Chapitre I

Rappels de probabilités

Le volume horaire du cours ne permettant pas de revisiter en profondeur les bases de probabilités, on ne présentera que quelques éléments utiles pour la suite. Vous êtes invités à revoir par vous mêmes certains ouvrages de référence.

I.1 Revoir dans les cours de Licence…

I.1.1 Expérience aléatoire – Événement – Probabilités – Indépendance – Probabilités conditionnelles

- Intuitivement, une probabilité décrit
  - une « fréquence limite » pour un nombre infini d'expériences aléatoires ;
  - un moyen de quantifier une incertitude / d'exprimer une croyance.
- Quelques mots-clés vers une formalisation mathématique
  - Expérience aléatoire – Résultat = Événement
  - Univers Ω – Tribu T de parties de Ω (famille non vide de parties de Ω stable par complémentation et union dénombrable)
  - Définition axiomatique de la probabilité, sur un espace probabilisable (Ω, T) :
    $\mathbb{P}: T \to [0;1]$ telle que
    · $\mathbb{P}(\Omega) = 1$;
    · si $\cap_i A_i = \emptyset$, alors $\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i)$ pour des unions finies ou dénombrables.

  $\hookrightarrow$ Il vient
    · $\mathbb{P}(\emptyset) = 0$;
    · $\forall A \in T,\ \mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$;
    · $\forall A, B \in T,\ \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
    · $\forall A, B \in T,\ (A \subset B) \Longrightarrow (\mathbb{P}(A) \leq \mathbb{P}(B))$;
    · Théorème des probabilités totales : si $\{B_i\}$ est une partition de Ω (*i.e.*, les événements B_i sont mutuellement exclusifs et exhaustifs), alors $\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i)$.
- Indépendance d'événements
  - Indépendance de 2 événements ;
  - Indépendance de N événements (qui implique mais n'est pas équivalente à leur indépendance 2 à 2 !).
- Probabilités jointes – Probabilités conditionnelles – Formule de Bayes – …

I.1.2 Notion de variable aléatoire

- $X \in \mathbb{R}$ variable aléatoire (VA) ; pour l'événement ω, $X(\omega)$ = réalisation x.
  - événements élémentaires $X \leq x$;
  - formation d'événements tels que $X \in I$ ou $X \in I_1 \cup I_2$ ou… par un ensemble dénombrable d'opérations union et intersection ;
  - définition des probabilités d'événements élémentaires :

$$\textit{fonction de répartition (cdf)}\ \mathrm{P}_X(x) = \mathbb{P}(X \leq x), \tag{I.1.1}$$

  où $\mathrm{P}_X(x) \geq 0$;
  - définition des probabilités d'autres événements au moyen des axiomes de probabilité ;
  - définition de la *densité de probabilité (probability density function –pdf)*

$$\mathrm{p}_X(x) = \left[\frac{\mathrm{dP}_X(\xi)}{\mathrm{d}\xi}\right]_{\xi=x} = \lim_{\mathrm{d}\xi \to 0} \frac{\mathbb{P}(x < X \leq x + \mathrm{d}\xi)}{\mathrm{d}\xi}, \tag{I.1.2}$$

où : $\mathrm{p}_X(x) \geq 0$; $\mathrm{P}_X(x) = \int_{-\infty}^{x} \mathrm{p}_X(\xi)\mathrm{d}\xi$; $\mathrm{P}_X(+\infty) = 1$; $\mathbb{P}(X \in I) = \int_I \mathrm{p}_X(\xi)\mathrm{d}\xi$; etc.

- si $X \in \mathbb{R}^N$ vecteur aléatoire (VA), alors : $\mathrm{d}x = \mathrm{d}x_1 \ldots \mathrm{d}x_N$; $X \leq x$ s'interprète vectoriellement ; etc.

- X est
  - *continue* si $\mathrm{P}_X(x)$ est continue
  - *discrète* si $\mathrm{P}_X(x)$ est en escalier, auquel cas

$$\mathrm{p}_X(x) = \sum_i p_i \delta(x - x_i) \quad \text{où} \quad p_i = \mathbb{P}(X = x_i), \tag{I.1.3}$$

et on note

$$\textit{fonction masse (probability mass function –pmf)}\ \mu_X(x_i) = \mathbb{P}(X = x_i) = p_i, \tag{I.1.4}$$

de sorte que $\mathbb{P}(X \in I) = \sum_{x_i \in I} \mu_X(x_i)$, etc.

I.2 Propriétés des VA

I.2.1 Moments

- *Opérateur espérance mathématique*

$$\mathbb{E}_X[g(X)] = \int g(x)\mathrm{p}_X(x)\mathrm{d}x \quad \text{ou} \quad \mathbb{E}_X[g(X)] = \sum_{x_i} g(x_i)\mu_X(x_i) = \sum_{x_i} g(x_i)p_i. \tag{I.2.1}$$

$\hookrightarrow$ c'est un opérateur *linéaire*

- *Espérance/Moyenne*

$$m_X = \mathbb{E}_X[X] = \int x\mathrm{p}_X(x)\mathrm{d}x \quad \text{ou} \quad \sum_{x_i} x_i\mu_X(x_i) = \sum_{x_i} x_i p_i. \tag{I.2.2}$$

$\hookrightarrow$ X *centrée* $\Leftrightarrow m_X = 0$.

- *Moments d'ordre n – Moments centrés d'ordre n*, dont
  - *Variance* $(X \in \mathbb{R})$: $\sigma_X^2 = \mathbb{E}_X\left[(X - m_X)^2\right]$;
  - *Covariance* $(X \in \mathbb{R}^N)$: $\mathrm{Cov}_X = \mathbb{E}_X\left[(X - m_X)(X - m_X)^T\right] = \mathbb{E}_X\left[XX^T\right] - m_X m_X^T$;

  $\hookrightarrow$ Nota : $\mathbb{E}_X\left[(X - m_X)^T(X - m_X)\right] = \mathrm{trace}(\mathrm{Cov}_X)$.

- Quelques exercices/propriétés à méditer…

$\forall\ X \in \mathbb{R}^N$ V.A., $X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}$, $\mathbb{E}_X\left[g(X)\right] = \int_{x_1 \ldots x_N} g(x)\,\mathrm{p}_X(x)\,\mathrm{d}x = \int_{x_1 \ldots x_N} \begin{bmatrix} g_1(x_1 \ldots x_N) \\ \vdots \\ \vdots \end{bmatrix} \mathrm{p}_{X_1 \ldots X_N}(x_1, \ldots, x_N)\ \mathrm{d}x_1 \ldots \mathrm{d}x_N$

• $\mathbb{E}[.]$ est linéaire $\Longrightarrow \forall(\lambda, \mu) \in \mathbb{R} \times \mathbb{R}$ déterministes, $\forall\ (X, Y) \in \mathbb{R} \times \mathbb{R}$ V.As, $\mathbb{E}_{X,Y}\left[\lambda X + \mu Y\right] = \lambda\,\mathbb{E}_X[X] + \mu\,\mathbb{E}_Y[Y]$

$\hookrightarrow$ se démontre immédiatement en développant $\int_{x,y} (\lambda x + \mu y)\,\mathrm{p}_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y$; en effet,

$= \lambda \int_{x,y} x\,\mathrm{p}_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y + \mu \int_{x,y} y\,\mathrm{p}_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = \lambda \int_x x \underbrace{\left[\int_y \mathrm{p}_{X,Y}(x,y)\mathrm{d}y\right]}_{= \mathrm{p}_X(x)} \mathrm{d}x + \mu \int_y y \underbrace{\left[\int_x \mathrm{p}_{X,Y}(x,y)\mathrm{d}x\right]}_{= \mathrm{p}_Y(y)} \mathrm{d}y$ ∎

• Pour les mêmes raisons, quelles que soient les matrices (de dimensions compatibles) M et N, déterministes, et X, aléatoire,

$\mathbb{E}[X^T] = \mathbb{E}[X]^T$, $\quad \mathbb{E}[MXN] = M\,\mathbb{E}[X]N$, etc.

• $x \in \mathbb{R} \Rightarrow \bar{x} = \mathbb{E}[X] = \int_{\mathbb{R}} x\,\mathrm{p}_X(x)\mathrm{d}x$ et $\sigma_X^2 = \mathbb{E}\left[(X - m_X)^2\right] = \int_{\mathbb{R}} (x - \bar{x})^2\,\mathrm{p}_X(x)\mathrm{d}x$

$\rightarrow$ NOTA : $\sigma_X^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right]$

$= \mathbb{E}[X^2] - \mathbb{E}\left[2\mathbb{E}[X]X\right] + \mathbb{E}\left[\mathbb{E}[X]^2\right]$

$= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2$

$\sigma_X^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ ∎

• $x \in \mathbb{R}^N \Rightarrow \bar{x} = \mathbb{E}[X] \in \mathbb{R}^N$ et $\mathrm{Cov}_X = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right] = \mathbb{E}\left[XX^T - X\,\mathbb{E}[X]^T - \mathbb{E}[X]X^T + \mathbb{E}[X]\mathbb{E}[X]^T\right]$

$= \mathbb{E}[XX^T] - \mathbb{E}\left[X\mathbb{E}[X]^T\right] - \mathbb{E}\left[\mathbb{E}[X]X^T\right] + \mathbb{E}\left[\mathbb{E}[X]\mathbb{E}[X]^T\right]$

$= \mathbb{E}[XX^T] - 2\mathbb{E}[X]\mathbb{E}[X]^T + \mathbb{E}[X]\mathbb{E}[X]^T$

$\mathrm{Cov}_X = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right] = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$

$\rightarrow$ NOTA : $\forall x \in \mathbb{R}^N$, $(x - \bar{x}) \in \mathbb{R}^N \Rightarrow (x - \bar{x})(x - \bar{x})^T \in \mathbb{R}^{N \times N}$ est symétrique semi-définie positive et de rang 1

$\Downarrow$

$\mathrm{Cov}_X \in \mathbb{R}^{N \times N}$ est une matrice symétrique semi-définie positive (voire définie positive !)

• Rappel : • $\forall M \in \mathbb{R}^{N \times N}$, $\forall \xi \in \mathbb{R}^N$, on définit par $\xi^T M \xi$ la forme quadratique associée à M. Or, $\xi^T M \xi$ est toujours égale à $\xi^T \frac{M + M^T}{2} \xi$.

$\Rightarrow$ SPDG, on suppose $M = M^T$ et on dit que $M \geq 0$ (semi-définie positive) $\Leftrightarrow \forall \xi,\ \xi^T M \xi \geq 0$

$M > 0$ (définie positive) $\Leftrightarrow \forall \xi \neq 0,\ \xi^T M \xi > 0$

⚠ M peut être indéfinie !

• $M = M^T \Rightarrow$ les v.p de M sont réelles et les $\vec{v}$p de M peuvent être choisis orthogonaux

et $M > 0 \Leftrightarrow$ toutes les v.p de M sont $> 0 \Leftrightarrow$ les dét$^{\circ}$ des matrices mineures successives de M sont > 0

$M \geq 0 \Leftrightarrow$ - - ——— - - - - - $\geq 0 \Leftrightarrow \exists$ N telle que $M = NN^T$ (en effet, $\xi^T M \xi = (N\xi)^T(N\xi) = \|N\xi\|^2$ est $\geq 0\ \forall \xi$!

• Nota : $X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \Rightarrow \mathbb{E}[X] = \int_{x \in \mathbb{R}^N} x\,\mathrm{p}_X(x)\,\mathrm{d}x = \int_{x_1} \cdots \int_{x_N} \begin{bmatrix} e_1^T x \\ \vdots \\ e_N^T x \end{bmatrix} \mathrm{p}_{X_1 \ldots X_N}(x_1 \ldots x_N)\,\mathrm{d}x_1 \ldots \mathrm{d}x_N$, où $e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \# i$

et bien sûr, $\int_{x_1 \ldots x_N} \int e_i^T x\ \mathrm{p}_{X_1 \ldots X_N}(x_1 \ldots x_N)\,\mathrm{d}x_1 \ldots \mathrm{d}x_N = \int_{x_i} x_i \underbrace{\left[\int\!\!\int\!\!\int_{\{x_j, j \neq i\}} \mathrm{p}_{X_1 \ldots X_N}(x_1 \ldots x_N) \prod_{j \neq i} \mathrm{d}x_j\right]}_{= \mathrm{p}_{X_i}(x_i)} \mathrm{d}x_i = \mathbb{E}[X_i]$

I.2.2 Loi jointe d'un couple de VAs – Lois marginales – Lois conditionnelles

- *Rappel* : loi jointe de X_1, X_2 (cas continu)

$$\mathrm{p}_{X_1 X_2}(x_1, x_2) = \frac{\partial \mathrm{P}_{X_1 X_2}(x_1, x_2)}{\partial x_1 \partial x_2}. \tag{I.2.3}$$

$\hookrightarrow$ $\mathbb{E}_{X_1 X_2}[g(X_1, X_2)] = \int\!\!\int g(x_1, x_2)\mathrm{p}_{X_1 X_2}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2$; etc.

- *Lois marginales*

$$\mathrm{p}_X(x) = \int \mathrm{p}_{XY}(x, y)\mathrm{d}y, \quad \text{ou} \quad \mu_X(x_i) = \sum_j \mu_{XY}(x_i, y_j).\ \text{(cf. théo. proba. totales)} \tag{I.2.4}$$

- *Lois conditionnelles*

$$\mathrm{p}_{X|Y}(x|y) = \frac{\mathrm{p}_{XY}(x,y)}{\mathrm{p}_Y(y)} = \frac{\mathrm{p}_{XY}(x,y)}{\int \mathrm{p}_{XY}(x,y)\mathrm{d}x} \text{ si } \mathrm{p}_Y(y) \neq 0, \text{ (Bayes pdf)} \tag{I.2.5}$$

ou

$$\mu_{X|Y}(x_i|y_j) = \frac{\mu_{XY}(x_i,y_j)}{\mu_Y(y_j)} = \frac{\mu_{XY}(x_i,y_j)}{\sum_i \mu_{XY}(x_i,y_j)} \text{ si } \mu_Y(y_j) \neq 0. \text{ (Bayes proba.)} \tag{I.2.6}$$

- *Indépendance – Corrélation*
  - $X_1,\ldots,X_N$ indépendantes $\Leftrightarrow \forall x_1,\ldots,x_N,\ \mathrm{p}_{X_1\ldots X_N}(x_1,\ldots,x_N) = \mathrm{p}_{X_1}(x_1)\ldots\mathrm{p}_{X_N}(x_N)$ ou $\forall x_1,\ldots,x_N,\ \mu_{X_1\ldots X_N}(x_1,\ldots,x_N) = \mu_{X_1}(x_1)\ldots\mu_{X_N}(x_N)$.
  - $X_1,\ldots,X_N$ i.i.d. $\Leftrightarrow X_1,\ldots,X_N$ indépendantes et $\mathrm{p}_{X_1}(x_1) = \ldots = \mathrm{p}_{X_N}(x_N)$.
  - X,Y non corrélées $\Leftrightarrow \mathbb{E}_{XY}\left[XY^T\right] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]^T \Leftrightarrow C_{XY} = 0$.

  $\hookrightarrow$ X,Y indépendantes $\Rightarrow X,Y$ non corrélées (et équivalence si $(X^T,Y^T)^T$ Gaussien).

  $\hookrightarrow$ Pour $X_1 \in \mathbb{R}$, $X_2 \in \mathbb{R}$, $\mathrm{cov}(X_1,X_2)$ est liée à $\sigma^2_{X_1},\sigma^2_{X_2}$ par $\rho_{12} = \frac{\mathrm{cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}}$, avec
    - $|\rho_{12}| \leq 1$;
    - $\rho_{12} = 0 \Leftrightarrow X_1, X_2$ non corrélées ;
    - $\rho_{12} = 1 \Leftrightarrow \exists(\alpha_1,\alpha_2) \neq (0,0),\ \alpha_1X_1 + \alpha_2X_2 = 0$, auquel cas $C_{\binom{X_1}{X_2}}$ est singulière...

I.2.3 Quelques autres résultats

- *Densité de probabilité de la somme de deux VA indépendantes*
Si $Y = X_1 + X_2$, avec X_1, X_2 indépendantes,
alors

$$\mathrm{p}_Y(y) = \int \mathrm{p}_{X_1}(y-x)\mathrm{p}_{X_2}(x)\mathrm{d}x. \tag{I.2.7}$$

- *Théorème central limite*
Si $X_1,\ldots,X_N$ indépendantes, de mêmes lois, mais avec moyennes m_i et covariances P_i possiblement distinctes ;
si $Y_N = \sum_{n=1}^N X_n$;
si $Z_N = C_{Y_N}^{-\frac{1}{2}}\Big(Y_N - \mathbb{E}[Y_N]\Big)$, où $L^{\frac{1}{2}}$ tq $L^{\frac{1}{2}}(L^{\frac{1}{2}})^T = L$;
(Nota : $\mathbb{E}[Y_N] = \sum_i m_i$ et $C_{Y_N} = \sum_i P_i$)
alors, lorsque $N \to +\infty$, Z_N converge (faiblement, en loi) vers $\mathcal{N}(0,\mathbb{I})$.

- *Loi des espérances itérées*

$$\mathbb{E}_Y\left[\mathbb{E}_{X|Y}[X|y]\right] = \mathbb{E}_X[X]. \tag{I.2.8}$$

- *Changement de variable*
Si $Y = \phi(X)$, où $\phi^{-1}(.)$ existe et $\phi(.),\phi^{-1}(.)$ sont C[1]
alors

$$\mathrm{p}_Y(y) = \Bigg|\underbrace{\det\frac{\partial\phi^{-1}(y)}{\partial y^T}}_{\text{Jacobienne de }\phi^{-1}(.)}\Bigg|\,\mathrm{p}_X\big(\phi^{-1}(y)\big). \tag{I.2.9}$$

I.2.4 Lois Gaussiennes et de chi-deux

- Soit $X \in \mathbb{R}$. $X \sim \mathcal{N}(\bar{x},\sigma^2)$ ssi

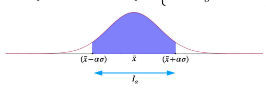
$$\mathrm{p}_X(x) = \mathcal{N}(x;\bar{x},\sigma^2) \triangleq \frac{1}{\sqrt{2\pi}\,\sigma}\exp -\frac{(x-\bar{x})^2}{2\sigma^2}. \tag{I.2.10}$$

- Soit $X \in \mathbb{R}^N$. $X \sim \mathcal{N}(\bar{x},P)$ ssi

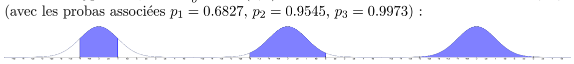
$$\mathrm{p}_X(x) = \mathcal{N}(x;\bar{x},P) \triangleq \frac{1}{\sqrt{(2\pi)^N\det(P)}}\exp\Big(-\frac{1}{2}(x-\bar{x})^TP^{-1}(x-\bar{x})\Big), \tag{I.2.11}$$

$$\triangleq \frac{1}{\sqrt{\det(2\pi P)}}\exp\Big(-\frac{1}{2}(x-\bar{x})^TP^{-1}(x-\bar{x})\Big). \tag{I.2.12}$$

- *Ensembles de confiance*
  - Pour $X \in \mathbb{R}$, $X \sim \mathcal{N}(\bar{x},\sigma^2)$, I_α de taille minimale tel que $\mathbb{P}(X \in I_\alpha) = p_\alpha$ donnée est l'intervalle de confiance $I_\alpha = [\bar{x}-\alpha\sigma;\bar{x}+\alpha\sigma] = \Big\{\xi : \frac{(\xi-\bar{x})^2}{\sigma^2} \leq \alpha^2\Big\}$



  - Sous cette hypothèse, $Y = \frac{X-\bar{x}}{\sigma} \sim \mathcal{N}(0,1)$ dont voici les intervalles de confiance I_1, I_2, I_3 (avec les probas associées $p_1 = 0.6827$, $p_2 = 0.9545$, $p_3 = 0.9973$) :



  - Pour $X \in \mathbb{R}^N$, $X \sim \mathcal{N}(\bar{x},P)$

• $\mathcal{E}_\alpha$ ensemble de taille minimale tq $\mathbb{P}(X \in I_\alpha) = p_\alpha$ donnée
est un ellipsoïde de confiance, $\mathcal{E}_\alpha = \{\xi : (\xi-\bar{x})^TP^{-1}(\xi-\bar{x}) \leq \alpha^2\}$
→ ses directions principales sont portées par les $\vec{v}p$ de P
→ son excursion selon la direction portée par le $\vec{v}p$ unitaire v_i tq $Pv_i = \lambda_i v_i$ est égale à $\alpha\sqrt{\lambda_i}$
↳ vérifier dans le cas où $P = \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}$ par exemple !

- *Fonction affine d'un VA Gaussien*
Soit $X \in \mathbb{R}^N$ un VA. Soient A,b une matrice et un vecteur déterministes.
Alors,

$$X \sim \mathcal{N}(\bar{x},P) \implies Y = AX+b \sim \mathcal{N}(A\bar{x}+b, APA^T). \tag{I.2.13}$$

- *Loi conditionnelle de deux VA conjointement Gaussiens* (ou : Gaussiens dans leur ensemble)
Soient X et Z Gaussiens dans leur ensemble ("jointly Gaussian").
Alors, conditionnellement à l'événement $Z = z$, la distribution de X est elle-même Gaussienne. La formule ci-dessous explicite ses moments a posteriori.

$$\binom{X}{Z} \text{ tel que } \mathrm{p}_{X,Z}(x,z) = \mathcal{N}\left(\binom{x}{z};\binom{m_X}{m_Z},\begin{pmatrix}P_{XX} & P_{XZ}\\ P_{ZX} & P_{ZZ}\end{pmatrix}\right)$$

$$\Downarrow$$

$$\mathrm{p}_{X|Z}(x|z) = \mathcal{N}(x;m_{X|Z},P_{X|Z}) \tag{I.2.14}$$

avec

$$\begin{cases} m_{X|Z} \triangleq \mathbb{E}[X|Z=z] = m_X + P_{XZ}P_{ZZ}^{-1}(z-m_Z)\\ P_{X|Z} \triangleq \mathbb{E}\Big[(X-m_{X|Z})(X-m_{X|Z})^T|Z=z\Big] = P_{XX} - P_{XZ}P_{ZZ}^{-1}P_{ZX}.\end{cases}$$

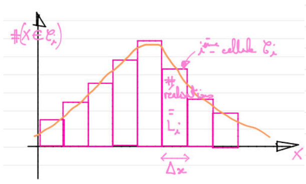
  - La preuve exploite la règle de Bayes (I.2.5). On obtient un quotient de deux lois Gaussiennes (lesquelles ?) qui se « simplifie » en la loi conditionnelle dont on recherche les moments.
  - Si X et Z sont non corrélés, alors les moments a posteriori $m_{X|Z}, P_{X|Z}$ de X conditionnellement à $Z = z$ sont égaux aux moments a priori m_X, P_{XX} de X.
  - Si $P_{ZZ} = \infty\mathbb{I}$, alors les moments a posteriori $m_{X|Z}, P_{X|Z}$ de $X|Z=z$ sont égaux aux moments a priori m_X, P_{XX} de X.
  - $P_{X|Z}$ ne dépend pas de la réalisation z de Z !

- *Loi de chi-deux à N degrés de liberté*
  - $X_1,\ldots,X_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $Y = X_1^2+\cdots+X_N^2 \implies Y \sim \chi^2_N$.
  - $X \in \mathbb{R}^N$, $X \sim \mathcal{N}(\bar{x},P)$, $Z = (X-\bar{x})^TP^{-1}(X-\bar{x}) \implies Z \sim \chi^2_N$.

I.2.5 Détermination empirique d'une densité de probabilité

- Soit $\mathcal{H}_X$ un histogramme de X, avec
  - Δx la taille des cellules ;
  - M le nombre total d'échantillons (réalisations) de X ;
  - L_i le nombre d'échantillons de X dans chaque cellule no i de centre x_i.

  Alors,



total échantillons $M = \sum_i L_i$
$\Rightarrow \mathbb{P}(X \in \mathcal{C}_i) = \frac{L_i}{M}$
$\Rightarrow \hat{\mathrm{p}}(x_i \text{ centre de } \mathcal{C}_i) = \frac{L_i/M}{\Delta x}$

en fait, $\hat{\mathrm{p}}_X(x_i) = \frac{1}{\Delta x}\int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}} \mathrm{p}(x)\mathrm{d}x$

Chapitre II

Problématique de l'estimation

II.1 Modélisation

Modéliser un phénomène ou un système réel consiste à établir une *représentation mathématique* de celui-ci. Cette représentation est conçue *dans un but précis*, par exemple :
- l'analyse du phénomène afin d'en approfondir sa compréhension (modèles de la physique,...),
- la prédiction de comportements,
- la commande,
- la détermination de grandeurs pour lesquelles aucun capteur n'est disponible, à partir de mesures indirectes,
- le test d'hypothèses (diagnostic médical, contrôle de sûreté de fonctionnement d'installations industrielles, ...),
- le traitement de signaux (suppression du bruit, compression de données, filtrage, interpolation,...),
- etc.

Les caractéristiques d'un modèle (sa représentativité, sa souplesse d'exploitation, sa fiabilité, sa complexité,...) sont donc adaptées au but recherché. Elles sont le fruit d'un compromis entre les moyens nécessaires à son obtention ou son exploitation (temps, calcul, expérimentations), et les retombées espérées. À titre d'exemple, les modèles pour l'analyse et la commande d'un système doivent d'une part permettre la représentation fidèle du comportement qualitatif et quantitatif de celui-ci, capturer ses propriétés cruciales sur un domaine de fonctionnement possiblement restreint, et, d'autre part, admettre une structure suffisamment simple pour permettre une procédure d'analyse relativement simple et une méthode de commande efficace.

Il existe différentes classes de modèles, parmi elles, nous trouvons :
- les modèles de connaissance *vs* de représentation,
- les modèles déterministes *vs* aléatoires,
- les modèles à temps continu *vs* à temps discret,
- les modèles linéaires *vs* non linéaires *par rapport aux paramètres*.

Un *modèle stochastique* est un modèle de la forme

$$Z = \mathcal{M}(\theta,V) \tag{II.1.1}$$

où :
- $\theta \in \mathbb{R}^M$ désigne le vecteur des paramètres « caché » à estimer ;

- V est un vecteur aléatoire modélisant les bruits/perturbations ; le modèle intègre une description probabiliste de V ; la valeur (réalisation) de V pour l'expérience en cours est inconnue ;
- $Z \in \mathbb{R}^N$ désigne le vecteur aléatoire d'observation (ou de mesure) ; pour l'expérience en cours, ce vecteur se réalise en le vecteur z des mesures effectuées†

Le problème d'estimation s'énonce alors comme suit.

> **Énoncé** *Disposant de la réalisation z de Z, quelle information peut-on obtenir sur le vecteur de paramètres θ ?* □

Préalablement à la mise en place de tout schéma d'estimation, il convient de recenser la *connaissance a priori* dont on dispose sur le vecteur de paramètres caché θ. Il sera vu plus loin que :

- si on ne sait rien sur θ, alors ce vecteur est considéré comme déterministe inconnu ;
- si, au contraire, on dispose d'une connaissance a priori sur les valeurs admissibles de θ, alors celle-ci est exprimée au moyen de la distribution de probabilité d'un vecteur aléatoire Θ ; θ est alors considéré comme la réalisation de Θ pour l'expérience en cours‡.

Parmi les problèmes potentiels en estimation stochastique, on peut citer :

- la sélection d'un modèle (II.1.1) inapproprié : structure ne permettant pas de rendre compte du phénomène considéré ; propriétés structurelles non satisfaites ; caractérisation erronée des statistiques de Θ, V ;
- la difficulté (voire l'impossibilité) de parvenir au calcul de l'estimé de θ ;
- la dégradation des performances de l'estimateur lorsque les hypothèses formulées dans le modèle ne sont pas parfaitement respectées ; on souhaite que le schéma utilisé soit doté d'une robustesse relative par rapport au non-respect de ces hypothèses.

II.2 Formalisation mathématique d'un problème d'estimation

Dans cette section, nous supposerons que le vecteur de paramètres θ appartient à un ensemble admissible $\mathcal{U}_{ad} \subset \mathbb{R}^M$) et que le vecteur z des mesures effectuées appartient à $\mathbb{R}^N$.

II.2.1 Estimateur / Estimé

Un problème d'estimation stochastique consiste à analyser les propriétés, voire à synthétiser, une fonction *estimateur* définie comme suit.

Définition II.2.1 *Soit une fonction g de l'espace des observations dans l'espace des paramètres, qui, à chaque vecteur de mesure z (réalisation de Z pour l'expérience en cours) associe $\hat{\theta} = g(z)$,* i.e.,

$$\begin{aligned} g : \mathbb{R}^N &\to \mathbb{R}^M \\ z &\mapsto \hat{\theta} = g(z). \end{aligned} \tag{II.2.1}$$

$\hat{\theta}$ est appelé estimé *de θ. La fonction g est appelée* estimateur. □

Dans tout le cours, la fonction g est déterministe, *i.e.*, pour un vecteur de mesures z donné, l'image $g(z)$ de z par g est parfaitement définie, et ne dépend pas de l'expérience§.

†On peut noter $z = Z(\omega)$, où ω symbolise le résultat de l'expérience (ou « événement »).
‡La densité de probabilité de Θ, ou « loi a priori », sera notée $p_\Theta(\theta)$.
§Il existe en effet des fonctions g aléatoires. En pratique, il s'agit d'estimateurs qui, pour calculer l'image $g(z)$ du vecteur d'observation z, reposent de manière interne sur la génération de nombres (pseudo-)aléatoires. Ces méthodes d'estimation, dites « de Monte Carlo » ne seront pas vues ici.

On note $\hat{\Theta} = g(Z)$, de sorte que $\hat{\Theta} \in \mathbb{R}^M$. Du fait que Z est une variable aléatoire vectorielle, $\hat{\Theta}$ est également un vecteur aléatoire, appelé *estimateur*†. Bien sûr, $\hat{\Theta}$ se réalise en $\hat{\theta}$ lorsque Z se réalise en z.

Il vient la conséquence fondamentale suivante, cf. Figure II.1.

> **Conséquence fondamentale** *Un estimateur $\hat{\Theta}$ étant une fonction* (ici, déterministe) *d'une variable aléatoire (la variable aléatoire de mesure Z), il est lui-même une variable aléatoire, de sorte que ses propriétés doivent être évaluées statistiquement.* □

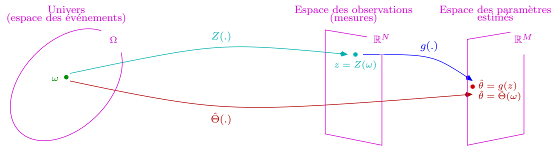


Figure II.1 : Principe de l'estimation stochastique.

L'exemple ci-dessous permet d'illustrer cette conclusion.

Exemple 1 *Soit a un niveau constant inconnu. On prélève la séquence de mesures $z[1], \ldots, z[N]$, indexées par $n = 1, \ldots, N$. Chaque $z[n]$ est la réalisation de la variable aléatoire $Z[n] = a + W[n]$, où W désigne du bruit. On suppose que la séquence $W[1], \ldots, W[N]$ est indépendante, identiquement distribuée (i.i.d.) selon $\mathcal{N}(0, \sigma^2)$.*

Disposant des observations $z[1], \ldots, z[N]$, comment estimer a ?

Soient les deux estimés

$$\hat{a} = \frac{1}{N}\sum_{n=1}^{N} z[n] \quad \textit{et} \quad \mathring{a} = z[1].$$

- *Sur une expérience, impossible de dire lequel est le plus proche de a.*
- *Si on répète l'expérience un nombre élevé de fois, alors intuitivement,*
  - *$\hat{a}$ et $\mathring{a}$ renvoient sans doute « en moyenne » la valeur correcte de a (plus exactement, les estimateurs associés $\hat{A}$ et $\mathring{A}$ sont centrés sur a) ;*
  - *grâce à un effet de « lissage » (à démontrer !), le fait que $\hat{a}$ soit obtenu sur la base d'un nombre plus important de mesures permet qu'il se disperse de manière moins importante autour de a (plus exactement, la dispersion de $\hat{A}$ autour de sa moyenne semble moins importante).*
- *En conclusion,*
  - *on peut avoir une idée sur la base de simulations « de Monte Carlo », mais un nombre insuffisant d'expériences peut conduire à des résultats trompeurs ;*
  - *il est nécessaire de mener une analyse statistique de l'estimateur (examen de sa pdf, etc.) pour conclure.* □

†En fait, le vocable « estimateur » se réfèrera préférentiellement à $\hat{\Theta}$. La fonction g ne sera que rarement référencée explicitement.

II.2.2 Les deux familles d'estimateurs en contexte stochastique

On distingue deux familles d'estimateurs, selon la connaissance a priori disponible sur le vecteur de paramètres caché $\theta \in \mathbb{R}^M$. Les notations utilisées pour désigner des densités de probabilités sont rappelées en bas de page†.

On ne dispose d'aucune connaissance a priori sur θ Comme indiqué plus haut, θ est alors supposé déterministe inconnu. On se situe dans le contexte des *estimateurs classiques*, ou *de Fisher*, (Figure II.2). L'estimation classique repose sur le *modèle d'observation*

$$p_{Z;\theta}(z;\theta). \tag{II.2.2}$$

On commettra l'abus de notation usuel‡ consistant à écrire (II.2.2) sous la forme

$$p_{Z|\theta}(z|\theta). \tag{II.2.3}$$

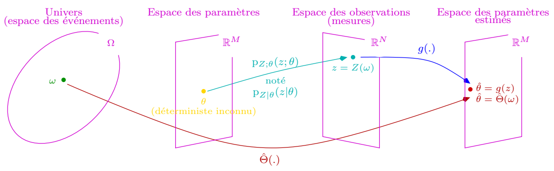


Figure II.2 : Principe de l'estimation stochastique classique (Fisher).

On dispose d'une une connaissance a priori sur θ Comme indiqué plus haut, θ est alors considéré comme la réalisation d'un vecteur aléatoire Θ. On parle d'*estimateurs bayésiens*, ou *de Bayes* (Figure II.3). L'estimation bayésienne repose sur le modèle

$$p_{Z,\Theta}(z,\theta), \tag{II.2.4}$$

c'est-à-dire sur la *loi a priori*

$$p_\Theta(\theta) \tag{II.2.5}$$

et le *modèle d'observation*

$$p_{Z|\Theta}(z|\theta). \tag{II.2.6}$$

†Les notations sont sans doute un peu lourdes, mais nous semblent nécessaires pour permettre le développement rigoureux des concepts.
- On désigne par $p_A(a)$ la densité de probabilité de la variable aléatoire A. C'est bien une fonction de la variable muette a, qui désigne une réalisation possible de A. Si on trace la représentation de $p_A(a)$, on porte en abscisse les valeurs (réalisations) possibles a de A et en ordonnée la valeur numérique de l'évaluation de $p_A(.)$ en a.
- On désigne par $p_{A|B}(a|b)$ la densité de probabilité de la variable aléatoire A conditionnellement à l'événement « la variable aléatoire B s'est réalisée en la valeur b ». C'est bien une fonction de la variable muette a, qui désigne une réalisation possible de A. Si on trace la représentation de $p_{A|B}(a|b)$ pour l'événement $B = b$, on porte en abscisse les valeurs (réalisations) possibles a de A et en ordonnée la valeur numérique de l'évaluation de $p_{A|B}(.|b)$ en a.

‡Cet abus de notation consiste naturellement à écrire $p_{\text{variable aléatoire}|\text{paramètre déterministe}}(.|.)$ au lieu de $p_{\text{variable aléatoire}|\text{variable aléatoire}}(.|.)$…

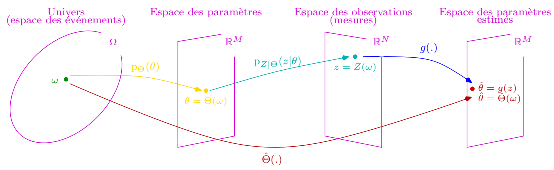


Figure II.3 : Principe de l'estimation stochastique bayésienne (Bayes).

Illustrons ces problématiques sur quelques exemples.

Exemple 2

1. *On dispose de N échantillons*

$$z[1], \ldots, z[N] \overset{i.i.d.}{\sim} p_{Z|(\mu,\sigma^2)}\big(z|(\mu,\sigma^2)\big) = \mathcal{N}(z;\mu,\sigma^2),$$

   où $\mathcal{N}(z;\mu,\sigma^2)$ désigne la loi Gaussienne scalaire réelle de moyenne μ et de variance σ^2. On se donne des estimés $\hat{\mu}$ et $\hat{\sigma}$ des paramètres μ et σ (supposés déterministes inconnus car pas de connaissance a priori à leur sujet). Par exemple,

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} z[n], \quad \textit{et} \quad \hat{\sigma} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(z[n] - \hat{\mu})^2}.$$

   ↪ *Il convient donc d'analyser les estimateurs associés.*

2. *Une personne immobile souhaite déterminer sa position (abscisse curviligne 1D) sur une route, sur la base de positions relevées par plusieurs capteurs (odomètre, localisation zigbee, GPS, etc.).*
   ↪ *Si on exploite une connaissance a priori, Bayes statique.*

3. *Un robot mobile souhaite se localiser sur la base de relevés relatifs à plusieurs amers, voire souhaite également construire une carte dans laquelle il estime la position de ces amers concomitamment à sa localisation : "SLAM" (Simultaneous Localization And Mapping).*
   ↪ *Bayes dynamique — Filtrage.*

4. *Suivi hors ligne d'animaux marins sur la base de relevés satellitaires mensuels et d'un modèle de dynamique a priori.*
   ↪ *Bayes dynamique — Lissage.*

5. *Mais aussi : radar, sonar, parole (reconnaissance de phonèmes → estimation des paramètres d'un modèle linéaire prédictif –LPC– pour déterminer l'enveloppe spectrale, qui est un attribut insensible au pitch), image (suivi d'objets / de personnes), médecine, communications, trajectographie, séismologie, météo, robotique, etc.* □

Chapitre III

Propriétés des estimateurs

III.1 Biais – Covariance – Erreur quadratique moyenne

Dans tout ce qui suit, $\hat{\Theta} = g(Z)$.

III.1.1 Biais d'un estimateur

- Fisher

$$b_{\hat{\Theta}}(\theta) = \mathbb{E}_{Z|\theta}\Big[\hat{\Theta} - \theta|\theta\Big] = \mathbb{E}_{Z|\theta}\Big[\hat{\Theta}|\theta\Big] - \theta = \mathbb{E}_{Z|\theta}[g(Z)|\theta] - \theta. \tag{III.1.1}$$

  - $\hat{\Theta}$ non biaisé $\Leftrightarrow$ *Quel que soit* θ, $b_{\hat{\Theta}}(\theta) = 0$.
  - Ci-dessus, $\mathbb{E}_{Z|\theta}[g(Z)|\theta] = \int g(z) \mathrm{p}_{Z|\theta}(z|\theta)\mathrm{d}z$.

- Bayes

$$b_{\hat{\Theta}} = \mathbb{E}_{Z,\Theta}\Big[\hat{\Theta} - \Theta\Big] = \mathbb{E}_{Z,\Theta}\Big[\hat{\Theta}\Big] - \mathbb{E}_{\Theta}[\Theta] = \mathbb{E}_{Z,\Theta}[g(Z)] - \mathbb{E}_{\Theta}[\Theta]. \tag{III.1.2}$$

  - $\hat{\Theta}$ non biaisé $\Leftrightarrow$ $b_{\hat{\Theta}} = 0$. $b_{\hat{\Theta}}$ est parfois appelé « biais moyen ».
  - On a aussi $b_{\hat{\Theta}} = \mathbb{E}_{\Theta}\big[b_{\hat{\Theta}}(\theta)\big] = \mathbb{E}_{\Theta}\Big[\mathbb{E}_{Z|\Theta}\Big[\hat{\Theta}|\theta\Big] - \theta\Big]$, car $\mathbb{E}_{Z,\Theta}[g(Z)] = \mathbb{E}_{\Theta}\Big[\mathbb{E}_{Z|\Theta}\Big[\hat{\Theta}|\theta\Big]\Big]$.
  - En effet, ci-dessus, $\mathbb{E}_{Z,\theta}[g(Z)] = \iint g(z)\mathrm{p}_{Z,\Theta}(z,\theta)\mathrm{d}z = \int\Big[\int g(z)\mathrm{p}_{Z|\Theta}(z|\theta)\mathrm{d}z\Big]\mathrm{p}_{\Theta}(\theta)\mathrm{d}\theta$.

III.1.2 Covariance d'un estimateur

- Fisher

$$\mathrm{Cov}_{\hat{\Theta}}(\theta) = \mathbb{E}_{Z|\theta}\bigg[\Big(\hat{\Theta} - \mathbb{E}_{Z|\theta}\Big[\hat{\Theta}|\theta\Big]\Big)\Big(\hat{\Theta} - \mathbb{E}_{Z|\theta}\Big[\hat{\Theta}|\theta\Big]\Big)^T \Big| \theta\bigg]. \tag{III.1.3}$$

- Bayes

$$\mathrm{Cov}_{\hat{\Theta}} = \mathbb{E}_{Z,\Theta}\bigg[\Big(\hat{\Theta} - \mathbb{E}_{Z,\Theta}\Big[\hat{\Theta}\Big]\Big)\Big(\hat{\Theta} - \mathbb{E}_{Z,\Theta}\Big[\hat{\Theta}\Big]\Big)^T\bigg]. \tag{III.1.4}$$

  - Et $\mathrm{Cov}_{\hat{\Theta}} = \mathbb{E}_{\Theta}\big[\mathrm{Cov}_{\hat{\Theta}}(\theta)\big]$.

III.1.3 Erreur quadratique moyenne d'un estimateur

- Fisher

$$\mathrm{EQM}_{\hat{\Theta}}(\theta) = \mathbb{E}_{Z|\theta}\Big[\Big(\hat{\Theta} - \theta\Big)^T\Big(\hat{\Theta} - \theta\Big)\Big|\theta\Big] \tag{III.1.5}$$

$$= \mathrm{trace}\Big(\mathrm{Cov}_{\hat{\Theta}}(\theta) + b_{\hat{\Theta}}(\theta)b_{\hat{\Theta}}(\theta)^T\Big). \tag{III.1.6}$$

  - L'équation ci-dessus montre que l'EQM constitue un *compromis biais-variance.*
  - Preuve :
    $\mathrm{EQM}_{\hat{\Theta}}(\theta) = \mathrm{trace}\Big(\mathbb{E}_{Z|\theta}\Big[(\hat{\Theta} - \theta)(\hat{\Theta} - \theta)^T\Big|\theta\Big]\Big)$;
    et $(\hat{\Theta} - \theta)(\hat{\Theta} - \theta)^T = (\hat{\Theta} - m_{\hat{\Theta}}(\theta) + m_{\hat{\Theta}}(\theta) - \theta)(\hat{\Theta} - m_{\hat{\Theta}}(\theta) + m_{\hat{\Theta}}(\theta) - \theta)^T$,
    où $m_{\hat{\Theta}}(\theta) = \mathbb{E}_{Z|\theta}\Big[\hat{\Theta}|\theta\Big]$ et $b_{\hat{\Theta}}(\theta) = m_{\hat{\Theta}}(\theta) - \theta$.

- Bayes

$$\mathrm{EQM}_{\hat{\Theta}} = \mathbb{E}_{Z,\Theta}\Big[\Big(\hat{\Theta} - \Theta\Big)^T\Big(\hat{\Theta} - \Theta\Big)\Big]. \tag{III.1.7}$$

III.1.4 Exercices : Calcul du biais et de la variance des estimateurs proposés dans les Exemples 1 et 2.1

$Z_n = a + W_n$, a dét° inconnu, $W_1 \dots W_N \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$

$\hat{A} = \frac{1}{N}\sum_{n=1}^{N} Z_n$ – biais et (co)variance de $\hat{A}$?

* Calcul de $\mathbb{E}[\hat{A}]$ (qu'on pourrait noter $\mathbb{E}_{Z|a}[\hat{A}|a]$

- $\mathbb{E}[\hat{A}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[Z_n] = \frac{1}{N}\sum_{n=1}^{N}\big[a + \mathbb{E}[W_n]\big] = \frac{1}{N}.[Na] = a$

$\Rightarrow \forall a$, $b_{\hat{A}}(a) = \mathbb{E}_{Z|a}\big[(\hat{A} - a)|a\big] = 0 \Leftrightarrow \hat{A}$ est non biaisé

* Calcul de $\mathrm{Var}_{\hat{A}}(a) = \mathbb{E}_{Z|a}\big[(\hat{A} - \mathbb{E}_{Z|a}[\hat{A}|a])^2 |a\big] = \mathbb{E}[\hat{A}^2] - \mathbb{E}[\hat{A}]^2$

- $\mathrm{Var}_{\hat{A}}(a) = \mathbb{E}\big[(\hat{A} - a)^2\big] = \mathbb{E}[\hat{A}^2] - a^2$

$= \mathbb{E}\Big[\Big(\frac{1}{N}\sum_{n=1}^{N} Z_n\Big)^2\Big] - a^2 = \frac{1}{N^2}\mathbb{E}\Big[\Big(\sum_n Z_n\Big)\Big(\sum_m Z_m\Big)\Big] - a^2$

$= \frac{1}{N^2}\mathbb{E}\Big[\sum_{n=1}^{N} Z_n^2 + \sum_{\substack{n=1 \\ n\neq m}}^{N}\sum_{m=1}^{N} Z_n Z_m\Big] - a^2$

- Or, comme $Z_1, \dots, Z_N \overset{iid}{\sim} \mathcal{N}(a,\sigma^2)$, on a : $\mathbb{E}[Z_n^2] = \sigma^2 + a^2$ et $\mathbb{E}\underset{n\neq m}{[Z_n Z_m]} = 0 + a^2$

d'où $\mathrm{Var}_{\hat{A}}(a) = \frac{1}{N^2}\Big(\sum_{n=1}^{N}\mathbb{E}[Z_n^2] + \sum_{n\neq m}\sum_{m}\mathbb{E}[Z_n Z_m]\Big) - a^2 = \frac{1}{N^2}\Big[(N\sigma^2 + Na^2) + (N^2 - N)a^2\Big] - a^2 = \frac{\sigma^2}{N}$.

III.1.5 Estimateur non biaisé à minimum de variance : Estimateur efficace

- Soit $\hat{\Theta} = g(Z)$ *de biais nul.* $\hat{\Theta}$ est un estimateur efficace si pour tout estimateur $\mathring{\Theta} = f(Z)$ *également non biaisé*, on a $\mathrm{Cov}_{\hat{\Theta}} \leq \mathrm{Cov}_{\mathring{\Theta}}$, *i.e.*, $(\mathrm{Cov}_{\mathring{\Theta}} - \mathrm{Cov}_{\hat{\Theta}})$ définie positive.
- $\hat{\Theta}$ efficace $\Rightarrow$ $\hat{\Theta}$ non biaisé à minimum de variance.
  - cf. cas scalaire, où l'inégalité matricielle devient une inégalité scalaire...

III.1.6 Inégalité de Cramér-Rao (Cas de l'estimation classique)

Soit $\theta \in \mathbb{R}^M$ déterministe inconnu (fixe), Z une variable aléatoire liée à θ, et $\mathrm{p}_{Z|\theta}(z|\theta)$ différentiable par rapport à θ.

- On définit la *Matrice d'Information de Fisher*

$$I(\theta) = \mathbb{E}_{Z|\theta}\bigg[\frac{\partial \ln \mathrm{p}_{Z|\theta}(z|\theta)}{\partial\theta}\frac{\partial \ln \mathrm{p}_{Z|\theta}(z|\theta)}{\partial\theta^T}\Big|\theta\bigg] \tag{III.1.8}$$

$$= -\mathbb{E}_{Z|\theta}\bigg[\frac{\partial^2 \ln \mathrm{p}_{Z|\theta}(z|\theta)}{\partial\theta\partial\theta^T}\Big|\theta\bigg]. \tag{III.1.9}$$

- Soit $\hat{\Theta} = g(Z)$ un estimateur *non biaisé* quelconque de θ, alors

$$\mathbb{E}_{Z|\theta}\Big[(\hat{\Theta} - \theta)(\hat{\Theta} - \theta)^T\Big|\theta\Big] \geq I^{-1}(\theta). \quad \text{(Inégalité de Cramér-Rao)} \tag{III.1.10}$$

  - Il s'agit d'une inégalité matricielle.
  - Cette borne inférieure peut ne jamais être atteinte.
  - Par conséquent, ($\hat{\Theta}$ non biaisé de variance $I^{-1}(\theta)$) $\Rightarrow$ ($\hat{\Theta}$ efficace) !
  - Il existe aussi une borne inférieure pour tout estimateur, éventuellement biaisé.

III.2 Propriétés asymptotiques des estimateurs

Étant donné un estimateur $\hat{\Theta}$ de θ obtenu sur la base d'un échantillon de taille N, il s'agit de caractériser ses propriétés pour $N \longrightarrow +\infty$.

Pour cela, on définit :

- Des notions de convergence stochastique : en probabilité ; en moyenne quadratique ; presque partout.
- La notion d'estimateur asymptotiquement non biaisé ; asymptotiquement efficace ; asymptotiquement efficace et normal.

Chapitre IV

Estimateurs de Fisher

θ est supposé déterministe inconnu. Le modèle d'observation est $\mathrm{p}_{Z|\theta}(z|\theta)$ ou $\mu_{Z|\theta}(z|\theta)$, selon que Z est un VA continu ou discret.

IV.1 Estimateur du maximum de vraisemblance (EMV, MLE)

- Notion de *fonction de vraisemblance (likelihood)*

$$\mathrm{L}(\theta; z) = \mathrm{p}_{Z|\theta}(z|\theta) \text{ ou } \mu_{Z|\theta}(z|\theta). \tag{IV.1.1}$$

  - Anti-log-vraisemblance : $\mathrm{NLL}(\theta; z) = -\ln \mathrm{L}(\theta; z)$.
  - ATTENTION ! (Dans le cas continu, mais les résultats se transposent au cas discret) $\int \mathrm{L}(\theta; z)\mathrm{d}z = 1$, mais rien de particulier pour $\int \mathrm{L}(\theta; z)\mathrm{d}\theta$!

- Définition de l'*estimateur du maximum de vraisemblance* $\hat{\Theta}_{\mathrm{MLE}}$

$$\hat{\Theta}_{\mathrm{MLE}} \text{ tel que } \hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \mathrm{p}_{Z|\theta}(z|\theta) \tag{IV.1.2}$$

$$= \arg\max_{\theta} \mathrm{L}(\theta; z) \tag{IV.1.3}$$

$$= \arg\min_{\theta} \mathrm{NLL}(\theta; z). \tag{IV.1.4}$$

- Illustration : qui de $\theta_1 = 1$ et $\theta_2 = 5$ explique le mieux le fait que $Z(\omega) = 2$ suit la loi $\mathcal{N}(0, 1)$?

- Propriétés importantes de $\hat{\Theta}_{\mathrm{MLE}}$
  - Invariance par reparamétrisation

$$(\phi = f(\theta)) \Rightarrow (\hat{\phi}_{\mathrm{MLE}} = f(\hat{\theta}_{\mathrm{MLE}})) \text{ si } f(.) \text{ bijectif.} \tag{IV.1.5}$$

◦ Sous quelques hypothèses de « régularité » de $p_{Z|\theta}(z|\theta)$, $\hat{\Theta}_{\text{MLE}}$ est asymptotiquement efficace et normal, *i.e.*,

$$\lim_{N\longrightarrow+\infty} \hat{\Theta}_{\text{MLE}} \sim \mathcal{N}(\theta, I^{-1}(\theta)). \tag{IV.1.6}$$

IV.2 Cas d'un modèle d'observation avec perturbations additives Gaussiennes

- Si $Z = h(\theta) + V$, où $Z \in \mathbb{R}^N$, $\theta \in \mathbb{R}^M$, $N \geq M$, $V \in \mathbb{R}^N$, $V \sim \mathcal{N}(0, R)$ avec $R > 0$ (*e.g.*, $R = \sigma^2 \mathbb{I}_{N\times N}$ si V_i i.i.d.), alors

$$\hat{\Theta}_{\text{MLE}} \text{ tel que } \hat{\theta}_{\text{MLE}} = \arg\min_{\theta} (z - h(\theta))^T R^{-1} (z - h(\theta)). \tag{IV.2.1}$$

◦ Preuve : à démontrer !

IV.3 Cas d'un modèle d'observation linéaire avec perturbations additives Gaussiennes

- Si $Z = H\theta + V$, où $Z \in \mathbb{R}^N$, $\theta \in \mathbb{R}^M$, $N \geq M$, $V \in \mathbb{R}^N$, $V \sim \mathcal{N}(0, R)$ avec $R > 0$ (*e.g.*, $R = \sigma^2 \mathbb{I}_{N\times N}$ si V_i i.i.d.), H déterministe de rang plein M, alors

$$\hat{\Theta}_{\text{MLE}} \text{ tel que } \hat{\theta}_{\text{MLE}} = (H^T R^{-1} H)^{-1} H^T R^{-1} z. \tag{IV.3.1}$$

◦ Preuve : à démontrer !
◦ Montrer que $\hat{\Theta}_{\text{MLE}}$ admet un biais nul, *i.e.*,

$$\forall \theta,\ b_{\hat{\Theta}_{\text{MLE}}}(\theta) = 0. \tag{IV.3.2}$$

$Z = H\theta + V$, $V \sim \mathcal{N}(0, R)$, $\hat{\Theta}_{\text{MLE}} = (H^T R^{-1} H)^{-1} H^T R^{-1} Z$, où $H^T R^{-1} H$ inversible > 0 d'après les hypothèses

* Biais $b_{\hat{\Theta}_{\text{MLE}}}(\theta)$ de $\hat{\Theta}_{\text{MLE}}$

- $b_{\hat{\Theta}_{\text{MLE}}}(\theta) = \mathbb{E}[\hat{\Theta}_{\text{MLE}}] - \theta = \left(\mathbb{E}\left[(H^T R^{-1} H)^{-1} H^T R^{-1} (H\theta + V)\right]\right) - \theta = \theta + (H^T R^{-1} H)^{-1} H^T R^{-1} \underbrace{\mathbb{E}[V]}_{0} - \theta = 0$

$\Rightarrow \hat{\Theta}_{\text{MLE}}$ est non biaisé

◦ Montrer que la covariance de l'estimateur et l'erreur quadratique moyenne satisfont

$$\forall \theta,\ \text{Cov}_{\hat{\Theta}_{\text{MLE}}}(\theta) = (H^T R^{-1} H)^{-1}. \tag{IV.3.3}$$

$$\forall \theta,\ \text{EQM}_{\hat{\Theta}_{\text{MLE}}}(\theta) = \text{trace}\left((H^T R^{-1} H)^{-1}\right). \tag{IV.3.4}$$

* Covariance $C_{\hat{\Theta}_{\text{MLE}}}(\theta)$ de $\hat{\Theta}_{\text{MLE}}$

- $C_{\hat{\Theta}_{\text{MLE}}}(\theta) = \mathbb{E}_{Z|\theta}\left[\left(\hat{\Theta}_{\text{MLE}} - \mathbb{E}_{Z|\theta}[\hat{\Theta}_{\text{MLE}}|\theta]\right)(\Diamond)^T \,\middle|\, \theta\right] = \mathbb{E}\left[(\hat{\Theta}_{\text{MLE}} - \theta)(\Diamond)^T\right] = \mathbb{E}\left[\left[(H^T R^{-1} H)^{-1} H^T R^{-1} V\right][\Diamond]^T\right]$

$= \mathbb{E}\left[(H^T R^{-1} H)^{-1} H^T R^{-1} V V^T R^{-1} H (H^T R^{-1} H)^{-1}\right]$

$= (H^T R^{-1} H)^{-1} H^T R^{-1} \mathbb{E}[V V^T] \ R^{-1} H (H^T R^{-1} H)^{-1}$

$= (H^T R^{-1} H)^{-1} H^T R^{-1} \ R \ R^{-1} H (H^T R^{-1} H)^{-1} = (H^T R^{-1} H)^{-1}$

1 2

$\Rightarrow \forall \theta,\ \text{Cov}_{\hat{\Theta}_{\text{MLE}}}(\theta) = (H^T R^{-1} H)^{-1}$

* $\hat{\Theta}_{\text{MLE}}$ est efficace ! En effet, il est non biaisé et

- $p_{Z|\theta}(z|\theta) = \mathcal{N}(z; H\theta, R) = \dfrac{1}{\sqrt{(2\pi)^d \det(R)}} \exp\left[-\frac{1}{2}(z - H\theta)^T R^{-1} (z - H\theta)\right]$
- $-\ln p_{Z|\theta}(z|\theta) = C + \frac{1}{2}(z - H\theta)^T R^{-1} (z - H\theta)$
- $\dfrac{\partial}{\partial\theta}\left[-\ln p_{Z|\theta}(z|\theta)\right] = -H^T R^{-1}(z - H\theta)$
- $\dfrac{\partial^2}{\partial\theta\partial\theta^T} = \dfrac{\partial}{\partial\theta}\left[\left(-H^T R^{-1}(z - H\theta)\right)^T\right] = H^T R^{-1} H$
- $I(\theta) = \mathbb{E}_{Z|\theta}\left[\dfrac{\partial^2[-\ln p_{Z|\theta}(z|\theta)]}{\partial\theta\partial\theta^T}\right] = \mathbb{E}_{Z|\theta}\left[H^T R^{-1} H \,\middle|\, \theta\right] = H^T R^{-1} H$
- $\text{CRLB} = [I(\theta)]^{-1} = (H^T R^{-1} H)^{-1}$ et $\text{Cov}_{\hat{\Theta}_{\text{MLE}}}(\theta) = \text{CRLB}$! (y compris si $N < \infty$!)

Chapitre V

Estimateurs Bayésiens

On dispose d'une connaissance a priori sur θ. Le modèle sous-jacent est $p_{Z,\Theta}(z, \theta) = p_\Theta(\theta) p_{Z|\Theta}(z|\theta)$, où la *loi a priori* $p_\Theta(\theta)$ et le modèle d'observation $p_{Z|\Theta}(z|\theta)$ sont donnés. (Dans le cas de VA discrètes, on exploite les fonctions de masse en lieu et place des densités de probabilités.)

V.1 Loi a posteriori

La règle de Bayes permet d'écrire

$$p_{\Theta|Z}(\theta|z) = \frac{p_{Z|\Theta}(z|\theta) p_\Theta(\theta)}{p_Z(z)}. \tag{V.1.1}$$

- *La* loi a posteriori $p_{\Theta|Z}(\theta|z)$ *capture toute l'information sur* θ *caché contenue dans l'observation* z !
- Nota : $p_Z(z)$ est simplement une constante de normalisation, de sorte qu'on écrit souvent $p_{\Theta|Z}(\theta|z) \propto p_{Z|\Theta}(z|\theta) p_\Theta(\theta)$.

V.2 Définition d'un estimateur Bayésien

- Le passage de $p_{\Theta|Z}(\theta|z)$ à $\hat{\Theta} = g(Z)$ s'effectue par minimisation d'un *risque de Bayes*, *i.e.*,

$$\hat{\Theta} = \arg\min_{\mathring{\Theta}} \mathbb{E}_{Z,\Theta}\left[C(\mathring{\Theta} - \Theta)\right], \text{ où } \mathring{\Theta} = \mathring{g}(Z), \tag{V.2.1}$$

$$= g(Z), \text{ où } g(.) = \arg\min_{\mathring{g}(.)} \iint C(\mathring{g}(z) - \theta) p_{Z,\Theta}(z, \theta) \mathrm{d}\theta \mathrm{d}z, \tag{V.2.2}$$

$$= g(Z), \text{ où } g(.) = \arg\min_{\mathring{g}(.)} \int \underbrace{\left[\int C(\mathring{g}(z) - \theta) p_{\Theta|Z}(\theta|z) \mathrm{d}\theta\right]}_{(\star)} p_Z(z) \mathrm{d}z. \tag{V.2.3}$$

Pour résoudre le problème, on recherche $\mathring{g}(.)$ qui minimise $(\star)$ *pour tout* z, et il vient

$$\hat{\Theta} = g(Z), \text{ où } g(.) = \arg\min_{\mathring{g}(.)} \mathbb{E}_{\Theta|Z}\left[C(\mathring{g}(Z) - \Theta)\middle| z\right] \text{ pour tout } z. \tag{V.2.4}$$

V.3 Estimateur du minimum d'erreur quadratique moyenne (EMQ, MMSEE)

- On pose $C(\mathring{\Theta} - \Theta) = (\mathring{\Theta} - \Theta)^T Q (\mathring{\Theta} - \Theta)$, où $Q \geq 0$ donnée.
Alors,

$$\hat{\Theta}_{\text{MMSEE}} \text{ tel que pour } Z(\omega) = z,\ \hat{\Theta}_{\text{MMSEE}}(\omega) = \hat{\theta}_{\text{MMSEE}} = \mathbb{E}_{\Theta|Z}[\Theta|z] \tag{V.3.1}$$

$$= \int \theta p_{\Theta|Z}(\theta|z) \mathrm{d}\theta. \tag{V.3.2}$$

◦ L'estimé $\hat{\theta}_{\text{MMSEE}}$ du minimum d'erreur quadratique moyenne est donc la *moyenne a posteriori* de Θ conditionnellement à $Z = z$.
◦ L'estimateur $\hat{\Theta}_{\text{MMSEE}}$ est *non biaisé* et *à minimum de variance.*
◦ Illustration

V.4 Estimateur du maximum a posteriori (MAP)

- Si $C(\mathring{\Theta} - \Theta) = 1 - \delta(\mathring{\Theta} - \Theta)$, alors

$$\hat{\Theta}_{\text{MAP}} \text{ tel que pour } Z(\omega) = z,\ \hat{\Theta}_{\text{MAP}}(\omega) = \hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p_{\Theta|Z}(\theta|z). \tag{V.4.1}$$

◦ L'estimé $\hat{\theta}_{\text{MAP}}$ du maximum a posteriori est donc le *mode de la loi a posteriori* de Θ conditionnellement à $Z = z$.
◦ $\hat{\theta}_{\text{MAP}} = \arg\max_\theta \ln p_{\Theta|Z}(\theta|z) = \arg\max_\theta \left(\ln p_\Theta(\theta) + \ln p_{Z|\Theta}(z|\theta)\right)$.
◦ Et si pas de connaissance a priori ($p_\Theta(\theta)$ uniforme), alors $\hat{\Theta}_{\text{MAP}} = \hat{\Theta}_{\text{MLE}}$!!!
◦ Illustration

V.5 Estimateur du meilleur estimé linéaire (BLUE)

...

V.6 Cas d'un modèle linéaire Gaussien

- Si $Z = HX + V$, où $Z \in \mathbb{R}^N$, $X \in \mathbb{R}^M$, $V \in \mathbb{R}^N$, $\binom{X}{V} \sim \mathcal{N}\left(\binom{\hat{x}^-}{0}, \begin{pmatrix} P^- & 0 \\ 0 & R \end{pmatrix}\right)$ avec $R > 0$, alors $p_{X|Z}(x|z) = \mathcal{N}(x; \hat{x}^+, P^+)$, et les moments $\hat{x}^+ = \mathbb{E}_{X|Z}[X|z]$ et

$P^+ = \mathbb{E}_{X|Z}\Big[(X-\hat{x}^+)(X-\hat{x}^+)^T|z\Big]$ sont donnés par

$$\begin{cases} \hat{x}^+ = \hat{x}^- + K(z - H\hat{x}^-) \\ P^+ = P^- - KHP^- \\ K = P^-H^T\big(R + HP^-H^T\big)^{-1}. \end{cases} \tag{V.6.1}$$

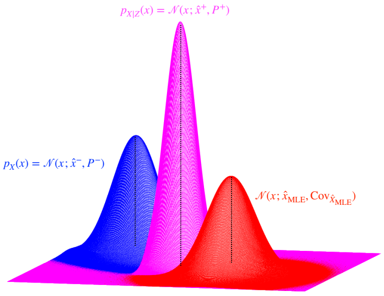
- Preuve : il suffit d'exprimer les moments de $\binom{X}{Z}$ et d'exploiter (I.2.14).
- Remarques : on montre (!) que

$$K = P^+H^TR^{-1}\,;\quad (P^+)^{-1} = (P^-)^{-1} + H^TR^{-1}H\,;\quad (P^+)^{-1}\hat{x}^+ = (P^-)^{-1}\hat{x}^- + H^TR^{-1}z. \tag{V.6.2}$$

- Interprétation en terme de *fusion d'estimés*
  Si $N \geq M$, H déterministe de rang plein M, alors il vient

$$\begin{cases} (P^+)^{-1}\hat{x}^+ = (P^-)^{-1}\hat{x}^- + (\mathrm{Cov}_{\hat{X}_{\mathrm{MLE}}})^{-1}\hat{x}_{\mathrm{MLE}} \\ (P^+)^{-1} = (P^-)^{-1} + (\mathrm{Cov}_{\hat{X}_{\mathrm{MLE}}})^{-1}, \end{cases} \tag{V.6.3}$$

  de sorte que $\hat{x}^-, P^-$ d'une part, et $\hat{x}_{\mathrm{MLE}}, \mathrm{Cov}_{\hat{X}_{\mathrm{MLE}}}$ d'autre part, jouent un rôle symétrique dans la constitution de $\hat{x}^+, (P^+)$!
- Illustration



  · Si $P^- = \infty\mathbb{I}$, on obtient une interprétation stochastique des moindres carrés récursifs...
  · $\hat{x}^+$ est d'autant plus proche de $\hat{x}^-$ (resp. de $\hat{x}_{\mathrm{MLE}}$) que P^- (resp. $\mathrm{Cov}_{\hat{X}_{\mathrm{MLE}}}$) est faible.
  · P^+ est à la fois plus petit que P^- et $\mathrm{Cov}_{\hat{X}_{\mathrm{MLE}}}$ (car apport d'information).