisemblance $\theta \mapsto p_\theta(x)$ a (au moins) un maximum $\widehat{\theta}(x)$ sur Θ pour tout $x \in \mathbb{R}^n$, alors en remplaçant x par X, la variable aléatoire $\widehat{\theta}(X)$ est appelé **estimateur du maximum de vraisemblance** (**estimateur ML** [4]).

4. maximum likelihood estimator

En pratique on peut garder la lettre majuscule X (désignant une quantité aléatoire) dans les calculs

et on peut écrire

$$\widehat{\theta}_{\mathrm{ML}}(X) = \arg\max_{\theta\in\Theta} p_\theta(X)$$

C'est équivalent et souvent plus commode de maximiser la log-vraisemblance LL :

$$\boxed{\widehat{\theta}_{\mathrm{ML}}(X) = \arg\max_{\theta\in\Theta} \log p_\theta(X)}$$

Là où $p_\theta(X)$ s'annule, la contribution du logarithme est $\log 0 = -\infty$ ce qui ne perturbe pas la recherche du maximum.

L'estimateur ML n'existe pas toujours et quand il existe, il n'est pas nécessairement unique [5]. Il n'est pas forcément efficace, il peut même très bien être biaisé.

5. voir exercice p. 87

Équation de vraisemblance

Si la log-vraisemblance LL est dérivable en θ et *concave*, son maximum s'obtient en annulant sa dérivée. Pour trouver $\widehat{\theta}_{\mathrm{ML}}$ on résout alors en θ l'**équation de vraisemblance** :

$$S_\theta(X) = \frac{\partial}{\partial\theta} \log p_\theta(X) = 0$$

dans le cas scalaire, ou $\nabla_\theta \log p_\theta(X) = 0$ dans le cas vectoriel.

Pour justifier que la solution est bien un maximum global, il est important de vérifier que la LL est concave, par exemple grâce à la condition (pour tout X) [6] :

$$\frac{\partial^2}{\partial\theta^2} \log p_\theta(X) \leqslant 0$$

dans le cas scalaire, ou $\nabla_\theta \nabla_\theta^t \log p_\theta(X) \leqslant 0$ dans le cas vectoriel.

S'il n'y a pas concavité, l'équation de vraisemblance n'est qu'une condition nécessaire et il faut montrer directement par une étude des variations que la solution obtenue correspond bien à un maximum (global) de vraisemblance.

6. La relation $\mathbb{E}\left\{\frac{\partial^2}{\partial\theta^2} \log p_\theta(X)\right\} = -J_\theta \leqslant 0$ sur l'information de Fisher montre que cette propriété est vraie en moyenne sur X, ce qui n'est pas suffisant.

Propriétés d'invariance

L'estimateur ML est robuste au changement :

• La vraisemblance dépend de la mesure dominante μ choisie : si on en change et qu'on choisit $\mu' \gg \mu$ à la place, alors la nouvelle densité est $p'_\theta(x) = p_\theta(x)\frac{\mathrm{d}\mu}{\mathrm{d}\mu'}(x)$. Mais comme c'est une multiplication par un facteur constant en θ qui ne change pas son maximum, *l'estimateur ML est invariant par changement de mesure dominante*.

• Si on veut estimer $\alpha = g(\theta)$ à la place de θ, où g est bijective (par exemple un difféomorphisme), alors cette reparamétrisation ne change pas $p_\theta(x) = p_{g^{-1}(\alpha)}(x)$ et par conséquent son maximum en α n'est autre que $\widehat{\alpha}_{\mathrm{ML}} = g(\widehat{\theta}_{\mathrm{ML}})$: *l'estimateur ML est invariant*[7] *par reparamétrisation*.

7. le terme plus exact est « covariant »

Propriétés asymptotiques pour un modèle i.i.d.

On se place dans le cas d'un modèle i.i.d., sous certaines conditions techniques (non précisées ici).

Notons $\widehat{\theta}_N$ l'estimateur ML pour N observations i.i.d. $X_1, X_2, \ldots, X_N \sim p_\theta$. Le grand intérêt de cet estimateur ML est qu'il jouit de propriétés asymptotiquement agréables quand $N \to +\infty$:

• *Consistance :* $\widehat{\theta}_N$ est **consistant** (on dit aussi **convergent**)[8] :

$$\widehat{\theta}_N \xrightarrow{\mathbb{P}_\theta} \theta \qquad (N \to +\infty)$$

(convergence en probabilité[9]). Autrement dit, l'estimateur converge vers le paramètre estimé. $\widehat{\theta}_N$ peut être aussi **fortement consistant** ce qui signifie que la convergence $\widehat{\theta}_N \to \theta$ est presque sûre.

8. voir exercice p. 95

9. $(\forall \varepsilon > 0) \quad \mathbb{P}_\theta(|\widehat{\theta}_N - \theta| > \varepsilon) \to 0$

On en déduit (sous certaines conditions[10]) :

10. voir exercice p. 89

• $\widehat{\theta}_N$ est **asymptotiquement non biaisé** :

$$\mathbb{E}(\widehat{\theta}_N) \to \theta \qquad (N \to +\infty)$$

Autrement dit, son biais tend vers zéro.

Pour un modèle **régulier** (avec certaines conditions techniques) :

- $\widehat{\theta}_N$ est **asymptotiquement normal** [11] :

$$\sqrt{N}\,(\widehat{\theta}_N - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{J_{\theta,1}})$$

(convergence en loi) où $J_{\theta,1}$ est l'information de Fisher pour un échantillon. Ceci précise la vitesse de convergence de $\widehat{\theta}_N$ vers θ (en $\frac{1}{\sqrt{N}}$).

11. voir exercice p. 97

On en déduit intuitivement que :

- $\widehat{\theta}_N$ est **asymptotiquement efficace** : Il est asymptotiquement non biaisé, et de plus,

$$\mathbb{V}(\widehat{\theta}_N) \sim \text{CRB} = \frac{1}{J_{\theta,N}} = \frac{1}{NJ_{\theta,1}}$$

Autrement dit, $\mathbb{V}(\widehat{\theta}_N) J_{\theta,N} \to 1$. Quand $N \to +\infty$, la variance de l'estimateur ML décroît en $\frac{1}{N}$ en atteignant asymptotiquement la borne de Cramér-Rao.

Pour tous ces résultats, il existe des théorèmes, mais avec des hypothèses très techniques et non optimales dont la vérification est pénible. C'est pourquoi en pratique on préfère démontrer directement ces propriétés sur l'estimateur étudié.

Typiquement, la consistance se démontre avec une *loi des grands nombres*, et la normalité asymptotique avec un *théorème limite central* [12].

12. éventuellement avec la méthode δ, voir les exercices

RÉSUMÉ: pour un paramètre $\theta \in \mathbb{R}$:

- *vraisemblance :* la fonction $\theta \mapsto p_\theta(X)$
- *log-vraisemblance :* la fonction $\theta \mapsto \log p_\theta(X)$
- *estimateur ML :* maximise la vraisemblance :

$$\widehat{\theta}_{\mathrm{ML}}(X) = \arg\max_{\theta \in \Theta} \log p_\theta(X)$$

- *équation de vraisemblance :* $\frac{\partial}{\partial \theta} \log p_\theta(X) = 0$

Pour un modèle i.i.d., quand $N \to +\infty$, l'estimateur ML : $\widehat{\theta}_N = \widehat{\theta}_{\mathrm{ML}}(X_1, X_2, \ldots, X_N)$ est (sous certaines conditions de régularité) :

- *consistant :* $\widehat{\theta}_N \xrightarrow{\mathbb{P}_\theta} \theta$
- *asymptotiquement non biaisé :* $\mathbb{E}(\widehat{\theta}_N) \to \theta$
- *asymptotiquement normal :* $\sqrt{N}(\widehat{\theta}_N - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{J_{\theta,1}})$
- *asymptotiquement efficace :* $\mathbb{V}(\widehat{\theta}_N) \sim \frac{1}{J_{\theta,N}} = \frac{1}{N J_{\theta,1}}$

Exercice: non existence et non unicité du maximum de vraisemblance

L'estimateur ML n'est pas nécessairement unique, quand il existe !

On considère le modèle i.i.d. laplacien :

$$p_\theta(x) = \frac{1}{4} e^{-\frac{|x-\theta|}{2}}$$

dont on cherche à estimer la moyenne θ.

1. Que donne la maximisation de la vraisemblance dans ce cas ?

On considère maintenant le modèle i.i.d. du χ^2 bilatéral :

$$p_\theta(x) = \frac{1}{2\sqrt{2\pi|x-\theta|}} e^{-\frac{|x-\theta|}{2}}$$

dont on cherche toujours à estimer la moyenne θ.

2. Que donne la maximisation de la vraisemblance dans ce cas ?

Exercice: non existence et non unicité du maximum de vraisemblance

Indications:

1. Minimiser $\sum_i |X_i - \theta|$ revient à trouver une *médiane*. Distinguer les cas N pair ou impair.
2. La vraisemblance est-elle bornée ? (considérer le cas où θ est arbitrairement proche d'un échantillon).

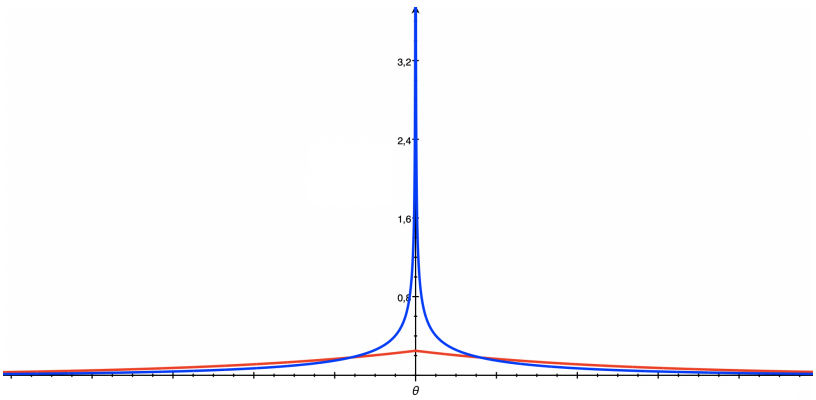


FIGURE 9.1: Densités bilatérales exponentielle (en rouge) et χ^2 à un degré de liberté (en bleu)

Exercice: estimateurs du maximum de vraisemblance

La plupart des modèles simples permettent de calculer facilement l'estimateur ML. On peut ensuite examiner ses propriétés, notamment asymptotiques...

Pour chacun des modèles i.i.d. suivants, trouver l'estimateur ML et déterminer s'il est (asymptotiquement) biaisé/normal/efficace et (fortement) consistant.

1. $\mathcal{B}(\theta)$ (Bernoulli)
2. $\mathcal{P}(\theta)$ (Poisson)
3. $\mathcal{N}(\theta, \sigma^2)$
4. $\mathcal{N}(\mu, \theta)$
5. $\mathcal{E}(1/\theta)$ (exponentielle de moyenne θ)
6. $\mathcal{E}(\theta)$
7. $\mathcal{U}[0, \theta]$ (uniforme)

8. Mêmes questions pour le modèle vectoriel (non i.i.d.) :
$$X = \mathbf{H}\,\theta + W$$
où $W \sim \mathcal{N}(0, \mathbf{C})$.

Exercice: estimateurs du maximum de vraisemblance

Indications:

1. Équation de vraisemblance $\frac{\sum_i X_i}{\theta} - \frac{N-\sum_i X_i}{1-\theta} = 0$, étudier son signe pour justifier d'un maximum de vraisemblance. Fortement consistant par la LLN [15] , asymptotiquement normal par le CLT [16].

15. *(strong) law of large numbers*, loi forte des grands nombres

16. *central limit theorem*, théorème limite central

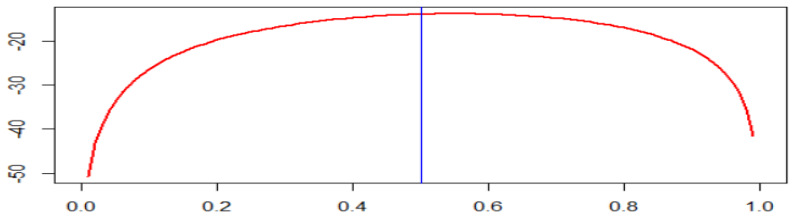


FIGURE 9.3: Maximum de log-vraisemblance pour le modèle de Bernoulli.

2. Équation de vraisemblance $\frac{\sum_i X_i}{\theta} - N = 0$, traitement similaire.
3. Équation de vraisemblance $\sum_i X_i - N\theta = 0$, traitement similaire. Ici $\widehat{\theta}$ est déjà normal !
4. Équation de vraisemblance $\frac{1}{2\theta^2}\sum_i (X_i - \mu)^2 - \frac{N}{2\theta} = 0$. Noter que le moment d'ordre 4 de $\mathcal{N}(0, \sigma^2)$ est $3\sigma^4$.
5. Équation de vraisemblance $\frac{\sum_i X_i}{\theta^2} - \frac{N}{\theta}$.
6. Traitement similaire, ou se déduit du cas précédent par reparamétrisation $\alpha = \frac{1}{\theta}$! [17]. Calcul du biais et de la variance : utiliser le fait que $\sum_{i=1}^N X_i \sim \Gamma(N, \theta)$ (loi Gamma).
7. Vraisemblance $\frac{1_{\max_i X_i \leqslant \theta}}{\theta^N}$ (étudier ses variations en θ). Noter que $\mathbb{P}(\max_i X_i \leqslant x) = \mathbb{P}(X_1 \leqslant x)^N$ pour déterminer la loi de l'estimateur. Le modèle est irrégulier, la loi limite de $N(\theta - \widehat{\theta}(X))$ est exponentielle (et non normale) avec une vitesse de convergence en $\frac{1}{N}$ (au lieu de $\frac{1}{\sqrt{N}}$).
8. C'est l'estimateur efficace [18], qui est normal.

17. ce qui donne consistance forte et normalité asymptotique, cf. exercice p. 89

18. voir exercice p. 79

Exercice: divergence de Kullback-Leibler et consistance de l'estimateur du maximum de vraisemblance

La divergence de Kullback-Leibler (1951), aussi appelée entropie relative, est une notion fondamentale en théorie de l'information. Elle permet de justifier l'intérêt de l'estimation ML dans le cas asymptotique, en particulier de démontrer sa consistance.

La divergence de Kullback-Leibler est définie par

$$D(p_\theta \| p_{\theta'}) = \mathbb{E}_\theta \log \frac{p_\theta(X)}{p_{\theta'}(X)}$$

1. Montrer que $D(p_\theta \| p_{\theta'}) \geqslant 0$ avec égalité $= 0$ si et seulement si $\theta = \theta'$.
2. *Lien avec l'information de Fisher :* Pour un modèle régulier, montrer que $D(p_\theta \| p'_\theta) \sim \frac{(\theta'-\theta)^2}{2} J_\theta$ quand $\theta' \to \theta$.

Soit $\widehat{\theta} = \widehat{\theta}(\underline{X})$ un estimateur ML pour un modèle i.i.d. $\underline{X} = (X_1, X_2, \ldots, X_N)$. On note X l'un quelconque des échantillons X_i.

3. Montrer les deux inégalités opposées

$$\begin{cases} \mathbb{E}_\theta \log p_\theta(X) \geqslant \mathbb{E}_\theta \log p_{\widehat{\theta}(\underline{X})}(X) \\ \log p_\theta(\underline{X}) \leqslant \log p_{\widehat{\theta}(\underline{X})}(\underline{X}) \end{cases}$$

4. Montrer par ailleurs que pour tout $\theta' \in \Theta$, $\frac{1}{N} \log p_{\theta'}(\underline{X}) \xrightarrow{\mathbb{P}_\theta} \mathbb{E}_\theta \log p_{\theta'}(X)$ quand $N \to +\infty$.

On admet qu'on peut se ramener au cas où cette dernière convergence est uniforme en θ' :

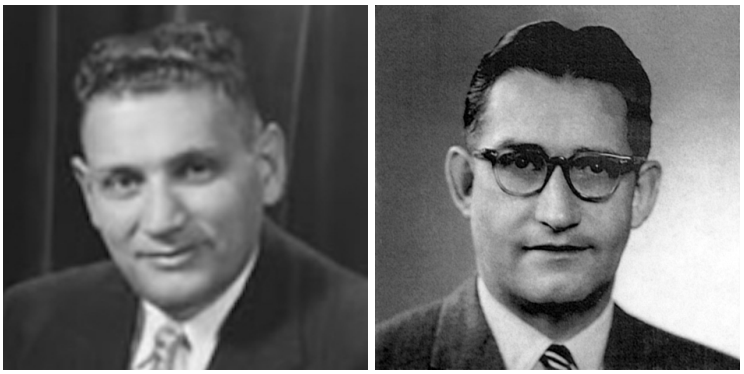
$$\sup_{\theta' \in \Theta} \left| \frac{1}{N} \log p_{\theta'}(\underline{X}) - \mathbb{E}_\theta \log p_{\theta'}(X) \right| \xrightarrow{\mathbb{P}_\theta} 0$$

5. En conclure que $D(p_\theta \| p_{\widehat{\theta}(\underline{X})}) \xrightarrow{\mathbb{P}_\theta} 0$.
6. En déduire que $\widehat{\theta}$ est consistant.

Exercice: divergence de Kullback-Leibler et consistance de l'estimateur du maximum de vraisemblance

Indications:

1. Appliquer l'inégalité de Jensen de (stricte) concavité du logarithme à $-D(p_\theta \| p_{\theta'}) = \mathbb{E}_\theta \log \frac{p_{\theta'}(X)}{p_\theta(X)}$. Le cas d'égalité utilise le fait que le modèle est identifiable.
2. Taylor à l'ordre 2 : calculer la dérivée première et seconde en $\theta' = \theta$ de $D(p_\theta \| p_{\theta'})$.
3. La première inégalité provient de la positivité de la divergence. La deuxième inégalité provient de la définition de l'estimateur ML.
4. Comme le modèle est i.i.d., $p_{\theta'}(\underline{X}) = \prod_{i=1}^N p_{\theta'}(X_i)$. Appliquer la loi (faible) des grands nombres.
5. Grâce à l'hypothèse d'uniformité, on a aussi $\frac{1}{N} \log p_{\widehat{\theta}(\underline{X})}(\underline{X}) \xrightarrow{\mathbb{P}_\theta} \mathbb{E}_\theta \log p_{\widehat{\theta}(\underline{X})}(X)$ et on conclut avec des deux inégalités opposées.
6. Par continuité en θ' de la divergence $D(p_\theta \| p_{\theta'})$, $|\widehat{\theta}(\underline{X}) - \theta| > \varepsilon > 0$ implique l'existence d'un $\eta > 0$ tel que $D(p_\theta \| p_{\widehat{\theta}(\underline{X})}) > \eta$.



(a) Solomon Kullback vers 1950 (b) Richard Arthur Leibler vers 1950

FIGURE 9.5: Kullback et Leibler

Exercice: normalité asymptotique de l'estimateur du maximum de vraisemblance

La normalité asymptotique de l'estimateur du maximum de vraisemblance $\widehat{\theta}$ est une propriété importante car elle montre que la vitesse de convergence de $\widehat{\theta}$ vers θ est en $\frac{1}{\sqrt{N}}$ lorsque le nombre d'échantillons N croît.

On considère un modèle régulier i.i.d. et un estimateur ML $\widehat{\theta} = \widehat{\theta}_N$ consistant.

1. Justifier que $S_{\widehat{\theta}(X)}(X) = 0$.
2. En déduire que $S_\theta(X) = r(X) \cdot (\widehat{\theta}(X) - \theta)$ où $r(X) = -\frac{\partial^2}{\partial\theta^2} \log p_\theta(X) + o(\widehat{\theta}(X) - \theta)$.
3. Montrer que $\frac{1}{\sqrt{N}} S_\theta(X) \xrightarrow{\mathcal{L}} \mathcal{N}(0, J_{\theta,1})$.
4. Montrer par ailleurs que $\frac{1}{N} r(X) \xrightarrow{\mathbb{P}_\theta} -J_{\theta,1}$.
5. Conclure : $\sqrt{N}\ (\widehat{\theta}(X) - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{J_{\theta,1}})$.

Exercice: normalité asymptotique de l'estimateur du maximum de vraisemblance

Indications:

1. $\widehat{\theta}(X)$ vérifie l'équation de vraisemblance.
2. Développement de Taylor de $S_{\widehat{\theta}}(X)$ en θ au premier ordre.
3. Théorème central limite appliqué à $S_\theta(X) = \sum_{i=1}^N S_{\theta,1}(X_i)$.
4. Loi des grands nombres appliquée à $\frac{\partial^2}{\partial\theta^2} \log p_\theta(X) = \sum_{i=1}^N \frac{\partial^2}{\partial\theta^2} \log p_{\theta,1}(X_i)$ pour le premier terme et consistance pour le deuxième.
5. On combine les deux convergences par le lemme de Slutsky [20].

20. voir cours de probabilités

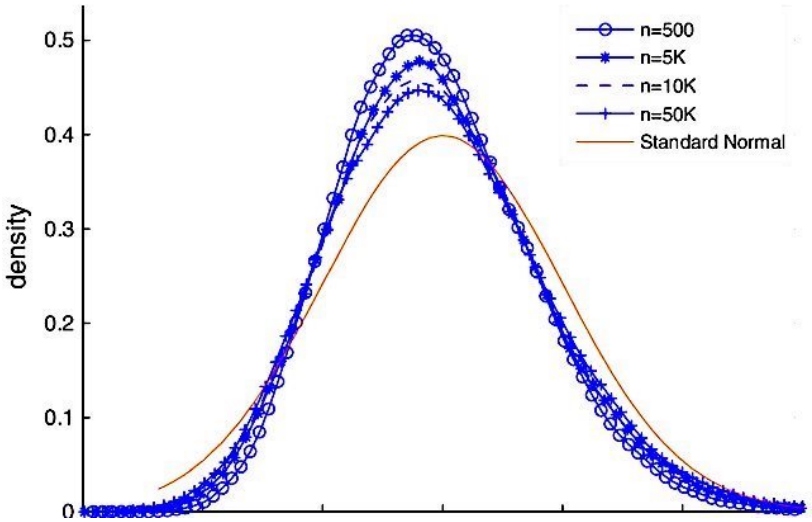


FIGURE 9.6: Illustration d'une normalité asymptotique.

10
Estimation linéaire

Estimateur BLUE

On se place toujours dans le classe des estimateurs *non biaisés*. L'estimateur optimal MVU est difficile à trouver et nécessite de connaître exactement la loi des données $p_\theta(x)$. Or, dans la pratique, cette loi n'est pas toujours connue.

Une solution pragmatique consiste à se restreindre à un estimateur (non biaisé) **linéaire**, de la forme

$$\widehat{\theta}(X) = a^t \cdot X \text{ avec } \mathbb{E}(\widehat{\theta}) = a^t \cdot \mathbb{E}(X) = \theta$$

dans le cas scalaire, où a (à déterminer) et X sont des vecteurs colonnes de N coefficients. Dans le cas vectoriel $\widehat{\theta}(X) = \mathbf{A}X$ où $\mathbf{A}$ est une matrice $n \times N$ et la condition de biais nul s'écrit $\mathbb{E}(\widehat{\theta}) = \mathbf{A}\mathbb{E}(X) = \theta$.

L'**estimateur linéaire non biaisé optimal** (**BLUE** [1]) est alors déterminé de sorte à minimiser la variance :

$$\boxed{\min_a \mathbb{V}(a^t \cdot X) = \min_a a^t \mathbf{C}_X a}$$

sous la contrainte de biais nul $a^t \cdot \mathbb{E}(X) = \theta$, où $\mathbf{C}_X$ désigne la **matrice de covariance** [2] des données X. (Résolution classique par la méthode du lagrangien).

1. best linear unbiased estimator

2. voir cours de probabilités

• L'intérêt est qu'on a pas besoin de connaître la loi p_θ, mais uniquement ses deux premiers moments : moyenne $\mathbb{E}(X)$ et covariance $\mathbf{C}_X$.

• L'inconvénient est que le BLUE est sous-optimal car on se restreint au cas linéaire. De plus, la condition $a^t \cdot \mathbb{E}(X) = \theta$ impose souvent un **modèle linéaire** en θ (pour que a ne dépende pas de θ). Sinon, il faut pouvoir « *linéariser* » les données avant d'appliquer un BLUE.

11
Moindres carrés

Estimateur LSE

L'estimateur BLUE nécessite encore un modèle probabiliste (même limité aux moments d'ordre 1 et 2). On peut aller encore plus loin et *s'affranchir de tout hypothèse probabiliste* pour s'appliquer à tout type de données X.

Pour cela, on se donne un modèle noté $S(\theta)$ et on pose

$$X = S(\theta) + W$$

où W est l'erreur de modèle. Autrement dit $X_i = S_i(\theta) + W_i$ pour $i = 1$ à N.

L'**estimateur aux moindres carrés** (**LSE**[1]) minimise la somme des carrés des erreurs :

1. least squares estimator

$$\boxed{\widehat{\theta}(X) = \arg\min_{\theta} \sum_{i=1}^{N} \big(X_i - S_i(\theta)\big)^2 = \arg\min_{\theta} \|X - S(\theta)\|^2}$$

À la place de la fonction de coût[2] quadratique $L(\theta) = \|X - S(\theta)\|^2$ on peut considérer des moindres carrés **pondérés** : $L(\theta) = \big(X - S(\theta)\big)^t \mathbf{P}\big(X - S(\theta)\big)$ où $\mathbf{P}$ est une matrice poids.

2. loss function

Typiquement, le modèle est *linéaire*, de la forme $S(\theta) = \mathbf{H}\,\theta$ où $\mathbf{H}$ est une matrice (de rang plein) de taille $N \times n$, et il s'agit de résoudre le système aux moindres carrés :

$$\min_{\theta} \|X - \mathbf{H}\,\theta\|^2$$

On aboutit alors à des **équations normales**[3] qui donnent l'estimateur linéaire appelé **LLSE**[4] :

3. soit par un calcul direct, soit par un argument géométrique, voir exercice p. 105

4. linear least squares estimator,

$$\boxed{\widehat{\theta}_{\text{LLSE}} = (\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}X}$$

RÉSUMÉ:

- *Estimateur BLUE* pour un paramètre $\theta \in \mathbb{R}$

$$\begin{cases} \widehat{\theta}_{\text{BLUE}}(X) = a^t \cdot X \text{ avec} \\ a = \arg\min \left\{ a^t \mathbf{C}_X a \;\middle|\; a^t \cdot \mathbb{E}(X) = \theta \right\} \end{cases}$$

où $\mathbf{C}_X$ est la matrice de covariance de X.

- *Estimateur LSE* pour un modèle $X = S(\theta) + W$

$$\widehat{\theta}_{\text{LLSE}}(X) = \arg\min_{\theta} \|X - S(\theta)\|^2$$

- *Estimateur LLSE* pour un modèle linéaire $X = \mathbf{H}\,\theta + W$ où $\theta \in \mathbb{R}^n$:

$$\begin{aligned} \widehat{\theta}_{\text{LLSE}}(X) &= \arg\min_{\theta} \|X - \mathbf{H}\,\theta\|^2 \\ &= (\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}X \end{aligned}$$

Exercice: moindres carrés et équations normales

Le terme « équations normales » a été introduit par Gauss en 1822 pour résoudre le problème des moindres carrés, méthode introduite par Legendre en 1805. On peut obtenir ces équations par un calcul direct ou par un argument géométrique.

On veut résoudre le problème des moindres carrés pour trouver l'estimateur LLSE :

$$\widehat{\theta}_{\text{LLSE}}(X) = \arg\min_{\theta} \|X - \mathbf{H}\,\theta\|^2$$

où $\mathbf{H}$ est un matrice $N \times n$ de rang plein et $N > n$.

Méthode directe :

1. Montrer que la solution est caractérisée par la condition $\nabla_\theta \|X - \mathbf{H}\,\theta\|^2 = 0$ et en déduire qu'elle vérifie le système d'*équations normales* :
$$\mathbf{H}^t\mathbf{H}\,\theta = \mathbf{H}^t X$$
(où la matrice du système $\mathbf{H}^t\mathbf{H}$ est une matrice *normale*). En déduire l'expression du LLSE.
2. Discuter l'optimalité du LLSE dans le cas d'un modèle $X = \mathbf{H}\,\theta + W$ *normal* (où W suit une loi normale $\mathcal{N}(0, \sigma^2\mathbf{I})$).
3. Calculer le coût minimum.

Approche géométrique : Soit $\mathcal{V}$ le sous-espace de l'espace euclidien $\mathbb{R}^N$ engendré par les colonnes de $\mathbf{H}$.

4. Montrer que résoudre le problème des moindres carrés revient à projeter orthogonalement X sur $\mathcal{V}$.
5. En déduire la *condition d'orthogonalité* nécessaire et suffisante pour la solution $\widehat{\theta}$: l'erreur de modèle $X - \mathbf{H}\,\theta$ doit être *normale* au sous-espace $\mathcal{V}$.
6. Retrouver les équations normales, l'expression du LLSE et celle du coût minimum.

Exercice: moindres carrés et équations normales

Indications:

1. Pour alléger les calculs, procéder directement en matriciel avec les formules
$$\begin{cases} \nabla_\theta\left(\theta^t \mathbf{A}\theta\right) = 2\mathbf{A}\theta \\ \nabla_\theta\left(\theta^t \mathbf{A}\right) = \mathbf{A} \end{cases}$$
2. Cf. exercice p. 79.
3. Réinjecter l'expression de $\widehat{\theta}$ dans la fonction de coût $L(\theta) = \|X - \mathbf{H}\,\theta\|^2$.
4. On cherche à minimiser la distance euclidienne d'un point de $\mathcal{V}$ à X.
5. Appliquer le théorème de projection orthogonale (sur le sous-espace $\mathcal{V}$).
6. $X - \mathbf{H}\,\theta \perp \mathcal{V}$ implique que $X - \mathbf{H}\,\theta$ est orthogonal à toute colonne de $\mathbf{H}$.
   Un des deux $\mathbf{H}\widehat{\theta}$ dans l'expression du coût $L(\theta) = (X - \mathbf{H}\,\theta)^t(X - \mathbf{H}\,\theta)$ disparaît avec la condition d'orthogonalité, ce qui simplifie le calcul.

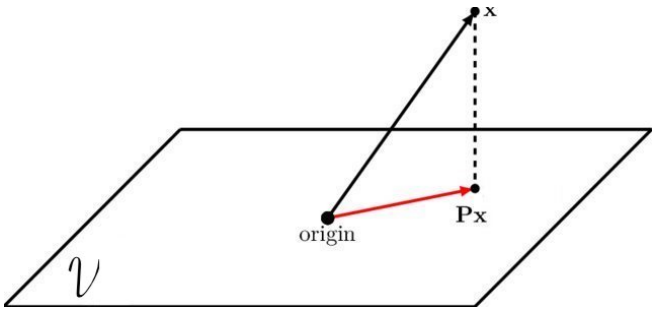


FIGURE 11.1: Projection orthogonale de l'observation sur un le sous-espace du modèle.

Exercice: estimation d'amplitude dans du bruit corrélé

On revient au modèle linéaire de signal dans du bruit (p. 27)

$$X = \theta S + W$$

Il est particulièrement instructif de comparer dans ce cas estimateurs MVU, ML, BLUE et (L)LSE.

On considère le modèle ci-dessus où $W \sim \mathcal{N}(0, \mathbf{C})$ est de matrice de covariance $\mathbf{C}$ inversible.

1. Justifier (sans calcul) l'expression de l'estimateur MVU et ML.
2. Calculer l'estimateur BLUE.
3. Calculer l'estimateur LSE pour la fonction de coût pondérée :

$$L(\theta) = (X - \theta S)^t \mathbf{C}^{-1} (X - \theta S)$$

4. Que donne le cas d'une fonction de coût non pondérée ?

Exercice: estimation d'amplitude dans du bruit corrélé

Indications:

1. Le modèle est le même que celui p. 79 où $N = 1$ et $\mathbf{H} = S$. Par conséquent l'estimateur efficace (MVU) est
$$\widehat{\theta}(X) = \frac{S^t \mathbf{C}^{-1} X}{S^t \mathbf{C}^{-1} S}$$
identique à l'estimateur ML (voir p. 91).
2. On doit minimiser $a^t \mathbf{C} a$ sous la contrainte de biais nul $a^t \mathbb{E}(X) = \theta$, par exemple avec la méthode du Lagrangien :
$$\mathcal{L}(a) = a^t \mathbf{C} a + \lambda a^t S$$
3. On dérive le coût quadratique $L(\theta)$ ce qui revient à calculer l'estimateur ML...
4. On obtient l'expression ci-dessus pour $\mathbf{C} = \mathbf{I}$. Ce LSE n'est pas optimal contrairement au précédent, mais le choix de $\mathbf{C}^{-1}$ comme matrice de pondération était arbitraire (ad-hoc).

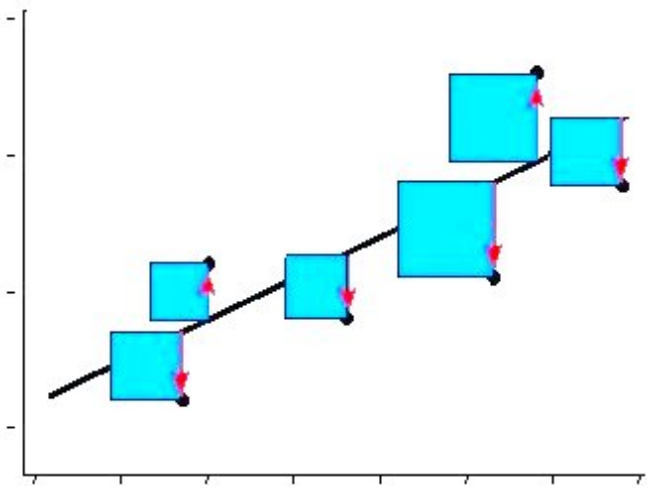


FIGURE 11.2: Illustration de l'estimation linéaire aux moindres carrés : on cherche la droite qui minimise la surface des carrés des erreurs.

12
Le point de vue bayesien

Loi a priori

En estimation bayesienne, la paramètre θ n'est plus déterministe, c'est une variable aléatoire. Sa loi est la loi (ou distribution) **a priori** [1], décrite par une densité $p(\theta)$ (par rapport à une mesure dominante ν).

1. *prior* en anglais

L'a priori n'est pas une probabilité au sens fréquentiste du terme, car il n'y a pas d'expérience aléatoire concernant θ ; c'est plutôt un degré de confiance (de plausibilité) qu'on lui accorde a priori.

Loi des données et vraisemblance

La loi de l'observation $X \sim p_\theta(x)$ correspondait en fréquentiste à une valeur de θ fixée. Elle devient en bayesien la loi conditionnelle de X sachant θ :

$$\boxed{p_\theta(x) = p(x|\theta)}$$

En résumé le modèle bayesien s'écrit

$$\begin{aligned} X \mid \theta &\sim p(x|\theta) \\ \theta &\sim p(\theta) \end{aligned}$$

La **vraisemblance** de θ est la fonction $\theta \mapsto p(x|\theta)$.

La loi **conjointe** $p(x, \theta) = p(\theta)p(x|\theta)$ est dominée par la mesure produit $\mu \otimes \nu$. La loi de X seule n'est plus $p_\theta(x)$, mais $p(x)$, une loi « inconditionnelle » (en moyenne sur l'a priori), qui s'obtient en marginalisant la loi conjointe :

$$p(x) = \int p(x, \theta) \, \mathrm{d}\nu(\theta) = \int p(x|\theta)p(\theta) \, \mathrm{d}\nu(\theta)$$

Loi a posteriori

L'inférence bayesienne consiste *a posteriori* (une fois les données X observées) à modifier l'a priori du modèle θ. La loi (ou distribution) **a posteriori** [2] est donnée par la *formule de Bayes* [3] :

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\,\mathrm{d}\nu(\theta)}$$

Il est important de noter qu'en pratique, il est souvent *inutile* de calculer l'intégrale figurant au dénominateur. On pose simplement que $p(\theta|x)$ est *proportionnelle à la loi conjointe* $p(x,\theta) = p(\theta)p(x|\theta)$:

$$\boxed{p(\theta|x) \propto p(\theta)p(x|\theta)}$$

où le facteur de proportionnalité ne dépend pas de θ, puis on normalise l'expression obtenue de façon à obtenir une distribution de probabilité en θ.

2. *posterior* en anglais

3. voir cours de probabilités : c'est la formule générale de Bayes sur des densités par rapport à la mesure dominante $\mu \otimes \nu$

Philiosophie bayesienne [4]

La donnée de l'**a priori** implique que θ n'est plus totalement inconnu : l'information a priori va changer (améliorer !) l'estimation. On obtient donc de « **meilleurs** » résultats qu'en fréquentiste (en particulier pour un nombre limité N d'observations) mais ces résultats dépendent forcément d'un a priori qui peut être plus ou moins **subjectif**...

L'approche bayesienne renverse en quelque sorte l'étude statistique : plutôt que d'observer l'« effet » X produit par une certaine « cause » θ (la **vraisemblance** $p_\theta(x) = p(x|\theta)$), on cherche la confiance à accorder à la cause hypothétique θ produite par une certain effet X (l'**a posteriori** $p(\theta|x)$).

Chaque nouvelle observation modifie notre a priori qui devient l'a posteriori, et à la limite quand le nombre d'observations est très grand, l'effet de l'a priori peut finir par s'évanouir. Les résultats de l'approche bayesienne, valables même lorsque N est relativement petit, finit souvent [5] par rejoindre les résultats donnés par l'approche fréquentiste lorsque $N \to +\infty$.

4. qui a fait couler beaucoup d'encre ! — on ne donne ici que quelques idées parmi les moins provocatrices

5. voir le théorème de Bernstein-von Mises p. 129

13
Erreur quadratique moyenne minimale

Dans l'approche bayesienne, l'estimateur $\widehat{\theta} = \widehat{\theta}(X)$ estime toujours θ, mais comme une réalisation d'une variable aléatoire : sa performance sera donc *moyennée sur l'a priori* $p(\theta)$ et ne dépendra plus de θ. Cela va considérablement simplifier la théorie mathématique des estimateurs optimaux.

Risque quadratique

L'erreur quadratique moyenne (MSE) d'un estimateur $\widehat{\theta}$ est, comme dans l'approche fréquentiste, donnée par l'expression (où « $|\cdot|$ » désigne la norme euclidienne) :

$$\boxed{R(\widehat{\theta}) = \mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\{|\widehat{\theta}(X) - \theta|^2\}}$$

sauf que l'espérance porte maintenant non seulement sur X, mais aussi sur θ :

$$\begin{aligned}\mathrm{MSE} &= \iint p(x,\theta)\ |\widehat{\theta}(x) - \theta|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\nu(\theta) \\ &= \int_\Theta p(\theta)\Big[\int p(x|\theta)|\theta - \widehat{\theta}(x)|^2 \,\mathrm{d}\mu(x)\Big]\,\mathrm{d}\nu(\theta)\end{aligned}$$

L'expression entre crochets correspond au MSE dans le cas fréquentiste où on a identifié $p_\theta(x) = p(x|\theta)$.

On peut réécrire le MSE à l'aide de la *distribution a posteriori* $p(\theta|x)$: puisque $p(x,\theta) = p(\theta|x)p(x)$ il vient [1] :

$$\mathrm{MSE} = \int p(x)\Big[\int_\Theta p(\theta|x)|\theta - \widehat{\theta}(x)|^2 \,\mathrm{d}\nu(\theta)\Big]\,\mathrm{d}\mu(x)$$

Cette réécriture fonctionne pour toute fonction de coût (pas seulement la norme euclidienne).

1. par Fubini-Tonelli

Estimateur MMSE

L'estimateur optimal minimise le MSE, il est appelé **estimateur MMSE**[2] $\widehat{\theta}_{\mathrm{MMSE}}(X)$. D'après l'expression ci-dessus du MSE, trouver $\widehat{\theta}_{\mathrm{MMSE}}(x)$ revient, pour tout x fixé, à minimiser en $\widehat{\theta}$ la forme quadratique

$$\int p(\theta|x)|\widehat{\theta}-\theta|^2 = |\widehat{\theta}|^2 - 2\,\widehat{\theta}\cdot\int \theta\, p(\theta|x) + \int |\theta|^2 p(\theta|x)$$

où « $\cdot$ » désigne le produit scalaire euclidien. Le minimum s'obtient donc en annulant le gradient, d'où

$$\boxed{\widehat{\theta}_{\mathrm{MMSE}}(X) = \int_\Theta \theta\, p(\theta|X)\,\mathrm{d}\nu(\theta) = \mathbb{E}(\theta|X)}$$

L'estimateur optimal (MMSE) *existe* toujours[3] et est *unique* : c'est l'**espérance conditionnelle** du paramètre θ sachant les données observées X.

En d'autres termes le MMSE est la **moyenne de l'a posteriori** : Pour le calculer il suffit de déterminer d'abord l'a posteriori, puis d'identifier sa moyenne comme une fonction de X.

2. *minimum mean-squared error*

3. à la seule condition (satisfaite en pratique) que la loi a priori $p(\theta)$ admette une espérance (c'est-à-dire que la v.a. θ soit intégrable), pour que l'espérance conditionnelle soit bien définie

RÉSUMÉ:

- *modèle bayesien* : $X \mid \theta \sim p(x|\theta)$
- *vraisemblance* $\theta \mapsto p(x|\theta)$
- *a priori* : $p(\theta)$
- *a posteriori* : $p(\theta|x)$ déterminé à un facteur près en multipliant a priori et vraisemblance :

$$p(\theta|x) \propto p(\theta)p(x|\theta)$$

- *estimateur MMSE (optimal)* $\mathbb{E}(\theta|X)$ moyenne de l'a posteriori.
- *MMSE (MSE minimum)* $\mathbb{E}(\mathbb{V}(\theta|X))$ variance (moyenne) de l'a posteriori [cf. exercice p. 121].

Exercice: loi de la variance totale

Au delà de la loi (ou formule) bien connue des probabilités totales, qui se généralise en loi de l'espérance totale, on a également une « loi de la variance totale » très importante pour identifier la proportion de variance expliquée et celle inexpliquée par un estimateur optimal. On l'appelle aussi loi d'$\mathbb{E}\mathbb{V}\mathbb{E}$ *à cause de la forme* $\mathbb{V} = \mathbb{E}\mathbb{V} + \mathbb{V}\mathbb{E}$.

On considère un modèle bayesien scalaire ($\theta \in \mathbb{R}$). Le MMSE $\widehat{\theta}(X)$ est égal à l'*espérance conditionnelle* $\mathbb{E}(\theta|X)$. On définit la *variance conditionnelle* :

$$\mathbb{V}(\theta|X) = \mathbb{E}\Big\{ \big[\theta - \mathbb{E}(\theta|X)\big]^2 \,\Big|\, X \Big\}$$

qui est, comme l'espérance conditionnelle, une variable aléatoire fonction de X.

1. *Loi de l'espérance totale* : Montrer que l'estimateur MMSE est de biais moyen nul.
2. Montrer que l'erreur quadratique moyenne minimale (également notée MMSE !) vaut

$$\text{MMSE} = \mathbb{E}\big\{\mathbb{V}(\theta|X)\big\}$$

3. *Loi de la variance totale ou loi d'Ève* : Montrer que

$$\mathbb{V}(\theta) = \mathbb{E}\big\{\mathbb{V}(\theta|X)\big\} + \mathbb{V}\big\{\mathbb{E}(\theta|X)\big\}$$

   et expliquer la signification de chacun des deux termes :

   - variance inexpliquée (intra-classes)
   - variance expliquée (inter-classes)

Exercice: loi de la variance totale

Indications:

1. Évaluer $\mathbb{E}(\mathbb{E}(\theta|X))$ d'après la loi de l'espérance totale.
2. Justifier que MMSE $= \mathbb{E}(\mathbb{E}((\theta - \widehat{\theta})^2|X))$ d'après la loi de l'espérance totale.
3. Développer $\mathbb{V}(\theta)$ en introduisant $\widehat{\theta}$ dans l'expression $\theta - \mathbb{E}(\theta) = \theta - \widehat{\theta} + \widehat{\theta} - \mathbb{E}(\theta)$, et utiliser les deux questions précédentes. La variance non expliquée (par l'estimateur) est le MSE résiduel.

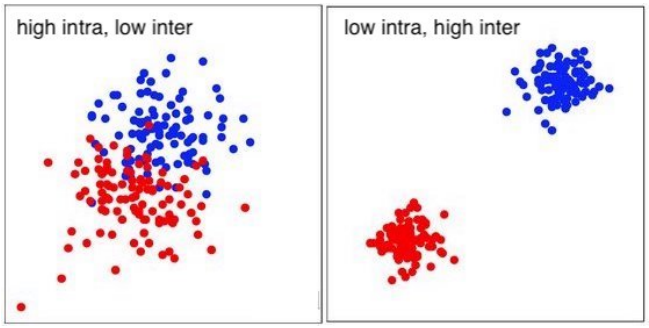


FIGURE 13.1: Variances intra-classes et inter-classes pour deux valeurs de X (classes) et une variance de θ donnée.

Exercice: a priori normal en estimation d'amplitude dans du bruit gaussien

On revient au modèle $\mathcal{N}\theta, \sigma^2)$ d'estimation d'amplitude dans du bruit additif gaussien (cf. p. 27) avec une propriété dite de conjugaison *qui simplifie les calculs en permettant une inférence exacte et analytique. Ici, les données font passer d'un a priori normal à un a posteriori également normal.*

On considère le modèle bayesien i.i.d. $X = \theta + W$ avec un a priori normal $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$ indépendant du bruit additif gaussien $W \sim \mathcal{N}(0, \sigma^2)$.

1. Montrer que l'a priori *normal* donne un a posteriori *normal*. (On dit alors que la loi normale est *conjuguée* du modèle.) Quel est l'intérêt ?
2. Trouver l'estimateur MMSE.
3. En déduire sa variance et le MMSE (MSE minimal).
4. Comparer ses performances à l'estimateur fréquentiste $\overline{X}_N$. Que peut-on en dire en "très faible a priori" ?.

Exercice: a priori normal en estimation d'amplitude dans du bruit gaussien

Indications:

1. Calculer $\log p(\theta)p(x|\theta)$ en ne gardant que les termes en θ. L'intérêt d'une loi conjuguée est qu'il suffit de mettre à jour les paramètres.

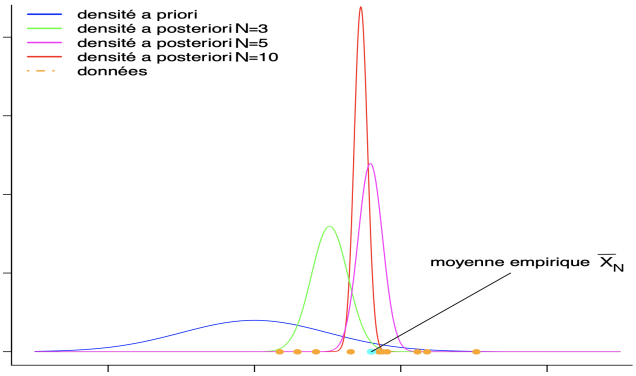


FIGURE 13.2: Densités normales a priori et a posteriori.

2. La moyenne de l'a posteriori donne l'estimateur MMSE.
3. La variance de l'a posteriori donne le MMSE (MSE minimal) puis pour simplifier le calcul on peut utiliser la loi de la variance totale (p. 121) pour trouver la variance de l'estimateur MMSE.
4. Que se passe-t-il si $\sigma_\theta \to \infty$?

14
Maximum a posteriori

Détection bayesienne

Toujours dans l'approche bayesienne, lorsque θ prend des valeurs discrètes (typiquement un ensemble fini de valeurs) on peut aussi utiliser le critère de **probabilité d'erreur moyenne** :

$$\boxed{R(\widehat{\theta}) = \mathbb{P}_e = \mathbb{P}_e(\widehat{\theta}(X) \neq \theta) = \mathbb{E}\{1_{\widehat{\theta}(X)\neq\theta}\}}$$

où le coût $1_{\widehat{\theta}\neq\theta}$ vaut 1 si $\widehat{\theta}(X) \neq \theta$ et 0 sinon, et où l'espérance l'espérance porte conjointement sur X *et* θ.

Comme dans le cas du MSE on peut réécrire [1] cette expression à l'aide de la distribution a posteriori $p(\theta|x)$:

1. par Fubini-Tonelli

$$\begin{aligned}\mathbb{P}_e &= \sum_\theta p(\theta) \int 1_{\widehat{\theta}(x)\neq\theta} p(x|\theta)\,\mathrm{d}\mu(x) \\ &= \int p(x) \underbrace{\Big[\sum_\theta 1_{\theta\neq\widehat{\theta}(x)} p(\theta|x)\Big]}_{1-p(\widehat{\theta}(x)|x)} \mathrm{d}\mu(x)\end{aligned}$$

Estimateur MAP

D'après l'expression ci-dessus de $\mathbb{P}_e$, trouver l'estimateur optimal revient, pour tout x fixé, à minimiser $1 - p(\theta|x)$, c'est-à-dire *maximiser la probabilité a posteriori* :

$$\boxed{\widehat{\theta}_{\mathrm{MAP}}(X) = \arg\max_\theta p(\theta|X)}$$

Cet estimateur est connu sous le nom d'estimateur MAP [2]. La probabilité de succès maximisée $\mathbb{P}_s$ a alors pour expression $\mathbb{P}_s = 1 - \mathbb{P}_e = \mathbb{E}\max_\theta p(\theta|X)$.

2. maximum a posteriori probability

Lorsque θ n'est plus discret mais continu, on peut obtenir l'estimateur bayesien MAP par un passage à la limite quand $\varepsilon \to 0$ du risque :

$$\begin{aligned} R_\varepsilon(\widehat{\theta}) &= \mathbb{E}\{1_{|\widehat{\theta}(X)-\theta|>\varepsilon}\} \\ &= \int p(x)\left[1 - \int_{|\theta-\widehat{\theta}|\leqslant\varepsilon} p(\theta|x)\,\mathrm{d}\nu(\theta)\right]\mathrm{d}\mu(x) \end{aligned}$$

Minimiser $R_\varepsilon(\widehat{\theta})$ revient alors à maximiser l'intégrale figurant entre crochets. Si la distribution a posteriori est continue, cela revient à la limite quand $\varepsilon \to 0$ à l'expression usuelle $\widehat{\theta}_{\mathrm{MAP}}(X) = \arg\max_\theta p(\theta|X)$ de l'estimateur MAP.

Mode vs. moyenne

Dans tous les cas l'estimateur MAP n'est autre que le **mode de la distribution a posteriori**, qui est souvent plus simple à calculer que le MMSE (la moyenne de la distribution a posteriori).

Naturellement, *si* la distribution a posteriori $p(\theta|x)$ est symétrique et unimodale (exemple : le cas gaussien) alors mode = moyenne et $\widehat{\theta}_{\mathrm{MAP}} = \widehat{\theta}_{\mathrm{MMSE}}$.

Estimateur MAP vs. ML

Dans le cas où l'a priori $p(\theta)$ est **uniforme** (ou plus généralement si $p(\theta)$ est constante autour du maximum de l'a posteriori pour tout X) alors maximiser

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

sur θ revient simplement à maximiser $p(X|\theta)$ et on retrouve la définition de l'estimateur (fréquentiste) du maximum de vraisemblance :

$$\boxed{\widehat{\theta}_{\mathrm{ML}}(X) = \arg\max_{\theta\in\Theta} p(X|\theta)}$$

qui ne dépend plus de l'a priori.

Il en résulte que comme dans le cas de l'estimateur ML, l'estimateur MAP peut ne pas exister ou ne pas être unique.

Contrairement à l'estimateur ML, l'estimateur MAP n'est pas toujours invariant par reparamétrisation $\alpha = g(\theta)$ à cause de la présence de l'a priori qui n'est pas nécessairement invariant par ce changement de variable [3].

3. voir exercice p. 115

Propriétés asymptotiques pour un modèle i.i.d.

On se place dans le cas d'un modèle i.i.d., sous certaines conditions techniques (non précisées ici).

En fréquentiste, on a vu que l'estimateur ML est consistant et asymptotiquement normal. En bayesien, on établit des propriétés analogues pour la distribution a posteriori de $\theta \mid X$ pour N observations $X = (X_1, X_2, \ldots, X_N)$ lorsque $N \to +\infty$.

Pour cela, on se place dans le cas **fréquentiste** en supposant que

$$X \sim p_{\theta^*}$$

où θ^* est le « **vrai** » paramètre et on suppose que la distribution a priori $p(\theta)$ choisie **contient θ^* dans son support** [4] :

$$\theta^* \in \operatorname{Supp} \theta$$

4. sinon, il n'est clairement pas possible d'estimer correctement θ^* !

- *Consistance :* Sous ces conditions, on a, pour presque tout X, la convergence en probabilité [5]

$$\theta \mid X \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*$$

5. voir l'exercice p. 139

Il en résulte en pratique que les estimateurs bayesiens (MMSE, MAP,...) seront consistants (convergents vers le vrai paramètre θ^*).

- *Normalité asymptotique :*

Théorème de Bernstein-von Mises (1917, 1931). [6] *Sous les conditions ci-dessus, on a approximativement en loi* [7]

$$\theta \mid X \approx \mathcal{N}(\widehat{\theta}_{\mathrm{ML}}, \frac{1}{J_{\theta^*}} = \frac{1}{N J_{\theta^*,1}})$$

6. voir l'exercice p. 141 pour une esquisse de preuve

7. plus précisément, la différence des deux lois tend vers zéro en variation totale quand $N \to +\infty$

Cela signifie qu'un estimateur bayesien va se comporter asymptotiquement comme l'estimateur ML : *plus* N augmente, *moins* l'a priori a d'influence sur l'estimation bayesienne qui devient asymptotiquement **optimale** (ce qui peut justifier « a posteriori » son intérêt !).

RÉSUMÉ:

- *estimateur MAP :* maximise la probabilité a posteriori :

$$\widehat{\theta}_{\text{MAP}}(X) = \arg\max_{\theta} p(\theta|X)$$

- *si θ est discret,* $\widehat{\theta}_{\text{MAP}}$ minimise la probabilité d'erreur moyenne $\mathbb{P}_e$.

- *lien avec l'estimateur MMSE :*
  $\widehat{\theta}_{\text{MAP}}$ = *mode* de l'a posteriori
  $\widehat{\theta}_{\text{MMSE}}$ = *moyenne* de l'a posteriori

- *lien avec l'estimateur ML :*
  $\widehat{\theta}_{\text{MAP}} = \widehat{\theta}_{\text{ML}}$ pour un a priori *uniforme*

Pour un modèle i.i.d., quand $N \to +\infty$, si le vrai paramètre θ^* est dans le support de la distribution a priori (+ certaines autres conditions), la distribution *a posteriori* est :

- *consistante :* $\theta \mid X \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*$ p.s.

- *asymptotiquement normale* et se comporte comme l'estimateur (fréquentiste) ML.

Exercice: comparaison entre estimateurs MMSE, MAP et ML

Il est très instructif de comparer les performances des estimateurs bayesiens à l'estimateur fréquentiste ML, notamment pour ses propriétés asymptotiques quand le nombre d'observations $N \to +\infty$.

Trouver les estimateurs MMSE, MAP et ML et comparer leurs performances asymptotiques pour les modèles i.i.d. bayesiens suivants :

1. $X \,|\, \theta \sim \mathcal{N}(\theta, \sigma^2)$ et $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$
   Justifier ici le théorème de Bernstein-von Mises.
2. $X \,|\, \theta \sim \mathcal{N}(\mu, \theta)$ et $\theta \sim \Gamma^{-1}(\alpha, \beta)$ (loi inverse gamma)
3. $X \,|\, \theta \sim \mathcal{E}(\theta)$ et $\theta \sim \Gamma(\alpha, \beta)$ (loi gamma)
4. $X \,|\, \theta \sim \mathcal{E}(\frac{1}{\theta})$ et $\theta \sim \Gamma^{-1}(\alpha, \beta)$
5. $X \,|\, \theta \sim \mathcal{B}(\theta)$ et $\theta \sim \mathrm{B}(\alpha, \beta)$ (loi bêta)
6. $X \,|\, \theta \sim \mathcal{P}(\theta)$ (loi de Poisson) et $\theta \sim \Gamma(\alpha, \beta)$
7. $X \,|\, \theta \sim \mathcal{N}(\mu, \theta)$ et $\theta \sim \mathcal{U}[-A, A]$.
   Vérifier sur cet exemple que :
   - le MAP est plus simple à calculer que le MMSE
   - le MAP est toujours meilleur que le ML.

Exercice: comparaison entre estimateurs MMSE, MAP et ML

Indications:

On calcule facilement pour les lois usuelles :

loi	moyenne	mode
$\mathcal{N}(\mu,\sigma^2)$	μ	μ
$\mathrm{B}(\alpha,\beta)$	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha-1}{\alpha+\beta-2}$

loi	moyenne	mode
$\Gamma(\alpha,\beta)$	$\dfrac{\alpha}{\beta}$	$\dfrac{\alpha-1}{\beta}$
$\Gamma^{-1}(\alpha,\beta)$	$\dfrac{\beta}{\alpha-1}$	$\dfrac{\beta}{\alpha+1}$

où α, β sont telles que moyennes et modes sont > 0.

1. Voir p. 123. Ici l'a posteriori est purement normal !
2. La loi inverse gamma est conjuguée du modèle.
   Voir p. 91 pour le ML.
3. La loi gamma est conjuguée du modèle.
4. La loi inverse gamma est conjuguée du modèle.
5. La loi bêta est conjuguée du modèle, voir p. 125.
6. La loi gamma est conjuguée du modèle.
7. Ici la loi uniforme n'est *pas* conjuguée du modèle et les calculs sont plus difficiles ! $\widehat{\theta}_{\mathrm{MAP}} = \widehat{\theta}_{\mathrm{ML}}$ dans l'intervalle $[-A, A]$, mais pas en dehors.

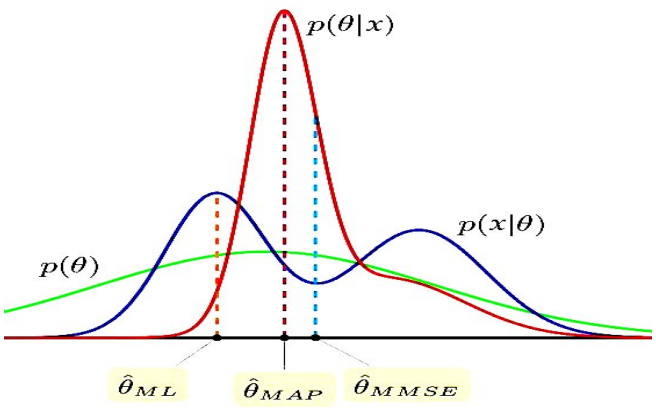


FIGURE 14.1: Comparaison entre MMSE, MAP et ML

Exercice: comparaison entre estimations fréquentiste et bayesienne

Un estimateur MAP peut être meilleur qu'un estimateur ML lorsque l'a priori est bien choisi. Mais si l'a priori est contraire à l'observation effective, l'estimation bayesienne peut être absurde.

On joue N fois à pile (1) ou face (0) et obtient k 'piles'. On veut estimer $\theta = \mathbb{P}(\text{pile})$.

1. Préciser l'estimateur ML.
2. Donner une expression de l'estimateur MMSE pour un a priori $\theta \in [\frac{1}{4}, \frac{3}{4}]$.
3. Trouver l'estimateur MAP pour le même a priori.
4. On observe que des pile. Comparer les valeurs exactes des MMSE, MAP et ML. Cela contredit-il le théorème de Bernstein-von Mises ?

Exercice: comparaison entre estimations fréquentiste et bayesienne

Indications:

1. Voir par exemple p. 91.
2. L'expression intégrale à trouver n'est pas explicitement calculable en général.
3. Ici le calcul est beaucoup plus simple et similaire à celui du ML.
4. Dans un cadre fréquentiste, que peut-on dire sur le support de l'a priori ?

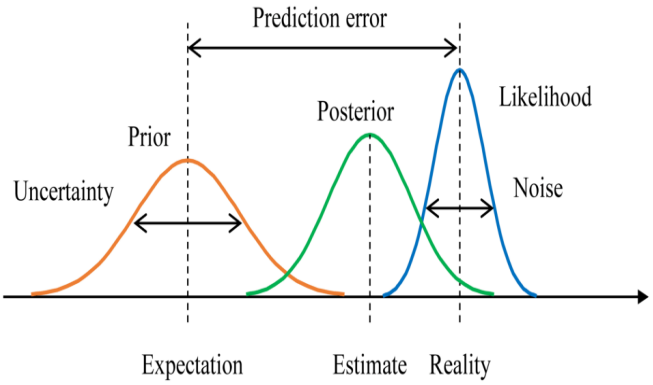


FIGURE 14.2: Effet d'une erreur de modèle a priori

Exercice: inégalité de van Trees

*La borne de van Trees (1968) est une version bayesienne de la borne de Cramér-Rao. On suppose toujours le modèle $p(x|\theta) = p_\theta(x)$ régulier en θ. L'**information de Fisher bayesienne** est*

$$J_\theta = \mathbb{V}\Big(\frac{\partial \log p(\theta, X)}{\partial \theta}\Big)$$

où $p(\theta, x)$ est la loi conjointe du modèle bayesien.

1. Justifier que J_θ est bien définie est vaut également $J_\theta = -\mathbb{E}\Big(\frac{\partial^2 \log p(X,\theta)}{\partial \theta^2}\Big)$.

On note $\mathbb{B}_\theta(\widehat{\theta})$ le biais d'un estimateur $\widehat{\theta}$, $\mathbb{B}(\widehat{\theta}) = \mathbb{E}_\theta \mathbb{B}_\theta(\widehat{\theta})$ son biais moyen, et $R(\widehat{\theta}) = \mathbb{E}(|\widehat{\theta} - \theta|^2)$ son risque quadratique moyen. On suppose dorénavant que $p(\theta)B_\theta$ *tend vers zéro à l'infini* ($|\theta| \to +\infty$).

2. Cette condition est-elle très restrictive pour un apriori gaussien ou uniforme ?
3. Démontrer que $\mathbb{E}\big(\frac{\partial \log p(x,\theta)}{\partial \theta} \cdot (\theta - \widehat{\theta})\big) = 1$.
4. En déduire l'inégalité de van Trees (ou borne de Cramér-Rao bayesienne) : $\boxed{R(\widehat{\theta}) \geqslant 1/J_\theta}$. Commenter en comparant au cas fréquentiste.

L'estimateur bayesien $\widehat{\theta}$ est **efficace** si $R(\widehat{\theta}) = 1/J_\theta$.

5. Montrer qu'un estimateur efficace est nécessairement l'estimateur MMSE.
6. En examinant le cas d'égalité $R(\widehat{\theta}) = 1/J_\theta$ montrer qu'un estimateur efficace est nécessairement l'estimateur MAP (et donc, dans ce cas, MAP = MMSE).
7. Démontrer que pour un estimateur efficace, l'a posteriori $p(\theta|x)$ est nécessairement *normal* (gaussien).

Exercice: inégalité de van Trees

Indications:

1. Même calcul qu'en fréquentiste, sur $p(x, \theta)$ au lieu de $p_\theta(x) = p(x|\theta)$.
2. Si l'a priori $p(\theta)$ est à décroissance rapide il suffit que le biais n'explose pas à l'infini, ce qui est très raisonnable.
3. Dériver par rapport à θ l'intégrale donnant $p(\theta)B_\theta$, puis l'intégrer entre $-\infty$ et $+\infty$.
4. Comme en fréquentiste, par l'inégalité de Cauchy-Schwarz. On a pas supposé que le biais est nul, et la condition sur le biais est peu restrictive.
5. L'estimateur MMSE minimise le risque moyen $R(\widehat{\theta})$ qui est toujours $\geqslant \frac{1}{J_\theta}$.
6. Cas d'égalité dans l'inégalité de Cauchy-Schwarz, à écrire sur l'a posteriori. Or l'estimateur MAP est le mode de l'a posteriori.
7. Dériver à nouveau la condition d'égalité pour obtenir une équation différentielle du 2e ordre.

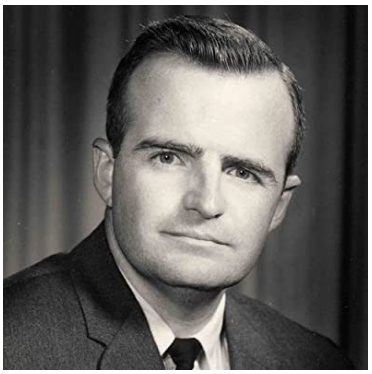


FIGURE 14.3: Harry Leslie Van Trees

Exercice: consistance de l'estimation bayesienne

Le théorème de consistance de Doob (1949) établit que l'a posteriori $\theta \mid X$ est consistant lorsque $X \sim p_{\theta^*}$ pour presque tout θ^* tirée au hasard selon la loi a priori $p(\theta)$. Ce théorème très général est une application de la théorie des martingales. On montre ici une version plus spécifique applicable à un modèle i.i.d. régulier en s'appuyant sur la consistance de l'estimateur ML.

On se place dans le cadre fréquentiste avec un modèle i.i.d. $X \sim p_{\theta^*}$ (θ^* est le vrai paramètre) et on considère une distribution (a priori) quelconque $p(\theta)$ qui contient θ^* dans son support.

1. Montrer que $p(\theta|X) \propto p(\theta)e^{-N\cdot\overline{\ell}_X(\theta)}$ où on a posé $\overline{\ell}_X(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}$.
2. En s'appuyant sur la consistance de l'estimateur ML (p. 95), montrer que $\overline{\ell}_X(\theta) \to D(p_{\theta^*}\|p_\theta)$ p.s. quand $N \to +\infty$.
3. En déduire que si $|\theta - \theta^*| > \varepsilon$, on a $\overline{\ell}_X(\theta) > \eta$ p.s. pour N assez grand et un certain $\eta > 0$; et que si $0 < \eta' < \eta$, on aura $\overline{\ell}_X(\theta) < \eta'$ p.s. si N est assez grand dès que $|\theta - \theta^*| < \varepsilon'$ pour $\varepsilon' < \varepsilon$ assez petit.
4. En déduire par majoration que le rapport $\dfrac{\mathbb{P}(|\theta - \theta^*| > \varepsilon \mid X)}{\mathbb{P}(|\theta - \theta^*| < \varepsilon' \mid X)} \to 0$ p.s. quand $N \to +\infty$.
5. Conclure : $\theta \mid X \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*$ p.s.

Exercice: consistance de l'estimation bayesienne

Indications:

1. Utiliser le fait que le modèle est i.i.d. ; proportionnalité à un facteur près ne dépendant pas de θ.
2. Loi (forte !) des grands nombres.
3. Utiliser la question précédente et le fait que, comme le modèle est identifiable, la divergence $D(p_{\theta^*} \| p_\theta) \geqslant 0$, qui est continue en θ, ne s'annule qu'en $\theta = \theta^*$.
4. Remarquer que comme θ^* est dans le support de θ, $\mathbb{P}(|\theta - \theta^*| < \varepsilon') > 0$.
5. Si $\beta = \frac{\alpha}{1-\alpha}$ alors $\alpha = \frac{1}{1+\frac{1}{\beta}}$.

FIGURE 14.4: Joseph Leo Doob

Exercice: théorème de Bernstein-von Mises

Cet exercice donne une esquisse de preuve informelle du théorème de Bernstein-vonMises. Ce théorème dit en substance que la distribution a priori influence à peine la distribution a posteriori quand il y a suffisamment de données observées. Ceci peut s'interpréter à la fois comme une affirmation de la supériorité du fréquentiste (le choix éventuel d'un a priori finit par être sans importance en pratique) et une défense du concept bayesien (dans un cadre fréquentiste, l'estimation bayesienne est asymptotiquement optimale).

On se place dans le cadre fréquentiste avec un modèle régulier i.i.d. $X \sim p_{\theta^*}$ (θ^* est le vrai paramètre) et on considère une distribution a priori quelconque $p(\theta)$ qui contient θ^* dans son support. On étudie des propriétés asymptotiques quand $N \to +\infty$.

1. On pose $h = \sqrt{N}(\theta - \theta^*)$. Montrer que $\log p_{\theta^* + \frac{h}{\sqrt{N}}}(x) = \log p_{\theta^*}(x) + h \cdot \frac{1}{\sqrt{N}} S_{\theta^*}(X) - \frac{h^2}{2} J_{\theta^*,1} + r_N$ où $r_N \xrightarrow{\mathbb{P}_{\theta^*}} 0$.
2. En déduire que la densité de $h|X$ est approximativement $\mathcal{N}(J_{\theta^*,1}^{-1} \frac{1}{\sqrt{N}} S_{\theta^*}(X), J_{\theta^*,1}^{-1})$.
3. En s'appuyant sur la normalité asymptotique de l'estimateur ML (p. 97), justifier que $\sqrt{N}(\widehat{\theta}_{\mathrm{ML}} - \theta^*) - J_{\theta^*,1}^{-1} \frac{1}{\sqrt{N}} S_{\theta^*}(X) \xrightarrow{\mathbb{P}_{\theta^*}} 0$.
4. Conclure : $\theta|X \approx \mathcal{N}(\widehat{\theta}_{\mathrm{ML}}, \frac{1}{N J_{\theta^*,1}})$ approximativement en loi.

Exercice: théorème de Bernstein-von Mises

Indications:

1. Développement de Taylor à l'ordre 2 (cf. p.97).
2. Pour $N \to +\infty$ le terme a priori $p(\theta = \theta^* + \frac{h}{\sqrt{N}})$ est approximativement constant $\approx p(\theta^*)$ sous des hypothèses de régularité.
3. Réexaminer le développement de Taylor fait à l'exercice p. 97.
4. On en conclut que $h|X \approx \mathcal{N}(\sqrt{N}(\widehat{\theta}_{\mathrm{ML}} - \theta^*), J^{-1}_{\theta^*,1})$ approximativement en loi, puis on divise par $\sqrt{N}$ et on rajoute θ^*.

(a) Sergueï Natanovitch Bernstein (b) Richard von Mises

FIGURE 14.5: Bernstein et von Mises

15
Estimation bayesienne linéaire

Dans l'approche classique (fréquentiste) le MVU est souvent difficile à déterminer et requiert la connaissance complète de la distribution de données $p_\theta(x)$; on obtient alors une approche plus pragmatique en se restreignant à un estimateur *linéaire*, le BLUE, qui ne nécessite que les moments d'ordre 1 et 2 des données.

Dans l'approche bayesienne, de façon similaire, le MMSE est souvent difficile à calculer et requiert la connaissance complète de la distribution conjointe $p(\theta, x)$; on obtient alors une approche plus pragmatique en se restreignant à un estimateur *linéaire* qui ne nécessite que les moments d'ordre 1 et 2 des données :

Estimateur LMMSE

On le cherche sous la forme linéaire[1] suivante :

1. Selon la tradition française on devrait dire « affine »

$$\widehat{\theta}(X) = a^t \cdot X + b$$

dans le cas scalaire, où a (à déterminer) et X sont des vecteurs colonnes de N coefficients. Dans le cas vectoriel $\widehat{\theta}(X) = \mathbf{A}X + b$ où $\mathbf{A}$ est une matrice $n \times N$ et b un vecteur colonne de n coefficients.

On se limite dans la suite au cas scalaire.

L'estimateur linéaire optimal qui minimise l'erreur quadratique moyenne :

$$\boxed{\min_{a,b} \mathrm{MSE} = \min_{a,b} \mathbb{E}\left\{ \left| \theta - a^t \cdot X - b \right|^2 \right\}}$$

est appelé **estimateur LMMSE**[2].

2. linear minimum mean-squared error

Cas centré

On peut facilement déterminer la constante optimale b. En effet, par un calcul similaire à celui fait pour le compromis biais-variance,

$$\begin{aligned}\mathrm{MSE} &= \mathbb{E}\left\{\left|\theta - \mathbb{E}(\theta) - a^t(X - \mathbb{E}(X)) + (\mathbb{E}(\theta) - a^t\mathbb{E}(X) - b)\right|^2\right\} \\ &= \mathbb{E}\left\{\left|\underbrace{\theta - \mathbb{E}(\theta)}_{\text{centré}} - a^t(\underbrace{X - \mathbb{E}(X)}_{\text{centré}})\right|^2\right\} + \left|\underbrace{\mathbb{E}(\theta) - a^t\mathbb{E}(X) - b}_{\text{biais moyen}}\right|^2\end{aligned}$$

où le terme croisé s'annule. Le premier terme ne dépend pas de b, et le deuxième (le « biais moyen ») s'annule pour

$$b = \mathbb{E}(\theta) - a^t \cdot \mathbb{E}(X)$$

On peut ainsi se ramener au cas **centré** où $\mathbb{E}(X) = 0$ et $\mathbb{E}(\theta) = 0$.

Il ne reste alors qu'à résoudre un système de **moindres carrés moyens** (LMS [3])

3. least mean squares

$$\boxed{\min_a \mathbb{E}\left\{\left|\theta - a^t \cdot X\right|^2\right\}}$$

similaire au système des moindres carrés pour le LLSE. On aboutit à des équations « normales » de type « **Wiener-Hopf** » [4]

4. soit par un calcul direct, soit par un argument géométrique, voir exercice p. 147

$$\mathbf{R}_{XX} \cdot a = \mathbf{R}_{X\theta}$$

qui donnent l'estimateur linéaire LMMSE :

$$\boxed{\widehat{\theta}_{\mathrm{LMMSE}} = \mathbf{R}_{\theta X} \cdot \mathbf{R}_{XX}^{-1} \cdot X}$$

où $\mathbf{R}_{XX} = \mathbb{E}(XX^t)$ est la matrice $N \times N$ d'autocorrélation des données (supposée inversible) et $\mathbf{R}_{\theta X} = (\mathbf{R}_{X\theta})^t = \mathbb{E}(\theta X^t)$ est un vecteur ligne d'intercorrélation.

Exercice: moindres carrés moyens et équations de Wiener-Hopf

Originellement résolues analytiquement par Wiener et Hopf en 1931 sous forme intégrale, les équations de Wiener-Hopf (discrétisées) interviennent dans de très nombreux domaines des sciences.

On veut résoudre le problème des moindres carrés moyens (pour $\theta \in \mathbb{R}$ scalaire) :

$$\min_a \mathbb{E}\left\{\left|\theta - a^t \cdot X\right|^2\right\}$$

où X, a sont des vecteurs colonne de taille N.

Méthode directe :

1. Montrer que la solution est caractérisée par la condition $\nabla_a \mathbb{E}\left\{\left|\theta - a^t \cdot X\right|^2\right\} = 0$ et en déduire qu'elle vérifie le système d'*équations de Wiener-Hopf* : $\mathbf{R}_{XX} \cdot a = \mathbf{R}_{X\theta}$.
   En déduire l'expression du LMMSE dans le cas où X et θ sont centrés, puis dans le cas général.
2. Calculer le MSE minimum correspondant.
3. Discuter l'optimalité du LMMSE dans le cas d'un modèle gaussien $X = \theta + W$ où $W \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ et $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$.

Approche géométrique : Soit $\mathcal{V}$ le sous-espace de L^2 engendré par $X_1, X_2, \ldots, X_N$ (*espace d'observation*).

4. Montrer que résoudre le problème des moindres carrés revient à projeter orthogonalement θ sur $\mathcal{V}$.
5. En déduire la *condition d'orthogonalité* nécessaire et suffisante pour la solution a : l'erreur $\theta - a^t \cdot X$ doit être *orthogonale* au sous-espace $\mathcal{V}$.
6. Retrouver les équations de Wiener-Hopf, l'expression du LMMSE et du MSE minimum correspondant.

Exercice: moindres carrés moyens et équations de Wiener-Hopf

Indications:

1. Écrire $\left|\theta - a^t \cdot X\right|^2 = (\theta - a^t \cdot X)(\theta - a^t \cdot X)^t$, développer et utiliser les formules
$$\begin{cases} \nabla_a\left(a^t \mathbf{R} a\right) = 2\mathbf{R}a \\ \nabla_a\left(a^t \mathbf{R}\right) = \mathbf{R} \end{cases}$$
Dans le cas général, il suffit de remplacer θ par $\theta - \mathbb{E}(\theta)$ et X par $X - \mathbb{E}(X)$.
2. Réinjecter l'expression de la solution a dans celle du MSE.
3. Voir p. 123.
4. On minimise la distance L^2 d'un point de $\mathcal{V}$ à θ.
5. Appliquer le théorème de projection orthogonale sur le sous-espace $\mathcal{V}$.
6. $\theta - a^t \cdot X \perp \mathcal{V}$ implique que $\theta - a^t \cdot X$ est orthogonal à toute composante de X.
Un des deux $a^t \cdot X$ dans l'expression du MSE $= \mathbb{E}\{(\theta - a^t \cdot X)(\theta - a^t \cdot X)^t\}$ disparaît avec la condition d'orthogonalité, ce qui simplifie le calcul.

(a) Norbert Wiener vers 1930 (b) Eberhard Frederich Ferdinand Hopf

FIGURE 15.1: Wiener et Hopf

Exercice: comparaison entre estimateurs LMMSE et ML

On considère le modèle i.i.d. d'amplitude dans du bruit additif gaussien $X = \theta + W$, $W \sim \mathcal{N}(0, \sigma^2)$, avec un a priori uniforme $\theta \sim \mathcal{U}[-A, A]$.

1. Rappeler l'expression de l'estimateur ML et le MSE associé.

On cherche à déterminer l'expression du LMMSE.

1re méthode : On applique la formule du cours $\widehat{\theta}_{\text{LMMSE}} = \mathbf{R}_{\theta X} \cdot \mathbf{R}_{XX}^{-1} \cdot X$.

2. Calculer le LMMSE par cette méthode.

2e méthode : On résout directement les équations de Wiener-Hopf $\mathbf{R}_{XX} \cdot a = \mathbf{R}_{X\theta}$.

3. Calculer le LLMSE par cette méthode en explicitant les N équations à résoudre.
4. Comparer sur le MSE les deux estimateurs LMMSE et ML.

Exercice: comparaison entre estimateurs LMMSE et ML

Indications:

1. Voir p. 91.
2. Utiliser le lemme d'inversion matricielle (formule de Woodbury, 1950) :

   $$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$$

$$\begin{aligned}(A+BCD)^{-1} &= A^{-1} - (A+BCD)^{-1}\big((A+BCD)A^{-1} - I\big)\\ &= A^{-1} - (A+BCD)^{-1}BCDA^{-1}\\ &= A^{-1} - A^{-1}(I+BCDA^{-1})^{-1}BCDA^{-1}\\ &= A^{-1} - A^{-1}B(I+CDA^{-1}B)^{-1}CDA^{-1}\\ &= A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}\end{aligned}$$

FIGURE 15.2: Preuve du lemme d'inversion matriciel de Woodbury

3. Déterminer d'abord chaque a_i en fonction de la somme $\sum_{i=1}^{N} a_i$, puis en déduire la valeur de $\sum_{i=1}^{N} a_i$.
4. Appliquer la formule du MSE établie dans l'exercice p. 147.
   Examiner les cas $A \to +\infty$; $N \to +\infty$.

Exercice: lissage et filtrage de Wiener, prédiction linéaire

Suivant le type de problème à résoudre, il existe de nombreuses variantes des équations de Wiener-Hopf (1931), Quand on cherche à éliminer du bruit additif, on obtient un filtrage de Wiener (1942) (également dû à Kolmogorov, 1941). Quand on cherche à prédire linéairement un signal on obtient des équations de Yule-Walker (1927, 1931) caractérisant par ailleurs un processus auto-régressif.

On considère le modèle i.i.d. $X = S + W$ (signal + bruit) où W est indépendant de S et tous les signaux sont centrés, les moments d'ordre 2 (matrices de corrélation $\mathbf{R}_{SS}$ et $\mathbf{R}_{WW}$) sont connus.

Trouver le LMMSE dans les cas suivants :

1. On estime $\theta = S$ (*lissage de Wiener*).
2. On estime $\theta = S_N$ (*filtrage de Wiener*). (Une généralisation pour un modèle dynamique de S est connue sous le nom de *filtrage de Kalman*.)
3. On estime $\theta = X_N$ à partir de $X_1, \ldots, X_{N-1}$ (*prédiction linéaire*). Ici les équations à résoudre sont les *équations de Yule-Walker*.

Exercice: lissage et filtrage de Wiener, prédiction linéaire

Indications:

1. Appliquer le formule du cours en déterminant $\mathbf{R}_{XX}$ et $\mathbf{R}_{SX}$ en fonction de $\mathbf{R}_{SS}$ et $\mathbf{R}_{WW}$.
2. Cas particulier du précédent où $\mathbf{R}_{S_N S}$ est un vecteur ligne.

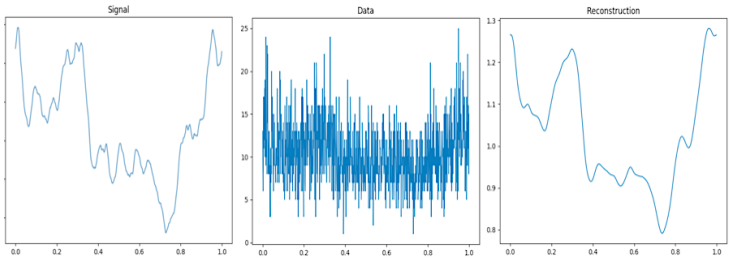


FIGURE 15.3: Filtrage de Wiener pour débruiter un signal

3. Exprimer le résultat en fonction de $\mathbf{R}_{X_N X}$ et $\mathbf{R}_{XX}$.

Exercice: révision sur l'estimation statistique

On modélise le trafic internet à l'aide d'une distribution de la taille des paquets envoyés, lorsque celle-ci dépasse un seuil u connu à l'avance, donnée par la loi de Pareto *de seuil u et d'indice θ définie par :*

$$p_\theta(x) = \frac{\theta u^\theta}{x^{\theta+1}} \mathbb{1}_{x>u}$$

(densité par rapport à la mesure de Lebesgue). Le modèle statistique est défini par $X = (X_1, X_2, \ldots, X_N)$ *i.i.d. selon la loi de Pareto* (u, θ) *où* $N \geqslant 3$.

1. Calculer l'estimateur ML noté $\widehat{\theta}(X)$.
2. Calculer la fonction de répartition complémentaire de la loi de Pareto : $\mathbb{P}\{X_i > x\}$ pour $x \geqslant u$ et en déduire la loi de $Y_i = \log \frac{X_i}{u}$. En déduire la loi de l'estimateur ML.
3. Trouver le biais et le risque quadratique de l'estimateur ML.
4. Montrer que l'estimateur corrigé $\widehat{\theta}'(X) = \frac{N-1}{N}\widehat{\theta}(X)$ est non biaisé et calculer sa variance.
5. Calculer l'information de Fisher J_θ pour le modèle statistique considéré (dont on admettra qu'il est régulier).
6. Les estimateurs $\widehat{\theta}$ et $\widehat{\theta}'$ sont-ils efficaces ? asymptotiquement efficaces ? Quel est le meilleur estimateur ?

On suppose maintenant que θ suit la distribution *a priori* $\theta \sim \Gamma(\alpha, \lambda)$

7. Calculer la loi *a posteriori*.
8. En déduire l'estimateur MMSE. Peut-on retrouver le cas limite sans a priori ?

Exercice: révision sur l'estimation statistique

Indications:

1. Verifier que la log-vraisemblance du modèle est concave sur le domaine considéré. On trouve $\widehat{\theta}(X) = \frac{N}{\sum_{i=1}^{N} \log \frac{X_i}{u}}$.
2. On trouve $Y_i \sim \mathcal{E}(\theta) = \Gamma(1, \theta)$ (loi Gamma).
3. Rappel : si $Y_i \sim \Gamma(\alpha_i, \theta)$ sont indépendantes alors leur somme $\sum_i Y_i \sim \Gamma(\sum_i \alpha_i, \theta)$. Si $X \sim \Gamma(\alpha, \lambda)$ alors $Y = 1/X$ suit la loi inverse Gamma $\Gamma^{-1}(\alpha, \lambda)$.
4. Rappel : si $Y \sim \Gamma^{-1}(\alpha, \lambda)$, son espérance (si $\alpha > 1$) et sa variance (si $\alpha > 2$) sont respectivement $\mathbb{E}(Y) = \frac{\lambda}{\alpha - 1}$ et $\mathbb{V}(Y) = \frac{\lambda^2}{(\alpha-1)^2(\alpha-2)}$.
5. Biais nul et variance multipliée par $\frac{(N-1)^2}{N^2}$.
6. On a $J_{\theta,N} = N J_{\theta,1}$ car le modèle est i.i.d., et pour une observation le score a déjà été calculé comme la dérivée de la log-vraisemblance.
7. Ils sont tous deux non efficaces mais asymptotiquement efficaces. Puisque MSE>MSE′ c'est $\widehat{\theta}'$ qui est meilleur.
8. La loi Gamma est conjuguée du modèle.
9. Pour le cas limite $\lambda \to 0$ (aucun a priori) et $\alpha = 0$ ou 1, le MMSE redonne $\widehat{\theta}$ et $\widehat{\theta}'$.

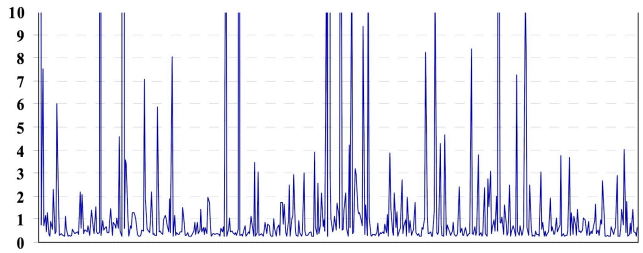


FIGURE 15.4: Flux de trafic entrant (en paquets mesurés) à un nœud d'un réseau en utilisant la distribution de Pareto pendant environ 500 secondes.