Introduction aux séries temporelles

Olivier Cappé, Maurice Charbit, Eric Moulines

30 mars 2007

Table des matières

Chapitre 1

Processus aléatoires stationnaires au second ordre

Le paragraphe 1.1 définit le formalisme probabiliste permettant de décrire les *processus aléatoires*. Les quelques exemples qui suivent illustrent la diversité des situations dans lesquelles la modélisation stochastique (ou aléatoire) des séries temporelles joue un rôle important.

Exemple 1.1 : Battements cardiaques
La figure 1.1 représente l'évolution, sur une durée totale de 900 *secondes, du rythme cardiaque d'un sujet au repos. Ce rythme est mesuré en nombre de battements par minute toutes les* 0.5 *secondes.*

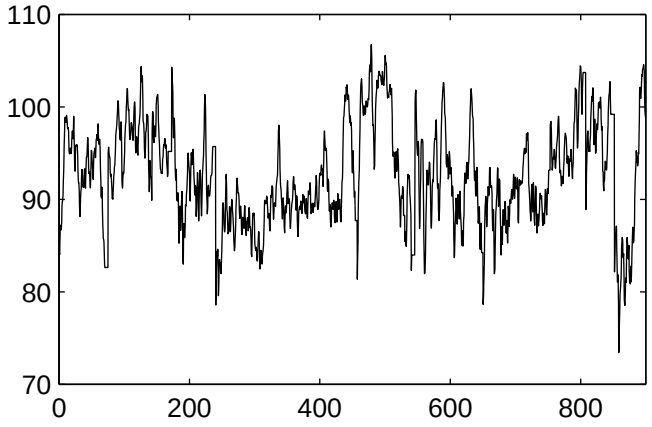

FIG. 1.1 – *Battements cardiaques : évolution du nombre de battements par minute en fonction du temps mesuré en seconde.*

Exemple 1.2 : Trafic internet
La figure 1.2 représente les temps d'inter-arrivées de paquets TCP, mesurés en secondes, sur la passerelle du laboratoire Lawrence Livermore. La trace représentée a été obtenue en enregistrant 2 heures de trafic. Pendant cette durée, environ 1.3 millions de paquets TCP, UDP, etc. ont été enregistrés, en utilisant la procédure tcpdump *sur une station Sun. D'autres séries de ce type peuvent être obtenues sur* The Internet Traffic Archive, http ://ita.ee.lbl.gov/.

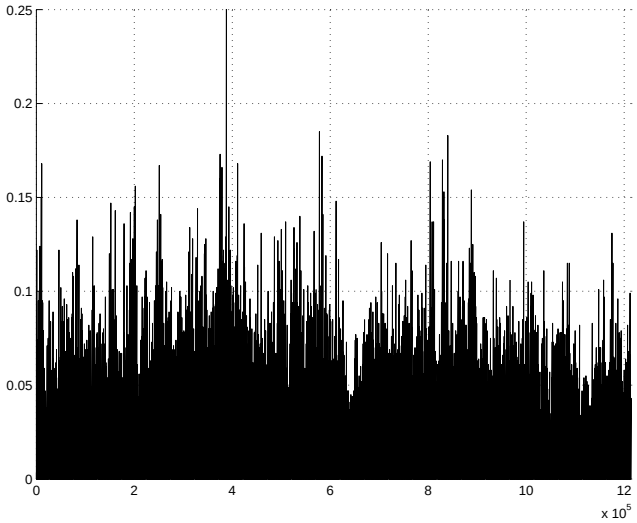

FIG. 1.2 – *Trace de trafic Internet : temps d'inter-arrivées de paquets TCP.*

Exemple 1.3 : Parole
La figure 1.3 représente un segment de signal vocal échantillonné (la fréquence d'échantillonnage est de 8000 Hz). Ce segment de signal correspond à la réalisation du phonème ch *(comme dans* ch*at) qui est un son dit* fricatif*, c'est-à-dire produit par les turbulences du flot d'air au voisinage d'une constriction (ou resserrement) du conduit vocal.*

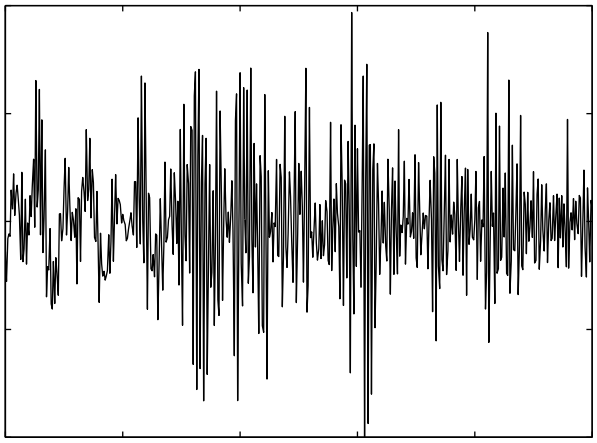

FIG. 1.3 – *Signal de parole échantillonné à* 8000 *Hz : son non voisé* ch.

Exemple 1.4 : Indice financier
La figure 1.4 représente les cours d'ouverture journaliers de l'indice Standard and Poor 500, du 2 *Janvier* 1990 *au 25 Août 2000. l'indice S&P*500 *est calculé à partir de* 500 *actions choisies parmi les valeurs cotées au New York Stock Exchange (NYSE) et au NASDAQ en fonction de leur capitalisation, leur liquidité, leur*

représentativité dans différents secteurs d'activité. Cet indice est obtenu en pondérant le prix des actions par le nombre total d'actions, le poids de chaque valeur dans l'indice composite étant proportionnel à la capitalisation.

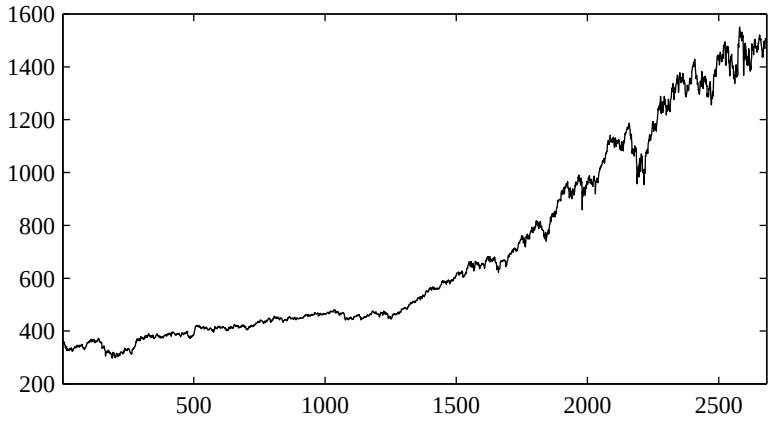

FIG. 1.4 – *Cours quotidien d'ouverture de l'indice S&P*500 *: entre Janvier 1990 et Août 2000.*

1.1 Propriétés générales

Définition 1.1 (Processus aléatoire)**.** *Soient* $(\Omega, \mathcal{F}, \mathbb{P})$ *un espace de probabilité,* T *un ensemble d'indices et* $(E, \mathcal{E})$ *un espace mesurable. On appelle processus aléatoire une famille* $\{X(t), t \in T\}$ *de v.a. à valeurs dans* $(E, \mathcal{E})$ *indexées par* $t \in T$*.*

Le paramètre t représente ici le temps. Lorsque $T \subset \mathbb{Z}$, nous dirons que le processus est à *temps discret* et, lorsque $T \subset \mathbb{R}$, que le processus est à *temps continu*. Dans la suite de cet ouvrage, nous nous intéresserons de façon prioritaire aux processus à temps discret $T \subset \mathbb{Z}$. Quant à $(E, \mathcal{E})$, nous considérerons le plus souvent $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (où $\mathcal{B}(\mathbb{R})$ est la tribu borélienne de $\mathbb{R}$) ou $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Dans le premier cas, on dira que le processus aléatoire est *scalaire*. Dans le second, nous dirons que le processus est *vectoriel*.

Notons qu'en fait un processus est une application $X : \Omega \times T \to E$ telle que :
- à chaque instant $t \in T$, l'application $\omega \mapsto X(t, \omega) \in (E, \mathcal{E})$ est une variable aléatoire,
- pour chaque épreuve $\omega \in \Omega$, l'application $t \mapsto X(t, \omega)$ est une fonction de $T \to E$ qui s'appelle la *trajectoire* associée à l'épreuve ω.

1.1.1 Répartitions finies

On note $\mathcal{I}$ l'ensemble des parties finies ordonnées de T. Un élément I de $\mathcal{I}$ s'écrit $I = \{t_1 < t_2 < \cdots < t_n\}$. On note $|I|$ le cardinal de I et $\mathbb{P}_I$ la loi du vecteur aléatoire $(X(t_1), X(t_2), \cdots, X(t_n))$, c'est-à-dire la mesure image par les variables aléatoires $(X(t_1), X(t_2), \cdots, X(t_n))$ de la probabilité $\mathbb{P}$: $\mathbb{P}_I$ est la probabilité sur $(E^{|I|}, \mathcal{E}^{\otimes |I|})$ définie par

$$\mathbb{P}_I A_1 \times A_2 \times \cdots \times A_n = \mathbb{P}X(t_1) \in A_1, X(t_2) \in A_2, \cdots, X(t_n) \in A_n \ , \tag{1.1}$$

où $\{A_1, \cdots, A_n\}$ sont des éléments quelconques de la tribu cE. La probabilité $\mathbb{P}_I$ est une *probabilité fini-dimensionnelle* du processus. Pour caractériser la loi d'un processus, il est nécessaire de disposer de la famille des répartitions finies, indexée par l'ensemble des parties finies ordonnées $\mathcal{I}$.

Définition 1.2. *On appelle* famille des répartitions finies *l'ensemble des répartitions finies,* $(\mathbb{P}_I, I \in \mathcal{I})$.

La spécification de la mesure image $\mathbb{P}_I$ permet de calculer la probabilité d'événements de la forme $\mathbb{P}\cap_{t\in I}\{X(t) \in A_t\}$ où $(A_t, t \in I)$ sont des éléments de la tribu $\mathcal{E}$, ou de manière équivalente, de calculer l'espérance $\mathbb{E}\left[\prod_{t\in I} f_t(X(t)\right]$ où $(f_t, t \in I)$ sont des fonctions boréliennes positives. Il est important de noter que, la donnée des répartitions finies ne permet pas *a priori* d'évaluer la probabilité d'un événement faisant intervenir un nombre infini d'indices de temps ; par exemple, pour un processus à temps discret indexé par $T = \mathbb{Z}$, les répartitions finies ne permettent pas, *a priori*, d'évaluer la probabilité d'un événement de la forme $\{\max_{t\in\mathbb{Z}} X(t) \geq a\}$. Soit $J \subset I$ deux parties finies ordonnées. Soit $\Pi_{I,J}$ la projection canonique de $E^{|I|}$ sur $E^{|J|}$, *i.e.*

$$\Pi_{I,J}(\{x(t_k), k \in I\}) = \{x(t_k), k \in J\}. \tag{1.2}$$

La projection canonique préserve uniquement les coordonnées du vecteur appartenant au sous ensemble d'indices J. L'équation (1.1) implique que :

$$\mathbb{P}_I \circ \Pi_{I,J} = \mathbb{P}_J \tag{1.3}$$

et donc, pour tout ensemble $A \in \mathcal{E}^{\otimes|J|}$, on a $\mathbb{P}_J(A) = \mathbb{P}_I(\Pi_{I,J}(A))$. Cette relation formalise le résultat intuitif que la distribution fini-dimensionnelle d'un sous-ensemble $J \subset I$ se déduit de la distribution fini-dimensionnelle P_I en "intégrant" par rapport aux variables $X(t_i)$ sur l'ensemble des indices appartenant au complémentaire de J dans I. Cette propriété montre que la famille des répartitions finies d'un processus est fortement structurée. En particulier, les répartitions finies doivent, au moins, vérifier les conditions de compatibilité (1.3). Nous allons voir dans la suite que cette condition est en fait aussi *suffisante*.

Soit Π_I la projection canonique de T sur I,

$$\Pi_I(\{x(t), t \in T\}) = \{x(t), t \in I\}. \tag{1.4}$$

Théorème 1.1 (Théorème de Kolmogorov)**.** *Soit* $\{\nu_I, I \in \mathcal{I}\}$ *une famille de probabilités indexées par l'ensemble des parties finies ordonnées de* T *telle, que pour tout* $I \in \mathcal{I}$, ν_I *est une probabilité sur* $(E^I, \mathcal{E}^{\otimes I})$. *Supposons de plus que la famille* $\{\nu_I, I \in \mathcal{I}\}$ *vérifie les conditions de compatibilité* (1.3)*, pour tout* $I, J \in \mathcal{I}$*, tel que* $I \subset J$, $\nu_I \circ \Pi_{I,J} = \nu_J$. *Il existe une probabilité unique* $\mathbb{P}$ *sur l'espace mesurable* $(E^T, \mathcal{E}^{\otimes T})$ *où* $\mathcal{E}^{\otimes T}$ *telle que, pour tout* $I \in \mathcal{I}$, $\nu_I = \mathbb{P} \circ \Pi_I$.

Soit $X = \{X_t, t \in T\}$ *le processus aléatoire défini sur* $(E^T, \mathcal{E}^T)$ *par* $X(t, \omega) = \omega(t)$. *Les répartitions finies du processus canonique* X *sur* $(E^T, \mathcal{E}^T, \mathbb{P})$ *sont données par* $\{\nu_I, I \in \mathcal{I}\}$.

On appellera ce processus le *processus canonique* de répartitions finies $(\nu_I, I \in \mathcal{I})$ et la probabilité $\mathbb{P}$ ainsi construite la *loi du processus* du processus canonique X. Cette loi est donc *entièrement* déterminée par la donnée des répartitions finies.

Exemple 1.5 : Suite de v.a. indépendantes
Soit $(\nu_n, n \in \mathbb{N})$ une suite de probabilités sur $(E, \mathcal{E})$. Pour $I = \{n_1 < n_2 < \cdots < n_p)$ on pose

$$\nu_I = \nu_{n_1} \otimes \cdots \otimes \nu_{n_p} \tag{1.5}$$

Il est clair que l'on définit ainsi une famille $(\nu_I, I \in \mathcal{I})$ compatible, c'est-à-dire, vérifiant la condition donnée par l'équation (1.3). Donc, si $\Omega = E^{\mathbb{N}}$, $X_n(\omega) = \omega_n$ et $\mathcal{F} = \sigma(X_n, n \in \mathbb{N})$, il existe une unique probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ telle que $(X_n, n \in \mathbb{N})$ soit une suite de v.a. indépendantes de loi ν_n.

1.1.2 Stationnarité stricte

La notion de stationnarité joue un rôle central dans la théorie des processus aléatoires. On distingue ci-dessous deux versions de cette propriété, la stationnarité "stricte" qui fait référence aux répartitions finies à l'invariance des répartitions finies par translation de l'origine des temps, et une notion plus faible, la stationnarité au second ordre, qui porte sur l'invariance par translation des moments d'ordre un et deux (lorsque ceux-ci existent).

Définition 1.3 (Stationnarité stricte)**.** *Un processus aléatoire est stationnaire au sens strict si les répartitions finies sont invariantes par translation de l'origine des temps, i.e. que, pour tout $\tau \in T$ et toute partie finie $I \in \mathcal{I}$, on a $\mathbb{P}_I = \mathbb{P}_{I+\tau}$ où $I + \tau = \{t + \tau, t \in I\}$.*

Exemple 1.6 : Processus i.i.d et transformations
Soit $\{Z(t)\}$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d). $\{Z(t)\}$ est un processus stationnaire au sens strict, car, pour toute partie finie ordonnée $I = \{t_1, < t_2 < \cdots < t_n\}$ nous avons :

$$\mathbb{P}Z(t_1) \in A_1, \cdots, Z(t_n) \in A_n = \prod_{j=1}^{n} \mathbb{P}Z(0) \in A_j$$

Soient k un entier et g une fonction borélienne de $\mathbb{R}^k$ dans $\mathbb{R}$. Il est facile de vérifier que le processus aléatoire $\{X(t)\}$ défini par

$$X(t) = g(Z(t), Z(t-1), \cdots, Z(t-k+1))$$

est encore un processus aléatoire stationnaire au sens strict. Par contre, ce processus obtenu par transformation n'est plus i.i.d dans la mesure où, dès que $k \geq 1$, $X(t), X(t+1), \ldots, X(t+k-1)$ bien qu'ils aient la même distribution marginale sont, en général, dépendants car fonctions de variables aléatoires communes. Un tel processus est dit k-dépendant dans la mesure où, par contre, $\tau \geq k$ implique que $X(t)$ et $X(t+\tau)$ sont indépendants (ils dépendent de deux groupes indépendants de k variables aléatoires).

Définition 1.4 (Processus du second ordre)**.** *Le processus $X = (X(t), t \in T)$ à valeurs dans $\mathbb{R}^d$ est dit du second ordre, si $\mathbb{E}\left[\|X(t)\|^2\right] < \infty$, où $\|x\|$ est la norme euclidienne de $x \in \mathbb{R}^d$.*

Notons que la *moyenne* $\mu(t) = \mathbb{E}[X(t)]$ est un vecteur de dimension d dépendant de t et que la *fonction d'autocovariance* définie par :

$$\Gamma(s,t) = \mathrm{cov}(X(s), X(t)) = \mathbb{E}\left[(X(t) - \mu(t))(X(s) - \mu(s))^T\right]$$

est une matrice de dimension $d \times d$ dépendant à la fois de s et de t.

Propriété 1.1. *Pour un processus du second ordre on a :*

1. *$\Gamma(s,s) \geq 0$, l'égalité ayant lieu si et seulement si $X(s)$ est presque sûrement égale à sa moyenne.*

2. *Symétrie hermitienne*[1]

$$\Gamma(s,t) = \Gamma(t,s)^T \tag{1.6}$$

3. *Type positif*
Pour tout n, pour toute suite d'instants $(t_1 < t_2 < \cdots < t_n)$ *et pour toute suite de vecteurs complexes* $(a_1, \cdots, a_n)$ *de dimension d, on a :*

$$\sum_{1 \leq k,m \leq n} a_k^H \Gamma(t_k, t_m) a_m \geq 0 \tag{1.7}$$

Démonstration. Formons la combinaison linéaire $Y = \sum_{k=1}^n a_k^H X(t_k)$. Y est une variable aléatoire complexe. Sa variance, qui est positive, s'écrit

$$\mathrm{var}(Y) = \mathbb{E}\left[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^*\right] \geq 0$$

On note $X_c(t) = X(t) - \mathbb{E}[X(t)]$ le processus centré. En développant $\mathrm{var}(Y)$ en fonction de $X_c(t_k)$, il vient :

$$\mathrm{var}(Y) = \mathbb{E}\left[\sum_{k=1}^n \lambda_k^H X_c(t_k) \sum_{m=1}^n X_c^T(t_k)\lambda_k\right] = \sum_{1 \leq k,m \leq n} \lambda_k^H \Gamma(t_k, t_m)\lambda_m$$

ce qui établit (1.7). ■

Dans le cas scalaire ($d = 1$), on note en général $\gamma(s,t)$ la covariance, en réservant la notation $\Gamma(s,T)$ au cas des processus vectoriels ($d > 1$).

1.1.3 Processus gaussiens

Définition 1.5 (Variable aléatoire gaussienne réelle)**.** *On dit que X est une variable aléatoire réelle gaussienne si sa loi de probabilité a pour fonction caractéristique :*

$$\phi_X(u) = \mathbb{E}\left[e^{iuX}\right] = \exp(i\mu u - \sigma^2 u^2/2)$$

où $\mu \in \mathbb{R}$ *et* $\sigma \in \mathbb{R}^+$.

On en déduit que $\mathbb{E}[X] = \mu$ et que $\mathrm{var}(X) = \sigma^2$. Si $\sigma \neq 0$, la loi possède une densité de probabilité qui a pour expression :

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Définition 1.6 (Vecteur gaussien réel)**.** *Un vecteur aléatoire réel de dimension n* $(X_1, \ldots, X_n)$ *est un vecteur gaussien si toute combinaison linéaire de* $X_1, \ldots, X_n$ *est une variable aléatoire gaussienne réelle.*

[1] L'exposant T sert à indiquer l'opération de transposition et l'exposant H l'opération de transposition et conjugaison.

Notons μ le vecteur moyenne de $(X_1, \ldots, X_n)$ et Γ la matrice de covariance. Par définition d'un vecteur aléatoire gaussien, la variable aléatoire $Y = \sum_{k=1}^{n} u_k X_k = u^T X$ est une variable aléatoire réelle gaussienne. Par conséquent, sa loi est complètement déterminée par sa moyenne et sa variance qui ont pour expressions respectives :

$$\mathbb{E}[Y] = \sum_{k=1}^{n} u_k \mathbb{E}[X_k] = u^T \mu \quad \text{et} \quad \text{var}(Y) = \sum_{j,k=1}^{n} u_j u_k \text{cov}(X_j, X_k) = u^T \Gamma u$$

On en déduit l'expression, en fonction de μ et de Γ, de la fonction caractéristique de la loi de probabilité d'un vecteur gaussien $X(1), \ldots, X(n)$:

$$\phi_X(u) = \mathbb{E}\left[\exp(iu^T X)\right] = \mathbb{E}[\exp(iY)] = \exp\left(iu^T \mu - \frac{1}{2} u^T \Gamma u\right) \tag{1.8}$$

De plus si Γ est de rang plein n, alors la loi de probabilité de X possède une densité dont l'expression est :

$$p_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Gamma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Gamma^{-1}(x-\mu)\right)$$

Dans le cas où Γ est de rang $r < n$, c'est à dire où Γ possède $n - r$ valeurs propres nulles, X se trouve, avec probabilité 1, dans un sous espace de dimension r de $\mathbb{R}^n$, dans la mesure où il existe $r - n$ combinaisons linéaires indépendantes a_i telles que $\text{cov}(a_i^T X) = 0$.

Définition 1.7 (Processus gaussien réel)**.** *On dit qu'un processus réel $X = \{X(t), t \in T\}$ est gaussien si, pour toute suite finie d'instants $\{t_1 < t_2 < \cdots < t_n\}$, $(X(t_1), X(t_2), \cdots, X(t_n))$ est un vecteur gaussien.*

D'après (1.8), la famille des répartitions finies est donc caractérisée par la donnée de la fonction moyenne $\mu : t \in T \mapsto \mu(t) \in \mathbb{R}$ et de la fonction de covariance $\gamma : (t, s) \in (T \times T) \mapsto \gamma(t, s) \in \mathbb{R}$. Réciproquement, donnons nous une fonction $\mu : t \in T \mapsto m(t) \in \mathbb{R}$ et une fonction de covariance $\gamma : (t, s) \in (T \times T) \mapsto \gamma(t, s) \in \mathbb{R}$ de type positif, c'est-à-dire telle que, pour tout n, toute suite $(u_1, \cdots, u_n)$ et toute suite $(t_1, \cdots, t_n)$ on ait :

$$\sum_{j=1}^{n} \sum_{k=1}^{n} u_j u_k \gamma(t_j, t_k) \geq 0 \tag{1.9}$$

On peut alors définir, pour $I = \{t_1 < \cdots < t_n\}$, une probabilité gaussienne ν_I sur $\mathbb{R}^n$ par :

$$\nu_I := \mathcal{N}_n(\mu_I, \Gamma_I) \tag{1.10}$$

où $\mu_I = (\mu(t_1), \cdots, \mu(t_n))$ et Γ_I est la matrice positive d'éléments $\gamma_I(m, k) = \gamma(t_m, t_k)$, où $1 \leq m, k \leq n$. La famille $(\nu_I, I \in \mathcal{I})$, ainsi définie, vérifie les conditions de compatibilité et l'on a ainsi établi, d'après le théorème 1.1, le résultat suivant :

Théorème 1.2. *Soit $r \mapsto \mu(t)$ une fonction et $(s, t) \mapsto \gamma(s, t)$ une fonction de type positif (vérifiant l'équation (1.9)). Il existe un espace de probability $(\Omega, \mathcal{F}, \mathbb{P})$ et un processus aléatoire $\{X(t), t \in T\}$ gaussien défini sur cet espace vérifiant*

$$\mu(t) = \mathbb{E}[X(t)] \quad \textit{et} \quad \gamma(s, t) = \mathbb{E}[(X(s) - \mu(s))(X(t) - \mu(t))]$$

1.2 Stationnarité au second ordre

Dans la suite du document, nous considérons principalement le cas de processus à temps discret (avec $T = \mathbb{Z}$) pour lesquels nous utiliserons la notation X_t plutôt que $X(t)$, cette dernière étant réservée aux processus à temps continus. Par ailleurs, et sauf indication du contraire, les processus considérés sont en général à valeur dans $\mathbb{R}$.

Définition 1.8 (Stationnarité au second ordre)**.** *Le processus $\{X_t, t \in T\}$ est dit stationnaire au second ordre si :*

- *X est un processus du second ordre, i.e. $\mathbb{E}\left[|X_t|^2\right] < +\infty$,*
- *pour tout $t \in T$, $\mathbb{E}[X_t] = \mu$,*
- *pour tout couple $(s,t) \in T \times T$,*

$$\gamma(s,t) = \mathbb{E}\left[(X_t - \mu)(X_s - \mu)^T\right] = \gamma(t-s)$$

1.2.1 Covariance d'un processus stationnaire au second ordre

Propriété 1.2. *La fonction d'autocovariance $\gamma : T \to \mathbb{R}$ d'un processus stationnaire au second ordre vérifie les propriétés suivantes qui sont une conséquence directe des propriétés 1.1.*

1. *Symétrie hermitienne :*
$$\gamma(h) = \gamma(-h)$$
2. *caractère positif Pour toute partie finie $I = \{t_1 < \cdots < t_n\}$ et toute suite $(a_1, \cdots, a_n)$ de valeurs complexes, $\lambda_k \in \mathbb{C}$,*
$$\sum_{k=1}^{n}\sum_{j=1}^{n} a_k^* \gamma(k-j) a_j \geq 0$$

Ces propriétés découlent immédiatement des propriétés de la fonction d'autocovariance d'un processus. Les matrices de covariance de sections de n valeurs consécutives du processus sont positives d'après le point 2 de la propriété 1.2. Elles possèdent de plus une structure particulière, dite de *Toëplitz* (caractérisée par le fait que $(\Gamma_n)_{ij} = \gamma(i-j)$) :

$$\begin{aligned}\Gamma_n &= \mathbb{E}\left[[(X_t - \mu_X)\dots(X_{t-n+1} - \mu_X)]^T[(X_t - \mu_X)\dots(X_{t-n+1} - \mu_X)]\right] \\ &= \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & & & \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix}\end{aligned} \tag{1.11}$$

Définition 1.9 (Fonction d'autocorrélation)**.** *Pour un processus stationnaire, on appelle fonction d'autocorrélation $\rho(h) = \gamma(h)/\gamma(0)$. Il s'agit d'une quantité normalisée dans le sens où $\rho(1) = 1$ et $|\rho(k)| \leq 1$.*

En effet, l'inégalité de Cauchy-Schwarz appliquée à $\gamma(k)$ s'écrit

$$|\gamma(h)| = |\mathbb{E}[(X_{t+h} - \mu_X)(X_t - \mu_X)]| \leq \sqrt{\mathbb{E}[(X_{t+h} - \mu_X)^2]\,\mathbb{E}[(X_t - \mu_X)^2]} = \gamma(0)$$

la dernière inégalité découlant de l'hypothèse de stationnarité. Attention, certaines références (livres et publications), en général anciennes, utilisent (incorrectement) le terme de "fonction d'autocorrélation" pour $\gamma(h)$. Dans la suite de ce document, le terme autocorrélation est réservée à la quantité normalisée $\rho(h)$.

Exemple 1.7 : Processus retourné temporel
Soit X_t un processus aléatoire stationnaire au second ordre de moyenne μ_X et de fonction d'autocovariance $\gamma_X(h)$. On note $X_t^r = X_{-t}$ le processus retourné temporel. Alors X_t^r est un processus stationnaire au second ordre de même moyenne et de même fonction d'autocovariance que le processus X_t. En effet on a :

$$\mathbb{E}[X_t^r] = \mathbb{E}[X_{-t}] = \mu_X$$
$$\mathrm{cov}(X_{t+h}^r, X_t^r) = \mathrm{cov}(X_{-t-h}, X_{-t}) = \gamma_X(-h) = \gamma_X(h)$$

Définition 1.10 (Bruit blanc)**.** *On appelle bruit blanc un processus aléatoire stationnaire au second ordre, centré, de fonction d'autocovariance, $\gamma(s,t) = \gamma(t-s) = \sigma^2\delta_{t,s}$. On le notera $\{X_t\} \sim \mathrm{BB}(0,\sigma^2)$.*

Définition 1.11 (Bruit blanc fort)**.** *On appelle bruit blanc fort une suite du second ordre de variables aléatoires $\{X_t\}$, centrées, indépendantes et identiquement distribuées (i.i.d.) de variance $\mathbb{E}\left[X_t^2\right] = \sigma^2 < \infty$. On le notera $\{X_t\} \sim \mathrm{IID}(0,\sigma^2)$.*

Par définition si $\{X_t\} \sim \mathrm{IID}(0,\sigma^2)$, $\mathbb{E}[X_t] = 0$, $\mathbb{E}\left[X_t^2\right] = \sigma^2$ et pour tout $h \neq 0$, $\mathbb{E}[X_{t+h}X_t] = \mathbb{E}[X_{t+h}]\,\mathbb{E}[X_t] = 0$. $\{X_t\}$ est donc également stationnaire au second ordre, de fonction d'autocovariance $\gamma(s,t) = \sigma^2\delta(t-s)$. La structure de bruit blanc fort est clairement plus contraignante que celle de simple bruit blanc. En général, il est tout à fait inutile de faire un telle hypothèse lorsque l'on s'intéresse à des modèles de signaux supposés stationnaires au second ordre. Il arrivera cependant dans la suite que nous adoptions cette hypothèse plus forte afin de simplifier les développements mathématiques. Notons que dans le cas d'une série temporelle gaussienne, ces deux notions sont confondues puisque la loi gaussienne est complètement caractérisée par les moments du premier et du second ordre (un bruit blanc gaussien est donc également un bruit blanc fort).

Exemple 1.8 : Processus MA(1)
Soit $\{X_t\}$ le processus stationnaire au second ordre défini par :

$$X_t = Z_t + \theta Z_{t-1} \tag{1.12}$$

où $\{Z_t\} \sim \mathrm{BB}(0,\sigma^2)$ et $\theta \in \mathbb{R}$. On vérifie aisément que $\mathbb{E}[X_t] = 0$ et que :

$$\gamma(t,s) = \begin{cases} \sigma^2(1+\theta^2) & t = s \\ \sigma^2\theta & |t-s| = 1 \\ 0 & |t-s| > 1 \end{cases}$$

Le processus X_t est donc bien stationnaire au second ordre. Un tel processus est appelé processus à moyenne ajusté d'ordre 1. Cette propriété se généralise, sans difficulté, à un processus MA(q). Nous reviendrons plus en détail, paragraphe 1.4, sur la définition et les propriétés de ces processus.

Exemple 1.9 : Processus harmonique
Soient $\{A_k\}_{1\leq k\leq N}$ N variables aléatoires vérifiant $\mathrm{cov}(A_k, A_l) = \sigma_k^2\delta(k-l)$ et $\{\Phi_k\}_{1\leq k\leq N}$, N variables aléatoires indépendantes et identiquement distribuées (i.i.d), de loi uniforme sur $[-\pi,\pi]$, et indépendantes de $\{A_k\}_{1\leq k\leq N}$. On définit :

$$X_t = \sum_{k=1}^{N} A_k \cos(\lambda_k t + \Phi_k) \tag{1.13}$$

où $\{\lambda_k\} \in [-\pi, \pi]$ sont N pulsations. Le processus X_t est appelé processus harmonique. On vérifie aisément que $\mathbb{E}[X_t] = 0$ et que sa fonction d'autocovariance est donnée par :

$$\gamma(h) = \mathbb{E}[X_{t+h}X_t] = \frac{1}{2}\sum_{k=1}^{N} \sigma_k^2 \cos(\lambda_k h)$$

Le processus harmonique est donc stationnaire au second ordre.

Exemple 1.10 : Marche aléatoire
Soit S_t le processus défini sur $t \in \mathbb{N}$ par $S_t = X_0 + X_1 + \cdots + X_t$, où X_t est un bruit blanc. Un tel processus est appelé une marche aléatoire. *On en déduit que $\mathbb{E}[S_t] = 0$, que $\gamma(t,t) = \mathbb{E}\left[X_t^2\right] = t\sigma^2$ et que, pour $h > 0$, on a :*

$$\gamma(t+h, t) = \mathbb{E}[(S_t + X_{t+1} + \cdots + X_{t+h})S_t] = t\sigma^2$$

Le processus $\{S_t\}$ n'est donc pas stationnaire au second ordre.

Exemple 1.11
Nous allons montrer que la suite définie, pour $h \in \mathbb{Z}$, par :

$$R(h) = \begin{cases} 1 & h = 0, \\ \rho & |h| = 1 \\ 0 & |h| \geq 2 \end{cases}$$

est la fonction d'autocovariance d'un processus stationnaire au second ordre si et seulement si $|\rho| \leq 1/2$. Nous avons déjà montré exemple 1.8 que la fonction d'autocovariance d'un processus MA(1) est donnée par :

$$\gamma(h) = \begin{cases} \sigma^2(1+\theta^2) & \textit{pour} \quad h = 0 \\ \sigma^2\theta & \textit{pour} \quad |h| = 1 \\ 0 & \textit{pour} \quad |h| > 1 \end{cases}$$

La suite $R(h)$ est donc la fonction d'autocovariance d'un processus MA(1) si et seulement si $\sigma^2(1+\theta^2) = 1$ et $\sigma^2\theta = \rho$. Lorsque $|\rho| \leq 1/2$, ce système d'équations admet comme solution :

$$\theta = (2\rho)^{-1}(1 \pm \sqrt{1 - 4\rho^2}) \quad \textit{et} \quad \sigma^2 = (1+\theta^2)^{-1}$$

Lorsque $|\rho| > 1/2$, ce système d'équations n'admet pas de solution réelles et la suite $R(h)$ n'est donc pas la fonction d'autocovariance d'un processus MA(1). On vérifie facilement que $R(h)$ ne vérifie pas dans ce cas la condition de positivité (en prenant $a_k = (-1)^k$ pour $\rho > 1/2$ et $a_k = 1$ dans le cas opposé). Pour $|\rho| > 1/2$, $R(h)$ n'est donc pas une séquence d'autocovariance.

1.2.2 Interprétation de la fonction d'autocovariance

Dans les exemples précédents, nous avons été amené à évaluer la fonction d'autocovariance de processus pour quelques exemples simples de séries temporelles. Dans la plupart des problèmes d'intérêt pratique, nous ne partons pas de modèles de série temporelle définis *a priori*, mais d'*observations*, $\{x_1, \cdots, x_n\}$ associées à une *réalisation* du processus. Afin de comprendre la structure de dépendance entre les différentes observations, nous serons amenés à *estimer* la loi du processus, ou du moins des caractéristiques de ces lois. Pour un processus stationnaire au second ordre, nous pourrons, à titre d'exemple, estimer sa moyenne par la *moyenne empirique* :

$$\hat{\mu}_n = n^{-1}\sum_{k=1}^{n} x_k$$

et les fonctions d'autocovariance et d'autocorrélation par les fonctions d'autocorrélation et d'autocovariance *empiriques*

$$\hat{\gamma}(h) = n^{-1} \sum_{k=1}^{n-|h|} (x_k - \hat{\mu}_n)(x_{k+|h|} - \hat{\mu}_n) \quad \text{et} \quad \hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$$

Lorsqu'il est *a priori* raisonnable de penser que la série considérée est stationnaire au second ordre, la moyenne empirique, la fonction d'autocovariance empirique et la fonction d'autocorrélation empirique sont de "bons" estimateurs, dans un sens que nous préciserons chapitre 2. L'analyse de la fonction d'autocovariance empirique est un élément permettant de guider le choix d'un modèle approprié pour les observations. Par exemple, le fait que la fonction d'autocovariance empirique soit *proche* de zéro pour tout $h \neq 0$ (proximité qu'il faudra définir dans un sens statistique précis) indique par exemple qu'un bruit blanc est un modèle adéquat pour les données. La figure 1.5 représente les 100 premières valeurs de la fonction d'autocorrélation empirique de la série des battements cardiaques représentés figure 1.1. On observe que cette série est *positivement corrélée* c'est-à-dire que les fonctions coefficients d'autocorrélation sont positifs et significativement non nuls. Nous avons, à titre de comparaison, représenté aussi la fonction d'autocorrélation empirique d'une trajectoire de même longueur d'un bruit blanc gaussien. Une forte corrélation peut être interprétée comme l'indice d'une

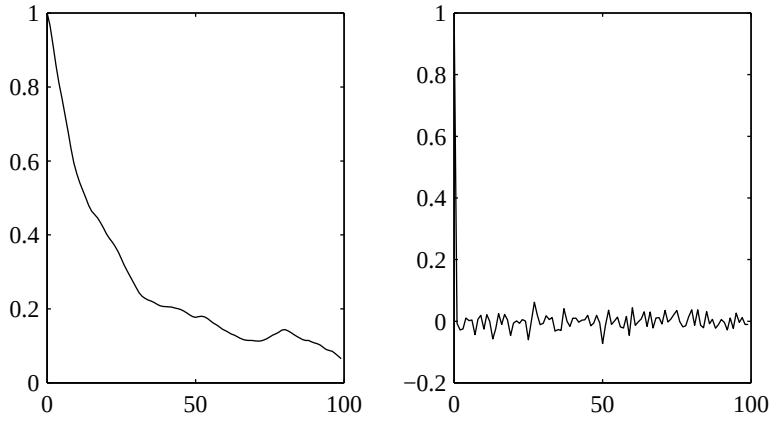

FIG. 1.5 – *Courbe de gauche : fonction d'autocorrélation empirique de la série des battements cardiaques (figure 1.1). Courbe de droite : fonction d'autocorrélation empirique d'une trajectoire de même longueur d'un bruit blanc gaussien.*

dépendance linéaire. Ainsi la figure 1.6 montre que le fait que $\hat{\rho}(1) = 0.966$ pour la série des battements cardiaques se traduit par une très forte prédictabilité de X_{t+1} en fonction de X_t (les couples de points successifs s'alignent quasiment sur une droite). Nous montrerons au chapitre 4, que dans un tel contexte, $\mathbb{E}\left[(X_{t+1} - \mu) - \rho(1)(X_t - \mu)\right] = (1 - \rho^2)\mathrm{cov}(X_t)$, c'est à dire, compte tenu de la valeur estimée pour $\rho(1)$, que la variance de "l'erreur de prédiction" $X_{t+1} - [\mu + \rho(1)(X_t - \mu)]$ est 15 fois plus faible que celle du signal original. L'indice S&P500 tracé (fig. 1.4) présente un cas de figure plus difficile, d'une part parce que la série de départ ne saurait être tenue pour stationnaire et qu'il

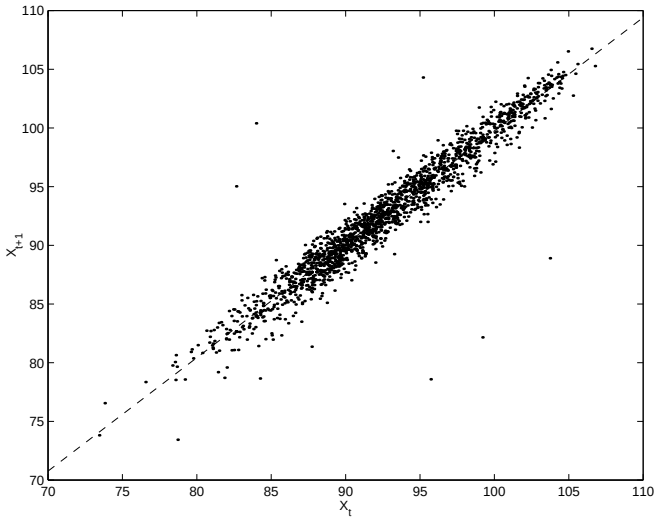

FIG. 1.6 – X_{t+1} *en fonction de* X_t *pour la série des battements cardiaques de la figure 1.1). Les tirets figurent la meilleure droite de régression linéaire de* X_{t+1} *sur* X_t.

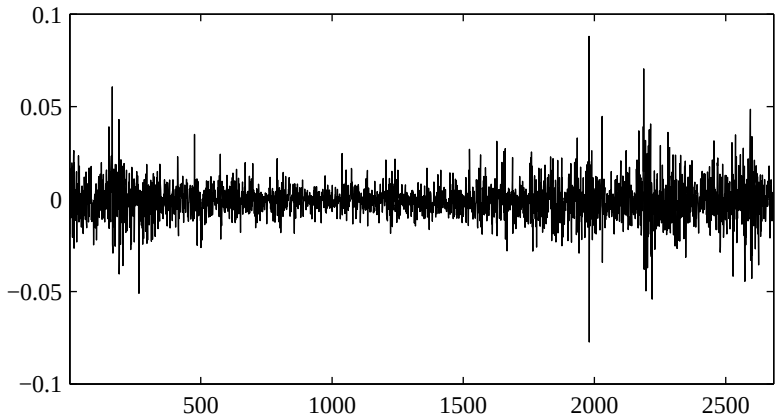

FIG. 1.7 – *Log-Retour de la série S&P 500 (figure 1.4).*

nous faudra considérer la série des évolutions journalières ; d'autre part, parce que selon le choix de la transformation des données considérées, la série transformée présente ou non des effets de corrélation. On définit tout d'abord les *log-retours* de l'indice S&P500 comme les différences des logarithmes de l'indice à deux dates successives :

$$X_t = \log(S_t) - \log(S_{t-1}) = \log\left(1 + \frac{S_t - S_{t-1}}{S_{t-1}}\right)$$

La série des log-retours de la série S&P 500 est représentée figure 1.7. Les coefficients d'autocorrélation

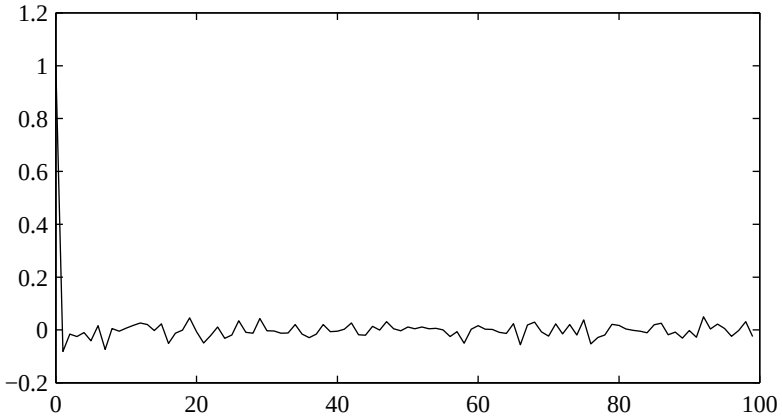

FIG. 1.8 – *Fonction d'autocorrélation empirique de la série des log-retours de l'indice S&P 500.*

empiriques de la série des log-retours sont représentés figure 1.8. On remarque qu'ils sont approximativement nuls pour $h \neq 0$ ce qui suggère de modéliser la série des log-retours par un bruit blanc (une suite de variables décorrélées). Il est intéressant d'étudier aussi la série des log-retours absolus, $A(t) = |X_t|$. On peut, de la même façon, déterminer la suite des coefficients d'autocorrélation empirique de cette série, qui est représentée dans la figure 1.9. On voit, qu'à l'inverse de la série des log-retours, la série des valeurs absolues des log-retours est positivement corrélée, les valeurs d'autocorrélation étant significativement non nuls pour $|h| \leq 100$. On en déduit, en particulier, que la suite des log-retours peut être modélisée comme un bruit blanc, mais pas un bruit blanc fort : en effet, pour un bruit blanc fort X_t, nous avons, pour toute fonction f telle que $\mathbb{E}\left[f(X_t)^2\right] = \sigma_f^2 < \infty$, $\mathrm{cov}(f(X_{t+h}), f(X_t) = 0$ pour $h \neq 0$ (les variables $f(X_{t+h})$ et $f(X_t)$ étant indépendantes, elles sont a fortiori non corrélées). Nous reviendrons dans la suite du cours sur des modèles possibles pour de telles séries.

1.2.3 Mesure spectrale d'un processus stationnaire au second ordre à temps discret

Dans toute la suite, I désigne l'intervalle $[-\pi, \pi]$ et $\mathcal{B}(I)$ la tribu de borélienne associée. Le théorème d'Herglotz ci dessous établit l'équivalence entre la fonction d'autocovariance et une mesure finie définie sur l'intervalle $\{I, \mathcal{B}(I)\}$. Cette mesure, appelée *mesure spectrale du processus*, joue un rôle analogue à celui de la représentation de Fourier pour les signaux déterministes. En particulier elle confère une expression simple aux formules de filtrage linéaire.

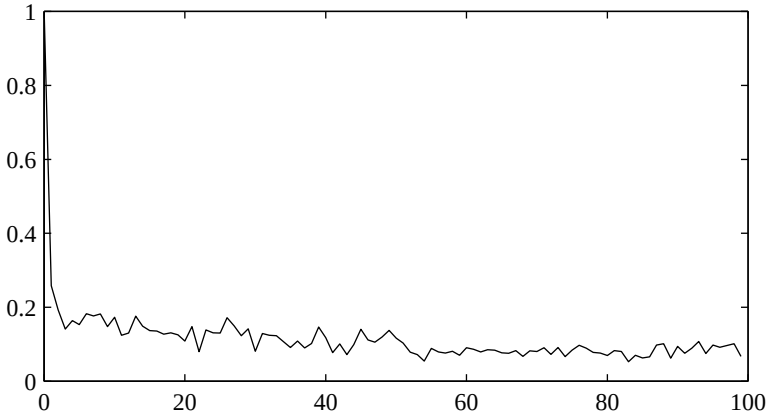

FIG. 1.9 – *Fonction d'autocorrélation empirique de la série des valeurs absolues des log-retours de l'indice S&P 500.*

Théorème 1.3 (Herglotz)**.** *Une suite $\{\gamma(h)\}_{h\in\mathbb{Z}}$ est de type positif si et seulement si il existe une mesure positive sur $\{I, \mathcal{B}(I)\}$ telle que :*

$$\gamma(h) = \int_I e^{ih\lambda}\nu(d\lambda) \tag{1.14}$$

Si la suite $\gamma(h)$ est sommable (i.e. $\sum_h |\gamma(h)| < \infty$), la mesure ν possède une densité f (fonction positive) par rapport à la mesure de Lebesgue sur $\{I, \mathcal{B}(I)\}$, donnée par la série entière uniformément convergente :

$$f(\lambda) = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}}\gamma(h)e^{-ih\lambda} \geq 0$$

Lorsque γ est la fonction d'autocovariance d'un processus stationnaire au second ordre, la mesure ν est appelée la mesure spectrale *et la fonction f, lorsque qu'elle existe, est dite* densité spectrale de puissance*.*

Démonstration. Tout d'abord si $\gamma(n)$ a la représentation (1.14), il est clair que $\gamma(n)$ est de type positif. En effet, pour tout n et toute suite $\{a_k \in \mathbb{C}\}_{1\leq k\leq n}$, on a :

$$\sum_{k,m} a_k a_m^* \gamma(k-m) = \int_I \sum_{k,m} a_k a_m^* e^{ik\lambda}e^{-im\lambda}\nu(d\lambda) = \int_I \left|\sum_k a_k e^{ik\lambda}\right|^2 \nu(d\lambda) \geq 0$$

Réciproquement, supposons que $\gamma(n)$ soit une suite de type positif et considérons la suite de fonctions indexée par n :

$$f_n(\lambda) = \frac{1}{2\pi n}\sum_{k=1}^{n}\sum_{m=1}^{n}\gamma(k-m)e^{-ik\lambda}e^{im\lambda} = \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1}\left(1-\frac{|k|}{n}\right)\gamma(k)e^{-ik\lambda} = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma_n(k)e^{-ik\lambda}$$

où nous avons posé :

$$\gamma_n(k) = \mathbf{I}_{\{-(n-1),\cdots,(n-1)\}}(k)\left(1 - \frac{|k|}{n}\right)\gamma(k)$$

qui vérifie $|\gamma_n(k)e^{-ik\lambda}| \leq |\gamma(k)|$ et $\lim_{n\to\infty}\gamma_n(k) = \gamma(k)$. Par construction, $f_n(\lambda)$ est une fonction positive (pour tout n) du fait de la positivité de la séquence d'autocovariance $\gamma(k)$. En supposant de plus[2] que $\sum_{k=-\infty}^{\infty}|\gamma(k)| < \infty$, une application directe du théorème de convergence dominé montre que :

$$\lim_{n\to\infty} f_n(\lambda) = \frac{1}{2\pi}\lim_{n\to\infty}\sum_{k=-\infty}^{\infty}\gamma_n(k)e^{-ik\lambda} = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\lim_{n\to\infty}\gamma_n(k)e^{-ik\lambda} = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)e^{-ikt} = f(\lambda)x$$

et donc $f(\lambda)$ est positive comme limite de fonctions positives. Une application directe du théorème de Fubini (la permutation étant légitime car $\int_I\sum_{k=-\infty}^{\infty}|\gamma(k)|d\lambda < \infty$), montre que, pour tout $h \in \mathbb{Z}$, on a :

$$\int_I f(\lambda)e^{ih\lambda}d\lambda = \sum_{k=-\infty}^{\infty}\gamma(k)\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{i(h-k)\lambda}d\lambda = \gamma(h)$$

■

Propriété 1.3 (Corollaire du théorème d'Herglotz)**.** *Une suite $R(h)$ à valeurs complexes absolument sommable est de type positif si et seulement si la fonction :*

$$f(\lambda) = \frac{1}{2\pi}\sum_{h=-\infty}^{+\infty}R(h)e^{-ih\lambda}$$

est positif pour tout $\lambda \in I$.

Exemple 1.12
En reprenant l'exemple 1.11, on vérifie immédiatement que $R(h)$ est de module sommable et que :

$$f(\lambda) = \frac{1}{2\pi}\sum_{k}R(h)e^{-ih\lambda} = \frac{1}{2\pi}(1 + 2\rho\cos(\pi\lambda))$$

et donc que la séquence est une fonction d'autocovariance uniquement lorsque $|\rho| \leq 1/2$.

Exemple 1.13 : Densité spectrale de puissance du bruit blanc
La fonction d'autocovariance d'un bruit blanc est donnée par $\gamma(h) = \sigma^2\delta(h)$, d'où l'expression de la densité spectrale correspondante

$$f(\lambda) = \frac{\sigma^2}{2\pi}$$

La densité spectrale d'un bruit blanc est donc constante. Cette propriété est à l'origine de la terminologie "bruit blanc" qui provient de l'analogie avec le spectre de la lumière blanche constant dans toute la bande de fréquences visibles.

[2]Nous donnons ici une preuve élémentaire grâce à l'hypothèse que la suite des coefficients d'autocovariance est absolument sommable. La démonstration, dans le cas général, requiert l'utilisation d'arguments plus complexes de théorie des probabilités. Elle est donnée annexe A.

Exemple 1.14 : Densité spectrale de puissance du processus MA(1)
Le processus MA(1) introduit dans l'exemple 1.8 possède une séquence d'autocovariance donnée par $\gamma(0) = \sigma^2(1+\theta^2)$, $\gamma(1) = \gamma(-1) = \sigma^2\theta$ *et* $\gamma(h) = 0$ *sinon (cf. exemple 1.8). D'où l'expression de sa densité spectrale :*

$$f(\lambda) = \frac{\sigma^2}{2\pi}(2\theta\cos(\lambda) + (1+\theta^2)) = \frac{\sigma^2}{2\pi}\left|1+\theta e^{-i\lambda}\right|^2$$

La densité spectrale d'un tel processus est représentée figure 1.10 pour $\theta = -0.9$ *et* $\sigma^2 = 1$ *avec une échelle logarithmique (dB).*

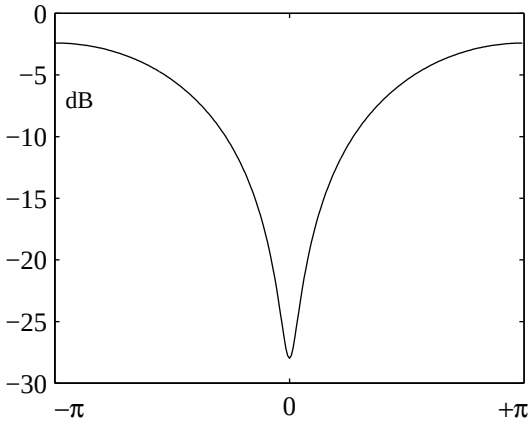

FIG. 1.10 – *Densité spectrale (en dB) d'un processus MA-1, défini par l'équation (1.12) pour* $\sigma = 1$ *et* $\theta = -0.9$.

Exemple 1.15 : Mesure spectrale du processus harmonique
La fonction d'autocovariance du processus harmonique $X_t = \sum_{k=1}^{N} A_k \cos(\lambda_k t + \Phi_k)$ *(voir exemple 1.9) est donnée par :*

$$\gamma(h) = \frac{1}{2}\sum_{k=1}^{N} \sigma_k^2 \cos(\lambda_k h) \tag{1.15}$$

où $\sigma_k^2 = \mathbb{E}\left[A_k^2\right]$. *Cette suite de coefficients d'autocovariance n'est pas sommable et la mesure spectrale n'admet pas de densité. En notant cependant que :*

$$\cos(\lambda_k h) = \frac{1}{2}\int_{-\pi}^{\pi} e^{ih\lambda}(\delta_{\lambda_k}(d\lambda) + \delta_{-\lambda_k}(d\lambda))$$

où $\delta_{x_0}(d\lambda)$ *désigne la mesure de Dirac au point* x_0 *(cette mesure associe la valeur* 1 *à tout borélien de* $[-\pi, \pi]$ *contenant* x_0 *et la valeur* 0 *sinon), la mesure spectrale du processus harmonique peut s'écrire :*

$$\nu(d\lambda) = \frac{1}{4}\sum_{k=1}^{N} \sigma_k^2 \delta_{\lambda_k}(d\lambda) + \frac{1}{4}\sum_{k=1}^{N} \sigma_k^2 \delta_{-\lambda_k}(d\lambda)$$

Elle apparaît donc comme une somme de mesures de Dirac, dont les masses σ_k^2 *sont localisées aux pulsations des différentes composantes harmoniques.*

Une remarque intéressante est que par rapport aux autres exemples étudiés, le processus harmonique est très particulier en ce qu'il possède une fonction d'autocovariance, donnée par 1.15, non absolument sommable ($\gamma(h)$ ne tend pas même vers 0 pour les grandes valeurs de h) et que par la suite, il admet une mesure spectrale mais pas une densité spectrale. La propriété suivante, à démontrer à titre d'exercice, implique que le processus harmonique est en fait entièrement prédictible à partir de quelques unes de ses valeurs passées.

Propriété 1.4. *S'il existe un rang n pour lequel la matrice de covariance Γ_n définie en (1.11) est non inversible, le processus correspondant X_t est prédictible dans le sens où il existe une combinaison linéaire $a_1, \dots a_l$ avec $l \leq n-1$ telle que $X_t = \sum_{k=1}^{l} a_k X_{t-k}$, l'égalité ayant lieu presque sûrement.*

L'expression de la fonction d'autocovariance, obtenue en (1.15) pour le processus harmonique, montre que les matrices de covariances associées s'écrivent comme la somme de $2N$ matrices complexes de rang 1. Par conséquent, les matrices Γ_n ne sont pas inversibles dès que $n > 2N$, ce qui implique que le processus harmonique est prédictible dès lors que l'on en a observé $2N$ valeurs. Ce résultat est sans surprise compte tenu du fait que les trajectoires de ce processus sont des sommes de sinusoïdes de fréquences $\lambda_1, \dots, \lambda_N$ dont seules les amplitudes et les phases sont aléatoires. La propriété suivante donne une condition suffisante simple pour éviter ce type de comportements "extrêmes". Cette propriété implique en particulier que, pour une fonction d'autocovariance absolument sommable (tous les exemples vus ci-dessus en dehors du processus harmoniques), les valeurs futures du processus correspondant ne sont pas prédictibles sans erreur à partir d'un ensemble fini de valeurs passées du processus. Nous reviendrons en détail sur ces problèmes de prédiction au chapitre 4.

Propriété 1.5. *Soit $\gamma(h)$ la fonction d'autocovariance d'un processus stationnaire au second ordre. On suppose que $\gamma(0) > 0$ et que $\gamma(h) \to 0$ quand $h \to \infty$. Alors, quel que soit n, la matrice de covariance définie en (1.11) est de rang plein et donc inversible .*

Démonstration. Supposons qu'il existe une suite de valeurs complexes $(a_1, \dots, a_n)$ non toutes nulles, telle que $\sum_{k=1}^{n} \sum_{m=1}^{n} a_k a_m^* \gamma(k-m) = 0$. En notant ν_X la mesure spectrale de X_t, on peut écrire :

$$0 = \sum_{k=1}^{n} \sum_{m=1}^{n} a_k a_m^* \int_I e^{i(k-m)\lambda} \nu_X(d\lambda) = \int_I \left| \sum_{k=1}^{n} a_k e^{ik\lambda} \right|^2 \nu_X(d\lambda)$$

Ce qui implique que $\left|\sum_{k=1}^{n} a_k e^{ik\lambda}\right|^2 = 0$ ν_X presque partout, c'est à dire que $\nu_X(\{\lambda : \left|\sum_{k=1}^{n} a_k e^{ik\lambda}\right|^2 \neq 0\}) = \nu_X(I - Z) = 0$ où $Z = \{\lambda_1, \dots, \lambda_M : \sum_{k=1}^{n} a_k e^{ik\lambda_m} = 0\}$ désigne l'ensemble *fini* ($M < n$) des racines $x \in I$ du polynôme trigonométrique $\sum_{k=1}^{n} a_k e^{ik\lambda}$. Par conséquent, les seuls éléments de $\mathcal{B}(I)$, qui peuvent être de mesure non nulle pour ν_X, sont les singletons $\{\lambda_m\}$. Ce qui implique que $\nu_X = \sum_{m=1}^{M} a_m \delta_{\lambda_m}$ (où $a_m \geq 0$ ne peuvent être tous nuls si $\gamma(0) \neq 0$). Mais, dans ce cas, $\gamma(h) = \sum_{m=1}^{M} a_m e^{ih\lambda_m}$, ce qui contredit l'hypothèse que $\gamma(h)$ tend vers 0 quand n tend vers l'infini. ■

Une autre preuve est donnée exercice **??**.

1.3 Filtrage des processus

Dans ce paragraphe, nous nous intéressons au filtrage des processus. On introduit tout d'abord l'*opérateur de retard*, noté B (comme *backshift* en anglais), dont l'effet sur le processus $\{X_t\}$ défini sur

$(\Omega, \mathcal{F}, \mathbb{P})$ est de retarder d'un échantillon les trajectoires dans le sens où $(BX_t)(\omega) = X_{t-1}(\omega)$ (l'égalité ayant lieu $\mathbb{P}$-presque partout). On note $B^k = B \circ B^{k-1}$ pour $k \geq 2$ les compositions successives de l'opérateur B. Avec cette notation, l'opérateur $\sum_k \psi_k B^k$, où $\{\psi_k\}$ est une séquence réelle, désigne l'opérateur de filtrage linéaire qui, au processus X_t, fait correspondre le processus

$$Y_t = \left(\sum_k \psi_k B^k \right) X_t = \sum_k \psi_k X_{t-k}$$

Pour plus de concision, on utilisera souvent les notations $\psi(B) = \sum_k \psi_k B^k$ et $Y_t = \psi(B) X_t$.
Le premier problème à résoudre est de déterminer les conditions sous lesquelles Y_t est stationnaire si X_t l'est. Il est clair que si $\psi(B) = B^k$ (le filtrage est un simple retard de k échantillon), Y_t est bien stationnaire de même fonction d'autocovariance que X_t. De même, si $\psi(B) = \sum_k \psi_k B^k$, où la suite $\{\psi_k\}$ est différente de 0 pour un *nombre fini d'indices* (filtre à réponse impulsionnelle finie), on a directement par linéarité de l'espérance :

$$\mu_Y = \mathbb{E}[Y_t] = \mu_X \sum_k \psi_k$$

et

$$\gamma_Y(h) = \mathbb{E}[(Y_{t+h} - \mu_Y)(Y_t - \mu_Y)] = \sum_j \sum_k \psi_k \psi_m \gamma_X(h + k - j)$$

où μ_X et $\gamma_X(h)$ sont respectivement la moyenne et la fonction d'autocovariance du processus $\{X_t\}$ (nous avions déjà traité le cas particulier d'un filtre causal d'ordre 1 dans l'exemple 1.14). Les expressions ci-dessus montrent que $\{Y_t\}$ est alors stationnaire au second ordre. La question devient plus délicate lorsque l'on considère des filtres à réponse impulsionnelle infinie puisque Y_t doit alors être défini comme la limite, dans un sens à préciser, d'une suite de variables aléatoires.

Théorème 1.4. *Soit $\{\psi_k\}_{k\in\mathbb{Z}}$ une suite absolument sommable,* i.e. $\sum_{k=-\infty}^{\infty} |\psi_k| < \infty$ *et soit $\{X_t\}$ un processus aléatoire tel que* $\sup_{t\in\mathbb{Z}} \mathbb{E}[|X_t|] < \infty$. *Alors, pour tout $t \in \mathbb{Z}$, la suite :*

$$Y_{n,t} = \sum_{s=-n}^{n} \psi_s X_{t-s}$$

converge presque sûrement, quand n tend vers l'infini, vers une limite Y_t que nous notons

$$Y_t = \sum_{s=-\infty}^{\infty} \psi_s X_{t-s} \ .$$

De plus, la variable aléatoire Y_t est intégrable, i.e. $\mathbb{E}[|Y_t|] < \infty$ *et la suite $\{Y_{n,t}\}_{n\geq 0}$ converge vers Y_t en norme $\mathcal{L}^1$,*

$$\lim_{n\to\infty} \mathbb{E}[|Y_{n,t} - Y_t|] = 0 \ .$$

Supposons que $\sup_{t\in\mathbb{Z}} \mathbb{E}\left[X_t^2\right] < \infty$. *Alors,* $\mathbb{E}\left[Y_t^2\right] < \infty$ *et la suite $\{Y_{n,t}\}_{n\geq 0}$ converge en moyenne quadratique vers la variable aléatoire Y_t, c'est à dire que*

$$\lim_{n\to\infty} \mathbb{E}\left[|Y_{n,t} - Y_t|^2\right] = 0 \ .$$

Démonstration. Voir le paragraphe 1.5 en fin de chapitre. ■

Le résultat suivant établi que le processus obtenu par filtrage linéaire d'un processus stationnaire du second ordre est lui-même stationnaire au second ordre, à condition que la réponse impulsionnelle $\{\psi_k\}$ soit de module sommable.

Théorème 1.5 (Filtrage des processus stationnaires au second ordre)**.** *Soit $\{\psi_k\}$ une suite telle que $\sum_{k=-\infty}^{\infty} |\psi_k| < \infty$ et soit $\{X_t\}$ un processus stationnaire au second ordre de moyenne $\mu_X = \mathbb{E}[X_t]$ et de fonction d'autocovariance $\gamma_X(h) = \mathrm{cov}(X_{t+h}, X_t)$. Alors le processus $Y_t = \sum_{s=-\infty}^{\infty} \psi_s X_{t-s}$ est stationnaire au second ordre de moyenne :*

$$\mu_Y = \mu_X \sum_{k=-\infty}^{\infty} \psi_k \tag{1.16}$$

de fonction d'autocovariance :

$$\gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h+k-j) \tag{1.17}$$

et de mesure spectrale :

$$\nu_Y(d\lambda) = |\psi(e^{-i\lambda})|^2 \nu_X(d\lambda) \tag{1.18}$$

où $\psi(e^{-i\lambda}) = \sum_k \psi_k e^{-ik\lambda}$ est la transformée de Fourier à temps discret de la suite $\{\psi_k\}_{k\in\mathbb{Z}}$.

Démonstration. Voir le paragraphe 1.5 à la fin de ce chapitre. ■

La relation (1.18) qui donne la mesure spectrale du processus filtré en fonction de la fonction de transfert du filtre et de la mesure d'entrée du processus d'entrée est particulièrement simple. Elle montre par exemple que la mise en série de deux filtres $\alpha(B)$, $\beta(B)$ de réponses impulsionnelles absolument sommables conduit à une mesure spectrale $|\alpha(e^{-i\lambda})|^2 |\beta(e^{-i\lambda})|^2 \nu_X(d\lambda)$ pour le processus de sortie (ce qui montre au passage que l'ordre d'application des filtres est indifférent).

Définition 1.12 (Processus linéaire)**.** *Nous dirons que $\{X_t\}$ est un processus linéaire s'il existe un bruit blanc $Z_t \sim \mathrm{BB}(0, \sigma^2)$ et une suite de coefficients $\{\psi_k\}_{k\in\mathbb{Z}}$ absolument sommable telle que :*

$$X_t = \mu + \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k} \tag{1.19}$$

où μ désigne une valeur arbitraire.

Il résulte directement de la discussion ci-dessus qu'un processus linéaire est stationnaire au second ordre, que sa moyenne est égale à μ, que sa fonction d'autocovariance est donnée par :

$$\gamma_X(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h}$$

et que sa mesure spectrale admet une densité dont l'expression est :

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} |\psi(e^{-i\lambda})|^2 \tag{1.20}$$

où $\psi(e^{-i\lambda}) = \sum_k \psi_k e^{-ik\lambda}$.

1.4 Processus ARMA

Dans ce paragraphe nous nous intéressons à une classe importante de processus du second ordre, les processus autorégressifs à moyenne ajustée ou processus ARMA. Il s'agit de restreindre la classe des processus linéaires en ne considérant que les filtres dont la fonction de transfert est rationnelle.

1.4.1 Processus MA(q)

Définition 1.13 (Processus MA(q))**.** *On dit que le processus $\{X_t\}$ est à moyenne ajustée d'ordre q (ou MA(q)) si $\{X_t\}$ est donné par :*

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{1.21}$$

où $Z_t \sim \mathrm{BB}(0, \sigma^2)$.

La terminologie "moyenne ajustée" est la traduction, assez malheureuse, du nom anglo-saxon "moving average" (moyenne mobile) En utilisant les résultats du théorème 1.5, on obtient $\mathbb{E}[X_t] = 0$, et

$$\gamma_X(h) = \begin{cases} \sigma^2 \sum_{t=0}^{t-|h|} \theta_k \theta_{k+|h|} & \text{si} \qquad 0 \leq |h| \leq q \\ 0 & \text{sinon} \end{cases} \tag{1.22}$$

Enfin, d'après la formule (1.20), le processus admet une densité spectrale dont l'expression est :

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \left| 1 + \sum_{k=1}^{q} \theta_k e^{-ik\lambda} \right|^2$$

Un exemple de densité spectrale pour le processus MA(1) est représenté figure 1.10. De manière générale, la densité spectrale d'un processus $MA(q)$ possède des anti-résonnances au voisinage des pulsations correspondant aux arguments des racines du polynôme $\theta(z) = \sum_{k=1}^{q} \theta_k z^k$. On démontrera, à titre d'exercice, la propriété suivante qui indique que toute suite de coefficients covariance $\{\gamma(h)\}$ non nulle sauf pour un nombre fini d'indices temporels (*i.e.* le cardinal de l'ensemble $\{h \in \mathbb{Z}, \gamma(h) \neq 0\}$) peut être considérée comme la suite des coefficients d'autocovariance d'un modèle linéaire à moyenne mobile.

Propriété 1.6. *Soit $\gamma(h)$ une fonction d'autocovariance telle que $\gamma(h) = 0$ pour $|h| > q$. Alors, il existe un bruit blanc $\{Z_t\}$ et un polynôme $\theta(z)$ de degré inférieur ou égal à q tels que $\gamma(h)$ soit la fonction d'autocovariance du processus $MA(q)$ défini par $X_t = Z_t + \sum_{k=1}^{q} \theta_k Z_{t-k}$.*

1.4.2 Processus AR(p)

Définition 1.14 (Processus AR(p))**.** *On dit que le processus $\{X_t\}$ est un processus autorégressif d'ordre p (ou AR(p)) si $\{X_t\}$ est un processus stationnaire au second-ordre et s'il est solution de l'équation de récurrence :*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t \tag{1.23}$$

où $Z_t \sim \mathrm{BB}(0, \sigma^2)$ *est un bruit blanc.*

Le terme “autorégressif” provient de la forme de l'équation (1.23) dans laquelle la valeur courante du processus s'exprime sous la forme d'une régression (terme synonyme de combinaison linéaire) des p valeurs précédentes du processus plus un bruit additif.

L'existence et l'unicité d'une solution stationnaire au second ordre de l'équation (1.23) constituent des questions délicates (qui ne se posaient pas lorsque nous avions défini les modèles MA). Nous détaillons ci-dessous la réponse à cette question dans le cas le cas $p = 1$.

Cas : $|\phi_1| < 1$

L'équation de récurrence s'écrit :

$$X_t = \phi_1 X_{t-1} + Z_t \tag{1.24}$$

Puisque $|\phi_1| < 1$, la fraction rationnelle $\psi(z) = (1 - \phi_1 z)^{-1}$ a un développement en série entière de la forme :

$$\psi(z) = \frac{1}{1 - \phi_1 z} = \sum_{k=0}^{+\infty} \phi_1^k z^k$$

qui converge sur le disque $\{z \in \mathbb{C} \ : |z| < |\phi_1|^{-1}\}$. Considérons alors le filtre linéaire de réponse impulsionnelle $\psi_k = \phi_1^k$ pour $k \geq 0$ et $\psi_k = 0$ sinon. Comme ψ_k est absolument sommable, le processus

$$Y_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k} = \sum_{k=0}^{\infty} \phi_1^k Z_{t-k}$$

est bien défini et est stationnaire au second ordre. Par construction Y_t est solution de (1.24) ce que l'on peut également vérifier directement en notant que :

$$X_t = Z_t + \phi_1 \sum_{k=0}^{+\infty} \phi_1^k Z_{t-1-k} = Z_t + \phi_1 X_{t-1}$$

L'unicité de la solution est garantie par l'hypothèse de stationnarité au second ordre. Supposons en effet que $\{X_t\}$ et $\{Y_t\}$ soient deux processus stationnaires au second-ordre et que ces deux processus soient solutions de l'équation de récurrence (1.24). On a alors par différence $(X_t - Y_t) = \phi_1(X_{t-1} - Y_{t-1})$, relation qui itérée k fois implique

$$(X_t - Y_t) = \phi_1^k (X_{t-k} - Y_{t-k}) \ .$$

Par suite,

$$\mathbb{E}\left[|X_t - Y_t|\right] = \phi_1^k \mathbb{E}\left[|X_{t-k} - Y_{t-k}|\right] \leq \phi_1^k (\mathbb{E}\left[|X_{t-k}|\right] + \mathbb{E}\left[|Y_{t-k}|\right]) \leq \phi_1^k (\mathbb{E}\left[X_0^2\right]^{\frac{1}{2}} + \mathbb{E}\left[Y_0^2\right]^{\frac{1}{2}})$$

où k peut être pris quelconque. Comme ϕ_1 est en module plus petit que 1, on en déduit que $\mathbb{E}\left[|X_t - Y_t|\right] = 0$ et donc que $X_t = Y_t$ presque sûrement. La fonction d'autocovariance de X_t solution stationnaire de (1.24) est donnée par la formule (1.17) qui s'écrit ;

$$\gamma_X(h) = \sigma^2 \sum_{k=0}^{\infty} \phi_1^k \phi_1^{k+|h|} = \sigma^2 \frac{\phi_1^{|h|}}{1 - \phi_1^2} \tag{1.25}$$

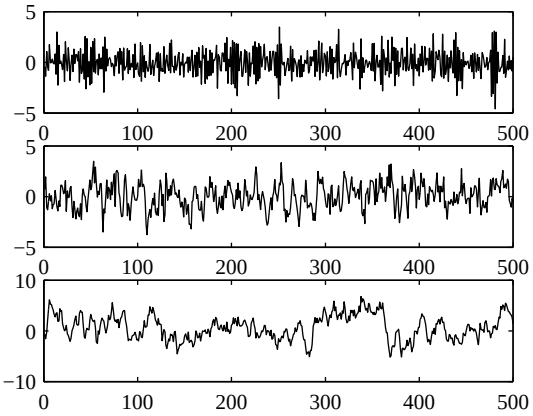

FIG. 1.11 – *Trajectoires de longueur* 500 *d'un processus* $AR(1)$*) gaussien. Courbe du haut :* $\phi_1 = -0.7$*. Courbe du milieu :* $\phi_1 = 0.5$*. Courbe du bas :* $\phi_1 = 0.9$

Lorsque $\phi_1 > 0$, le processus X_t est positivement corrélé, dans le sens où tous ses coefficients d'autocovariance sont positifs. Les exemples de trajectoires représentées sur la figure 1.11 montrent que des valeurs de ϕ_1 proches de 1 correspondent à des trajectoires "persistantes" (dont, par exemple, les temps successifs de passage par zéro sont relativement espacés). Inversement, des valeurs de ϕ_1 négatives conduisent à des trajectoires où une valeur positive a tendance à être suivie par une valeur négative. La densité spectrale de X_t est donnée par

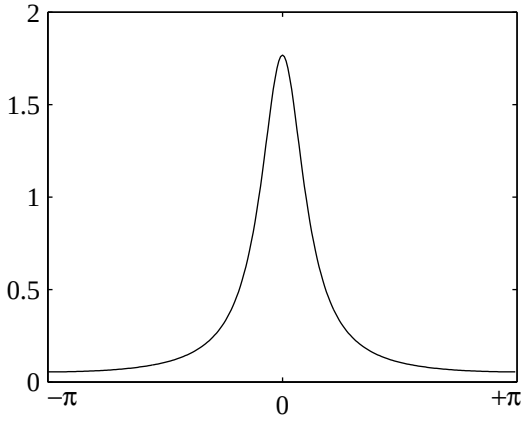

FIG. 1.12 – *Densité spectrale d'un processus AR(1), défini par (1.24) pour* $\sigma = 1$ *et* $\phi_1 = 0.7$*.*

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \left| \sum_{k=0}^{\infty} \phi_1^k e^{-ik\lambda} \right|^2 = \frac{\sigma^2}{2\pi} \frac{1}{|1 - \phi_1 e^{-i\lambda}|^2} \tag{1.26}$$

La figure 1.12 donne la forme de cette densité spectrale pour $\phi_1 = 0.7$.

Cas $|\phi_1| > 1$

Nous allons montrer que le processus retourné temporel vérifie une équation récurrente qui nous ramène au cas précédent. Pour cela posons $X_t^r = X_{-t}$. En portant X_t^r dans l'équation (1.24), on obtient

$$X_t^r = X_{-t} = \phi_1 X_{-t-1} + Z_{-t} = \phi_1 X_{t+1}^r + Z_{-t}$$

qui peut encore s'écrire :

$$X_t^r = \phi_1^{-1} X_{t-1}^r + W_t \tag{1.27}$$

où $W_t = -\phi_1^{-1} Z_{-t-1}$ est un bruit blanc de variance $\sigma_W^2 = \sigma^2/\phi_1^2$. L'équation (1.27) est maintenant du type que (1.23) puisque $|\phi_1^{-1}| < 1$. Par conséquent il existe un unique processus stationnaire solution de l'équation 1.27 donné par

$$X_t^r = \sum_{k=0}^{\infty} \phi_1^{-k} W_{t-k} \tag{1.28}$$

Comme $\{X_t^r\}$ est stationnaire au second ordre, le processus

$$X_t = X_{-t}^r = \sum_{k=0}^{\infty} \phi_1^{-k} W_{-t+k} = -\sum_{k=1}^{\infty} \phi_1^{-k} Z_{t+k} \tag{1.29}$$

l'est également (cf. exemple 1.7) avec la même moyenne et la même fonction d'autocovariance. Les expressions de la fonction d'autocovariance et de la densité spectrale du processus sont donc données respectivement par (1.25) et (1.26) à condition de substituer ϕ_1 par $1/\phi_1$. Un point remarquable à propos de l'expression de la solution stationnaire donnée par (1.29) est que celle ci est entièrement anti-causale, dans le sens où elle ne dépend que des valeurs futures du bruit Z_t. Cette remarque montre qu'il ne faut pas se laisser tromper par l'apparence de la relation de récurrence (1.27) : la solution stationnaire ne s'exprime par forcément comme un filtrage causal du bruit Z_t, point que nous développerons au paragraphe 1.4.2.

Cas $|\phi_1| = 1$

Nous avons déjà montré à propos de l'exemple 1.10 que lorsque $\phi_1 = 1$, un processus X_t vérifiant $X_t = X_{t-1} + Z_t$ ne peut avoir une variance constante au cours du temps (on a montré que $\mathbb{E}\left[X_t^2 | X_0\right] = t\sigma^2$, où σ^2 est la variance de Z_t, et donc $\mathbb{E}\left[X_t^2\right] = t\sigma^2$). A fortiori, un tel processus ne peut être stationnaire au second ordre. En utilisant la même technique, on montre aisément que l'équation de récurrence (1.24) ne peut avoir de solution stationnaire lorsque $|\phi_1| = 1$. Une remarque intéressante est que dans le cas où $\phi_1 = 1$, le processus $Z_t = X_t - X_{t-1}$ est par hypothèse stationnaire. On peut donc utiliser le modèle $X_t - X_{t-1} = Z_t$ pour un signal X_t non-stationnaire dont les incréments sont supposés stationnaires. C'est implicitement la stratégie que nous avons adoptée pour analyser la série de l'indice S&P500 représentée figure 1.4 au paragraphe 1.2.2 (en utilisant en plus une transformation logarithmique des données).

Cas général

Le théorème suivant étend les résultats précédents à un processus AR(p).

Théorème 1.6 (Existence des processus AR(p))**.** *L'équation récurrente :*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t \tag{1.30}$$

où $Z_t \sim \mathrm{BB}(0, \sigma^2)$ admet une solution stationnaire au second ordre si et seulement si le polynôme :

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \text{ pour } |z| = 1$$

et cette solution est unique. Elle a pour expression :

$$X_t = \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k} \tag{1.31}$$

où ψ_k est la suite des coefficients du développement en série de Laurent de $1/\phi(z)$ au voisinage du cercle unité.

Démonstration. La condition $\phi(z) \neq 0$ pour $|z| = 1$ implique que $\phi(z) \neq 0$ dans une couronne $1 - \delta \leq |z| \leq 1 + \delta$ et donc que la fonction $\psi(z) = 1/\phi(z)$ est analytique dans cette couronne. Il s'en suit que $1/\phi(z)$ admet, pour $1 - \delta \leq |z| \leq 1 + \delta$, un *développement en série de Laurent* qui s'écrit :

$$\frac{1}{\phi(z)} = \sum_{k=-\infty}^{\infty} \psi_k z^k = \psi(z) \tag{1.32}$$

où la suite $\{\psi_k\}$ est de module sommable et vérifie $\psi_0 = 1$. Nous pouvons alors considérer le filtre de réponse impulsionnelle $\{\psi_k\}$. D'après le théorème 1.5, nous pouvons appliquer ce filtre aux deux membres de l'équation récurrente $\phi(B)X_t = Z_t$. Nous obtenons $(\psi(B)\phi(B))X_t = X_t = \psi(B)Z_t$. On en déduit que l'unique solution stationnaire de l'équation (1.30) est donnée par (1.31). ■

AR(p) causal

On peut distinguer trois cas suivant la position des racines de $\phi(z)$ par rapport au cercle unité :

– Les racines du polynôme $\phi(z)$ sont strictement à l'extérieur du cercle unité. Alors la fonction $\psi(z) = 1/\phi(z)$ est analytique sur le disque $\{z : |z| < \rho_m\}$, où $\rho_m > 1$ est le module de la racine de $\phi(z)$ de module le plus petit. En particulier $\psi(z)$ est analytique en 0 et donc $\psi_k = 0$ pour $k < 0$. Il s'en suit que :

$$X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k}$$

On note que X_t s'exprime causalement en fonction de Z_t dans le sens où X_t dépend uniquement des valeurs présente et passées de Z_t. On dit dans ce cas que le modèle autorégressif est causal.

– Les racines du polynôme $\phi(z)$ sont strictement à l'intérieur du cercle unité. Alors la fonction $1/\phi(z)$ est analytique dans la couronne $\{z : |z| > \rho_M\}$, où $\rho_M < 1$ est le module de la racine de $\phi(z)$ de module le plus grand. On en déduit que $\psi_k = 0$ pour $k \geq 0$ et donc que X_t s'exprime anti-causalement en fonction de Z_t, dans le sens où X_t dépend uniquement des valeurs futures de Z_t. On dit dans ce cas que le modèle autorégressif est anti-causal.

– Le polynôme $\phi(z)$ a des racines de part et d'autre du cercle unité. La suite ψ_k est alors bilatérale. Dans ce cas X_t dépend à la fois des valeurs passées, présente et futures de Z_t. On dit dans ce cas que le modèle autorégressif est bilatérale.

Théorème 1.7 (AR(p) causal)**.** *L'équation récurrente :*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$$

où $Z_t \sim \mathrm{BB}(0, \sigma^2)$ admet une solution stationnaire au second ordre causale si et seulement si $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ pour $|z| \leq 1$. Cette solution est unique et a pour expression :

$$X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k} \tag{1.33}$$

où ψ_k est la suite des coefficients du développement en série de Laurent de $1/\phi(z)$ dans le disque $\{z : |z| \leq 1\}$.

Démonstration. Il nous reste à montrer que, si l'équation récurrente possède une solution stationnaire au second ordre *causale* c'est-à-dire telle que $X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k}$ avec ψ_k de module sommable, alors $\phi(z) \neq 0$ pour $|z| \leq 1$. En effet partons de $\phi(B)X_t = Z_t$ et remplaçons X_t par $\psi(B)Z_t$, où nous supposons que $\psi(z) = \sum_{k=0}^{\infty} \psi_k z^k$ est analytique pour $|z| \leq 1$. Alors on a $(\phi(B)\psi(B))Z_t = Z_t$ et donc

$$\phi(z)\psi(z) = 1 \text{ pour } |z| \leq 1$$

qui implique que $\phi(z) \neq 0$ pour $|z| \leq 1$. ■

Sauf indication contraire nous ne considérons, dans la suite, que des processus *autorégressifs causaux*. La propriété de causalité joue en effet un rôle essentiel pour l'estimation des paramètres (cf. les équations de Yule-Walker ci-dessous) ainsi que dans les problèmes de prédiction étudiés au chapitre 4. Par ailleurs, cette restriction n'en est pas vraiment une comme le montre l'exercice suivant :

Exercice 1.1
Soit $\mathcal{M}(p)$ un modèle AR(p) de paramètres $\sigma^2, \phi_1, \ldots \phi_p$ qui admet une solution stationnaire ($\phi(z) \neq 0$ pour $|z| = 1$). Montrer qu'il existe toujours un modèle $\mathcal{M}'(p)$ AR(p) stable et *causal* possédant la même fonction d'autocovariance que $\mathcal{M}(p)$ (indication : utiliser des facteurs passe-tout de la forme $(a_1 - z)/(1 - a_1^* z)$ où $\phi(a_1) = 0$).

Equations de Yule-Walker

Les équations de Yule-Walker fournissent une relation linéaire entre les paramètres $\phi_1, \ldots, \phi_p$ et σ^2 de l'équation (1.23), définissant un processus AR(p), et la fonction d'autocovariance de ce processus. Nous nous plaçons dans le cas où le processus AR(p) est *causal* et donc, pour $k > 0$ $\mathbb{E}[Z_t X_{t-k}] = 0$ d'après (1.33). On en déduit que :

$$\mathbb{E}[Z_t X_t] = \mathbb{E}[Z_t Z_t] + \sum_{j=1}^{p} \phi_j \mathbb{E}[Z_t X_{t-j}] = \sigma^2$$

et par suite en remplaçant, dans $\mathbb{E}[Z_t X_t]$, Z_t par $X_t - \sum_{j=1}^{p} \phi_j X_{t-j}$ il vient :

$$\sigma^2 = \mathbb{E}[Z_t X_t] = \mathbb{E}\left[(X_t - \sum_{j=1}^{p} \phi_j X_{t-j}) X_t\right] = \gamma(0) - \sum_{k=1}^{p} \phi_k \gamma(k) \tag{1.34}$$

En multipliant, pour $k > 0$, les deux membres de l'équation (1.23) par X_{t-k} et en en prenant l'espérance, on obtient $0 = \mathbb{E}[Z_t X_{t-k}] = \mathbb{E}\left[(X_t - \sum_{j=1}^{p} \phi_j X_{t-j}) X_{t-k}\right]$. On en déduit que la fonction d'autocovariance vérifie, pour tout $k > 0$, l'équation de récurrence :

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = 0 \tag{1.35}$$

En regroupant, sous forme matricielle, les p équations (1.35) pour $1 \leq k \leq p$, on obtient :

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & & \ddots & \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{bmatrix} \tag{1.36}$$

Les équations (1.34) et (1.36) sont appelées *équations de Yule-Walker*. Nous retrouverons ces équations, dans le cadre de la prédiction linéaire au chapitre 4 (équations (4.8) et (4.9)). Ces équations permettent également de déterminer les valeurs des paramètres du modèle à partir d'estimation de la fonction d'autocovariance (cf. chapitre 5).

Calcul des covariances d'un processus AR(p) causal

Partant des paramètres du modèle, il est également possible de calculer la fonction d'autocovariance du processus à partir des équations (1.34) et (1.36) en les réécrivant sous la forme

$$\left(\begin{bmatrix} 1 & -\phi_1 & \cdots & -\phi_p \\ -\phi_1 & & -\phi_p & 0 \\ \vdots & \iddots & \iddots & \vdots \\ -\phi_p & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\phi_1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ -\phi_p & \cdots & -\phi_1 & 1 \end{bmatrix} \right) \begin{bmatrix} \gamma(0)/2 \\ \gamma(1) \\ \vdots \\ \gamma(p) \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{1.37}$$

Partant alors de $\phi_1, \ldots, \phi_p, \sigma^2$, on calcule $\gamma(0), \ldots, \gamma(p)$ puis, en utilisant (1.35), on calcule $\gamma(k)$ pour tout $k > p$. Une autre façon de procéder consiste à calculer récursivement la suite ψ_k en remarquant que $1 = \psi(z)\phi(z) = (\psi_0 + \psi_1 z + \ldots)(1 - \phi_1 z - \cdots - \phi_p z^p)$ et donc, par identification, que :

$$\psi_0 = 1, \quad \psi_1 = \phi_1 \psi_0, \quad \psi_2 = \phi_2 \psi_1 + \phi_1 \psi_1, \quad \text{etc.}$$

puis d'appliquer la formule (1.17) pour un processus d'entrée de fonction d'autocovariance $\sigma^2 \delta(h)$ qui s'écrit

$$\gamma(h) = \sigma^2 \sum_{k=0}^{\infty} \psi_k \psi_{k+|h|}$$

Densité spectrale

Réécrivons l'équation (1.23) sous la forme $X_t - \sum_{k=1}^{p} \phi_k X_{t-k} = Z_t$. Le premier membre est un processus stationnaire au second ordre puisque il représente le filtrage, par un filtre de réponse impulsionnelle finie, du processus X_t. Ce processus possède donc une densité spectrale qui a pour expression $\left|1 - \sum_{k=1}^{p} \phi_k e^{-ik\lambda}\right|^2 f_X(\lambda)$ où $f_X(\lambda)$ désigne la densité spectrale de X_t. Cette densité spectrale est aussi égale à celle du second membre Z_t, c'est à dire à $\sigma^2/2\pi$. Par conséquent,

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{1}{\left|1 - \sum_{k=1}^{p} \phi_k e^{-ik\lambda}\right|^2} \tag{1.38}$$

1.4.3 Processus ARMA

La notion de processus ARMA généralise les notions de processus MA et AR.

Théorème 1.8 (Existence des processus ARMA(p, q))**.** *Soit l'équation récurrente :*

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{1.39}$$

où $Z_t \sim \mathrm{BB}(0, \sigma^2)$. On pose $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ et $\theta(z) = 1 + \theta_1 z + \cdots + \theta_p z^p$. On suppose que $\phi(z)$ et $\theta(z)$ n'ont pas de zéros communs. Alors l'équation (1.39) admet une solution stationnaire au second ordre si et seulement si le polynôme $\phi(z) \neq 0$ pour $|z| = 1$. Cette solution est unique et a pour expression :

$$X_t = \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k} \tag{1.40}$$

où ψ_k est la suite des coefficients du développement en série de Laurent de $\theta(z)/\phi(z)$ au voisinage du cercle unité.

Démonstration. Comme $\phi(z) \neq 0$ pour $|z| = 1$, $1/\phi(z)$ est développable en série de Laurent au voisinage du cercle unité, suivant :

$$\xi(z) = \frac{1}{\phi(z)} = \sum_{k=-\infty}^{\infty} \xi_k z^k$$

où la suite $\{\xi_k\}$ est de module sommable et vérifie $\xi_0 = 1$. D'après le théorème 1.5, nous pouvons donc appliquer le filtre de réponse impulsionnelle $\{\xi_k\}$ aux deux membres de l'équation récurrente $\phi(B)X_t = \theta(B)Z_t$. Nous obtenons $(\xi(B)\phi(B))X_t = X_t = \psi(B)Z_t$ où $\psi(B) = \xi(B)\theta(B)$. On en déduit que $\psi(z) = \sum_k \psi_k z^k$ avec :

$$\psi_k = \xi_k + \sum_{j=1}^{q} \theta_j \xi_{k-j}$$

où $\{\psi_k\}$ est absolument sommable. ■

Dans le cas où $\phi(z)$ et $\theta(z)$ ont des zéros communs, deux configurations sont possibles :
– Les zéros communs ne sont pas sur le cercle unité. Dans ce cas on se ramène au cas sans zéro commun en annulant les facteurs communs.

– Certains des zéros communs se trouvent sur le cercle unité. L'équation (1.39) admet une infinité de solutions stationnaires au second ordre.

Du point de vue de la modélisation, la présence de zéros communs ne présente aucun intérêt puisqu'elle est sans influence sur la densité spectrale de puissance. Elle conduit de plus à une ambiguïté sur l'ordre réel des parties AR et MA.

ARMA(p,q) causal

Comme dans le cas d'un processus AR(p), on peut distinguer trois cas, suivant que les zéros de $\phi(z)$ sont à l'extérieur, à l'intérieur ou de part et d'autre du cercle unité. Dans le cas où les zéros de $\phi(z)$ sont à l'extérieur du cercle unité, la suite ξ_k est causale ($\xi_k = 0$ pour $k < 0$) et donc $\psi_k = \xi_k + \sum_{j=1}^{q} \theta_j \xi_{k-j}$ est aussi causale. Par conséquent le processus X_t s'exprime causalement en fonction de Z_t.

Théorème 1.9 (ARMA(p,q) causal)**.**

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{1.41}$$

où $Z_t \sim \mathrm{BB}(0,\sigma^2)$. On pose $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ et $\theta(z) = 1 + \theta_1 z + \cdots + \theta_p z^p$. On suppose que $\phi(z)$ et $\theta(z)$ n'ont pas de zéros communs. Alors l'équation (1.41) admet une solution stationnaire causale au second ordre si et seulement si le polynôme $\phi(z) \neq 0$ pour $|z| \leq 1$. Cette solution est unique et a pour expression :

$$X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k} \tag{1.42}$$

où ψ_k est la suite des coefficients du développement en série de Laurent de $\theta(z)/\phi(z)$ dans le disque $\{z : |z| \leq 1\}$.

Démonstration. Il suffit de remarquer que la condition sur $\phi(z)$ implique que $1/\phi(z)$ possède un développement causal au voisinage du cercle unité. $\xi(B)$ correspond donc à une opération de filtrage causal (voir preuve du théorème 1.8 pour les notations), ce qui implique qu'il en va de même pour $\xi(B)\phi(B)$. ■

Calcul des covariances d'un processus ARMA(p,q) causal

Une première méthode consiste à utiliser l'expression (1.17) qui s'écrit, compte tenu du fait que $\{Z_t\}$ est un bruit blanc,

$$\gamma(h) = \sigma^2 \sum_{k=0}^{\infty} \psi_k \psi_{k+|h|}$$

où la suite $\{\psi_k\}$ se détermine de façon récurrente à partir de l'égalité $\psi(z)\theta(z) = \phi(z)$ par identification du terme en z^k. Pour les premiers termes on trouve :

$$\begin{aligned}
&\psi_0 = 1\\
&\psi_1 = \theta_1 + \psi_0\phi_1\\
&\psi_2 = \theta_2 + \psi_0\phi_2 + \psi_1\phi_1\\
&\cdots
\end{aligned}$$

La seconde méthode utilise une formule de récurrence, vérifiée par la fonction d'autocovariance d'un processus ARMA(p,q), qui s'obtient en multipliant les deux membres de (1.39) par X_{t-k} et en en prenant l'espérance. On obtient :

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = \sigma^2 \sum_{k\le j\le q} \theta_j\psi_{j-k} \qquad \text{pour } 0 \le k < \max(p, q+1) \tag{1.43}$$

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0 \qquad \text{pour } k \ge \max(p, q+1) \tag{1.44}$$

où nous avons utilisé la causalité du processus pour écrire que $\mathbb{E}[Z_t X_{t-k}] = 0$ pour tout $k \ge 1$. Le calcul de la suite $\{\psi_k\}$ pour $k = 1, \ldots, p$ se fait comme précédemment. En reportant ces valeurs dans (1.43) pour $0 \le k \le p$, on obtient $(p+1)$ équations linéaires aux $(p+1)$ inconnues $(\gamma(0), \ldots, \gamma(p))$ que l'on peut résoudre. Pour déterminer les valeurs suivantes on utilise l'expression (1.44).

Inversibilité d'un processus ARMA(p,q)

Théorème 1.10 (ARMA(p,q) inversible)**.** *Soit X_t un processus ARMA(p,q). On suppose que $\phi(z)$ et $\theta(z)$ n'ont pas de zéros communs. Alors il existe une suite $\{\pi_k\}$ causale absolument sommable telle que :*

$$Z_t = \sum_{k=0}^{\infty} \pi_k X_{t-k} \tag{1.45}$$

si et seulement si $\theta(z) \neq 0$ pour $z \le 1$. On dit alors que le modèle ARMA(p,q) est inversible. La suite π_k est la suite des coefficients du développement en série de $\phi(z)/\theta(z)$ dans le disque $\{z : |z| \le 1\}$.

La preuve de ce théorème est tout à fait analogue à celle du théorème 1.9. Remarquons que la notion d'inversibilité, comme celle de causalité, est bien relative au modèle ARMA(p,q) lui-même et pas uniquement au processus X_t comme le montre l'exercice suivant.

Exercice 1.2
Soit X_t un processus stationnaire au second ordre solution de l'équation de récurrence (1.41) où le modèle ARMA(p,q) correspondant est supposé sans zéro commun mais pas nécessairement inversible. Montrer qu'il existe un bruit blanc $\tilde{Z}_t$ tel que X_t soit solution de

$$\phi(B)X_t = \tilde{\theta}(B)\tilde{Z}_t$$

où le modèle ARMA(p,q) défini par $\phi_1, \ldots \phi_p$ et $\tilde{\theta}_1, \ldots \tilde{\theta}_q$ est inversible (indication : considérer des facteurs passe-tout).

Un modèle ARMA(p,q) est causal *et* inversible lorsque les racines des polynômes $\phi(z)$ et $\theta(z)$ sont toutes situées à l'extérieur du filtre unité. Dans ce cas, X_t et Z_t se déduisent mutuellement l'un de l'autre par des opérations de filtrage causal, la réponse impulsionnelle de chacun de ces filtres étant *à phase minimale* (c'est à dire inversible causalement).

Densité spectrale d'un processus ARMA(p,q)

Théorème 1.11 (Densité spectrale d'un processus ARMA(p,q))**.** *Soit X_t un processus ARMA(p,q) (pas nécessairement causal ou inversible) défini par $\phi(B)X_t = \theta(B)Z_t$ où $Z_t \sim BB(0,\sigma^2)$ et où $\theta(z)$ et*

$\phi(z)$ sont des polynômes de degré q et p n'ayant pas de zéros communs. Alors X_t possède une densité spectrale qui a pour expression :

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{\left|1 + \sum_{k=1}^{q} \theta_k e^{-ik\lambda}\right|^2}{\left|1 - \sum_{k=1}^{p} \phi_k e^{-ik\lambda}\right|^2} \tag{1.46}$$

1.5 Preuves des théorèmes 1.4 et 1.5

Théorème 1.4. *Soit $\{\psi_k\}_{k\in\mathbb{Z}}$ une suite telle que $\sum_{k=-\infty}^{\infty}|\psi_k|<\infty$ et soit $\{X_t\}$ un processus aléatoire tel que $\sup_{t\in\mathbb{Z}}\mathbb{E}\left[|X_t|\right]<\infty$. Alors, pour tout $t\in\mathbb{Z}$, la suite :*

$$Y_{n,t}=\sum_{s=-n}^{n}\psi_s X_{t-s}$$

converge presque sûrement, quand n tend vers l'infini, vers une limite que nous notons $Y_t=\sum_{s=-\infty}^{\infty}\psi_s X_{t-s}$ et $\lim_{n\to\infty}\mathbb{E}\left[|Y_{n,t}-Y_t|\right]=0$. Si de plus $\sup_t\mathbb{E}\left[X^2(t)\right]<\infty$, alors $\mathbb{E}\left[Y_t^2\right]<\infty$ et $Y_{n,t}$ converge en moyenne quadratique vers Y_t, c'est à dire que $\lim_{n\to\infty}\mathbb{E}\left[|Y_{n,t}-Y_t|^2\right]=0$.

Démonstration. Notons pour tout $t\in\mathbb{Z}$ et $n\in\mathbb{N}$, $|Y|_{n,t}=\sum_{s=-n}^{+n}|\psi_s||X_{t-s}$. La suite $\{|Y|_{n,t}\}_{n\geq 0}$ est une suite de variables aléatoires intégrables. Le théorème de convergence dominé (see Proposition **??**) montre que

$$\lim_{n\to\infty}\mathbb{E}\left[|Y|_{n,t}\right]=\mathbb{E}\left[|Y|_t\right]$$

où $|Y|_t=\sum_{s=-\infty}^{\infty}|\psi_s||X_{t-s}|$. Comme,

$$\mathbb{E}\left[|Y|_{n,t}\right]=\sum_{s=-n}^{+n}|\psi_s|\mathbb{E}\left[|X_{t-s}|\right]\leq\sup_{t\in\mathbb{Z}}\mathbb{E}\left[|X_t|\right]\sum_{s=-\infty}^{\infty}|\psi_s|\,,$$

on a donc

$$\mathbb{E}\left[\sum_{s=-\infty}^{\infty}|\psi_s||X_{t-s}|\right]<\infty\,.$$

Par conséquent, il existe un ensemble $A\in\mathcal{F}$, vérifiant $\mathbb{P}A=1$ tel que, pour tout $\omega\in A$, nous ayons

$$\sum_{s=-\infty}^{\infty}|\psi_s||X_{t-s}(\omega)|<\infty$$

Pour $\omega\in A$, la série de terme générique $s\mapsto\psi_s X_{t-s}(\omega)$ est normalement sommable, ce qui implique que, pour tout $\omega\in A$, la suite $n\mapsto Y_{n,t}(\omega)$ converge.

Notons, pour tout $\omega\in\Omega$, $Y_t(\omega)=\limsup Y_{n,t}(\omega)$. $\omega\mapsto Y_t(\omega)$ est une variable aléatoire comme limite supérieure de variables aléatoires et pour tout $\omega\in A$, nous avons $\lim_{n\to\infty}Y_{n,t}(\omega)=Y_t(\omega)$ et donc la suite $n\mapsto Y_{n,t}$ converge $\mathbb{P}$-p.s vers Y_t.

Remarquons également que la suite $n\mapsto Y_{n,t}$ est une suite de Cauchy dans $\mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$. En effet, pour tout $p\geq q$, nous avons :

$$\mathbb{E}\left[|Y_{p,t}-Y_{q,t}|\right]\leq\sup_{t\in\mathbb{Z}}\mathbb{E}\left[|X_t|\right]\sum_{s=q+1}^{p}|\psi_s|\underset{q,p\to\infty}{\longrightarrow}0$$

Fixons $\epsilon>0$ et choisissons n tel que

$$\sup_{p,q\geq n}\mathbb{E}\left[|Y_{p,t}-Y_{q,t}|\right]\leq\epsilon$$

Par application du lemme de Fatou nous avons alors, pour tout $q \geq n$,

$$\mathbb{E}\left[\liminf_{p\to\infty} |Y_{p,t} - Y_{q,t}|\right] = \mathbb{E}\left[|Y_t - Y_{q,t}|\right] \leq \liminf_{p\to\infty} \mathbb{E}\left[|Y_{p,t} - Y_{q,t}|\right] \leq \epsilon$$

et donc $\limsup_{q\to\infty} \mathbb{E}\left[|Y_{q,t} - Y_t|\right] \leq \epsilon$. Comme ϵ est arbitraire, nous avons donc $\lim_{q\to\infty} \mathbb{E}\left[|Y_{q,t} - Y_t|\right] = 0$. L'inégalité triangulaire

$$\mathbb{E}\left[|Y_t|\right] \leq \mathbb{E}\left[|Y_t - Y_{n,t}|\right] + \mathbb{E}\left[|Y_{n,t}|\right]$$

montre enfin que $Y_t \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Considérons maintenant le cas où $\sup_{t\in\mathbb{Z}} \mathbb{E}\left[X_t^2\right] < \infty$. Remarquons tout d'abord que $\mathbb{E}\left[|X_t|\right] \leq (\mathbb{E}\left[X_t^2\right])^{1/2}$ et donc que cette condition implique que $\sup_{t\in\mathbb{Z}} \mathbb{E}\left[|X_t|\right] < \infty$. La suite $m \mapsto Y_{m,t}$ est une suite de Cauchy dans $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. En effet, pour $p \geq q$, nous avons

$$\begin{aligned}\mathbb{E}\left[(Y_{p,t} - Y_{q,t})^2\right] &= \mathbb{E}\left[\sum_{s=q+1}^{p} \psi_s X_{t-s}\right]^2 = \sum_{j,k=q+1}^{p} \psi_j \psi_k \mathbb{E}\left[X_{t-j} X_{t-k}\right] \\ &\leq \sum_{j,k=q+1}^{p} |\psi_j||\psi_k| \sup_{t\in\mathbb{Z}} \mathbb{E}\left[X_t^2\right] = \sup_{t\in\mathbb{Z}} \mathbb{E}\left[X_t^2\right] \left(\sum_{j=q+1}^{p} |\psi_j|\right)^2\end{aligned}$$

Comme précédemment fixons $\epsilon > 0$ et choisissons n tel que :

$$\sup_{p,q\geq n} \mathbb{E}\left[|Y_{p,t} - Y_{q,t}|^2\right] \leq \epsilon.$$

Par application du lemme de Fatou, nous avons :

$$\mathbb{E}\left[\liminf_{p\to\infty} (Y_{p,t} - Y_{q,t})^2\right] = \mathbb{E}\left[(Y_t - Y_{q,t})^2\right] \leq \liminf_{p\to\infty} \mathbb{E}\left[(Y_{p,t} - Y_{q,t})^2\right] \leq \epsilon$$

et donc : $\limsup_{q\to\infty} \mathbb{E}\left[(Y_t - Y_{q,t})^2\right] \leq \epsilon$. Comme ϵ est arbitraire, $\limsup_{q\to\infty} \mathbb{E}\left[(Y_t - Y_{q,t})^2\right] = 0$, en d'autres termes, la suite $\{Y_{q,t}\}_{q\geq 0}$ converge en moyenne quadratique vers Y_t. Finalement, nous avons :

$$\mathbb{E}\left[Y_t^2\right] \leq 2(\mathbb{E}\left[(Y_t - Y_{q,t})^2\right] + \mathbb{E}\left[Y_{q,t}^2\right]) < \infty$$

et Y_t est donc une variable de carré intégrable. ■

Théorème 1.5. *Soit $\{\psi_k\}$ une suite telle que $\sum_{k=-\infty}^{\infty} |\psi_k| < \infty$ et soit $\{X_t\}$ un processus stationnaire au second ordre de moyenne $\mu_X = \mathbb{E}\left[X_t\right]$ et de fonction d'autocovariance $\gamma_X(h) = \mathrm{cov}(X_{t+h}, X_t)$. Alors le processus $Y_t = \sum_{s=-\infty}^{\infty} \psi_s X_{t-s}$ est stationnaire au second ordre de moyenne :*

$$\mu_Y = \mu_X \sum_{k=-\infty}^{\infty} \psi_k \tag{1.47}$$

de fonction d'autocovariance :

$$\gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h + k - j) \tag{1.48}$$

et de mesure spectrale :

$$\nu_Y(d\lambda) = |\psi(e^{-i\lambda})|^2 \nu_X(d\lambda) \tag{1.49}$$

où $\psi(e^{-i\lambda}) = \sum_k \psi_k e^{-ik\lambda}$ *est la fonction de transfert du filtre. Enfin l'intercovariance entre les processus* Y_t *et* X_t *a pour expression :*

$$\gamma_{YX}(h) = \mathbb{E}\left[(Y_{t+h} - \mu_Y)(X_t - \mu_X)\right] = \sum_{k=-\infty}^{\infty} \psi_k \gamma_X(h-k) \tag{1.50}$$

Démonstration. Comme $\mathbb{E}\left[\sum_{s=-\infty}^{\infty} |\psi_s| \mathbb{E}\left[|X_{t-s}|\right]\right] < \infty$, le théorème de Fubini implique

$$\mathbb{E}\left[\sum_{s=-\infty}^{\infty} \psi_s X_{t-s}\right] = \sum_{s=-\infty}^{\infty} \psi_s \mathbb{E}\left[X_{t-s}\right]$$

ce qui établit (1.47). Pour la fonction d'autocovariance, notons tout d'abord que, pour tout n, le processus $Y_{n,t} = \sum_{s=-n}^{n} \psi_s X_{t-s}$ est stationnaire au second ordre et que nous avons

$$\mathrm{cov}(Y_{n,t}, Y_{n,t+h}) = \sum_{j=-n}^{n} \sum_{k=-n}^{n} \psi_j \psi_k \gamma_X(h+k-j)$$

Remarquons ensuite que

$$\begin{aligned}
\mathrm{cov}(Y_t, Y_{t+h}) &= \mathrm{cov}(Y_{n,t} + (Y_t - Y_{n,t}), Y_{n,t+h} + (Y_{t+h} - Y_{n,t+h})) \\
&= \mathrm{cov}(Y_{n,t}, Y_{n,t+h}) + \mathrm{cov}(Y_t - Y_{n,t}, Y_{n,t+h}) \\
&+ \mathrm{cov}(Y_{n,t}, Y_{t+h} - Y_{n,t+h}) + \mathrm{cov}(Y_t - Y_{n,t}, Y_{t+h} - Y_{n,t+h}) \\
&= A + B + C + D
\end{aligned}$$

L'inégalité :

$$\mathrm{var}(Y_{n,t} - Y_t) = \lim_{p\to\infty} \mathrm{var}(Y_{n,t} - Y_{p,t}) \leq \left(\sum_{j=n+1}^{\infty} |\psi_j|\right)^2 \gamma_X(0)$$

permet ensuite de déduire, quand n tend vers l'infini, les limites suivantes

$$\begin{aligned}
|B| &\leq (\mathrm{var}(Y_t - Y_{n,t}))^{1/2} (\mathrm{var}(Y_{n,t+h}))^{1/2} \to 0 \\
|C| &\leq (\mathrm{var}(Y_{t+h} - Y_{n,t+h}))^{1/2} (\mathrm{var}(Y_{n,t}))^{1/2} \to 0 \\
|D| &\leq (\mathrm{var}(Y_{t+h} - Y_{n,t+h}))^{1/2} (\mathrm{var}(Y_t - Y_{n,t}))^{1/2} \to 0
\end{aligned}$$

et donc $\mathrm{cov}(Y_t, Y_{t+h}) = \lim_{n\to\infty} \mathrm{cov}(Y_{n,t}, Y_{n,t+h})$, ce qui démontre l'expression (1.48) [3]. En reportant dans cette expression $\gamma_X(h) = \int_I e^{ih\lambda} \nu_X(d\lambda)$ où ν_X désigne la mesure spectrale du processus $\{X_t\}$, nous obtenons

$$\gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \int_I e^{i(h+k-j)\lambda} \nu_X(d\lambda)$$

[3] Nous venons ici de démontrer directement la propriété de continuité de la covariance dans L^2 que nous verrons comme une conséquence de la structure d'espace de Hilbert au chapitre 4.

En remarquant ensuite que

$$\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\int_I |\psi_j||\psi_k|\nu_X(d\lambda) \leq \gamma_X(0)\left(\sum_{j=-\infty}^{\infty}|\psi_j|\right)^2$$

nous pouvons appliquer le théorème de Fubini et permuter les signes somme et intégrale dans l'expression de $\gamma_Y(h)$. Ce qui donne :

$$\gamma_Y(h) = \int_I e^{ih\lambda}\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\psi_j\psi_k e^{ik\lambda}e^{-ij\lambda} = \int_I e^{ih\lambda}|\psi(e^{-i\lambda})|^2\nu_X(d\lambda)$$

On en déduit que $\nu_Y(d\lambda) = |\psi(e^{-i\lambda})|^2\nu_X(d\lambda)$. Pour déterminer l'expression de l'intercovariance entre les processus entre les processus Y_t et X_t, il suffit de noter $|\text{cov}(Y_{t+h}, X_t)|^2 \leq \gamma_Y(0)\gamma_X(0) < +\infty$ et que :

$$\begin{aligned}\mathbb{E}\left[(Y_{t+h}-\mu_Y)(X_t-\mu_X)\right] &= \lim_{n\to\infty}\text{cov}(Y_{n,t+h}, X_t) = \lim_{n\to\infty}\sum_{k=-n}^{n}\psi_k\text{cov}(X_{t+h-k}X_t)\\ &= \sum_{k=-\infty}^{\infty}\psi_k\gamma_X(h-k)\end{aligned}$$

Ce qui conclut la preuve. ■

Chapitre 2

Estimation de la moyenne et des covariances

2.1 Estimation de la moyenne

Soit $\{X_t\}$ un processus aléatoire à temps discret stationnaire au second ordre, de moyenne $\mathbb{E}[X_t] = \mu$, et de fonction d'autocovariance $\gamma(h)$. On suppose avoir observé n échantillons consécutifs $X_1, \dots X_n$ du processus. L'estimateur de μ que nous considérons est la *moyenne empirique* définie par :

$$\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t \tag{2.1}$$

On constate tout d'abord que $\hat{\mu}_n$ est un estimateur *sans biais* de la moyenne μ car

$$\mathbb{E}[\hat{\mu}_n] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[X_t] = \mu \tag{2.2}$$

du fait de la stationnarité. Le *risque quadratique* de l'estimateur, qui mesure sa dispersion autour de la valeur inconnue μ de la moyenne, a pour expression

$$\begin{aligned} R(\hat{\mu}_n, \mu) &= \mathbb{E}\left[(\hat{\mu}_n - \mu)^2\right] \\ &= \mathbb{E}\left[\frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}(X_t-\mu)(X_s-\mu)\right] = \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\gamma(t-s) = \frac{1}{n}\sum_{h=-n+1}^{n-1}\left(1-\frac{|h|}{n}\right)\gamma(h) \end{aligned} \tag{2.3}$$

D'où la proposition suivante :

Proposition 2.1. *Soit $\{X_t\}$ un processus stationnaire au second ordre de moyenne μ et de fonction d'autocovariance $\gamma(h)$ avec $\sum|\gamma(h)| < \infty$. Alors, $\hat{\mu}_n = n^{-1}\sum_{t=1}^{n} X_t$ vérifie*

$$\lim_{n\to\infty} n\mathbb{E}\left[(\hat{\mu}_n-\mu)^2\right] = 2\pi f(0) \qquad \text{où} \quad f(\lambda) = \frac{1}{2\ pi}\sum_{\tau=-\infty}^{\infty}\gamma(\tau)e^{-i\tau\lambda}\,. \tag{2.4}$$

c'est à dire que $\hat{\mu}_n$ converge en moyenne quadratique vers μ, à la vitesse $\sqrt{n}$. De plus $\lim_{n\to\infty}\hat{\mu}_n = \mu$ $\mathbb{P}$-p.s.

Démonstration. Lorsque $\gamma(h)$ est absolument sommable, le théorème de la convergence dominée appliquée à (2.3) montre que

$$\lim_{n\to\infty} nR(\hat{\mu}_n, \mu) = \sum_{h=-\infty}^{\infty} \lim_{n\to\infty} \left(1 - \frac{|h|}{n}\right) \gamma(h) = \sum_{h=-\infty}^{\infty} \gamma(h) = 2\pi f(0)$$

où $f(\lambda) = (2\pi)^{-1} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-ih\lambda}$ est la densité spectrale du processus $\{X_t\}$. La preuve de la convergence presque sûre de $\hat{\mu}_n$ est traitée par l'exercice **??**. ■

Cette proposition montre que la loi des grands nombres, établie classiquement pour des variables aléatoires indépendantes, est également valable pour un processus stationnaire au second ordre, du moment que la fonction d'autocovariance décroît suffisamment rapidement à l'infini. Sous cette condition, il est possible d'estimer la moyenne à partir d'une seule réalisation de celui-ci. La proposition 2.1 nous donne accès à la valeur limite de $\mathbb{E}\left[(\sqrt{n}(\hat{\mu}_n - \mu))^2\right]$. Cependant pour construire des intervalles de confiance pour les paramètres estimés (cf. définition A.27) ou pour tester des hypothèses concernant la valeur des paramètres (voir définition A.28), il est nécessaire d'obtenir un résultat plus précis portant sur la distribution limite de $\sqrt{n}(\hat{\mu}_n - \mu)$. L'obtention de théorèmes de type limite centrale pour des suites de variables aléatoires *dépendantes* est un sujet délicat, qui a donné lieu à une vaste littérature. Il n'est bien entendu pas question ici de présenter une théorie générale et nous nous contentons donc d'énoncer un résultat valable dans le cas de processus linéaire fort. Le fait de devoir émettre une hypothèse aussi contraignante sur la loi du processus dans un contexte où, en fait, seules les propriétés au second ordre nous intéressent est bien sûr frustrant, mais il traduit la (relative) difficulté technique d'un tel résultat (la preuve de ce théorème est omise).

Théorème 2.1. *Soit $\{X_t\}$ un processus linéaire fort de moyenne μ. On a $X_t = \mu + \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k}$ avec $\sum_k |\psi_k| < \infty$ et $Z_t \sim \mathrm{IID}(0, \sigma^2)$. On pose $\hat{\mu}_n = n^{-1} \sum_{t=1}^{n} X_t$. Alors :*

$$\sqrt{n}(\hat{\mu}_n - \mu) \to_d \mathcal{N}(0, 2\pi f(0)) \tag{2.5}$$

où $f(0) = \sigma^2 |\hat{\psi}(0)|^2/(2\pi)$, $\hat{\psi}(\lambda) = \sum_{j=-\infty}^{\infty} \psi_j e^{ij\lambda}$, est la densité spectrale de X_t en 0.

Exemple 2.1 : Moyenne empirique pour un processus AR(1) (fort)
Soit X_t un processus autorégressif d'ordre 1 fort, de moyenne μ, solution stationnaire au second ordre défini par l'équation de récurrence

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t$$

où $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$ et $|\phi| < 1$. Nous rappelons que la fonction d'autocovariance d'un processus $AR(1)$ pour $|\phi| < 1$ est donnée par

$$\gamma_X(k) = \frac{\sigma^2}{(1 - \phi^2)} \phi^{|k|}$$

et que la densité spectrale de ce processus a pour expression

$$f(\lambda) = \frac{\sigma^2}{2\pi \left|1 - \phi e^{-i\lambda}\right|^2}$$

Dans ce cas, la variance limite qui intervient dans l'équation (2.5), est donnée par $2\pi f(0) = \sigma^2/(1-\phi)^2$. Cette valeur est à comparer avec la variance de X_t donnée par $\gamma(0) = \sigma^2/(1-\phi^2)$. On constate que le rapport $2\pi f(0)/\gamma(0) = (1+\phi)/(1-\phi)$ tend vers 0 lorsque $\phi \to -1$ et vers $+\infty$ lorsque $\phi \to 1$. Ce qui implique

par exemple lorsque l'on considère l'intervalle de confiance asymptotique de niveau 95% pour la moyenne μ donné par $[\hat{\mu}_n - 1.96\sigma n^{-1/2}/(1-\phi), \hat{\mu}_n + 1.96\sigma n^{-1/2}/(1-\phi)]$ que l'estimation de la moyenne est bien meilleure (plus précise) que si les données étaient iid lorsque ϕ est proche de -1. Inversement, lorsque ϕ est proche de 1, l'intervalle de confiance est beaucoup plus large, c'est à dire l'estimation est significativement moins précise, pour un nombre n d'échantillons comparable, que si les données étaient indépendantes. Cette constatation somme toute assez logique est à mettre en rapport avec l'allure des trajectoires représentées sur la figure 1.11.

2.2 Estimation des coefficients d'autocovariance et d'autocorrélation

Considérons à nouveau un processus $\{X_t\}$ stationnaire au second ordre, de moyenne μ et de fonction d'autocovariance $\gamma(h)$ supposée de module sommable. Pour estimer la suite $\gamma(h)$, nous considérons les estimateurs, dits de covariances empiriques, définis par :

$$\hat{\gamma}_n(h) = \begin{cases} n^{-1}\sum_{t=1}^{n-|h|}(X_{t+|h|} - \hat{\mu}_n)(X_t - \hat{\mu}_n) & \text{si } |h| \leq n-1 \\ 0 & \text{sinon} \end{cases} \tag{2.6}$$

où $\hat{\mu}_n = n^{-1}\sum_{t=1}^{n} X_t$. Remarquons que le nombre d'observations, dont nous disposons, étant précisément égal à n, il n'existe pas de paires d'observations séparées de plus de $n-1$ intervalles de temps et donc l'expression (2.6) ne permet pas d'estimer les valeurs de $\gamma(h)$ pour $|h| \geq n$. De plus, lorsque $|h|$ est proche de n, il est clair que l'estimateur (2.6) de la covariance n'est pas fiable, dans la mesure où on ne dispose que de peu de paires d'observations $(X_t, X_{t+|h|})$, ce qui implique que l'effet de moyennage statistique ne peut pas jouer. La partie la plus utile de la fonction d'autocovariance empirique est celle qui correspond au valeurs du décalage h significativement plus faibles que le nombre d'observations n. A échantillon fini, $\hat{\gamma}_n(h)$ est un estimateur biaisé de $\gamma(h)$. Un calcul simple montre par exemple que

$$\mathbb{E}\left[\hat{\gamma}_n(0)\right] = \gamma(0) - \frac{1}{n}\sum_{k=-(n-1)}^{(n-1)}\left(1 - \frac{|k|}{n}\right)\gamma(k)$$

Toutefois on peut montrer que, pour tout h, l'estimateur donné par (2.6) est asymptotiquement sans biais dans le sens où $\lim_{n\to\infty}\mathbb{E}\left[\hat{\gamma}_n(h)\right] = \gamma(h)$ à la vitesse $1/n$. Une propriété importante de cet estimateur est que la suite $\hat{\gamma}_n(h)$ est de type positif. En effet, si on définit le *périodogramme* par[1]

$$I_n(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n}(X_t - \hat{\mu}_n)e^{-it\lambda}\right|^2 \tag{2.7}$$

Par construction, $I_n(\lambda)$ est une fonction positive pour $\lambda \in [-\pi, \pi]$. Par ailleurs,

$$\int_{-\pi}^{\pi} e^{i\lambda h} I_n(\lambda)d\lambda = \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}(X_t - \hat{\mu}_n)(X_s - \hat{\mu}_n)\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda(h-t+s)} = \hat{\gamma}_n(h)$$

Par conséquent, d'après le théorème d'Herglotz 1.3, la suite $\hat{\gamma}_n(h)$ est de type positif.

[1] Le périodogramme joue un rôle fondamental pour l'estimation de la densité spectrale étudiée au chapitre 3.

Propriété 2.1. *Si $\hat{\gamma}_n(0) > 0$ alors, pour tout $p \leq n$, la matrice $\hat{\Gamma}_{n,p}$ définie par*

$$\hat{\Gamma}_n = \begin{bmatrix} \hat{\gamma}_n(0) & \hat{\gamma}_n(1) & \cdots & \hat{\gamma}_n(p-1) \\ \hat{\gamma}_n(1) & \hat{\gamma}_n(0) & \cdots & \hat{\gamma}_n(p-2) \\ \vdots & & & \\ \hat{\gamma}_n(p-1) & \hat{\gamma}_n(p-2) & \cdots & \hat{\gamma}_n(0) \end{bmatrix} \tag{2.8}$$

est de rang plein et est donc inversible.

Démonstration. La suite $\hat{\gamma}_n(h)$ est de type positif, $\hat{\gamma}_n(0) > 0$ et $\hat{\gamma}_n(h)$ tend vers 0 quand n tend vers l'infini. On en déduit, d'après la propriété 1.5, que, pour tout p, la matrice est inversible. ■

L'estimateur dit "non biaisé" de la fonction d'autocovariance obtenu en remplaçant n^{-1} par $(n - |h|)^{-1}$ dans l'expression (2.6) ne possède pas cette propriété. Ajouté au fait que ces deux estimateurs sont asymptotiquement équivalent, l'estimateur non biaisé présente peu d'intérêt dans le cas des séries temporelles. Les coefficients d'autocovariance empiriques interviennent quasiment dans tous les problèmes d'inférence statistique portant sur les processus stationnaires. A l'instar de la moyenne empirique, il est donc indispensable de disposer de résultats concernant leur distribution. Cependant, même pour les modèles de processus les plus simples, il est en général impossible de préciser la distribution exacte de la suite de variables aléatoires $\hat{\gamma}_n(0), \ldots, \hat{\gamma}_n(k)$ à n fini. Nous ne considérons ici que des résultats asymptotiques concernant la distribution limite jointe de $\hat{\gamma}_n(0), \ldots, \hat{\gamma}_n(k)$, pour k fixé, lorsque n tends vers l'infini. Il s'avère que le résultat le plus simple à utiliser (dans le cas général) est celui qui concerne la fonction d'autocorrélation empirique plutôt que la covariance. On rappelle que les coefficients d'autocorrélation sont définis par

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

et qu'ils vérifient $|\rho(h)| \leq \rho(0) = 1$ (cf. paragraphe 1.2). On définit les coefficients d'autocorrélation empiriques par

$$\hat{\rho}_n(h) = \frac{\hat{\gamma}_n(h)}{\hat{\gamma}_n(0)} \tag{2.9}$$

où $\hat{\gamma}(h)$ est donné par (2.6).

Théorème 2.2. *Soit $\{X_t\}$ un processus linéaire défini par $X_t - \mu = \sum_{s=-\infty}^{\infty} \psi_s Z_{t-s}$ avec $\sum_s |\psi_s| < \infty$. On suppose que $Z_t \sim \mathrm{IID}(0, \sigma^2)$ vérifie $\mathbb{E}\left[Z_t^4\right] < \infty$. Pour $k \geq 1$, on note $\widehat{\boldsymbol{\rho}}_n = (\hat{\rho}_n(1), \ldots, \hat{\rho}_n(k))^T$, $\boldsymbol{\rho} = (\rho(1), \ldots, \rho(k))^T$ et W la matrice de dimension $k \times k$ définie, pour $1 \leq \ell, m \leq k$, par l'élément :*

$$W_{\ell,m} = \sum_{s=1}^{\infty} (\rho(s+\ell) + \rho(s-\ell) - 2\rho(s)\rho(\ell))(\rho(s+m) + \rho(s-m) - 2\rho(s)\rho(m)) \tag{2.10}$$

Alors :

$$\sqrt{n}(\widehat{\boldsymbol{\rho}}_n - \boldsymbol{\rho}) \to_d \mathcal{N}(0, W) \tag{2.11}$$

Il est remarquable de noter que la distribution des coefficients d'autocorrélation ne dépend pas des moments du processus Z_t (on a uniquement supposé que $Z_t \sim \mathrm{IID}(0, \sigma^2)$ avec un moment du 4ème ordre fini). Comme dans le cas du théorème 2.1, on constate qu'il est nécessaire d'admettre des hypothèses relativement fortes pour garantir ce résultat dont nous omettons la démonstration.

Exemple 2.2 : Bruit blanc fort
Soit $\{X_t\} \sim \text{IID}(0, \sigma^2)$. *Dans ce cas* $\rho(h) = 0$ *pour tout* $h \neq 0$ *et la matrice de covariance asymptotique* W *est égale à la matrice identité. L'expression* (2.11) *montre que, pour* n *suffisamment grand, les coefficients d'autocorrélation empiriques* $\hat{\rho}_n(1), \cdots, \hat{\rho}_n(k)$ *sont indépendants, gaussiens de moyenne nulle et de variance égale à* $1/n$. *On en déduit que, pour tout* $h \neq 0$ *:*

$$\mathbb{P}-1.96n^{-1/2} \leq \hat{\rho}_n(h) \leq 1.96n^{-1/2} \approx 0.95 \tag{2.12}$$

Ce résultat peut être utilisé pour tester l'hypothèse que les valeurs observées sont celles d'un bruit blanc fort. En effet si $\hat{\rho}_n(1)$ *sort de l'intervalle* $(-1.96n^{-1/2}, 1.96n^{-1/2})$, *alors on peut, avec confiance, rejeter une telle hypothèse. Nous avons représenté figure 2.1 les* 60 *premiers coefficients d'autocorrélation empiriques d'un*

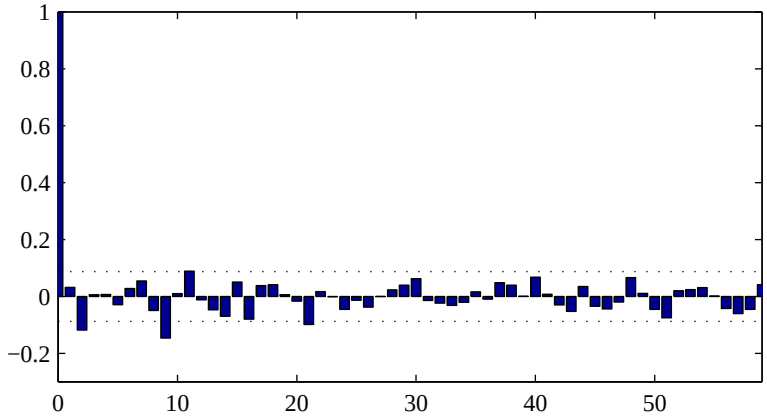

FIG. 2.1 – *Fonction d'autocorrélation empirique pour un échantillon de bruit blanc fort, gaussien, centré, de variance* $\sigma^2 = 1$ *et de longueur* $n = 500$. *Les droites en pointillé représentent la plage asymptotique autour de la vraie valeur* $\rho(h) = 0$, *plage où il y a* 95% *de chance de trouver* $\hat{\rho}_n(h)$.

échantillon de longueur $n = 500$, *d'un bruit blanc fort, gaussien, centré, de variance* $\sigma^2 = 1$. *A partir de la formule* (2.12), *nous avons reporté l'intervalle asymptotique* $[-1.96n^{-1/2}, 1.96n^{-1/2}]$ *autour de la vraie valeur* $\rho(h) = 0$ *où il y a* 95% *de chance de trouver* $\hat{\rho}_n(h)$ *sous l'hypothèse que l'observation est un bruit* IID. *Sur la réalisation considérée, cette hypothèse est très vraisemblable puisque seules quelques valeurs, sur les* 60 *coefficients empiriques calculés, sortent de cet intervalle. Ce type de tracé où l'on représente les coefficients d'autocorrélation empiriques ainsi que la limite de la zone crédible (à 95% par exemple) pour les estimateurs correspondants dans le cas du bruit blanc (fort) est très classique dans le domaine des séries temporelles où il est désigné sous le nom de corrélogramme. Il permet de détecter visuellement les décalages temporels pour lesquels l'hypothèse de décorrélation n'est pas admissible (comme dans le cas de la figure 2.2 par exemple). Il ne constitue cependant pas un test formel du caractère blanc dans la mesure où il ignore les éventuels effets conjoints concernant plusieurs décalages temporels. Un test de blancheur suggéré par 2.11 consiste par exemple à vérifier que la valeur de* $\sum_{l=1}^{k} \hat{\rho}_n(l)^2$ *correspond bien à une valeur inférieure à 95% pour la fonction de répartition de la loi* ξ_k^2 *du chi carré à* k *degrés de liberté.*

Exemple 2.3 : Processus MA(1)
On considère le processus MA(1) défini par $X_t = Z_t + \theta_1 Z_{t-1}$ *où* Z_t *est un bruit blanc fort, centré, de variance*

σ^2. *Ici, la suite des coefficients d'autocorrélation est donnée par :*

$$\rho(h) = \begin{cases} 1 & pour \quad h = 0 \\ \dfrac{\theta_1}{1+\theta_1^2} & pour \quad |h| = 1 \\ 0 & pour \quad |h| \geq 2 \end{cases}$$

On en déduit, d'après (2.10)*, que les éléments diagonaux de la matrice de covariance de la distribution limite des coefficients d'autocovariance empiriques ont pour expression :*

$$W_{h,h} = \begin{cases} 1 - 3\rho^2(1) + 4\rho^4(1) & pour \quad |h| = 1 \\ 1 + 2\rho(1)^2 & pour \quad |h| \geq 2 \end{cases}$$

Par conséquent la zone crédible à 95% *pour les coefficients d'autocorrélation empiriques sont donnés, pour* $h = 1$*, par :*

$$\hat{\rho}_n(1) \in \left[\rho(1) - 1.96 W_{1,1}^{1/2} n^{-1/2} \quad \rho(1) + 1.96 W_{1,1}^{1/2} n^{-1/2}\right]$$

et, pour $h \geq 2$*, par :*

$$\hat{\rho}_n(h) \in \left[-1.96 W_{2,2}^{1/2} n^{-1/2} \quad + 1.96 W_{2,2}^{1/2} n^{-1/2}\right]$$

Notons ici que ces régions dépendent, par l'intermédiaire de $\rho(1)$*, de la quantité a priori inconnue* θ_1*. Nous avons représenté figure 2.2 les* 60 *premiers coefficients d'autocorrélation empiriques d'un échantillon de longueur* $n = 500$ *d'un processus MA(1) défini par* $\theta_1 = -0.8$ *et* $\sigma = 1$*. Les traits en pointillé représentent les bornes asymptotiques autour des vraies valeurs au niveau 95%. Pour une réalisation particulière, nous*

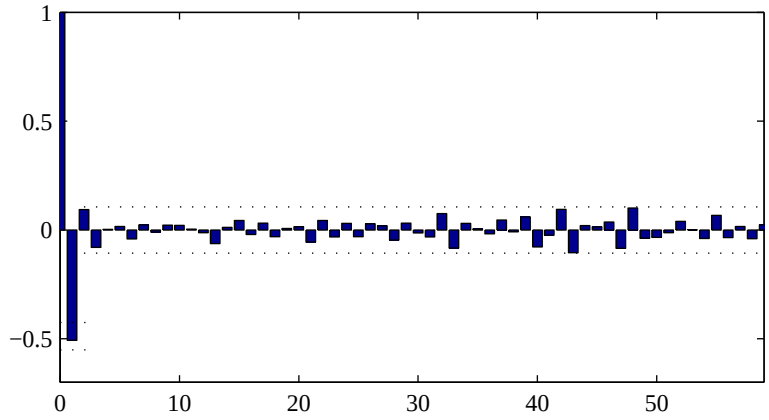

FIG. 2.2 – *Fonction d'autocorrélation empirique d'un échantillon de longueur* $n = 500$*) d'un processus MA(1) pour* $\theta_1 = -0.8$ *et donc* $\rho(1) = -0.4878$*. Les traits en pointillé représentent les plages où il y a* 95% *de chance de trouver* $\hat{\rho}_n(h)$ *si* $h \geq 2$*.*

avons obtenu $\hat{\rho}_n(1) = -0.4924$*. Cela permet d'affirmer avec une grande confiance que le processus n'est pas un bruit blanc car cette valeur est très en dehors de la plage* $\pm 1.96 n^{-1/2} = \pm 0.0877$ *correspondant à l'hypothèse que* X_t *soit un bruit blanc (cf. exemple 2.2). D'autre part, les résultats reportés figure 2.2 montrent que l'hypothèse que le processus observé est* $MA(1)$ *de paramètre* $\theta_1 = -0.8$ *est vraisemblable. En effet, les*

coefficients d'autocorrélation empiriques sont clairement à l'intérieur des plages théoriques déduites du calcul asymptotique.

Exemple 2.4 : Processus autorégressif fort d'ordre 1
On considère le processus aléatoire X_t défini par :

$$X_t = \phi X_{t-1} + Z_t$$

où $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$ et où $|\phi| < 1$. La fonction d'autocorrélation d'un tel processus est donnée par $\rho(h) = \phi^{|h|}$ et les éléments diagonaux de la matrice de covariance W sont donnés par

$$\begin{aligned} W_{h,h} &= \sum_{m=1}^{h} \phi^{2h}(\phi^{-m} - \phi^{m})^2 + \sum_{m=h+1}^{\infty} \phi^{2m}(\phi^{-i} - \phi^{i})^2 \\ &= (1-\phi^{2h})(1+\phi^2)(1-\phi^2)^{-1} - 2h\phi^{2h} \end{aligned}$$

Considérons la séquence, de longueur $n = 1800$, des battements cardiaques représentés figure 1.1 (chapitre 1). La figure 1.6 qui représente les couples (X_t, X_{t-1}) suggère fortement la présence d'une relation linéaire entre les variables X_t et X_{t-1} et invite donc à tester la validité d'un modèle autorégressif d'ordre 1. Pour estimer le paramètre ϕ du modèle autorégressif, une méthode naturelle, compte tenu de l'allure de la fonction d'autocorrélation de l'$AR(1)$, consiste à utiliser comme estimateur $\hat{\phi}_n = \hat{\rho}_n(1)$ qui donne $\hat{\phi}_n = 0.966$. Pour tester la validité du modèle, deux solutions s'offrent à nous : (i) tester que les résidus de prédiction donnés par $\hat{Z}_t = X_t - \hat{\mu}_n - \hat{\phi}_n(X_{t-1} - \hat{\mu}_n)$ sont compatibles avec un modèle de bruit blanc, (ii) vérifier directement que les coefficients d'autocorrélation empiriques sont compatibles avec ceux d'un modèle AR(1). Les résidus de prédiction sont reportés figure 2.3 et la fonction d'autocorrélation de ces résidus figure 2.4, où nous avons aussi indiqué les bornes de la zone crédible à 95% pour le bruit blanc avec un nombre d'observations $n = 1800$. Les corrélations empiriques, en particulier pour $h = 2$, sont significativement à l'extérieur des intervalles de confiance du bruit blanc, ce qui conduit à rejeter le modèle de bruit blanc pour les résidus et donc le modèle autorégressif d'ordre 1 pour les observations. Les résultats de l'analyse de la suite des coefficients d'autocorrélation empiriques du processus et des zones crédibles à 95% sous l'hypothèse d'un modèle AR(1) avec $\phi = 0.966$ sont reportés figure 2.5. On observe que les premières valeurs des coefficients de corrélation sont nettement à l'extérieur de cette zone, ce qui contribue ici encore à rejeter le modèle $AR(1)$.

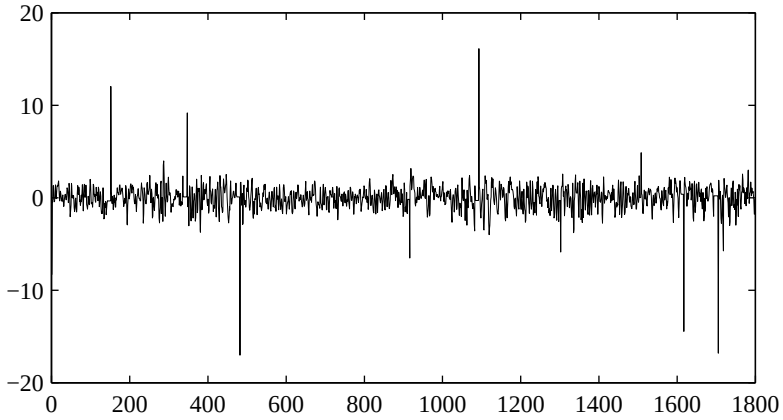

FIG. 2.3 – *Série des battements cardiaques : Résidu de prédiction* $\hat{Z}_t = (X_t - \hat{\mu}_n) - \hat{\phi}_n(X_{t-1} - \hat{\mu}_n)$.

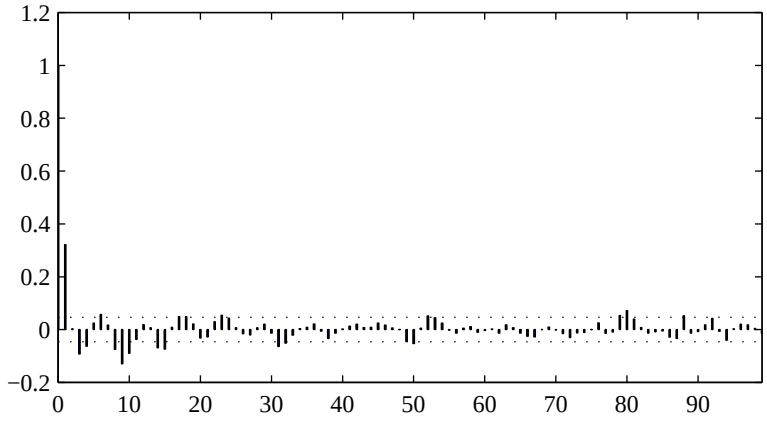

FIG. 2.4 – *Série des battements cardiaques : coefficients d'autocorrélation empiriques des résidus de prédiction* $\hat{Z}_t = (X_t - \hat{\mu}_n) - \hat{\phi}_n(X_{t-1} - \hat{\mu}_n)$ *et zones crédibles à* 95% *pour le bruit blanc (*$n = 1800$*).*

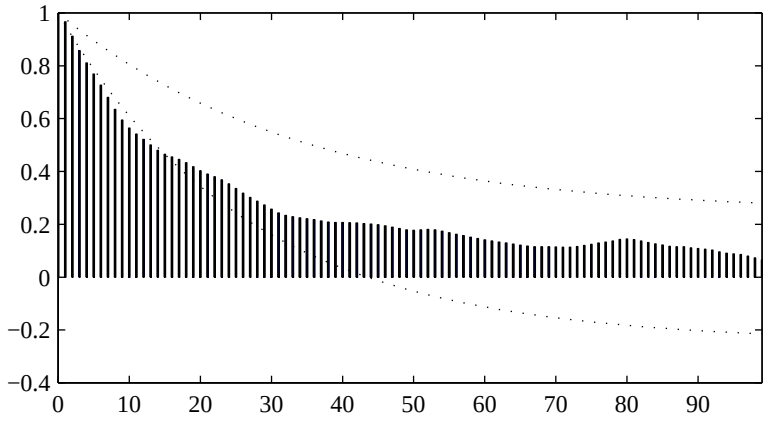

FIG. 2.5 – *Série des battements cardiaques : coefficients d'autocorrélation empiriques de la série et bornes des zones crédibles à* 95% *pour un modèle AR(1) de paramètre* $\phi = 0.966$.

Chapitre 3

Estimation spectrale non paramétrique

Dans le chapitre précédent, nous nous sommes intéressés à l'estimation de la fonction d'autocovariance. Dans certaines applications, il est plus pertinent d'essayer de modéliser la densité spectrale, qui décrit la façon dont l'énergie du processus se répartit en fréquence. L'information spectrale est souvent plus riche et plus facile à interpréter que la fonction d'autocovariance, révélant des structures (par exemple, cycles ou pseudo-cycles) qui ne sont pas directement visibles sur la forme d'onde ni même sur la suite des corrélations. Pour nous en convaincre considérons l'exemple de la forme d'onde représentée figure 3.1. Il s'agit d'un segment d'environ 40 millisecondes extrait d'un enregistrement d'un son produit par un harmonica. La forme d'onde est complexe, reflétant les deux caractéristiques essentielles du signal produit par cet instrument : des composantes cycliques liées aux vibrations des lames métalliques modulant de façon quasi-périodique le flux d'air et un bruit de friction. La fonction d'autocorrélation, que nous avons représentée à gauche figure 3.2, révèle en effet des structures temporelles complexes mais cette représentation n'est pas apte à réellement mettre en évidence la présence de (pseudo)-cycles. Ceux-ci apparaissent, par contre, clairement quand on observe le module de la transformée de Fourier du signal (à droite figure 3.2). Cette représentation fréquentielle n'est toutefois pas tout à fait satisfaisante, car elle est très "bruitée", ce qui rend difficile son interprétation. Cette variabilité est simplement la traduction, dans le domaine de Fourier, de la variabilité que nous observons dans la forme d'onde. L'objet de ce chapitre est de trouver une méthode d'estimation spectrale qui, tout en préservant les structures cycliques, soit capable de lisser les fluctuations.

3.1 Le périodogramme

Nous supposons dans cette partie que $\{X_t\}$ est un processus stationnaire au second-ordre de moyenne μ et de fonction de covariance $\gamma(h) \triangleq \mathbb{E}\left[(X_{t+h}-\mu)(X_t-\mu)\right]$ absolument sommable : $\sum|\gamma(h)| < \infty$. Sous ces hypothèses, le processus $\{X_t\}$ admet une densité spectrale donnée par :

$$f_X(\lambda) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\gamma(h)e^{-ih\lambda}$$

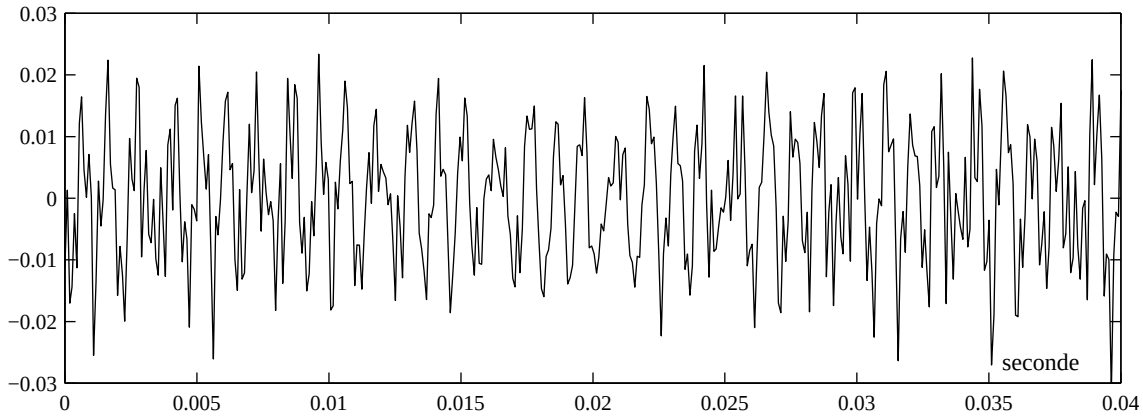

FIG. 3.1 – *Signal d'harmonica échantillonné à* 11.025 *kHz (temps en seconde).*

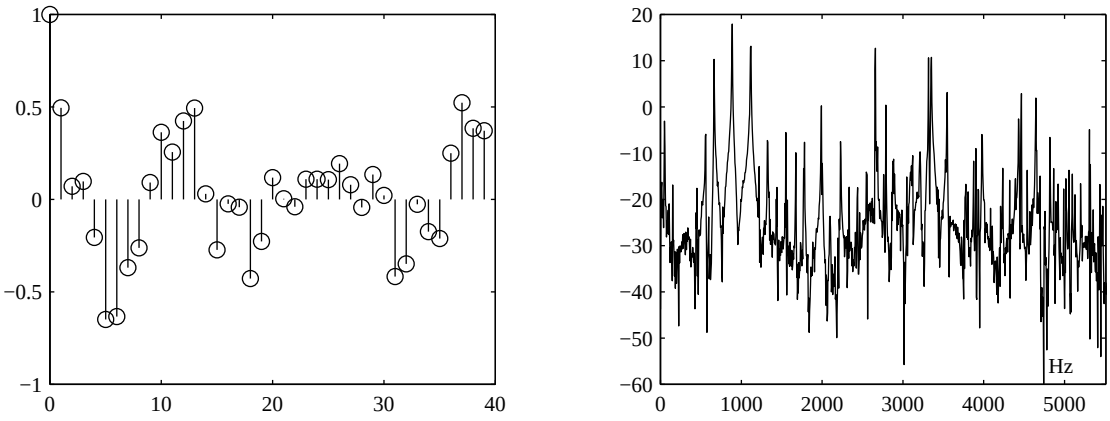

FIG. 3.2 – *A gauche, suite des* 40 *premiers coefficients de corrélation du signal représenté figure 3.1. A droite, transformée de Fourier (en dB) de ce signal (fréquence en Hz).*

Pour estimer la densité spectrale de $\{X_t\}$, il est naturel de s'intéresser au périodogramme, défini comme le module au carré de la transformée de Fourier discrète des observations $\{X_1, X_2, \cdots, X_n\}$:

$$I_n^X(\lambda_k) = |d_n^X(\lambda_k)|^2 \quad \text{où} \quad d_n^X(\lambda_k) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t e^{-it\lambda_k} \tag{3.1}$$

où $\lambda_k = 2\pi k/n$ sont les fréquences de Fourier. Remarquons ici que la relation :

$$\sum_{t=0}^{n-1} e^{-it\lambda_k} = 0 \quad \text{pour} \quad \lambda_k = 2\pi k/n \text{ et } k \in \{1, \ldots, (n-1)\}$$

montre que le périodogramme aux fréquences de Fourier λ_k, non nulles modulo 2π, est invariant par ajout d'une constante. Le périodogramme a été introduit par Sir Arthur Schuster (1898) pour étudier les "périodes cachées" apparaissant dans la série de tâches solaires. L'analyse spectrale des séries temporelles s'est ensuite considérablement développée avec l'apparition de moyens de calculs performants, et la découverte d'algorithmes de transformée de Fourier rapides (voir Brillinger, 1981).
Malheureusement nous allons voir dans la suite que le périodogramme n'est pas un "bon" estimateur de la densité spectrale, dans le sens où cet estimateur n'est pas consistant (il ne converge pas vers la vraie densité quand n tend vers l'infini). Néanmoins, il est à la base de la construction de la plupart des estimateurs de densité spectrale.
Rappelons tout d'abord que, comme nous l'avons déjà noté dans le chapitre 2 (voir expression (2.7)), le périodogramme est aussi égal à la transformée de Fourier discrète de la suite des coefficients d'autocovariance empiriques. En effet partant de :

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-|h|} (X_t - \hat{\mu}_n)(X_{t+|h|} - \hat{\mu}_n) \quad \text{où} \quad \hat{\mu}_n = n^{-1} \sum_{t=1}^{n} X_t$$

on vérifie aisément que

$$I_n^X(0) = \frac{1}{2\pi}\, n|\hat{\mu}_n|^2 \tag{3.2}$$

$$I_n^X(\lambda_k) = \frac{1}{2\pi} \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) \exp(-ih\lambda_k) \quad \text{pour } \lambda_k \neq 0 \tag{3.3}$$

Pour estimer la densité spectrale $f_X(\lambda)$ à toutes les fréquences, il est pratique d'étendre le périodogramme pour les valeurs de fréquences normalisées ne coïncidant pas avec les fréquences de Fourier. Ceci peut être fait de différentes manières ; nous suivrons l'extension adoptée par Fuller (1976) qui consiste à définir le périodogramme comme la fonction constante par morceaux donnée par :

$$I_n^X(\lambda) = \begin{cases} I_n^X(\lambda_k) & \text{si} \quad \lambda_k - \pi/n < \lambda \leq \lambda_k + \pi/n \quad \text{et} \quad 0 \leq \lambda \leq \pi \\ I_n^X(-\lambda) & \text{si} \quad -\pi \leq \lambda < 0 \end{cases} \tag{3.4}$$

Par construction, cette définition garantit que le périodogramme est une fonction paire, qui coïncide avec l'équation (3.1) aux fréquences $\lambda_k = 2\pi k/n$. De façon plus concise on peut alors écrire que :

$$I_n^X(\lambda) = I_n^X(g(n, \lambda))$$

où $g(n,\lambda)$ désigne, pour $\lambda \in [0,\pi]$, le multiple de $2\pi/n$ le plus proche de λ et, pour $\lambda \in [-\pi,0)$, $g(n,\lambda) = g(n,-\lambda)$. La proposition suivante établit que le périodogramme est asymptotiquement sans biais.

Théorème 3.1. *Soit $\{X_t\}$ un processus stationnaire de moyenne μ et de fonction d'autocovariance $\gamma(h)$ absolument sommable. Alors quand $n \to +\infty$ on a :*

$$\mathbb{E}\left[I_n^X(0)\right] - \frac{1}{2\pi} n\mu^2 \longrightarrow f_X(0)$$
$$\text{et} \qquad \mathbb{E}\left[I_n^X(\lambda)\right] \longrightarrow f_X(\lambda) \quad \text{pour} \quad \lambda \neq 0$$

Démonstration. Remarquons que, pour $\lambda \neq 0$, on a :

$$\mathbb{E}\left[I_n^X(g(n,\lambda))\right] = \frac{1}{2\pi} \sum_{h=-(n+1)}^{(n-1)} \left(1 - \frac{|h|}{n}\right) \gamma(\tau) e^{-ihg(n,\lambda)}$$

Posons $\gamma_n(h,\lambda) = (2\pi)^{-1}\mathbf{I}_{[-n,n]}(h)(1-|h|/n)\gamma(h)e^{-ihg(n,\lambda)}$. Nous avons $|\gamma_n(h,\lambda)| \leq |\gamma(h)|$ et $\lim_{n\to\infty} \gamma_n(h,x) = \gamma(h)e^{-ih\lambda}$. On conclut en appliquant le théorème de convergence dominée. ■

Pour comprendre les propriétés statistiques du périodogramme, nous allons tout d'abord nous intéresser à la distribution statistique du périodogramme d'un bruit blanc fort, c'est-à-dire d'une suite de variables aléatoires indépendantes et identiquement distribuées, de moyenne nulle et de variance finie.

Théorème 3.2. *Soit $\{Z_t\}$ une suite de variables aléatoires i.i.d., de moyenne nulle et de variance $\sigma^2 < \infty$. Sa distribution spectrale a pour densité $f_Z(\lambda) = \sigma^2/2\pi$.*

1. *Soient $0 < \omega_1 < \ldots < \omega_m < \pi$, m fréquences fixes. Le vecteur aléatoire $[I_n^Z(\omega_1), \cdots, I_n^Z(\omega_m)]$ converge en loi vers un vecteur de variables aléatoires indépendantes, distribuées suivant une loi exponentielle, de moyenne $\sigma^2/2\pi$.*
2. *Supposons que $\mathbb{E}\left[Z_t^4\right] < \infty$, alors :*

$$\text{var}\{I_n^Z(\lambda_k)\} = \begin{cases} 2f_Z^2(\lambda_k) + \kappa_4/4\pi^2 n & \lambda_k \in \{0,\pi\} \\ f_Z^2(\lambda_k) + \kappa_4/4\pi^2 n & 0 < \lambda_k < \pi \end{cases} \tag{3.5}$$

$$\text{et} \qquad \text{cov}\{I_n^Z(\lambda_j), I_n^Z(\lambda_k)\} = \kappa_4/4\pi^2 n \quad \text{pour} \quad \lambda_j \neq \lambda_k \tag{3.6}$$

   où $\lambda_k = 2\pi k/n$ sont les fréquences de Fourier et où κ_4 est le cumulant d'ordre 4 de la variable Z_1 défini par :

$$\kappa_4 = \mathbb{E}\left[Z_1^4\right] - 3(\mathbb{E}\left[Z_1^2\right])^2$$

3. *Supposons que les variables aléatoires Z_t soient gaussiennes. Alors $\kappa_4 = 0$ et, pour tout n, les variables aléatoires $I_n^Z(\lambda_k)/f_Z(\lambda)$, $k \in \{1,\cdots,(n-1)/2\}$ sont indépendantes et identiquement distribuées suivant une loi exponentielle[1] de moyenne 1.*

[1] Cette loi a pour densité $p(u) = e^{-u}\mathbf{I}(u \geq 0)$.

Démonstration. Elle est donnée en fin de chapitre. ■

La relation (3.5) du théorème 3.2 montre que *la variance de l'estimateur du périodogramme ne tend pas vers* 0 *lorsque le nombre d'échantillons tend vers l'infini.* Le périodogramme est bien un estimateur asymptotiquement sans biais de la densité spectrale du bruit blanc, mais *n'est pas consistant.* On voit même que $\sqrt{\mathrm{var}(I_n^Z(\lambda_k))}$ est de l'ordre de σ^2 et donc les fluctuations autour de la vraie valeur sont de l'ordre de grandeur de ce que l'on cherche à estimer. C'est ce que montre la figure 3.3 où nous avons représenté le périodogramme en dB d'un bruit blanc pour différentes valeurs de n. On observe sur ces réalisations qu'à certaines fréquences de Fourier les écarts avec la vraie valeur $\sigma^2/2\pi$ restent très importants même lorsque n augmente. Nous avons aussi reporté (droite en pointillé) le seuil de confiance à $\alpha = 90\%$ de la loi asymptotique de $I_n(\lambda_k)/f_Z(\lambda_k)$. Ce seuil a pour expression $s = -\log(1-\alpha)$.

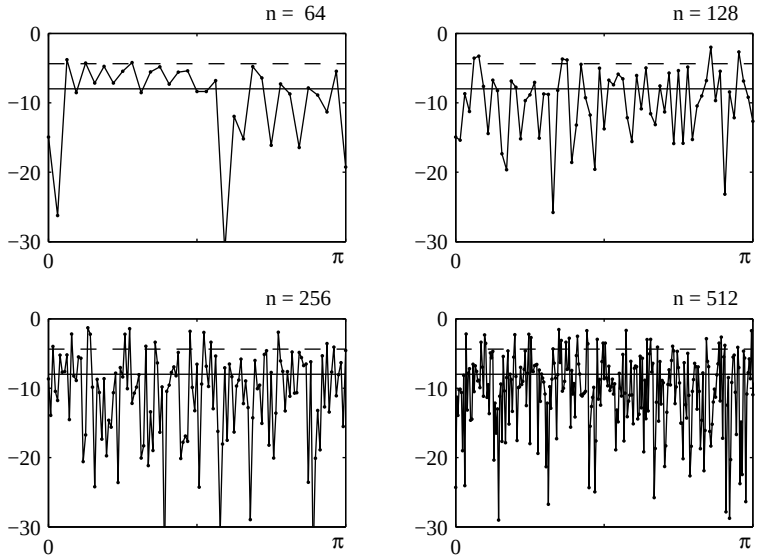

FIG. 3.3 – *Périodogramme en dB d'un bruit blanc de variance* 1 *en fonction de la fréquence* $\lambda \in (0,\pi)$*, pour différentes valeurs de* n*. La droite en trait plein représente la densité spectrale théorique* $\sigma^2/2\pi$ *et la droite en pointillé le seuil de confiance à* 90%.

Partant du théorème 3.2, valable pour les processus i.i.d., nous allons voir qu'il est encore possible d'étendre ce théorème à la classe plus large des processus linéaires forts centrés dont nous rappelons la définition.

Définition 3.1 (Processus linéaire fort)**.** *Le processus* $\{X_t\}$ *est linéaire fort, s'il existe un bruit blanc fort* $Z_t \sim IID(0,\sigma^2)$ *et une suite de coefficients* $\{\psi_k\}_{k\in\mathbb{Z}}$ *absolument sommable telle que :*

$$X_t = \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k} \tag{3.7}$$

On rappelle que X_t est stationnaire au second ordre, que $\mathbb{E}[X_t] = 0$ et que sa densité spectrale est donnée par :

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}|\psi(e^{-i\lambda})|^2 \tag{3.8}$$

Le théorème 3.3 montre qu'il existe une relation analogue à (3.8) entre le périodogramme $I_n^X(\lambda)$ du processus $\{X_t\}$ et le périodogramme $I_n^Z(\lambda)$ du bruit blanc fort $\{Z_t\}$ qui définit X_t.

Théorème 3.3. *Soit $\{X_t\}$ un processus linéaire fort. Supposons que $\sum_{j=-\infty}^{\infty} |\psi_j||j|^{1/2} < \infty$ et que $\mathbb{E}\left[Z_t^4\right] < \infty$. On a alors :*

$$I_n^X(\lambda_k) = |\psi(e^{-i\lambda_k})|^2 I_n^Z(\lambda_k) + R_n(\lambda_k)$$

où le terme $R_n(\lambda_k)$ vérifie[2] :

$$\max_{k\in\{1,\cdots,\lfloor(n-1)/2\rfloor\}} \mathbb{E}\left[|R_n(\lambda_k)|^2\right] = O(n^{-1})$$

Démonstration. Elle est donnée en fin de chapitre. ■

On comprend alors qu'en utilisant l'"approximation" donnée par le théorème 3.3 on puisse étendre le théorème 3.2 aux processus linéaires forts.

Théorème 3.4. *Soit $\{X_t\}$ un processus linéaire défini par :*

$$X_t = \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k}$$

où $\{Z_t\}$ est un bruit blanc fort $IID(0, \sigma^2)$ vérifiant $\mathbb{E}\left[Z_t^4\right] < \infty$. On suppose que $\sum_k |k|^{1/2}|\psi_k| < \infty$ et que $\psi(e^{-i\lambda}) = \sum_k \psi_k e^{-ik\lambda} \neq 0$. On note :

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\left|\psi(e^{-i\lambda})\right|^2$$

1. *Soient $0 < \omega_1 < \cdots < \omega_m < \pi$, m fréquences fixes. Le vecteur aléatoire $[I_n^X(\omega_1)/f_X(\omega_1), \cdots, I_n^X(\omega_m)/f_X(\omega_m)]$ converge en loi vers un vecteur de variables aléatoires indépendantes, distribuées suivant une loi exponentielle, de moyenne 1.*
2. *On a :*

$$\mathrm{var}(I_n^X(\lambda_k)) = \begin{cases} 2f_X^2(\lambda_k) + O(n^{-1/2}) & \lambda_k \in \{0, \pi\} \\ f_X^2(\lambda_k) + O(n^{-1/2}) & 0 < \lambda_k < \pi \end{cases}$$

$$\mathrm{cov}(I_n^X(\lambda_j), I_n^X(\lambda_k)) = O(n^{-1}) \quad \lambda_j \neq \lambda_k$$

Démonstration. La preuve est une conséquence directe des théorèmes 3.3 et 3.2. ■

[2]Notation : $O(n^{-\alpha})$ désigne une suite dépendant de n qui vérifie, quand $n \to \infty$, $O(n^{-\alpha})/n^{-\alpha} \to c \neq 0$ et $o(n^{-\alpha})$ vérifie $o(n^{-\alpha})/n^{-\alpha} \to 0$.

En conséquence, comme pour le bruit blanc fort, la variance du périodogramme d'un processus linéaire fort est, à une fréquence de Fourier, de l'ordre de grandeur du carré de la densité spectrale à cette fréquence. La figure 3.4 illustre ce résultat : elle montre le périodogramme, évalué sur 1024 échantillons, d'un processus AR(2) gaussien. L'écart-type du périodogramme est proportionnelle à la densité spectrale, ce qui rend bien entendu l'interprétation du périodogramme difficile. Le théorème

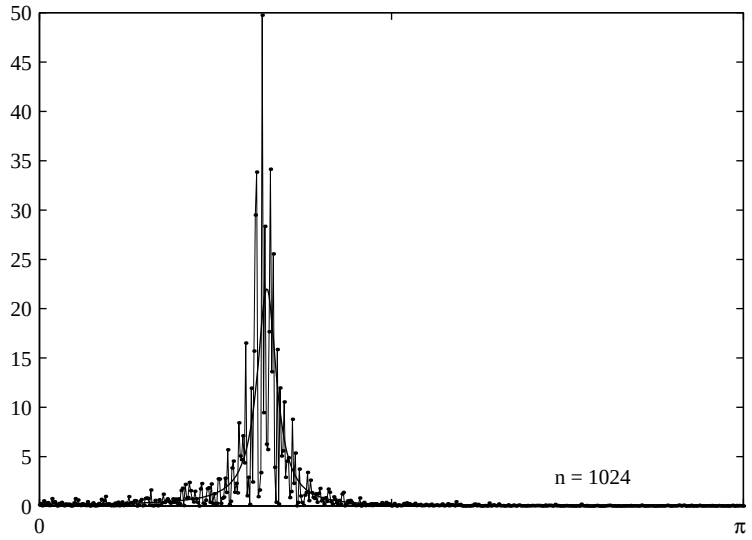

FIG. 3.4 – *Périodogramme pour un AR(2) de paramètres* $[1, -1, 0.9]$ *et* $\sigma^2 = 1$ *calculé sur* $n = 1024$ *échantillons, en fonction de la fréquence* $\lambda \in (0, \pi)$.

3.4 implique qu'asymptotiquement les variables aléatoires $[I_n(\lambda_1), \ldots, I_n(\lambda_{N/2})]$ se comportent comme un tableau de variables indépendantes distribuées marginalement comme $W f_X(\lambda_k)$ où W suit une loi exponentielle. Il s'agit donc d'une structure de bruit de type multiplicatif, où le paramètre d'intérêt, à savoir la densité spectrale, est *multipliée* par le "bruit" W. L'application d'une transformation logarithmique conduit naturellement à une structure de bruit additif : asymptotiquement le log-périodogramme est égal à la log-densité spectrale observée dans un bruit approximativement additif et de variance constante. Figure 3.4, nous avons représenté le spectre évalué en dB ainsi que l'intervalle de confiance à $\alpha = 90\%$ de la loi asymptotique de $I_n^X(\lambda_k)/f_X(\lambda_k)$ soit :

$$\lim_{n\to\infty} \mathbb{P}\left\{I_n^X(\lambda_k)/f_X(\lambda_k) > c\right\} = 1 - e^{-c} = \alpha$$

qui donne $c = -\log(1-\alpha)$.

3.2 Estimateur à noyau

Nous présentons ici une technique permettant de construire un estimateur non paramétrique de la densité spectrale, l'estimateur à noyau. Cette approche, qui effectue un lissage du périodogramme en fréquence, exploite les propriétés du périodogramme que nous avons mises en évidence dans le

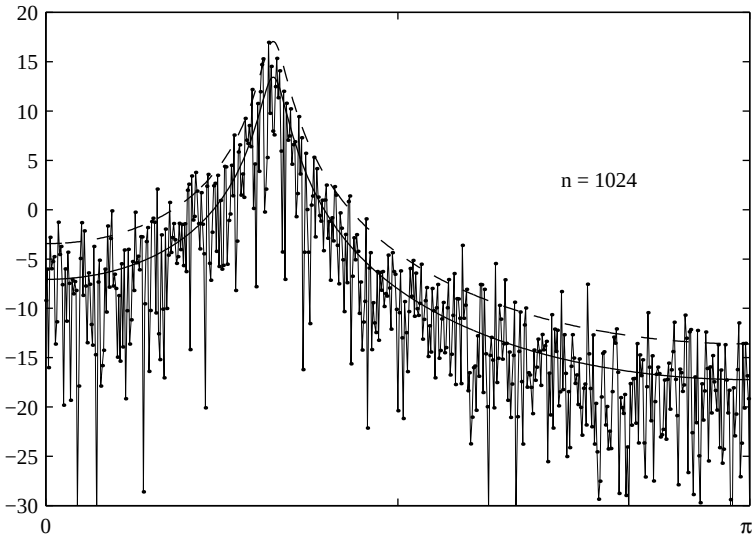

FIG. 3.5 – *Périodogramme en dB pour un AR(2) de paramètres* $[1, -1, 0.9]$ *et* $\sigma^2 = 1$ *calculé sur* $n = 1024$ *échantillons, en fonction de la fréquence* $\lambda \in (0, \pi)$. *La courbe en pointillé donne le seuil de confiance à* 90%.

paragraphe précédent. Nous supposons dans toute cette partie que $\{X_t\}$ est un processus linéaire fort, satisfaisant les conditions d'applications du théorème 3.4.

D'après le théorème 3.4, à la limite des grands échantillons, les coordonnées du périodogramme aux fréquences de Fourier $\lambda_k = 2\pi k/n$ sont des variables décorrélées d'écart type $\sigma^2|\psi(e^{-i\lambda_k})|^2/(2\pi)$. La fonction $\lambda \to |\psi(e^{-i\lambda})|^2$ est continue, elle varie donc "peu" sur de "petits" intervalles de fréquence. Ceci suggère de construire un estimateur de la densité spectrale à la fréquence λ en moyennant les coordonnées du périodogramme aux fréquences de Fourier dans un "voisinage" de la fréquence λ.

Nous appelons un *noyau* une fonction $W : \mathbb{R} \to \mathbb{R}^+$ satisfaisant les propriétés suivantes :

- $W(u) = 0$ pour $|u| > 1$, *i.e.* le noyau a un support compact
- $\int_{-1}^{1} W(u)du = 1$ et $\int_{-1}^{1} uW(u)du = 0$,
- W est deux fois continûment différentiables et $W'(-1) = \lim_{u\to -1^+} W'(u) = 0$ et $W'(1) = \lim_{u\to 1^-} W'(u) = 0$.

Soit $\{b_n\}_{n\geq 0}$ une suite décroissante au sens large de réels positifs, satisfaisant

$$\lim_{n\to\infty} b_n = 0 \, . \tag{3.9}$$

Nous considérons l'estimateur à noyau de la densité spectrale, défini par

$$\hat{f}_n^X(\lambda) = \frac{2\pi}{nb_n} \sum_{k=1}^{n} W\left[b_n^{-1}(\lambda - \lambda_k)\right] I_n^X(\lambda_k) \, . \tag{3.10}$$

Le paramètre b_n est appelé *largeur de bande*, *i.e.* en modifiant b_n nous agissons sur la "largeur" du noyau $b_n^{-1}W(b_n^{-1}\cdot)$. Nous allons, de façon informelle, caractériser la façon dont le paramètre b_n influe

sur la qualité de l'estimateur et essayer de déduire de ce comportement heuristique, des procédures permettant de choisir de manière automatique ce paramètre. Nous allons tout d'abord étudier le biais de cet estimateur, à savoir la différence entre la moyenne de l'estimateur $\mathbb{E}\left[\hat{f}_n^X(\lambda)\right]$ et $f_X(\lambda)$, à une fréquence $\lambda \neq 0, \pi \pmod{2\pi}$ (pour traiter ces valeurs limites, il conviendrait d'utiliser d'autres noyaux). En utilisant le théorème 3.3, nous savons que $\mathbb{E}\left[I_n^X(\lambda_k)\right] = f_X(\lambda_k) + O(n^{-1})$. Par conséquent

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_n^X(\lambda)\right] &= \frac{2\pi}{nb_n} \sum_{k=1}^{n} W\left[b_n^{-1}(\lambda - \lambda_k)\right] f(\lambda_k) + O(n^{-1}) \,, \\
&= \frac{1}{b_n} \int_0^{2\pi} W[b_n^{-1}(\lambda - \mu)] f(\mu) d\mu + O(n^{-1}) \,, \\
&= \int_{-b_n^{-1}(2\pi-\lambda)}^{b_n^{-1}\lambda} W(\nu) f(\lambda + b_n\nu) d\nu \to f_X(\lambda) \,.
\end{aligned}$$

Ceci montre que $\lim_{n\to\infty} \mathbb{E}\left[\hat{f}_n^X(\lambda)\right] = f(\lambda)$, *i.e.* $\hat{f}_{n,b}(\lambda)$ est un estimateur asymptotiquement sans biais de la densité spectrale $f(\lambda)$. Pour comprendre de façon plus précise la façon dont le biais dépend de la largeur de bande b_n, nous supposons dans la suite que la densité spectrale f_X est deux fois continûment différentiable. Nous avons donc, pour tout $\lambda \in [-\pi, \pi]$ et $\nu \in [-1, +1]$,

$$f_X(\lambda + b_n\nu) = f_X(\lambda) + b_n f_X'(\lambda)\nu + \frac{1}{2} b_n^2 f_X''(\lambda)\nu^2 + o(b_n^2)$$

où le terme $o(b_n^2)$ est uniforme en λ et en ν. En utilisant le fait que, pour $\int_{-1}^{+1} \nu W(\nu) d\nu = 0$, nous aurons donc, pour tout n tel que $-b_n^{-1}(2\pi - \lambda) < -1$ et $b_n^{-1}\lambda > 0$,

$$\mathbb{E}\left[\hat{f}_n^X(\lambda)\right] = f_X(\lambda) + \frac{1}{2} b_n^2 f_X''(\lambda) \int_{-1}^{1} \nu^2 W(\nu) d\nu + o(b_n^2), \tag{3.11}$$

ce qui montre que le biais de l'estimateur $\hat{f}_n^X(\lambda)$ est une fonction qui croît comme le carré de la largeur de bande b_n et qui est proportionnelle à la dérivée seconde de la densité spectrale en λ. Notons que comme nous avons supposé que le noyau a exactement un moment nul, $\int_{-1}^{1} \nu W(\nu) d\nu = 0$, le biais ne dépend pas de la dérivée de la densité spectrale $f'(\lambda)$ en λ. Il est facile de voir qu'il est possible de réduire le terme de biais en considérant des noyaux d'ordre supérieur.

Pour comprendre les performances de cet estimateur de la densité spectrale, nous allons évaluer son biais et sa variance. Pour simplifier l'analyse, nous supposerons dans la suite que la fonction $\lambda \to |\psi(e^{-i\lambda})|^2$ est trois fois différentiable sur $[-\pi, \pi]$ et que la dérivée troisième est bornée. En utilisant les résultats du théorème 3.3 nous avons :

$$\mathbb{E}\left[\hat{f}_n^X(\lambda)\right] = \sum_{|k|\leq m} W_{m,n}(k) f_X\left(g(n,\lambda) + 2\pi k/n\right) + O(n^{-1}) \tag{3.12}$$

où $f_X(\lambda) = (2\pi)^{-1}\sigma^2 |\psi(e^{-i\lambda})|^2$ est la densité spectrale du processus $\{X_t\}$. Comme la fonction f_X est deux fois continûment différentiables, nous avons, pour $|k| \leq m$,

$$f_X(g(n,\lambda) + 2\pi k/n) = f_X(g(n,\lambda)) + f_X'(g(n,\lambda)(2\pi k/n) + (1/2) f_X''(g(n,\lambda)(2\pi k/n)^2 + R_{k,m,n}$$

où $R_{k,m,n} \leq c \max |f_X'''(\lambda)|(m/n)^3$ pour $|k| \leq m$. Comme la fenêtre de pondération est symétrique, nous avons $\sum_{|k|\leq m} W_{m,n}(k)k = 0$, ce qui implique en utilisant (**??**)(ii) :

$$\sum_{|k|\leq m} W_{m,n}(k) f_X\left(g(n,\lambda) + 2\pi k/n\right) = f_X(g(n,\lambda)) + (1/2) f_X''(g(n,\lambda))\overline{W}_{m,n} + R_{m,n}$$

$$\text{où} \quad \overline{W}_{m,n} = \frac{4\pi^2}{n^2} \sum_{|k|\leq m} k^2 W_{m,n}(k)$$

et où $|R_{m,n}| \leq c \max |f_X'''(\lambda)|(m/n)^3$. En prenant par exemple la fenêtre de pondération rectangulaire, nous avons $\overline{W}_{m,n} \propto m^2/n^2$ ce qui montre que le biais de l'estimateur varie comme le carré du nombre de points de fréquence pris en compte dans le calcul de la moyenne pondérée. Le calcul de la variance de cet estimateur s'écrit :

$$\mathbb{E}\left[\left(\hat{f}_n^X(\lambda) - \mathbb{E}\left[\hat{f}_{X,n}(\lambda)\right]\right)^2\right] = \widetilde{W}_{m,n} f_X^2(g(n,\lambda)) + Q_{m,n}$$

$$\text{où} \quad \widetilde{W}_{m,n} = \frac{1}{4\pi^2} \sum_{|k|\leq m} W_{m,n}^2(k)$$

et où $|Q_{m,n}| \leq c \max[|f_X'(\lambda)|] \sum_{|k|\leq m} W_{m,n}^2(k)(m/n)$. On voit ici que la troisième des conditions (**??**) assure que la variance tend vers 0 quand n tend vers l'infini. En s'appuyant encore sur l'exemple de la fenêtre rectangulaire, nous avons $\widetilde{W}_{m,n} \propto 1/m$ ce qui montre que la variance de l'estimateur est inversement proportionnelle au nombre de points pris en compte dans le calcul de la moyenne locale. En conclusion dans le cas d'une fenêtre rectangulaire, le paramètre m (qui détermine le nombre de coordonnées de périodogramme moyennées) a un effet *néfaste* pour le biais et *bénéfique* pour la variance de l'estimateur. Le risque quadratique de l'estimateur (qui prend en compte ces deux effets) a pour expression :

$$\mathbb{E}\left[\left(\hat{f}_{X,n}(\lambda) - f_X(\lambda)\right)^2\right] \approx (1/4)\left(f_X''(g(n,\lambda)\overline{W}_{m,n}\right)^2 + \widetilde{W}_{m,n} f_{X,m}^2(g(n,\lambda))$$

Il est naturel de choisir le paramètre m de façon à minimiser l'erreur quadratique moyenne. Dans le cas où $W_{m,n}(k) = 1/(2m+1)$, cette optimisation peut être effectuée de façon explicite. Une autre fenêtre couramment utilisée est la fenêtre triangulaire définie par :

$$W_{m,n}(k) = \begin{cases} \frac{1}{m}\left(1 - \frac{|k|}{m}\right) & \text{pour} \quad |k| \leq m \\ 0 & \text{sinon} \end{cases}$$

Elle vérifie les conditions (**??**) et présente l'avantage d'assurer au spectre estimé d'être positif. Les résultats obtenus avec la fenêtre rectangulaire ont un caractère général : l'utilisation de fenêtre de pondération permet d'obtenir un risque qui tend vers 0 quand n tend vers l'infini. Ce résultat s'accompagne en général d'un biais asymptotiquement non nul. En règle générale, la valeur de m, qui détermine la largeur de la fenêtre, doit tendre vers l'infini, quand $n \to +\infty$, mais suffisamment lentement pour que le rapport n/m tende aussi vers l'infini. Il faut donc ajouter aux conditions (**??**) la condition suivante :

$$m(n) \to \infty \text{ et } m(n)/n \to 0 \text{ quand } n \to \infty$$

Typiquement on aura $m(n) = n^\alpha$ avec $0 < \alpha < 1$.

3.3 Preuves des théorèmes 3.2, 3.3

Théorème 3.2. *Soit $\{Z_t\}$ une suite de variables aléatoires i.i.d., de moyenne nulle et de variance $\sigma^2 < \infty$.*

1. *Soient $0 < \lambda_1 < \ldots < \lambda_m < \pi$, m fréquences fixes. Le vecteur aléatoire $[I_n^Z(\lambda_1), \cdots, I_n^Z(\lambda_m)]$ converge en loi vers un vecteur de variables aléatoires indépendantes, distribuées suivant une loi exponentielle, de moyenne σ^2.*
2. *Supposons que $\mathbb{E}\left[Z_t^4\right] < \infty$, alors :*

$$\mathrm{var}(I_n^Z(\lambda_k)) = \begin{cases} 2\left(\dfrac{\sigma^2}{2\pi}\right)^2 + \kappa_4 n^{-1} & \lambda_k \in \{0, \pi\} \\ \left(\dfrac{\sigma^2}{2\pi}\right)^2 + \kappa_4 n^{-1} & 0 < \lambda_k < \pi \end{cases} \tag{3.13}$$

$$et \qquad 4\pi^2 \mathrm{cov}(I_n^Z(\lambda_j), I_n^Z(\lambda_k)) = \kappa_4 n^{-1} \quad pour \quad \lambda_j \neq \lambda_k \tag{3.14}$$

   où $\lambda_k = 2\pi k/n$ sont les fréquences de Fourier et où κ_4 est le cumulant d'ordre 4 de la variable Z_1 défini par :

$$\kappa_4 = \mathbb{E}\left[Z_1^4\right] - 3(\mathbb{E}\left[Z_1^2\right])^2$$

3. *Supposons que les variables aléatoires Z_t soient gaussiennes. Alors $\kappa_4 = 0$ et, pour tout n, les variables aléatoires $(4\pi/\sigma^2) I_n^Z(\lambda_k)$, $k \in \{1, \cdots, (n-1)/2\}$ sont indépendantes et identiquement distribuées suivant une loi du χ^2 centrée à deux degrés de liberté.*

Démonstration. (i). Notons :

$$\begin{cases} \alpha_n^Z(\lambda_k) = (1/2\pi n)^{-1/2} \sum_{t=1}^n Z_t \cos(\lambda_k t) \\ \beta_n^Z(\lambda_k) = (1/2\pi n)^{-1/2} \sum_{t=1}^n Z_t \sin(\lambda_k t) \end{cases} \tag{3.15}$$

les parties réelles et imaginaire de la transformée de Fourier discrète de $\{Z_t\}$ aux points de fréquences $\lambda_k = 2\pi k/n$. Pour une fréquence arbitraire λ, nous avons :

$$I_n^Z(\lambda) = \frac{1}{2}\left(\alpha_n^Z(g(n,\lambda))^2 + \beta_n^Z(g(n,\lambda))^2\right)$$

Rappelons que si une suite de vecteurs aléatoires Y_n converge en loi vers une variable aléatoire Y et que ϕ est une fonction continue, alors $\phi(Y_n)$ converge en loi vers $\phi(Y)$. Il suffit donc de montrer que le vecteur aléatoire :

$$(\alpha_n^Z(\lambda_1), \beta_n^Z(\lambda_1), \cdots, \alpha_n^Z(\lambda_m), \beta_n^Z(\lambda_m)) \tag{3.16}$$

converge en loi vers une distribution normale de moyenne nulle et de matrice de covariance asymptotique $(\sigma^2/4\pi) I_{2m}$, où I_{2m} est la matrice identité $(2m \times 2m)$. Nous allons tout d'abord nous intéresser au cas $m = 1$. La preuve découle alors du théorème suivant :

Théorème 3.5 (Lindeberg). *Soit $U_{n,t}$, où $t = 1, \ldots, n$ et $n = 1, 2, \ldots$, une suite triangulaire de variables aléatoires centrées de variance finies. Pour tout n, les variables $\{U_{n,1}, \ldots, U_{n,n}\}$ sont indépendantes. On pose $Y_n = \sum_{t=1}^n U_{n,t}$ et $w_n^2 = \sum_{t=1}^n \mathrm{var}(U_{n,t})$. Alors si pour tout $\epsilon > 0$:*

$$\lim_{n\to\infty} \sum_{t=1}^{n} \frac{1}{w_n^2} \mathbb{E}\left[U_{n,t}^2 \mathbf{I}(|U_{n,t}| \geq \epsilon w_n)\right] = 0$$

on a :

$$Y_n / w_n \to_d \mathcal{N}(0,1)$$

Soit u et v deux réels quelconques fixés et $\lambda \in (0, \pi)$. Considérons la variable $Y_n = u\alpha_n^Z(g(n,\lambda)) + v\beta_n^Z(g(n,\lambda))$ que nous pouvons encore écrire :

$$Y_n = \sum_{t=1}^{n} U_{n,t} \quad \text{où} \quad U_{n,t} = \frac{1}{\sqrt{2\pi n}}(u\cos(g(n,\lambda)t) + v\sin(g(n,\lambda)t))Z_t$$

Notons que, pour n fixé les variables aléatoires $\{U_{n,t}\}$ sont indépendantes. D'autre part, pour tout $\lambda \neq 0$, on vérifie aisément que :

$$\sum_{t=1}^{n} \cos^2(g(n,\lambda)t) = \sum_{t=1}^{n} \sin^2(g(n,\lambda)t) = \frac{n}{2} \quad \text{et} \quad \sum_{t=1}^{n} \cos((g(n,\lambda)t)\sin(g(n,\lambda)t) = 0$$

Par suite, on peut écrire que :

$$\begin{aligned} w_n^2 = \textstyle\sum_{t=1}^n \mathrm{var}(U_{n,t}) &= \frac{1}{2\pi n} \sum_{t=1}^{n} (u^2\cos^2(g(n,\lambda)t) + v^2\sin^2(g(n,\lambda)t) + 2uv\cos((g(n,\lambda)t)\sin(g(n,\lambda)t))) \\ &= \frac{1}{4\pi}(u^2+v^2) = w_1^2 \end{aligned}$$

Par suite, en posant $c_0 = (|u| + |v|)/2\pi w_1$ et $\epsilon' = \epsilon\sqrt{2\pi} w_1/(|u| + |v|)$, on a :

$$\sum_{t=1}^{n} \frac{1}{w_n^2} \mathbb{E}\left[U_{n,t}^2 \mathbf{I}(|U_{n,t}| \geq \epsilon w_n)\right] \leq \frac{c_0}{n} \sum_{t=1}^{n} \mathbb{E}\left[Z_t^2 \mathbf{I}(|Z_t| \geq \epsilon'\sqrt{n})\right] = c_0 \mathbb{E}\left[Z_1^2 \mathbf{I}(|Z_1| \geq \epsilon'\sqrt{n})\right]$$

Le dernier terme tend vers 0 puisque on a $\mathbb{E}\left[Z_1^2 \mathbf{I}(|Z_1| \geq \epsilon'\sqrt{n})\right] \leq \mathbb{E}\left[|Z_1|^3\right]/\epsilon'\sqrt{n}$ et que $\mathbb{E}\left[|Z_1|^3\right] < \infty$ puisque $\mathbb{E}\left[|Z_1|^4\right] < \infty$. La preuve s'étend aisément à un ensemble de fréquences $\lambda_1, \ldots, \lambda_m$ en utilisant la méthode de Cramer-Wold que nous rappelons :

Proposition 3.1 (Cramér-Wold). *Soit $\{V_n\}_{n\geq 0}$ une suite de vecteurs aléatoires réels de dimension m. $V_n \to_d W$ si et seulement si, pour toute suite $\{\lambda_1, \cdots, \lambda_m\} \in \mathbb{R}^m$, la variable aléatoire $Y_n = \lambda_1 V_{n,1} + \cdots + \lambda_m V_{n,m} \to_d \lambda_1 W_1 + \cdots + \lambda_m W_m$.*

(ii). Par définition de $I_n^Z(\lambda_k)$, nous avons au premier ordre :

$$\mathbb{E}\left[I_n^Z(\lambda_k)\right] = (2\pi n)^{-1} \sum_{s,t=1}^{n} \mathbb{E}\left[Z_s Z_t\right] e^{i\lambda_k(t-s)} = (2\pi)^{-1}\sigma^2 \tag{3.17}$$

Au second ordre nous avons :

$$\mathbb{E}\left[I_n^Z(\lambda_j)I_n^Z(\lambda_k)\right] = (2\pi n)^{-2}\sum_{s,t,u,v=1}^{n}\mathbb{E}\left[Z_sZ_tZ_uZ_v\right]e^{i(\lambda_j(t-s)+\lambda_k(v-u))} \tag{3.18}$$

En utilisant que les variables aléatoires Z_t sont indépendantes, centrées, de même variance σ^2 et de moment d'ordre 4 fini et en posant $\mathbb{E}\left[Z_1^4\right] = \kappa_4 + 3\sigma^4$, on obtient :

$$\mathbb{E}\left[Z_sZ_tZ_uZ_v\right] = \kappa_4\delta_{s,t,u,v} + \sigma^4(\delta_{s,t}\delta_{u,v} + \delta_{s,u}\delta_{t,v} + \delta_{s,v}\delta_{t,u}) \tag{3.19}$$

En portant cette expression dans (3.18), nous avons :

$$\mathbb{E}\left[I_n^Z(\lambda_j)I_n^Z(\lambda_k)\right] = (2\pi)^{-2}n^{-1}\kappa_4 + (2\pi)^{-2}n^{-2}\sigma^4\left(n^2 + \left|\sum_{t=1}^{n}e^{i(\lambda_j+\lambda_k)t}\right|^2 + \left|\sum_{t=1}^{n}e^{i(\lambda_k-\lambda_j)t}\right|^2\right)$$

et donc :

$$\begin{aligned}\operatorname{cov}(I_n^Z(\lambda_j), I_n^Z(\lambda_k)) &= \mathbb{E}\left[I_n^Z(\lambda_j)I_n^Z(\lambda_k)\right] - \mathbb{E}\left[I_n^Z(\lambda_j)\right]\mathbb{E}\left[I_n^Z(\lambda_k)\right] \\ &= (2\pi)^{-2}n^{-1}\kappa_4 + (2\pi)^{-2}n^{-2}\sigma^4\left(\left|\sum_{t=1}^{n}e^{i(\lambda_j+\lambda_k)t}\right|^2 + \left|\sum_{t=1}^{n}e^{i(\lambda_k-\lambda_j)t}\right|^2\right)\end{aligned}$$

ce qui permet de conclure.

(iii). Lorsque $\{Z_t\}$ est une variable gaussienne centrée, le vecteur :

$$Q_n = \begin{bmatrix}\alpha_n^Z(\lambda_1) & \beta_n^Z(\lambda_1) & \cdots & \alpha_n^Z(\lambda_{\tilde{n}}) & \beta_n^Z(\lambda_{\tilde{n}})\end{bmatrix}$$

est gaussien comme transformée linéaire d'un vecteur gaussien. Il suffit donc de calculer le vecteur-moyenne et sa matrice de covariance. Il est facile de vérifier que le vecteur-moyenne est nul et que, pour $0 < \lambda_k \neq \lambda_j < \pi$, nous avons :

$$\begin{aligned}&\mathbb{E}\left[(\alpha_n^Z(\lambda_k))^2\right] = \mathbb{E}\left[(\beta_n^Z(\lambda_k))^2\right] = (4\pi)^{-1} \\ &\mathbb{E}\left[\alpha_n^Z(\lambda_k)\beta_n^Z(\lambda_k)\right] = 0 \\ &\mathbb{E}\left[\alpha_n^Z(\lambda_k)\alpha_n^Z(\lambda_j)\right] = \mathbb{E}\left[\beta_n^Z(\lambda_k)\beta_n^Z(\lambda_j)\right] = 0 \\ &\mathbb{E}\left[\alpha_n^Z(\lambda_k)\beta_n^Z(\lambda_j)\right] = 0\end{aligned}$$

La matrice de covariance est donc $\sigma^2 I_{\tilde{n}}/4\pi$ où $I_{\tilde{n}}$ est la matrice identité de taille $\tilde{n}$. Par conséquent les composantes de Q_n sont indépendantes. Rappelons que :

$$I_n^Z(\lambda_k) = (\alpha_n^Z(\lambda_k))^2 + (\beta_n^Z(\lambda_k))^2$$

De l'indépendance des composantes de Q_n, on déduit que les variables aléatoires $I_n^Z(\lambda_k)$ sont elles-même indépendantes et que $4\pi I_n^Z(\lambda_k)/\sigma^2$ est la somme du carré de deux variables gaussiennes centrées, indépendantes, de même variance 1, dont la distribution de probabilité est la loi dite du χ^2 à deux degrés de liberté. Ce qui conclut la preuve.

■

Théorème 3.3. *Soit $\{X_t\}$ un processus linéaire. Supposons que $\sum_{j=-\infty}^{\infty} |\psi_j||j|^{1/2} < \infty$ et que $\mathbb{E}\left[Z_t^4\right] < \infty$. On a alors :*

$$I_n^X(\lambda_k) = |\psi(e^{-i\lambda_k})|^2 I_n^Z(\lambda_k) + R_n(\lambda_k)$$

où le terme $R_n(\lambda_k)$ vérifie :

$$\max_{k\in\{1,\cdots,\lfloor(n-1)/2\rfloor\}} \mathbb{E}\left[|R_n(\lambda_k)|^2\right] = O(n^{-1})^3$$

Démonstration. Notons respectivement $d_n^X(\lambda_k)$ et $d_n^Z(\lambda_k)$ les transformées de Fourier discrètes des suites $\{X_1, \cdots, X_n\}$ et de $\{Z_1, \cdots, Z_n\}$ au point de fréquence $2\pi k/n$ avec $k \in \{1, \ldots, \lfloor(n-1)/2\rfloor$. Nous pouvons écrire successivement :

$$\begin{aligned}
d_n^X(\lambda_k) &= (2\pi n)^{-1/2} \sum_{t=1}^{n} X_t e^{-i\lambda_k t} \\
&= (2\pi n)^{-1/2} \sum_{j=-\infty}^{\infty} \psi_j e^{-i\lambda_k j} \left(\sum_{t=1}^{n} Z_{t-j} e^{-i\lambda_k (t-j)} \right) \\
&= (2\pi n)^{-1/2} \sum_{j=-\infty}^{\infty} \psi_j e^{-i\lambda_k j} \left(\sum_{t=1-j}^{n-j} Z_t e^{-i\lambda_k t} \right) \\
&= (2\pi n)^{-1/2} \sum_{j=-\infty}^{\infty} \psi_j e^{-i\lambda_k j} \left(\sum_{t=1}^{n} Z_t e^{-i\lambda_k t} + U_{n,j}(\lambda_k) \right) \\
&= \psi(e^{-i\lambda_k}) d_n^Z(\lambda_k) + Y_n(\lambda_k)
\end{aligned}$$

où nous avons posé :

$$U_{n,j}(\lambda_k) = \sum_{t=1-j}^{n-j} Z_t e^{-i\lambda_k t} - \sum_{t=1}^{n} Z_t e^{-i\lambda_k t} \tag{3.20}$$

$$\text{et } Y_n(\lambda_k) = (2\pi n)^{-1/2} \sum_{j=-\infty}^{\infty} \psi_j e^{-i\lambda_k j} U_{n,j}(\lambda_k) \tag{3.21}$$

On remarque que, pour $|j| < n$, $U_{n,j}(\lambda_k)$ est une somme de $2|j|$ variables indépendantes centrées de variance σ^2 tandis que, pour $|j| \geq n$, $U_{n,j}(\lambda_k)$ est la somme de $2n$ variables centrées indépendantes de variance σ^2. Par conséquent, partant de (3.20), on a :

$$\mathbb{E}\left[|U_{n,j}(\lambda_k)|^2\right] \leq 2\sigma^2 \min(|j|, n) \tag{3.22}$$

ainsi que :

$$\mathbb{E}\left[|U_{n,j}(\lambda_k)|^4\right] \leq C_R \sigma^4 (\min(|j|, n))^2 \tag{3.23}$$

où $C_R < \infty$ est une constante. Pour montrer (3.23), il suffit de poser $\mathbb{E}\left[Z_t^4\right] = \eta\sigma^4$ et d'utiliser l'inégalité (3.24) pour $p = 4$.

[3] Notation : quand $n \to \infty$, $O(n^{-\alpha})/n^{-\alpha} \to c \neq 0$ tandis que $o(n^{-\alpha})/n^{-\alpha} \to 0$.

Propriété 3.1 (Inégalité de Rosenthal (Petrov, 1985)). *Soient $(X_1, \ldots, X_n)$ des variables indépendantes (mais pas nécessairement identiquement distribuées) et soit $p \geq 2$. Alors il existe une constante universelle $C(p) < \infty$ telle que :*

$$\mathbb{E}\left[\left|\sum_{k=1}^{n} X_k\right|^p\right] \leq C(p)\left(\left(\sum_{k=1}^{n}\mathbb{E}\left[X_k^2\right]\right)^{p/2} + \sum_{k=1}^{n}\mathbb{E}\left[|X_k|^p\right]\right) \tag{3.24}$$

Utilisons à présent (3.23) pour majorer $\mathbb{E}\left[|Y_n(\lambda_k)|^4\right]$. En adoptant la notation $\|X\|_p = (\mathbb{E}\left[|X|^p\right])^{1/p}$ (pour $p > 0$) on a, d'après l'inégalité triangulaire (inégalité de Minkovski) $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$:

$$\sup_{k\in\{1,\cdots,\lfloor(n-1)/2\rfloor\}} \|Y_n(\lambda_k)\|_4 \leq \sup_{k\in\{1,\cdots,\lfloor(n-1)/2\rfloor\}} (2\pi n)^{-1/2} \sum_{j=-\infty}^{\infty} |\psi_j| \|U_{n,j}(\lambda_k)\|_4$$

D'après (3.23), $\|U_{n,j}(\lambda_k)\|_4 \leq c\sigma \min(|j|, n)^{1/2}$. Par conséquent :

$$\sup_{k\in\{1,\cdots,\lfloor(n-1)/2\rfloor\}} \|Y_n(\lambda_k)\|_4 \leq c\sigma(2\pi n)^{-1/2} \sum_{j=-\infty}^{\infty} |\psi_j| \min(|j|, n)^{1/2}$$

Maintenant on peut écrire :

$$\sum_{j=-\infty}^{\infty} |\psi_j| \min(|j|, n)^{1/2} \leq \sum_{j=-\infty}^{\infty} |\psi_j||j|^{1/2}$$

Par conséquent $\|Y_n(\lambda_k)\|_4$ est d'un ordre égal à $O(n^{-1/2})$.
Nous pouvons à présent exprimer $R_n(\lambda_k) = I_n^X(\lambda_k) - |\psi(e^{-i\lambda_k})|^2 I_n^Z(\lambda_k)$ en fonction de $Y_n(\lambda_k) = d_n^X(\lambda_k) - \psi(e^{-i\lambda_k}) d_n^Z(\lambda_k)$. Il vient :

$$\begin{aligned} R_n(\lambda_k) &= |\psi(e^{-i\lambda_k}) d_n^Z(\lambda_k) + Y_n(\lambda_k)|^2 - |\psi(e^{-i\lambda_k})|^2 I_n^Z(\lambda_k) \\ &= \psi(e^{-i\lambda_k}) d_n^Z(\lambda_k) Y_n(-\lambda_k) + \psi(e^{i\lambda_k}) d_n^Z(-\lambda_k) Y_n(\lambda_k) + |Y_n(\lambda_k)|^2 \end{aligned}$$

D'après l'inégalité de Hölder, $\|XY\|_r \leq \|X\|_p \|Y\|_q$ si $p^{-1} + q^{-1} = r^{-1}$. En faisant $p = q = 4$ et $r = 2$, il vient :

$$(\mathbb{E}\left[|R_n(\lambda_k)|^2\right])^{1/2} = \|R_n(\lambda_k)\|_2 \leq 2\sum_j |\psi_j| \|d_n^Z(\lambda_k)\|_4 \|Y_n(\lambda_k)\|_4 + \|Y_n(\lambda_k)\|_4$$

D'après le théorème 3.2, $\|d_n^Z(\lambda_k)\|_4$ est de l'ordre de $\sigma/\sqrt{2\pi}$. Par conséquent $\|R_n(\lambda_k)\|_2$ est de l'ordre de $n^{-1/2}$ et $\mathbb{E}\left[|R_n(\lambda_k)|^2\right] = \|R_n(\lambda_k)\|_2^2$ de l'ordre de $1/n$. Ce qui conclut la preuve. ■

Chapitre 4

Prédiction linéaire. Décomposition de Wold

4.1 Eléments de géométrie Hilbertienne

Définition 4.1 (Espace pré-hilbertien)**.** *Soit $\mathcal{H}$ un espace vectoriel sur l'ensemble des nombres complexes $\mathbb{C}$. L'espace $\mathcal{H}$ est appelé* pré-hilbertien *si $\mathcal{H}$ est muni d'un produit scalaire :*

$$(\boldsymbol{.},\boldsymbol{.}) : x, y \in \mathcal{H} \times \mathcal{H} \mapsto (x, y) \in \mathbb{R}$$

qui vérifie les propriétés suivantes :

(i). $(x, y) = (y, x)^*$

(ii). $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$

(iii). $(x, x) \geq 0$*, l'égalité ayant lieu si et seulement si* $x = 0$*.*

L'application :

$$\| \boldsymbol{.} \| : x \in \mathcal{H} \mapsto \sqrt{(x, x)} \geq 0$$

définit une norme pour tout vecteur x.

Exemple 4.1 : Espace $\mathbb{R}^n$
L'ensemble des vecteurs colonnes $x = [x_1 \quad \cdots \quad x_n]^T$, où $x_k \in \mathbb{R}$, est un espace vectoriel dans lequel la relation :

$$(x, y) = \sum_{k=1}^{n} x_k y_k$$

définit par un produit scalaire.

Exemple 4.2 : Espace $l^2(\mathbb{Z})$
L'ensemble des suites numériques complexes $\{x_k\}_{k\in\mathbb{Z}}$ vérifiant $\sum_{k=-\infty}^{\infty} |x_k|^2 < \infty$ est un espace vectoriel sur $\mathbb{C}$. On munit cet espace du produit intérieur :

$$(x, y) = \sum_{k=-\infty}^{\infty} x_k y_k^* \leq (1/2) \sum_{k=-\infty}^{\infty} (|x_k|^2 + |y_k|^2) < \infty$$

On vérifie aisément les propriétés (i-iii) de la définition 4.1. L'espace ainsi défini est donc un espace pré-Hilbertien, que l'on note $l^2(\mathbb{Z})$.

Exemple 4.3 : Fonctions de carré intégrable
L'ensemble $\mathcal{H}$ des fonctions boréliennes définies sur un intervalle T de $\mathbb{R}$, à valeurs complexes et de carré intégrable par rapport à la mesure de Lebesgue ($f \in \mathcal{H} : \int_T |f(t)|^2 dt < \infty$) est un espace vectoriel. Considérons alors le produit intérieur :

$$(f,g) \in \mathcal{H} \times \mathcal{H} \mapsto \int_T f(t)g^*(t)dt$$

On montre aisément que $(f,g) < \infty$ ainsi que les propriétés (i) et (ii) de la définition 4.1. Par contre la propriété (iii) nést pas vérifiée puisque :

$$(f,f) = 0 \not\Rightarrow \forall t \in T \ f(t) = 0$$

En effet une fonction f qui est nulle sauf sur un ensemble de mesure nulle pour la mesure de Lebesgue, vérifie $(f,f) = 0$. L'space $\mathcal{H}$ muni du produit (f,g) nést donc pas un espace pré-Hilbertienne. Nous verrons dans la suite qu'il est possible de lever cette difficulté en considérant les classes d'équivalence des fonctions égales presque partout.

On montre aisément les propriétés suivantes :

Théorème 4.1. *Pour tout $x, y \in \mathcal{H} \times \mathcal{H}$, nous avons :*

(i). Inégalité de Cauchy-Schwarz : $|(x,y)| \leq \|x\|\|y\|$,

(ii). Inégalité triangulaire : $|\|x\| - \|y\|| \leq \|x-y\| \leq \|x\| + \|y\|$,

(iii). Identité du parallélogramme :

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$

Définition 4.2 (Convergence dans $\mathcal{H}$)**.** *Soit x_n une suite de vecteurs et x un vecteur d'un espace $\mathcal{H}$ muni d'un produit scalaire. On dit que x_n tend vers x si et seulement si $\|x_n - x\| \to 0$ quand $n \to +\infty$. On note $x_n \to x$.*

Propriété 4.1. *Si dans un espace de Hilbert la suite $x_n \to x$, alors x_n est bornée.*

Démonstration. D'après l'inégalité triangulaire, on a :

$$\|x_n\| = \|(x_n - x) + x\| \leq \|x_n - x\| + \|x\|$$

■

Proposition 4.1 (Continuité du produit scalaire)**.** *Soit $x_n \to x$ et $y_n \to y$ deux suites convergentes de vecteurs d'un espace pré-hilbertien $\mathcal{H}$. Alors quand $n \to +\infty$: $(x_n, y_n) \to (x,y)$. En particulier, si $x_n \to x$, $\|x_n\| \to \|x\|$.*

Démonstration. D'après l'inégalité triangulaire puis l'inégalité de Schwarz, nous avons :

$$\begin{aligned}(x,y) - (x_n,y_n) &= ((x-x_n) + x_n, (y-y_n) + y_n) - (x_n,y_n) \\ &= (x-x_n, y-y_n) + (x-x_n, y_n) + (x_n, y-y_n) \\ &\leq \|x_n - x\|\|y_n - y\| + \|x_n - x\|\|y_n\| + \|y_n - x\|\|x_n\|\end{aligned}$$

Il suffit ensuite d'évoquer la convergence et la bornitude des suites x_n et y_n. ■

Définition 4.3 (Suite de Cauchy)**.** *Soit x_n une suite de vecteurs d'un espace pré-hilbertien $\mathcal{H}$. On dit que x_n est une suite de Cauchy si et seulement si :*

$$\|x_n - x_m\| \to 0$$

quand $n, m \to +\infty$.

Notons quén vertu de l'inégalité triangulaire toute suite convergente est une suite de Cauchy. La réciproque est fausse : une suite de Cauchy peut ne pas être convergente. En voici un contre-exemple :

Exemple 4.4 : Suite de Cauchy non convergente
Soit $\mathcal{C}([-\pi,\pi])$ l'espace des fonctions continues sur $[-\pi,\pi]$. L'espace $\mathcal{C}([-\pi,\pi])$, muni du produit $\int_{-\pi}^{\pi} f(x)g^*(x)dx$, est un espace pré-hilbertien. Considérons la suite de fonctions :

$$f_n(x) = \sum_{k=1}^{n} \frac{1}{k}\cos(kx)$$

Les fonctions $f_n(x)$, qui sont indéfiniment continûment différentiables, appartiennent à $\mathcal{C}(-\pi,\pi)$. Montrons que cette suite est une suite de Cauchy. En effet, pour $m > n$, on a :

$$\|f_n - f_m\|^2 = \pi \sum_{k=n+1}^{m} \frac{1}{k^2} \longrightarrow 0 \quad \textit{quand} \quad (n,m) \to \infty$$

D'autre part on montre aisément que la limite de cette suite $f_\infty(x) = \sum_{k=1}^{\infty} k^{-1}\cos(kx) = \log|\sin(x/2)|$ nést pas continue et n'appartient donc pas à $\mathcal{C}([-\pi,\pi])$.

Définition 4.4 (Espace de Hilbert)**.** *On dit qu'un espace vectoriel est complet si toute suite de suite de Cauchy de $\mathcal{H}$ converge dans $\mathcal{H}$. On dit $\mathcal{H}$ est un espace de Hilbert si $\mathcal{H}$ est pré-hilbertien et complet.*

Proposition 4.2 ($L^2([-\pi,\pi],dx)$)**.** *L'espace des fonctions de carré intégrable pour la mesure de Lebesgue, définie sur l'intervalle $[-\pi,\pi]$ muni de sa tribu de Borel $\mathcal{B}([-\pi,\pi])$, est un espace de Hilbert.*

Définition 4.5 (Sous espace vectoriel)**.** *Un sous-espace $\mathcal{E}$ d'un espace vectoriel $\mathcal{H}$ est un sous-ensemble de $\mathcal{H}$ tel que, pour tout $x, y \in \mathcal{E}$ et tout scalaire α, β, $\alpha x + \beta y \in \mathcal{E}$. Un sous-espace vectoriel est dit* propre *si $\mathcal{E} \neq \mathcal{H}$.*

Définition 4.6 (Sous-espace fermé)**.** *Soit $\mathcal{E}$ un sous-espace d'un espace de Hilbert $\mathcal{H}$. On dit que $\mathcal{E}$ est fermé, si toute suite $\{x_n\}$ de $\mathcal{E}$, qui converge, converge dans $\mathcal{E}$.*

Exemple 4.5 : Contre-exemple
Soit $L^2([-\pi,\pi],dx)$ l'espace de Hilbert des fonctions de carré intégrable pour la mesure de Lebesgue sur $[-\pi,\pi]$. Comme le montre l'exemple 4.4, l'ensemble des fonctions continues sur $[-\pi,\pi]$ est un sous-espace vectoriel de $L^2([-\pi,\pi],dx)$ mais nést pas fermé.

Définition 4.7 (Sous espace engendré par un sous-ensembe)**.** *Soit $\mathcal{X}$ un sous-ensemble de $\mathcal{H}$. Nous notons* $\mathrm{span}\{\mathcal{X}\}$ *le sous-espace vectoriel des combinaisons linéaires finies d'éléments de $\mathcal{X}$ et* $\overline{\mathrm{span}}\{\mathcal{X}\}$ *la fermeture de* $\mathrm{span}(\mathcal{X})$ *dans $\mathcal{H}$.*

Définition 4.8 (Orthogonalité)**.** *Deux vecteurs $x, y \in \mathcal{H}$ sont dit* orthogonaux*, si $(x,y) = 0$, ce que nous notons $x \perp y$. Si $\mathcal{S}$ est un sous-ensemble de $\mathcal{H}$, la notation $x \perp \mathcal{S}$, signifie que $x \perp s$ pour tout $s \in \mathcal{S}$. Nous notons $\mathcal{S} \perp \mathcal{T}$ si tout élément de $\mathcal{S}$ est orthogonal à tout élément de $\mathcal{T}$.*

Supposons qu'il existe deux sous-espaces $\mathcal{A}$ et $\mathcal{B}$ tels que $\mathcal{H} = \mathcal{A} + \mathcal{B}$, dans le sens où, pour tout vecteur $h \in \mathcal{H}$, il existe $a \in \mathcal{A}$ et $b \in \mathcal{B}$, tel que $h = a + b$. Si en plus $\mathcal{A} \perp \mathcal{B}$ nous dirons que $\mathcal{H}$ est la *somme directe* de $\mathcal{A}$ et $\mathcal{B}$, ce que nous notons $\mathcal{H} = \mathcal{A} \oplus \mathcal{B}$.

Définition 4.9 (Complément orthogonal)**.** *Soit $\mathcal{E}$ un sous-ensemble d'un espace de Hilbert $\mathcal{H}$. On appelle ensemble orthogonal de $\mathcal{E}$, l'ensemble défini par :*

$$\mathcal{E}^{\perp} = \{x \in \mathcal{H} : \forall y \in \mathcal{E} \ \ (x, y) = 0\}$$

Le théorème suivant, appelé *théorème de projection*, joue un rôle central en analyse Hilbertienne. Nous en donnons une démonstration complète en fin de chapitre, et nous encourageons le lecteur à s'arrêter sur cette démonstration pour comprendre l'essence de la construction.

Théorème 4.2 (De projection)**.** *Soit $\mathcal{E}$ est un sous-espace fermé d'un espace de Hilbert $\mathcal{H}$ et soit x un élément quelconque de $\mathcal{H}$, alors :*

(i). il existe un unique élément noté $(x|\mathcal{E}) \in \mathcal{E}$ tel que :

$$\|x - (x|\mathcal{E})\| = \inf_{w \in \mathcal{E}} \|x - w\|$$

(ii). $(x|\mathcal{E}) \in \mathcal{E}$ et $\|x - (x|\mathcal{E})\| = \inf_{w\in\mathcal{E}} \|x - w\|$ si et seulement si $(x|\mathcal{E}) \in \mathcal{E}$ et $x - (x|\mathcal{E}) \perp \mathcal{E}$.

Démonstration. Elle est donnée en fin de chapitre. ■

Proposition 4.3. *Soit $\mathcal{H}$ un espace de Hilbert et $(.|\mathcal{E})$ la projection orthogonale sur le sous-espace fermé $\mathcal{E}$. On a :*

1. *l'application $x \in \mathcal{H} \mapsto (x|\mathcal{E}) \in \mathcal{E}$ est linéaire :*

$$\forall \alpha, \beta \in \mathbb{C}, \quad (\alpha x + \beta y|\mathcal{E}) = \alpha(x|\mathcal{E}) + \beta(y|\mathcal{E})$$

2. $\|x\|^2 = \|(x|\mathcal{E})\|^2 + \|x - (x|\mathcal{E})\|^2$ *(Pythagore),*
3. *La fonction $(.|\mathcal{E}) : \mathcal{H} \to \mathcal{H}$ est continue,*
4. *$x \in \mathcal{E}$ si et seulement si $(x|\mathcal{E}) = x$,*
5. *$x \in \mathcal{E}^{\perp}$ si et seulement si $(x|\mathcal{E}) = 0$,*
6. *Soient $\mathcal{E}_1$ et $\mathcal{E}_2$ deux sous espaces vectoriels fermés de $\mathcal{H}$, tels que $\mathcal{E}_1 \subset \mathcal{E}_2$. Alors :*

$$\forall x \in \mathcal{H}, \qquad ((x|\mathcal{E}_2)|\mathcal{E}_1) = (x|\mathcal{E}_1)$$

7. *Soient $\mathcal{E}_1$ et $\mathcal{E}_2$ deux sous-espaces vectoriels fermés de $\mathcal{H}$, tels que $\mathcal{E}_1 \perp \mathcal{E}_2$. Alors :*

$$\forall x \in \mathcal{H}, \qquad (x|\mathcal{E}_1 \oplus \mathcal{E}_2) = (x|\mathcal{E}_1) + (x|\mathcal{E}_2).$$

Exemple 4.6 : Projection sur un vecteur
Soit $\mathcal{H}$ un espace de Hilbert, $\mathcal{C} = \text{span}\{v\}$ le sous-espace engendré par un vecteur $v \in \mathcal{H}$ et x un vecteur quelconque de $\mathcal{H}$. On a alors $(x|\mathcal{C}) = \alpha v$ avec $\alpha = (x, v)/\|v\|^2$. Si on note $\epsilon = x - (x|\mathcal{C})$, on a :

$$\|\epsilon\|^2 = \|x\|^2 \left(1 - \|\rho\|^2\right) \quad \text{où} \quad \rho = \frac{(x, v)}{\|x\|\|v\|} \quad \text{avec} \quad |\rho| \leq 1$$

Appliquons ce résultat à $\mathcal{H} = \mathbb{C}^n$ et au vecteur $v(\lambda_0)$ de composantes $v_t = n^{-1/2}e^{i\lambda_0 t}$ où $t \in \{1, \ldots, n\}$ et où la pulsation de Fourier $\lambda_0 \in (-\pi, \pi)$. On vérifie que $\|v(\lambda_0)\| = 1$. Soit $x = (x_1, \ldots, x_n)^T$ un vecteur quelconque de $\mathbb{C}^n$. La projection orthogonale de x sur $\text{span}\{v(\lambda_0)\}$ s'écrit $\alpha v(\lambda_0)$ avec :

$$\alpha = \sum_{t=1}^{n} x_t v_t^* = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-i\lambda_0 t}$$

qui est la transformée de Fourier à temps discret de la suite x_t calculée précisément à la pulsation λ_0.

Exemple 4.7 : Droite de régression
On est parfois conduit à chercher une relation linéaire entre deux suites de valeurs $\{x_t\}_{1\leq t\leq n}$ et $\{y_t\}_{1\leq t\leq n}$. Cela revient à trouver la suite $\hat{y}_t = \alpha_1 + \alpha_2 x_t$ qui s'approche quadratiquement au plus près de la suite y_t. D'après le théorème de projection, il suffit déxprimer que le vecteur $\hat{y} \in \mathbb{R}^n$ de composantes $\hat{y}_n$ est la projection orthogonale de $y = (y_1, \ldots, y_n)^T$ sur $\mathcal{E} = \text{span}\{u, x\}$ où $u = (1, \ldots, 1)^T$ et $x = (x_1, \ldots, x_n)^T$. Par conséquent α_1 et α_2 sont solutions du système de deux équations :

$$(y - (\alpha_1 + \alpha_2 x), 1) = 0 \quad et \quad (y - (\alpha_1 + \alpha_2 x), x) = 0$$

qui s'écrit encore :

$$\begin{bmatrix} n & \sum_t x_t \\ \sum_t x_t & \sum_t x_t^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \sum_t y_t \\ \sum_t x_t y_t \end{bmatrix}$$

Si la matrice est inversible la solution est unique.

Exemple 4.8 : Modèle linéaire et méthode des moindres carrés
On considère, pour $1 \leq t \leq n$, la suite d'observations :

$$x_t = \sum_{k=1}^{P} a_{t,k}\theta_k + z_t$$

où $\{a_{t,k}\}$, avec $1 \leq k \leq P$, $1 \leq t \leq n$ et $n > P$, sont des valeurs connues. $\{\theta_k\}$ est une suite de paramètres à estimer et z_t est un terme d'incertitude qui modélise par exemple des erreurs de mesure. Avec des notations matricielles évidentes on peut écrire $X = A\theta + Z$. On note $\mathcal{A}$ le sous-espace de $\mathbb{R}^n$ engendré par les colonnes de A. L'estimation, dite des moindres carrés, consiste à trouver θ qui minimise $\sum_{t=1}^{n} z_t^2$. Ce problème peut alors se formaliser de la façon suivante : déterminer le vecteur de $\mathcal{A}$ le plus proche de X. La solution est la projection orthogonale $(X|\mathcal{A})$ qui, d'après le point (ii) du théorème de projection, vérifie :

$$A^T(X - (X|\mathcal{A})) = 0 \Leftrightarrow A^T(X|\mathcal{A}) = A^T X$$

On sait que le vecteur $(X|\mathcal{A})$ est unique. Par contre la résolution, par rapport à θ, de l'équation $(X|\mathcal{A}) = A\theta$ n'a pas nécessairement une solution unique. Elle dépend du rang de la matrice A.

- *Si A est de rang plein P, A^TA est inversible et $\theta = (A^TA)^{-1}A^TX$ qui est alors unique.*
- *Si A est de rang strictement inférieur à P, alors il existe une infinité de valeurs de θ telle que $A^TA\theta = A^TX$. Elles diffèrent toutes par un vecteur u de l'espace nul de A défini par $Au = 0$.*

4.2 Espace des variables aléatoires de carré intégrables

Les espaces de Hilbert donnent un cadre théorique pratique pour l'analyse des processus du second-ordre. Soit $\{\Omega, \mathcal{F}, \mathbb{P}\}$ un espace de probabilité. Considérons $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ l'espace des variables aléatoires réelles, de carré intégrable sur $\{\Omega, \mathcal{F}, \mathbb{P}\}$, cést à dire toutes les variables aléatoires réelles vérifiant

$\mathbb{E}\left[X^2\right] < \infty$. Il est facile de vérifier que si X et Y sont deux éléments de $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ alors, pour tout $\alpha, \beta \in \mathbb{R}$, nous avons $\alpha X + \beta Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, et que $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ est un espace vectoriel sur $\mathbb{R}$. Considérons alors le produit intérieur défini par :

$$X, Y \in \mathcal{L}^2 \times \mathcal{L}^2 \mapsto \mathbb{E}\left[XY\right] = \int_\Omega X(\omega)Y(\omega)\mathbb{P}(d\omega)$$

ainsi que la forme positive :

$$X \in \mathcal{L}^2 \mapsto \mathbb{E}^{1/2}\left[X^2\right] \geq 0$$

Bien que cette forme soit positive et vérifie l'inégalité triangulaire, ce nést pas une norme, car la relation $\mathbb{E}\left[X^2\right] = 0$ implique seulement que $X = 0$ $\mathbb{P}$-p.s. (voir annexe A.1.3), et donc que X peut être différent de 0 sur un sous-ensemble de Ω de mesure nulle pour $\mathbb{P}$. Pour lever cette difficulté, considérons dans $\mathcal{L}^2$ la relation d'égalité presque sûre définie par :

$$X = Y\ (\mathbb{P}\text{-p.s.}) \Leftrightarrow \mathbb{P}\{\omega \in \Omega : X(\omega) \neq Y(\omega)\} = 0$$

On vérifie aisément que cette relation est réflexive, symétrique et transitive, ce qui définit une relation d'équivalence. Définissons alors $L^2(\Omega, \mathcal{F}, \mathbb{P})$ comme l'espace quotient de $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ par la relation d'équivalence définie ci-dessus. Les éléments de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ sont à présent des *classes d'équivalence*. Soient $\bar{X}$ et $\bar{Y}$ deux éléments de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ et soient X, X' deux représentants (éléments) de $\bar{X}$ et Y, Y' deux représentants de $\bar{Y}$. Nous avons d'après les égalités presque sûres :

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X'Y'\right]$$

ce qui nous permet de définir un produit intérieur dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$ par :

$$(\bar{X}, \bar{Y}) = \mathbb{E}\left[XY\right]$$

où X et Y sont respectivement deux représentants *quelconques* de $\bar{X}$ et de $\bar{Y}$. A présent le produit intérieur $(\bar{X}, \bar{Y})$ munit $L^2(\Omega, \mathcal{F}, \mathbb{P})$ d'une structure pré-hilbertienne. En effet $(\bar{X}, \bar{X}) = 0 \Leftrightarrow \bar{X} = \bar{0}$.
Dans la suite, pour simplifier l'écriture, nous noterons de la même manière les classes et les représentants des classes et confondrons $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ et sa classe d'équivalence $\bar{X} \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Ainsi nous noterons le produit scalaire dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$ sous la forme :

$$(X, Y) = \mathbb{E}\left[XY\right]$$

étant sous-entendu que (X, Y) fait référence au produit intérieur dans l'espace quotient.
Le résultat suivant est central.

Proposition 4.4. *L'espace $L^2(\Omega, \mathcal{F}, \mathbb{P})$ est un espace de Hilbert.*

Ce résultat est une conséquence immédiate de la propriété A.10 donnée annexe A.

Définition 4.10 (Convergence en moyenne quadratique)**.** *Soit $\{X_n\}$ une suite de $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Nous dirons que X_n converge en moyenne quadratique vers $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, si et seulement si :*

$$\lim_{n\to\infty} \|X_n - X\| = \lim_{n\to\infty} \mathbb{E}^{1/2}\left[(X_n - X)^2\right] = 0$$

Notons ici que $\mathbb{E}[X] = (X, 1)$. La propriété suivante est alors une conséquence directe de la continuité du produit scalaire.

Propriété 4.2. *Soit $\{X_n\}$ une suite de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ qui converge vers X. Alors $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$ et $\mathbb{E}\left[X^2\right] = \lim_{n\to\infty} \mathbb{E}\left[X_n^2\right]$.*

Exemple 4.9
Considérons un bruit blanc $\{Z_t\}_{t\in\mathbb{Z}}$, cést-à-dire une suite de variables centrées et orthonormées de $L^2(\Omega, \mathcal{F}, \mathbb{P})$. On a $\mathbb{E}[Z_t] = (Z_t, 1) = 0$ et $(Z_t, Z_s) = \delta_{t,s}$ pour tout couple $(t, s) \in \mathbb{Z} \times \mathbb{Z}$. Soit $\{a_t\}$ une suite réelle telle que $\sum_{t\geq 0} a_t^2 < +\infty$. Alors la suite :

$$X_n = \sum_{t=0}^{n} a_t Z_t$$

est une suite de variables aléatoires de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ centrées. Cette suite converge en moyenne quadratique dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$. En effet pour tout $m \geq n$:

$$\|X_n - X_m\|^2 = \mathbb{E}\left[\left|\sum_{t=n+1}^{m} a_t Z_t\right|^2\right] = \sum_{t=n+1}^{m}\sum_{s=n+1}^{m} a_t a_s (Z_t, Z_s) = \sum_{t=n+1}^{m} a_t^2$$

Comme $\sum_{t\geq 0} a_t^2 < +\infty$, $\sum_{t=n+1}^{m} a_t^2$ tend vers 0 quand n, m tendent vers l'infini et X_n est une suite de Cauchy dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Elle admet donc, en vertu de la proposition 4.4, une limite dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$ que nous notons X. D'après la propriété 4.2, $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n] = 0$ et $\mathrm{var}(X) = \lim_{n\to\infty} \mathrm{var}(X_n) = \sum_{t\geq 0} |a_t|^2$.

4.3 Prédiction linéaire

4.3.1 Estimation linéaire en moyenne quadratique

Soient X et $\{Y_1, \cdots, Y_p\}$ des variables aléatoires réelles de $L^2(\Omega, \mathcal{F}, \mathbb{P})$. On cherche à déterminer la meilleure approximation de X par une combinaison linéaire des variables Y_k. Nous supposons ici que nous connaissons les quantités $\mu = \mathbb{E}[X]$, $\nu_k = \mathbb{E}[Y_k]$ ainsi que les coefficients de covariance $\mathrm{cov}(X, Y_k)$ et $\mathrm{cov}(Y_k, Y_\ell)$, pour tout $1 \leq k, \ell \leq p$. En pratique, rappelons que nous avons vu chapitre 2 comment il est possible, sous certaines hypothèses, de les estimer "correctement" à partir d'une suite d'observations.
On considère l'espace fermé de dimension finie $\mathcal{Y} = \mathrm{span}(\{1, Y_1, \cdots, Y_p\})$ et on cherche l'élément $Y \in \mathcal{Y}$ qui minimise la norme de l'erreur déstimation $\|X - Y\|^2$. Il découle immédiatement du théorème de projection que le prédicteur linéaire optimal est la projection orthogonale $(X|\mathcal{Y})$ de X sur $\mathcal{Y}$ qui vérifie $(X - (X|\mathcal{Y})) \perp \mathcal{Y}$. On en déduit que :

$$\begin{cases} (X - (X|\mathcal{Y}), 1) = 0 \\ (X - (X|\mathcal{Y}), Y_k) = 0 \quad \text{pour} \quad k \in \{1, \cdots, p\} \end{cases} \tag{4.1}$$

Ce sont ces $(p+1)$ équations qui vont nous donner la solution cherchée. En effet $(X|\mathcal{Y}) \in \mathcal{Y}$ implique ($\mathcal{Y}$ est de dimension finie) qu'il se met sous la forme $(X|\mathcal{Y}) = a_0 + \sum_{k=1}^{p} a_k(Y_k - \nu_k)$. Reste à déterminer $a_0, a_1, \ldots, a_p$. Partant de la première expression de (4.1), on obtient :

$$(X - a_0 - \sum_{k=1}^{p} a_k(Y_k - \nu_k), 1) = (X, 1) - a_0 = 0 \tag{4.2}$$

qui donne $a_0 = \mu$. En faisant $a_0 = \mu$ dans la seconde expression de (4.1), on a alors pour $k \in \{1, \ldots, p\}$:

$$(X - \mu - \sum_{j=1}^{p} a_j(Y_j - \nu_j), Y_k - \nu_k) = (X - \mu, Y_k - \nu_k) - \sum_{j=1}^{p} a_j(Y_j - \nu_j, Y_k - \nu_k) = 0 \tag{4.3}$$

qui montrent que $\{a_1, \cdots, a_p\}$ sont solution d'un système de p équations linéaires à p inconnues. Ce système d'équations peut se mettre sous forme plus compacte en utilisant la matrice $\Gamma = [\mathrm{cov}(Y_k, Y_\ell)]_{1 \leq k, \ell \leq p}$ des coefficients de covariance de $(Y_1, \cdots, Y_p)$ et le vecteur $\boldsymbol{\gamma} = [\mathrm{cov}(X, Y_1), \cdots, \mathrm{cov}(X, Y_p)]^T$ des coefficients de covariance entre X et les composantes Y_k. Avec ces notations, le vecteur $\boldsymbol{\alpha} = [a_1, \cdots, a_p]^T$ est solution de l'équation :

$$\Gamma \boldsymbol{\alpha} = \boldsymbol{\gamma} \tag{4.4}$$

Ce système linéaire admet une unique solution si la matrice Γ est inversible. Notons enfin quén vertu de l'identité de Pythagore, nous avons :

$$\|X\|^2 = \|(X|\mathcal{Y})\|^2 + \|X - (X|\mathcal{Y})\|^2$$

et donc la norme minimale de l'erreur de prédiction a pour expression :

$$\|X - (X|\mathcal{Y})\|^2 = \|X\|^2 - \|(X|\mathcal{Y})\|^2$$

Nous allons à présent appliquer ce résultat à la prédiction d'un processus stationnaire au second-ordre à partir de son passé immédiat en prenant $X = X_t$ et $Y_k = X_{t-k}$ avec $k = \{1, \ldots, p\}$.

4.3.2 Prédiction linéaire d'un processus stationnaire au second-ordre

Soit $\{X_t, t \in \mathbb{Z}\}$ un processus stationnaire au second-ordre, de moyenne $\mathbb{E}[X_0] = \mu$ et de fonction d'autocovariance $\gamma(h) = \mathrm{cov}(X_h, X_0)$. On cherche à "prédire" la valeur du processus à la date t à partir d'une combinaison linéaire des p derniers échantillons du passé $\{X_{t-1}, \cdots, X_{t-p}\}$. Ce problème est bien entendu un cas particulier du problème précédent où nous avons $X = X_t$ et $Y_k = X_{t-k}$, pour $k \in \{1, \ldots, p\}$ et où :

$$\mathcal{H}_{t-1,p} = \mathrm{span}\{1, X_{t-1}, X_{t-2}, \cdots, X_{t-p}\} \tag{4.5}$$

Formons la matrice de covariance Γ_p du vecteur $[X_{t-1}, \cdots, X_{t-p}]$:

$$\Gamma_p = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & & & \gamma(1) \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(1) & \gamma(0) \end{bmatrix} \tag{4.6}$$

Cette matrice est dite de Toëplitz, ses éléments étant égaux le long de ses diagonales. Notons $\boldsymbol{\gamma}_p$ le vecteur $[\gamma(1), \gamma(2), \cdots, \gamma(p)]^T$ le vecteur des coefficients de corrélation. D'après l'équation (4.4), les coefficients $\{\phi_{k,p}\}_{1 \leq k \leq p}$ du prédicteur linéaire optimal défini par :

$$(X_t|\mathcal{H}_{t-1,p}) - \mu = \sum_{k=1}^{p} \phi_{k,p}(X_{t-k} - \mu) \tag{4.7}$$

sont solutions du système d'équations :

$$\Gamma_p \boldsymbol{\phi}_p = \boldsymbol{\gamma}_p \tag{4.8}$$

D'autre part l'erreur de prédiction minimale a pour expression :

$$\begin{aligned}\sigma_p^2 &= \|X_t - (X_t|\mathcal{H}_{t-1,p})\|^2 = (X_t, X_t - (X_t|\mathcal{H}_{t-1,p})) \\ &= (X_t, X_t) - (X_t - \mu, (X_t|\mathcal{H}_{t-1,p})) - (\mu, (X_t|\mathcal{H}_{t-1,p})) \\ &= \gamma(0) - \sum_{k=1}^{p} \phi_{k,p}\gamma(k) = \gamma(0) - \boldsymbol{\phi}_p^T \boldsymbol{\gamma}_p \end{aligned} \tag{4.9}$$

Les équations (4.8) et (4.9) sont appelées *équations de Yule-Walker*. Notons la propriété importante suivante : pour p fixé, la suite des coefficients $\{\phi_{k,p}\}_{1\leq k\leq p}$ du prédicteur linéaire optimal et la variance de l'erreur minimale de prédiction *ne dépendent pas de* t. Les équations (4.8) et (4.9) peuvent encore être réécrites à partir des coefficients de corrélation $\rho(h) = \gamma(h)/\gamma(0)$. Il vient :

$$\begin{bmatrix} \rho(0) & \rho(1) & \cdots & & \rho(p-1) \\ \rho(1) & \rho(0) & \rho(1) & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & & & \rho(1) \\ \rho(p-1) & \rho(p-2) & \cdots & \rho(1) & \rho(0) \end{bmatrix} \begin{bmatrix} \phi_{1,p} \\ \phi_{2,p} \\ \vdots \\ \vdots \\ \phi_{p,p} \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \vdots \\ \rho(p) \end{bmatrix} \tag{4.10}$$

Exemple 4.10 : Prédiction avant/arrière
Soit $X_t = Z_t + \theta_1 Z_{t-1}$ *où* $Z_t \sim \mathrm{BB}(0, \sigma^2)$. *On note* $\rho(h)$ *la fonction d'autocorrélation de* X_t.

1. $\rho(0) = (1+\theta_1^2)$, $\rho(\pm 1) = \theta_1$ *et* $\rho(h) = 0$ *pour* $|h| \geq 2$.
2. *Déterminons la prédiction de* X_3 *en fonction de* X_2 *et* X_1. *D'après le théorème de projection* $(X_3|\mathrm{span}\{X_2, X_1\}) = \alpha_1 X_1 + \alpha_1 X_2$ *vérifie* $(X_3 - \alpha_2 X_2 - \alpha_1 X_1, X_j) = 0$ *pour* $j = 1, 2$. *On en déduit que :*

$$\begin{bmatrix} 1+\theta_1^2 & \theta_1 \\ \theta_1 & 1+\theta_1^2 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} \theta_1 \\ 0 \end{bmatrix}$$

3. *Déterminons la prédiction de* X_3 *en fonction de* X_4 *et* X_5. *D'après le théorème de projection* $(X_3|\mathrm{span}\{X_4, X_5\}) = \alpha_4 X_4 + \alpha_5 X_5$ *vérifie* $(X_3 - \alpha_4 X_4 - \alpha_5 X_5, X_j) = 0$ *pour* $j = 4, 5$. *On en déduit que :*

$$\begin{bmatrix} 1+\theta_1^2 & \theta_1 \\ \theta_1 & 1+\theta_1^2 \end{bmatrix} \begin{bmatrix} \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} \theta_1 \\ 0 \end{bmatrix}$$

   Par conséquent $\alpha_1 = \alpha_5$ *et* $\alpha_2 = \alpha_4$.
4. *Déterminons la prédiction de* X_3 *en fonction de* X_1, X_2, X_4 *et* X_5. *Pour déterminer* $(X_3|\mathrm{span}\{X_1, X_2, X_4, X_5\}) = \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5$ *Il suffit de remarquer que* $\mathrm{span}\{X_1, X_2\} \perp \mathrm{span}\{X_3, X_5\}$ *et donc :*

$$(X_3|\mathrm{span}\{X_1, X_2, X_4, X_5\}) = (X_3|\mathrm{span}\{X_1, X_2\}) + (X_3|\mathrm{span}\{X_4, X_5\})$$

Exemple 4.11 : Cas d'un processus AR(m) causal
Soit le processus AR(m) causal solution stationnaire de l'équation récurrente :

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_m X_{t-m} + Z_t$$

où $Z_t \sim B(0,\sigma^2)$ et où $\phi(z) = 1 - \sum_{k=1}^m \phi_k z^k \neq 0$ pour $|z| \leq 1$. Comme la solution est causale on a, pour tout $h \geq 1$, $\mathbb{E}[Z_t X_{t-h}] = 0$ et donc $\mathbb{E}[(X_t - \sum_{k=1}^m \phi_k X_{t-k})X_{t-h}] = 0$ qui signifie que, pour tout $p \geq m$, (i) $(X_t - \sum_{k=1}^m \phi_k X_{t-k}) \perp \mathcal{H}_{t-1,p}$ et (ii) $\sum_{k=1}^m \phi_k X_{t-k} \in \mathcal{H}_{t-1,p}$. Par conséquent, d'après le théorème de projection, $\sum_{k=1}^m \phi_k X_{t-k} = (X_t|\mathcal{H}_{t-1,p})$ et donc, pour tout $p \geq m$:

$$\phi_{k,p} = \begin{cases} \phi_k & pour \quad 1 \leq k \leq m \\ 0 & pour \quad k > m \end{cases}$$

La projection orthogonale d'un $AR(m)$ causal sur son passé immédiat de longueur $p \geq m$ coïncide avec la projection orthogonale sur les m dernières valeurs et les coefficients de prédiction sont précisément les coefficients de l'équation récurrente.

Dans le cas où la matrice de covariance Γ_p, supposée *connue*, est inversible, le problème de la détermination des coefficients de prédiction $\boldsymbol{\phi}_p$ et de la variance de l'erreur de prédiction σ_p^2 a une solution unique. Rappelons que, d'après la propriété 1.5, si $\gamma(0) > 0$ et si $\lim_{n\to\infty} \gamma(n) = 0$, alors la matrice Γ_p est inversible à tout ordre.
Il est facile de démontrer que :

$$(X_t|\text{span}\{1, X_{t-1}, \ldots, X_{t-p}\}) = \mu + (X_t - \mu|\text{span}\{X_{t-1} - \mu, \ldots, X_{t-p} - \mu\}) \tag{4.11}$$

Par conséquent, dans le problème de la prédiction, il n'y a aucune perte de généralité à considérer que le processus est centré. S'il ne l'était pas, il suffirait, d'après l'équation (4.11), déffectuer le calcul des prédicteurs sur le processus centré $X_t^c = X_t - \mu$ puis d'ajouter μ. *Dans la suite, sauf indication contraire, les processus sont supposés centrés.*
Les coefficients de prédiction d'un processus stationnaire au second ordre fournissent une décomposition particulière de la matrice de covariance Γ_{p+1} sous la forme d'un produit de matrice triangulaire.

Théorème 4.3. *Soit $\{X_t\}$ un processus stationnaire au second ordre, centré, de fonction d'autocovariance $\gamma(h)$. On note :*

$$A_{p+1} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\phi_{1,1} & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ -\phi_{p,p} & -\phi_{p-1,p} & \cdots & -\phi_{1,p} & 1 \end{bmatrix} \qquad D_{p+1} = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & & \cdots & \sigma_p^2 \end{bmatrix}$$

On a alors :

$$\Gamma_{p+1} = A_{p+1}^{-1} D_{p+1} A_{p+1}^{-T} \tag{4.12}$$

Démonstration. Posons $\mathcal{F}_k = \text{span}\{X_k, \cdots, X_1\}$ et montrons tout d'abord que, pour $k \neq \ell$, nous avons :

$$(X_k - (X_k|\mathcal{F}_{k-1}), X_\ell - (X_\ell|\mathcal{F}_{\ell-1})) = 0 \tag{4.13}$$

En effet, pour $k < \ell$, on a $X_k - (X_k|\mathcal{F}_{k-1}) \in \mathcal{F}_k \subseteq \mathcal{F}_{\ell-1}$. On a aussi $X_\ell - (X_\ell|\mathcal{F}_{\ell-1}) \perp \mathcal{F}_{\ell-1}$ et donc $X_\ell - (X_\ell|\mathcal{F}_{\ell-1}) \perp X_k - (X_k|\mathcal{F}_{k-1})$, ce qui démontre (4.13). D'autre part, par définition des coefficients

de prédiction, on peut écrire successivement :

$$A_{p+1}\mathbf{X}_{p+1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\phi_{1,1} & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ -\phi_{p,p} & -\phi_{p-1,p} & \cdots & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{p+1} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 - (X_2|\mathcal{F}_1) \\ \vdots \\ X_{p+1} - (X_{p+1}|\mathcal{F}_p) \end{bmatrix}$$

qui donne :

$$\mathbb{E}\left[A_{p+1}\mathbf{X}_{p+1}\mathbf{X}_{p+1}^T A_{p+1}^T\right] = A_{p+1}\Gamma_{p+1}A_{p+1}^T = D_{p+1}$$

où, par définition, $\sigma_k^2 = \|X_k - (X_k|\mathcal{F}_{k-1})\|^2$, ce qui démontre (4.12) puisque la matrice A_{p+1} est inversible, son déterminant étant égal à 1. Ajoutons que l'inverse d'une matrice triangulaire supérieure est elle-même triangulaire supérieure. ■

Dans la suite nous notons $\mathcal{H}_{t-1,p} = \text{span}\{X_{t-1}, \ldots, X_{t-p}\}$ et nous appelons *erreur de prédiction directe* d'ordre p ou *innovation partielle* d'ordre p le processus :

$$\epsilon_{t,p}^+ = X_t - (X_t|\mathcal{H}_{t-1,p}) = X_t - \sum_{k=1}^{p} \phi_{k,p} X_{t-k} \tag{4.14}$$

D'après l'équation (4.12) lorsque la matrice Γ_{p+1} est inversible, la variance $\sigma_p^2 = \|\epsilon_{t,p}^+\|^2$ est strictement positive. Il est clair, d'autre part, que la suite σ_p^2 est décroissante et donc que σ_p^2 possède une limite quand p tend vers l'infini. Cela conduit à la définition suivante, dont nous verrons paragraphe 4.6 quélle joue un rôle fondamental dans la décomposition des processus stationnaires au second ordre.

Définition 4.11 (Processus régulier/déterministe)**.** *Soit $\{X_t\}$ un processus aléatoire stationnaire au second ordre. On note σ_p^2 la variance de l'innovation partielle d'ordre p et $\sigma^2 = \lim_{p\to+\infty} \sigma_p^2$. On dit que le processus $\{X_t\}$ est* régulier *si $\sigma^2 \neq 0$ et* déterministe *si $\sigma^2 = 0$.*

Nous avons déjà noté (voir équation (4.8)) que, pour p fixé, la suite $\{\phi_{k,p}\}$ ne dépend pas de t et donc que le processus $\epsilon_{t,p}^+$ (relativement à l'indice t) est stationnaire au second ordre, centré. On a aussi la formule suivante :

$$(\epsilon_{t,p}^+, \epsilon_{t,q}^+) = \sigma_{\max(p,q)}^2 \tag{4.15}$$

En effet soit $q > p$. Par construction, nous avons $\epsilon_{t,q}^+ \perp \mathcal{H}_{t-1,q}$, et comme $\mathcal{H}_{t-1,p} \subseteq \mathcal{H}_{t-1,q}$, $\epsilon_{t,q}^+ \perp \mathcal{H}_{t-1,p}$ et en particulier $\epsilon_{t,q}^+ \perp (X_t|\mathcal{H}_{t-1,p})$ puisque $(X_t|\mathcal{H}_{t-1,p}) \in \mathcal{H}_{t-1,p}$. Par conséquent, pour $q > p$, on a :

$$(\epsilon_{t,p}^+, \epsilon_{t,q}^+) = (X_t - (X_t|\mathcal{H}_{t-1,p}), \epsilon_{t,q}^+) = (X_t, X_t - (X_t|\mathcal{H}_{t-1,q})) = (X_t, X_t - (X_t|\mathcal{H}_{t-1,q})) = \sigma_q^2$$

ce qui démontre (4.15).
Notons ici que le problème de la recherche des coefficients de prédiction pour un processus stationnaire au second ordre se ramène à celui de la minimisation de l'intégrale :

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\psi(e^{-i\lambda})|^2 \nu_X(d\lambda)$$

sur l'ensemble $\mathcal{P}_p$ des polynômes à coefficients réels de degré p de la forme $\psi(z) = 1+\psi_1 z+\cdots+\psi_p z^p$. En effet, en utilisant la relation (1.18) de filtrage des mesures spectrales, on peut écrire que la variance de $\|\epsilon^+_{t,p}\|^2$, qui minimise de l'erreur de prédiction, a pour expression :

$$\sigma_p^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\phi_p(e^{-i\lambda})|^2 \nu_X(d\lambda) \tag{4.16}$$

où :

$$\phi_p(z) = 1 - \sum_{k=1}^{p} \phi_{k,p} z^k$$

désigne le *polynôme prédicteur d'ordre p.*

Théorème 4.4. *Si $\{X_t\}$ est un processus régulier, alors, pour tout p, $\phi_p(z) \neq 0$ pour $|z| \leq 1$. Tous les zéros des polynômes prédicteurs sont à l'extérieur du cercle unité.*

Démonstration. Elle est donnée en fin de chapitre. ■

Une conséquence directe du théorème 4.4 est qu'à toute matrice de covariance de type défini positif, de dimension $(p+1)\times(p+1)$, on peut associer un processus AR(p) causal dont les $(p+1)$ premiers coefficients de covariance sont précisément la première ligne de cette matrice. Ce résultat nést pas général. Ainsi il existe bien un processus AR(2) causal ayant $\gamma(0) = 1$ et $\gamma(1) = \rho$, comme premiers coefficients de covariance, à condition toutefois que la matrice de covariance soit positive cést-à-dire que $|\rho| < 1$, tandis qu'il néxiste pas, pour cette même matrice de processus MA(2). Il faut en effet, en plus du caractère positif, que $|\rho| \geq 1/2$ (voir exemple 1.11).

4.4 Algorithme de Levinson-Durbin

La solution directe du système des équations de Yule-Walker requiert de l'ordre de p^3 opérations : la résolution classique de ce système implique en effet la décomposition de la matrice Γ_p sous la forme du produit d'une matrice triangulaire inférieure et de sa transposée, $\Gamma_p = L_p L_p^T$ (décomposition de Choleski) et la résolution par substitution de deux systèmes triangulaires. Cette procédure peut s'avérer coûteuse lorsque l'ordre de prédiction est grand (on utilise généralement des ordres de prédiction de l'ordre de quelques dizaines à quelques centaines), ou lorsque, à des fins de modélisation, on est amené à évaluer la qualité de prédiction pour différents horizons de prédiction. L'algorithme de Levinson-Durbin exploite la structure géométrique particulière des processus stationnaires au second ordre pour établir une formule de récurrence donnant les coefficients de prédiction à l'ordre $(p+1)$ à partir des coefficients de prédiction obtenus à l'ordre p. Supposons que nous connaissions les coefficients de prédiction linéaire et la variance de l'erreur de prédiction à l'ordre p, pour $p \geq 0$:

$$(X_t|\mathcal{H}_{t-1,p}) = \sum_{k=1}^{p} \phi_{k,p} X_{t-k} \quad \text{et} \quad \sigma_p^2 = \|X_t - (X_t|\mathcal{H}_{t-1,p})\|^2$$

Nous avons besoin ici d'introduire l'*erreur de prédiction rétrograde à l'ordre p* définie par :

$$\epsilon^-_{t,p} = X_t - (X_t|\mathcal{H}_{t+p,p}) = X_t - (X_t|\text{span}\{X_{t+1},\cdots,X_{t+p}\})$$

Elle représente la différence entre l'échantillon courant X_t et la projection orthogonale de X_t sur les p échantillons $\{X_{t+1}, \cdots, X_{t+p}\}$ *qui suivent* l'instant courant. Le qualificatif *rétrograde* est clair : il traduit le fait que l'on cherche à prédire la valeur courante en fonction des valeurs futures. Indiquons que l'erreur rétrograde joue un rôle absolument essentiel dans tous les algorithmes rapides de résolution des équations de Yule-Walker. Remarquer tout d'abord que les coefficients de prédiction rétrograde coïncident avec les coefficients de prédiction directe. Cette propriété, que nous avons rencontrée exemple 4.10, est fondamentalement due à la *propriété de réversibilité* des processus stationnaires au second ordre. En effet, si $Y_t = X_{-t}$, alors Y_t a même moyenne et même fonction de covariance que X_t (voir exemple 1.7 chapitre 1) et par conséquent, en utilisant aussi l'hypothèse de stationnarité, on a simultanément pour tout $u, v \in \mathbb{Z}$:

$$(X_{t+u}|\mathcal{H}_{t+u-1,p}) = \sum_{k=1}^{p} \phi_{k,p} X_{t+u-k} \quad \text{et} \quad (X_{t+v}|\mathcal{H}_{t+v+p,p}) = \sum_{k=1}^{p} \phi_{k,p} X_{t+v+k}$$

ainsi que :

$$\sigma_p^2 = \|\epsilon^+_{t+u,p}\|^2 = \|\epsilon^-_{t+v,p}\|^2 \tag{4.17}$$

En particulier on a :

$$\begin{cases} (X_t|\mathcal{H}_{t-1,p}) = \sum_{k=1}^{p} \phi_{k,p} X_{t-k} \\ (X_{t-p-1}|\mathcal{H}_{t-1,p}) = \sum_{k=1}^{p} \phi_{k,p} X_{t-p-1+k} = \sum_{k=1}^{p} \phi_{p+1-k,p} X_{t-p-1+k} \end{cases} \tag{4.18}$$

Cherchons maintenant à déterminer, à partir de ces projections à l'ordre p, la projection de X_t à l'ordre $p+1$ sur le sous-espace $\mathcal{H}_{t-1,p+1} = \text{span}\{X_{t-1}, \cdots, X_{t-p-1}\}$. Pour cela décomposons cet espace en somme directe de la façon suivante :

$$\mathcal{H}_{t-1,p+1} = \mathcal{H}_{t-1,p} \oplus \text{span}\{X_{t-p-1} - (X_{t-p-1}|\mathcal{H}_{t-1,p})\} = \mathcal{H}_{t-1,p} \oplus \text{span}\{\epsilon^-_{t-p-1,p}\}$$

Un calcul simple montre (voir exemple 4.6) que

$$(X_t|\epsilon^-_{t-p-1,p}) = \alpha \epsilon^-_{t-p-1,p} \quad \text{avec} \quad \alpha = (X_t, \epsilon^-_{t-p-1,p})/\|\epsilon^-_{t-p-1,p}\|^2$$

et donc que

$$(X_t|\mathcal{H}_{t-1,p+1}) = (X_t|\mathcal{H}_{t-1,p}) + k_{p+1}(X_{t-p-1} - (X_{t-p-1}|\mathcal{H}_{t-1,p})) \tag{4.19}$$

où, en utilisant aussi (4.17), on peut écrire :

$$k_{p+1} = \frac{(X_t, \epsilon^-_{t-p-1,p})}{\sigma_p^2} = \frac{(X_t, \epsilon^-_{t-p-1,p})}{\|\epsilon^+_{t+u,p}\|\|\epsilon^-_{t+v,p}\|} \tag{4.20}$$

En portant à présent (4.18) dans (4.19), on obtient l'expression :

$$(X_t|\mathcal{H}_{t-1,p+1}) = \sum_{k=1}^{p+1} \phi_{k,p+1} X_{t-k} = \sum_{k=1}^{p} (\phi_{k,p} - k_{p+1}\phi_{p+1-k,p}) X_{t-k} + k_{p+1} X_{t-p-1}$$

On en déduit les formules de récurrence donnant les coefficients de prédiction à l'ordre $p+1$ à partir de ceux à l'ordre p :

$$\left\{ \begin{array}{rcl} \phi_{k,p+1} & = & \phi_{k,p} - k_{p+1}\phi_{p+1-k,p} \quad \text{pour} \quad k \in \{1, \cdots, p\} \\ \phi_{p+1,p+1} & = & k_{p+1} \end{array} \right. \tag{4.21}$$

Déterminons maintenant la formule de récurrence donnant k_{p+1}. En utilisant encore (4.18) et (4.19), on obtient :

$$(X_t, (X_{t-p-1}|\mathcal{H}_{t-1,p})) = \sum_{k=1}^{p} \phi_{k,p}\mathbb{E}\left[X_t X_{t-p-1+k}\right] = \sum_{k=1}^{p} \phi_{k,p}\gamma(p+1-k)$$

Partant de l'expression de $(X_t, \epsilon^-_{t-p-1,p})$ on en déduit que :

$$(X_t, \epsilon^-_{t-p-1,p}) = (X_t, X_{t-p-1} - (X_{t-p-1}|\mathcal{H}_{t-1,p})) = \gamma(p+1) - \sum_{k=1}^{p} \phi_{k,p}\gamma(p+1-k)$$

et donc d'après (4.20) :

$$k_{p+1} = \frac{\gamma(p+1) - \sum_{k=1}^{p} \phi_{k,p}\gamma(p+1-k)}{\sigma_p^2}$$

Il nous reste maintenant à déterminer l'erreur de prédiction σ^2_{p+1} à l'ordre $(p+1)$. En utilisant l'équation (4.19), on a

$$\epsilon^+_{t,p+1} = X_t - (X_t|\mathcal{H}_{t-1,p+1}) = X_t - (X_t|\mathcal{H}_{t-1,p}) - k_{p+1}(X_{t-p-1} - (X_{t-p-1}|\mathcal{H}_{t-1,p}))$$

dont on déduit d'après (4.20) :

$$\sigma^2_{p+1} = \|\epsilon^+_{t,p+1}\|^2 = \sigma_p^2 + k^2_{p+1}\sigma_p^2 - 2k_{p+1}(X_t - (X_t|\mathcal{H}_{t-1,p}), X_{t-p-1} - (X_{t-p-1}|\mathcal{H}_{t-1,p})) = \sigma_p^2(1-k^2_{p+1})$$

Pour initialiser l'algorithme, nous faisons $p=0$. Dans ce cas la meilleure prédiction de X_t est $\mathbb{E}\left[X_t\right] = 0$ et la variance de l'erreur de prédiction est alors donnée par $\sigma_0^2 = \mathbb{E}\left[(X_t - 0)^2\right] = \gamma(0)$. Au pas suivant on a $k_1 = \gamma(1)/\gamma(0)$, $\phi_{1,1} = \gamma(1)/\gamma(0)$ et $\sigma_1^2 = \gamma(0)(1-k_1^2)$.
Partant d'une suite de $(K+1)$ coefficients de covariance $\gamma(0), \ldots, \gamma(K)$, l'*algorithme de Levinson* détermine les coefficients de prédiction $\{\phi_{m,p}\}_{1\leq m\leq p, 1\leq p\leq K}$:

Valeurs initiales :

· $k_1 = \gamma(1)/\gamma(0)$, $\phi_{1,1} = \gamma(1)/\gamma(0)$ et $\sigma_1^2 = \gamma(0)(1-k_1^2)$

Pour $p = \{2, \ldots, K\}$ répéter :

· $k_p = \sigma_{p-1}^{-2}\left(\gamma(p) - \sum_{k=1}^{p-1}\phi_{k,p-1}\gamma(p-k)\right)$

· $\phi_{p,p} = k_p$

· pour $m \in \{1, \cdots, p-1\}$ faire :

$$\phi_{m,p} = \phi_{m,p-1} - k_p\phi_{p-m,p-1}$$

· $\sigma_p^2 = \sigma_{p-1}^2(1-k_p^2)$

Le coefficient k_p possède la propriété remarquable d'être de module inférieur à 1. Notons tout d'abord que $(X_t|\mathcal{H}_{t-1,p}) \perp \epsilon^-_{t-p-1,p}$ puisque $(X_t|\mathcal{H}_{t-1,p}) \in \mathcal{H}_{t-1,p}$ et que $\epsilon^-_{t-p-1,p} \perp \mathcal{H}_{t-1,p}$. Partant de (4.20) on peut écrire que :

$$k_{p+1} = \frac{(X_t - (X_t|\mathcal{H}_{t-1,p}), X_{t-p-1} - (X_{t-p-1}|\mathcal{H}_{t-1,p}))}{\|\epsilon^+_{t,p}\| \ \|\epsilon^-_{t-p-1,p}\|} = \frac{(\epsilon^+_{t,p}, \epsilon^-_{t-p-1,p})}{\|\epsilon^+_{t,p}\| \ \|\epsilon^-_{t-p-1,p}\|} \tag{4.22}$$

En utilisant l'inégalité de Schwarz, on montre que $|k_{p+1}| \leq 1$. Remarquons aussi que k_{p+1} apparaît comme le coefficient de corrélation entre l'erreur de prédiction directe et l'erreur de prédiction rétrograde. Dans la littérature ce coefficient est appelé coefficient d'autocorrélation partielle.

Définition 4.12 (Fonction d'autocorrélation partielle)**.** *Soit X_t un processus aléatoire, stationnaire au second ordre, de fonction de covariance $\gamma(h)$. On appelle fonction d'autocorrélation partielle la suite k_p définie par :*

$$k_p = \begin{cases} \text{Corr}(X_t, X_{t-1}) = \dfrac{(X_t, X_{t-1})}{\|X_t\| \ \|X_{t-1}\|} & \textit{pour} \quad p = 1 \\ \text{Corr}(\epsilon^+_{t,p-1}, \epsilon^-_{t-p,p-1}) = \dfrac{(X_t - (X_t|\mathcal{H}_{t-1,p-1}), X_{t-p} - (X_{t-p}|\mathcal{H}_{t-1,p-1}))}{\|X_t - (X_t|\mathcal{H}_{t-1,p-1})\| \ \|X_{t-p} - (X_{t-p}|\mathcal{H}_{t-1,p-1})\|} & \textit{pour} \quad p \geq 2 \end{cases} \tag{4.23}$$

Dans (4.23), l'expression pour $p = 1$ est en accord avec celle pour $p \geq 2$ dans la mesure où on peut noter que $\epsilon^+_{t,0} = X_t$ et que $\epsilon^-_{t-1,0} = X_{t-1}$. Notons aussi que, dans l'expression de k_p, X_t et X_{t-p} sont projetés sur le même sous-espace $\text{span}\{X_{t-1}, \ldots, X_{t-p+1}\}$. Le résultat remarquable est que la suite des coefficients de corrélation partielle est donnée par :

$$k_p = \phi_{p,p} \tag{4.24}$$

où $\phi_{p,p}$ est défini au moyen des équations de Yule-Walker (4.10). Dans le cas particulier d'un processus AR(m) causal, on a alors :

$$k_p = \begin{cases} \phi_{p,p} & \text{pour} \quad 1 \leq p < m \\ \phi_m & \text{pour} \quad p = m \\ 0 & \text{pour} \quad p > m \end{cases}$$

Notons enfin que contrairement à la fonction d'autocorrélation partielle d'un AR(m) qui est nulle pour un intervalle de temps supérieur à m, celle d'un MA(q) ne va pas à 0. Elle est cependant bornée en valeur absolue par une exponentielle décroissante.

4.5 Algorithme de Schur

Partant des coefficients d'autocorrélation, l'algorithme de Levinson-Durbin évalue à la fois les coefficients des prédicteurs linéaires optimaux et les coefficients d'autocorrélation partielle. Dans certains cas, seuls les coefficients d'autocorrélation partielle sont nécessaires. Il en est ainsi, par exemple, lorsque l'on cherche à calculer les erreurs de prédiction directe et rétrograde à partir du processus X_t. Montrons, en effet, que les erreurs de prédiction à l'ordre $(p+1)$ s'expriment, en fonction des erreurs

de prédictions à l'ordre p, à l'aide d'une formule de récurrence ne faisant intervenir que la valeur du coefficient de corrélation partielle :

$$\begin{cases} \epsilon^+_{t,p+1} &= \epsilon^+_{t,p} - k_{p+1}\epsilon^-_{(t-1)-p,p} \\ \epsilon^-_{t-(p+1),p+1} &= \epsilon^-_{(t-1)-p,p} - k_{p+1}\epsilon^+_{t,p} \end{cases} \tag{4.25}$$

Reprenons les expressions de l'erreur de prédiction directe et de l'erreur de prédiction rétrograde :

$$\epsilon^+_{t,p} = X_t - \sum_{k=1}^{p} \phi_{k,p} X_{t-k} \quad \text{et} \quad \epsilon^-_{t-p-1,p} = X_{t-p-1} - \sum_{k=1}^{p} \phi_{k,p} X_{t-p-1+k}$$

En utilisant directement la récursion de Levinson-Durbin, équations (4.21), dans l'expression de l'erreur de prédiction directe à l'ordre $p+1$, nous obtenons :

$$\begin{aligned} \epsilon^+_{t,p+1} &= X_t - \sum_{k=1}^{p+1} \phi_{k,p+1} X_{t-k} \\ &= \left(X_t - \sum_{k=1}^{p} \phi_{k,p} X_{t-k} \right) - k_{p+1} \left(X_{t-p-1} - \sum_{k=1}^{p} \phi_{k,p} X_{t-p-1+k} \right) \\ &= \epsilon^+_{t,p} - k_{p+1}\epsilon^-_{t-p-1,p} \end{aligned} \tag{4.26}$$

De façon similaire, nous avons :

$$\begin{aligned} \epsilon^-_{t-p-1,p+1} &= X_{t-p-1} - \sum_{k=1}^{p+1} \phi_{k,p+1} X_{t-p-1+k} \\ &= \left(X_{t-p-1} - \sum_{k=1}^{p} \phi_{k,p} X_{t-p-1+k} \right) - k_{p+1} \left(X_t - \sum_{k=1}^{p} \phi_{k,p} X_{t-k} \right) \\ &= \epsilon^-_{t-p-1,p} - k_{p+1}\epsilon^+_{t,p} \end{aligned} \tag{4.27}$$

Partant de la suite des autocorrélations, l'algorithme de Schur calcule récursivement les coefficients de corrélation partielle, sans avoir à déterminer les valeurs des coefficients de prédiction. Historiquement, l'algorithme de Schur a été introduit pour tester le caractère défini positif d'une suite (ou de façon équivalente, la positivité des matrices de Toëplitz construites à partir de cette suite). En effet, comme nous l'avons montré ci-dessus, une suite de coefficients de covariance est définie positive si et seulement si les coefficients de corrélation partielle sont de module strictement inférieur à 1. Déterminons à présent cet algorithme. En faisant $t = 0$ dans l'équation (4.26), en multipliant à gauche par X_m et en utilisant la stationnarité, il vient :

$$(X_m, \epsilon^+_{0,p+1}) = (X_m, \epsilon^+_{0,p}) - k_{p+1}(X_m, \epsilon^-_{-p-1,p}) = (X_m, \epsilon^+_{0,p}) - k_{p+1}(X_{m+p+1}, \epsilon^-_{0,p}) \tag{4.28}$$

En faisant $t = p+1$ dans l'équation (4.27), en multipliant à gauche par X_{m+p+1} et en utilisant la stationnarité, il vient :

$$(X_{m+p+1}, \epsilon^-_{0,p+1}) = (X_{m+p+1}, \epsilon^-_{0,p}) - k_{p+1}(X_{m+p+1}, \epsilon^+_{p+1,p}) = (X_{m+p+1}, \epsilon^-_{0,p}) - k_{p+1}(X_m, \epsilon^+_{0,p}) \tag{4.29}$$

En faisant $m = 0$ dans (4.29), il vient :

$$(X_{p+1}, \epsilon^-_{0,p+1}) = (X_{p+1}, \epsilon^-_{0,p}) - k_{p+1}(X_{p+1}, \epsilon^+_{p+1,p}) = (X_{p+1}, \epsilon^-_{0,p}) - k_{p+1}(X_0, \epsilon^+_{0,p}) \tag{4.30}$$

Mais on a aussi :

$$(X_{p+1}, \epsilon^-_{0,p+1}) = (X_{p+1}, X_0 - (X_0|\text{span}\{X_1, \cdots, X_{p+1}\})) = 0$$

Nous pouvons donc déduire de l'équation (4.30) :

$$k_{p+1} = \frac{(X_{p+1}, \epsilon^-_{0,p})}{(X_0, \epsilon^+_{0,p})} \tag{4.31}$$

En couplant les équations (4.28), (4.29) et (4.31) et en partant des conditions initiales :

$$(X_m, \epsilon^+_{0,0}) = \gamma(m) \quad \text{et} \quad (X_{m+1}, \epsilon^-_{0,0}) = \gamma(m+1)$$

on peut déterminer les coefficients de corrélation partielle directement, sans avoir à évaluer explicitement les coefficients de prédiction.
On note $u(m,p) = (X_m, \epsilon^+_{0,p})$ et $v(m,p) = (X_{m+p+1}, \epsilon^-_{0,p})$. Partant des $(K+1)$ coefficients de covariance $\{\gamma(0), \ldots, \gamma(K)\}$, l'*algorithme de Schur* calcule les K premiers coefficients de corrélation partielle :

Initialement faire pour $m = \{0, \ldots, K-1\}$:

- $u(m,0) = \gamma(m)$
- $v(m,0) = \gamma(m+1)$

Puis répéter pour $p = \{1, \ldots, K\}$:

- $k(p) = \dfrac{v(0,p-1)}{u(0,p-1)}$
- et pour $m = \{0, \ldots, K-p-1\}$ faire :

$$\begin{cases} u(m,p) = u(m,p-1) - k(p)v(m,p-1) \\ v(m,p) = v(m+1,p-1) - k(p)u(m+1,p-1) \end{cases}$$

La complexité de l'algorithme de Schur est équivalente à l'algorithme de Levinson.

Filtres en treillis

En notant $e(t,p) = [\epsilon^+_{t,p} \quad \epsilon^-_{t-p,p}]^T$ et en utilisant l'opérateur de retard D, les expressions (4.25) peuvent se mettre sous la forme matricielle :

$$e(t,p+1) = \begin{bmatrix} 1 & -k_{p+1}D \\ -k_{p+1}D & 1 \end{bmatrix} e(t,p)$$

Les erreurs initiales ($p = 0$) sont $e(t,0) = [X_t \quad X_t]^T$. Ces équations débouchent sur une structure de filtrage dite en treillis qui calcule, au moyen des coefficients de corrélation partielle, les erreurs de

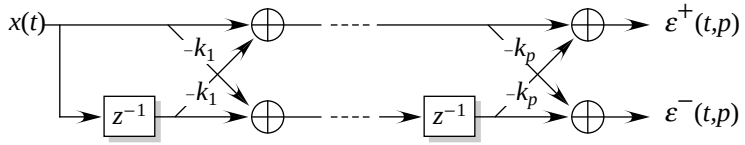

FIG. 4.1 – *Filtre d'analyse en treillis. Ce filtre permet de construire les erreurs de prédiction directe et rétrograde à partir du processus et de la donnée des coefficients de corrélation partielle.*

prédiction directe et rétrograde à partir du signal X_t. Ce filtre d'analyse est représenté figure 4.1. Les équations (4.25) peuvent encore s'écrire :

$$\begin{cases} \epsilon^+_{t,p} &= \epsilon^+_{t,p+1} + k_{p+1}\epsilon^-_{(t-1)-p,p} \\ \epsilon^-_{t-(p+1),p+1} &= \epsilon^-_{(t-1)-p,p} - k_{p+1}\epsilon^+_{t,p} \end{cases}$$

qui donne le schéma de filtrage de la figure 4.2.

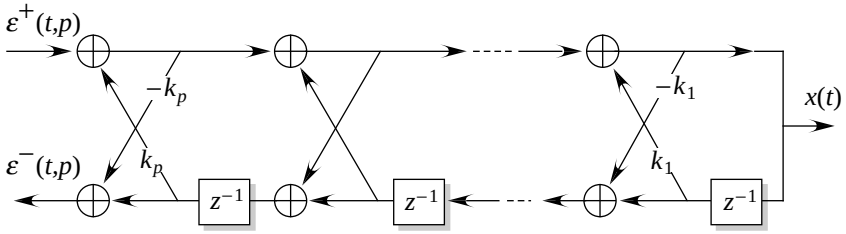

FIG. 4.2 – *Filtre de synthèse en treillis. Ce filtre permet de reconstruire le processus à partir de la suite des erreurs de prédiction directe et de la donnée des coefficients de corrélation partielle.*

4.6 Décomposition de Wold

Un des résultats fondamentaux de la théorie des processus stationnaires au second-ordre est la décomposition de Wold. Cette décomposition permet de décomposer n'importe quel processus stationnaire au second-ordre comme la somme de la sortie d'un filtre linéaire invariant dans le temps excité par un bruit blanc et d'un processus déterministe (définition 4.11). La preuve de ce résultat est de nature géométrique. L'idée de base est la suivante. Soit $\mathcal{H}_t^X = \overline{\text{span}}\{X_s, s \leq t\}$. $\mathcal{H}_t^X$ est appelé le *passé linéaire* du processus à la date t. Par construction, $\mathcal{H}_t^X \subset \mathcal{H}_{t+1}^X$, et nous disposons ainsi d'une famille de sous-espace emboîtés de $\mathcal{H}_\infty^X = \overline{\cup}_{t\in\mathbb{Z}}\mathcal{H}_t^X$. $\mathcal{H}_\infty^X$ est l'enveloppe linéaire du processus. L'espace $\bigcap_{t\in\mathbb{Z}}\mathcal{H}_t^X$, appelé le *passé infini du processus* (X) jouera aussi un rôle particulier. Par définition X_t appartient à $\mathcal{H}_t^X$, mais il n'appartient généralement pas à $\mathcal{H}_{t-1}^X$. Le théorème de projection dit qu'il existe un unique élément noté $(X_t|\mathcal{H}_{t-1}^X)$ et appartenant à $\mathcal{H}_{t-1}^X$ tel que :

$$\epsilon_t = X_t - (X_t|\mathcal{H}_{t-1}^X) \perp \mathcal{H}_{t-1}^X$$

Dans ce contexte ϵ_t s'appelle l'*innovation* (linéaire) du processus. Il découle de cette construction géométrique que le processus d'innovation est un *processus orthogonal* dans le sens où :

$$\forall s \neq t, \quad \epsilon_s \perp \epsilon_t \tag{4.32}$$

En effet, pour $s < t$, nous pouvons écrire $\epsilon_s \in \mathcal{H}_s^X \subset \mathcal{H}_{t-1}^X$ et $\epsilon_t \perp \mathcal{H}_{t-1}^X$. Et donc $\epsilon_s \perp \epsilon_t$.
La proposition qui suit montre que le processus d'innovation est la limite des processus d'innovations partielles à l'ordre p.

Proposition 4.5. *Pour tout $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ et tout $t \in \mathbb{Z}$ nous avons :*

$$\lim_{p\to\infty} (Y|\mathcal{H}_{t,p}^X) = (Y|\mathcal{H}_t^X)$$

où $\mathcal{H}_{t,p}^X = \text{span}\{X_t, X_{t-1}, \cdots, X_{t-p+1}\}$.

Exemple 4.12 : Bruit blanc
Supposons que $\{X_t\}$ soit un bruit blanc. Nous avons $(X_t|\mathcal{H}_{t-1,p}^X) = 0$ pour tout p et donc $(X_t|\mathcal{H}_{t-1}^X) = 0$. Nous avons donc $\epsilon_t = X_t - (X_t|\mathcal{H}_{t-1}^X) = X_t$: le processus X_t coïncide avec son innovation. Ceci signifie qu'un bruit blanc ne peut être prédit de façon linéaire à partir de son passé.

Exemple 4.13 : Prédiction d'un processus AR(p) causal
On considère le processus AR(p) causal défini par l'équation récurrente $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$ où $Z_t \sim \text{BB}(0, \sigma^2)$. Nous avons vu que $\mathcal{H}_t^X = \mathcal{H}_Z^t$ et que, pour tout $k \geq 1$, on avait $\mathbb{E}[X_{t-k} Z_t]$ (confère équation (??)). Par conséquent $Z_t \perp \mathcal{H}_{t-1}^X$ et $\mathcal{H}_t^X = \mathcal{H}_{t-1}^X \oplus \overline{\text{span}}\{Z_t\}$. On en déduit que :

$$(X_t|\mathcal{H}_{t-1}^X) = \sum_{k=1}^{p} \phi_k (X_{t-k}|\mathcal{H}_{t-1}^X) + (Z_t|\mathcal{H}_{t-1}^X) = \sum_{k=1}^{p} \phi_k X_{t-k}$$

et donc $X_t - (X_t|\mathcal{H}_{t-1}^X) = X_t - \sum_{k=1}^{p} \phi_k X_{t-k} = Z_t$. Par conséquent le bruit blanc Z_t, qui intervient dans l'équation récurrente d'un AR causal, est précisément l'innovation du processus AR. Ce résultat montre que $\sum_{k=1}^{p} \phi_k X_{t-k}$ est la projection de $X(t)$ sur tout le passé $\mathcal{H}_{t-1}$ et quélle coïncide avec la projection orthogonale sur le passé $\mathcal{H}_{t-1,p}$ de durée p. Par conséquent, pour tout $m \geq p$, la suite des coefficients de prédiction est $\{\phi_1, \ldots, \phi_p, \underbrace{0, \ldots, 0}_{m-p}\}$. Ce résultat est faux pour un AR non causal.

Exemple 4.14 : Processus harmonique
Soit le processus harmonique $X_t = A\cos(\lambda_0 t + \Phi)$ où A est une variable aléatoire, centrée, de variance σ_A^2 et Φ une variable aléatoire, indépendante de A et distribuée suivant une loi uniforme sur $[-\pi, \pi]$. Le processus X_t est stationnaire au second-ordre, centré, de fonction d'autocovariance $\gamma(\tau) = (\sigma_A^2/2)\cos(\lambda_0 \tau)$. Les coefficients du prédicteur linéaire optimal à l'ordre 2 sont donnés par :

$$\begin{bmatrix} \phi_{1,2} \\ \phi_{2,2} \end{bmatrix} = \begin{bmatrix} 1 & \cos(\lambda_0) \\ \cos(\lambda_0) & 1 \end{bmatrix}^{-1} \begin{bmatrix} \cos(\lambda_0) \\ \cos(2\lambda_0) \end{bmatrix} = \begin{bmatrix} \cos(\lambda_0) \\ -1 \end{bmatrix}$$

On vérifie facilement que $\sigma_2^2 = \|X_t - (X_t|\mathcal{H}_{t-1,2}^X)\|^2 = 0$. Par conséquent, on a :

$$X_t = (X_t|\mathcal{H}_{t-1,2}^X) = 2\cos(\lambda_0) X_{t-1} - X_{t-2} \in \mathcal{H}_{t-1}^X$$

et donc la projection $(X_t|\mathcal{H}_{t-1}^X) = X_t$, ce qui implique que $\epsilon_t = 0$. A l'inverse du bruit blanc, le processus est entièrement prédictible à partir de son passé.

En appliquant la proposition 4.5 à X_t, nous pouvons écrire :

$$\lim_{p\to\infty}(X_t|\mathcal{H}^X_{t-1,p}) = (X_t|\mathcal{H}^X_{t-1}) \quad \text{et} \quad \lim_{p\to\infty}\epsilon^+_{t,p} = \epsilon_t \tag{4.33}$$

Le processus d'innovation ϵ_t est donc la limite en moyenne quadratique de la suite des innovations partielles $\epsilon^+_{t,p} = X_t - (X_t|\mathcal{H}^X_{t-1,p})$. Une conséquence immédiate est que le processus d'innovation est un processus stationnaire au second ordre. En utilisant, en effet, la continuité du produit scalaire et la stationnarité au second ordre de l'innovation partielle d'ordre p, on peut écrire :

$$(\epsilon_{t+\tau}, \epsilon_t) = \lim_{p\to\infty}(\epsilon^+_{t+\tau,p}, \epsilon^+_{t,p}) = \lim_{p\to\infty}(\epsilon^+_{\tau,p}, \epsilon^+_{0,p}) \tag{4.34}$$

qui ne dépend que de τ. En particulier nous avons :

$$\sigma^2 = \|\epsilon_t\|^2 = \lim_{p\to\infty}\|X_t - (X_t|\mathcal{H}^X_{t,p})\|^2 = \lim_{p\to\infty}\sigma^2_p$$

Dans le cas du bruit blanc on obtient $\sigma^2 = \mathbb{E}\left[X_t^2\right] \neq 0$ et donc, d'après la définition 4.11, le bruit blanc est un processus régulier. D'un autre côté, le processus harmonique, pour lequel $\sigma^2 = 0$, est déterministe. Nous remarquons aussi que la somme d'un bruit blanc et d'un processus harmonique est un processus régulier.
La structure géométrique emboîtée des espaces $\{\mathcal{H}^X_t\}$ et l'orthogonalité des innovations fournissent, pour tout $s < t$, la formule suivante de décomposition en somme directe :

$$\mathcal{H}^X_t = \mathcal{H}^X_s \oplus \text{span}\{\epsilon_{s+1}, \cdots, \epsilon_t\} \tag{4.35}$$

Notons, tout d'abord, que $\epsilon_t = X_t - (X_t|\mathcal{H}^X_{t-1}) \in \mathcal{H}^X_t$ et que $\epsilon_t \perp \mathcal{H}^X_{t-1}$, ce qui implique que $\mathcal{H}^X_{t-1} \oplus \text{span}\{\epsilon_t\} \subseteq \mathcal{H}^X_t$. D'un autre côté, puisque $X_t = \epsilon_t + (X_t|\mathcal{H}^X_{t-1})$, $\mathcal{H}^X_t = \overline{\text{span}}\{\epsilon_t + (X_t|\mathcal{H}^X_{t-1}), \{X_s, s \leq t-1\}\} = \overline{\text{span}}\{\epsilon_t, \{X_s, s \leq t-1\}\}$, ce qui entraîne que $\mathcal{H}^X_t \subseteq \mathcal{H}^X_{t-1} \oplus \text{span}\{\epsilon_t\}$. En conclusion $\mathcal{H}^X_t = \mathcal{H}^X_{t-1} \oplus \text{span}\{Z_t\}$. En réitérant ce raisonnement, on en déduit la décomposition (4.35). Cette décomposition orthogonale de l'espace $\mathcal{H}^X_t$ nést pas sans rappeler la décomposition de Gram-Schmidt. Notons qu'à l'inverse de la décomposition de Gram-Schmidt classique, nous procédons ici dans le sens rétrograde. Définissons pour tout $s \geq 0$:

$$\psi_s = \frac{(X_t, \epsilon_{t-s})}{\sigma^2} \tag{4.36}$$

Remarquons que ψ_s ne dépend pas de t. En effet, la continuité du produit scalaire et la stationnarité conjointe du processus X_t et de l'innovation partielle impliquent que :

$$(X_t, \epsilon_{t-s}) = \lim_{p\to\infty}(X_t, \epsilon^+_{t-s,p}) = \lim_{p\to\infty}(X_0, \epsilon^+_{-s,p})$$

Lemme 4.1. *La suite $\{\psi_s\}$ est de carré sommable et $\psi_0 = 1$.*

Démonstration. Remarquons, tout d'abord, que la relation $((X_t|\mathcal{H}^X_{t-1}), \epsilon_t) = 0$ entraîne que :

$$\psi_0 = \frac{(X_t, \epsilon_t)}{\sigma^2} = \frac{(X_t - (X_t|\mathcal{H}^X_{t-1}), \epsilon_t)}{\sigma^2} = 1$$

D'autre part, pour tout $s \geq 0$, la projection orthogonale de X_t sur $\mathcal{H}^{\epsilon}_{t,s} = \text{span}\{\epsilon_t, \epsilon_{t-1}, \cdots, \epsilon_{t-s+1}\}$ s'écrit, du fait de l'orthogonalité du processus d'innovation, $(X_t|\mathcal{H}^{\epsilon}_{t,s}) = \sum_{k=0}^{s-1} \psi_k \epsilon_{t-k}$. On en déduit que $\|(X_t|\mathcal{H}^{\epsilon}_{t,s})\|^2 = \sigma^2 \sum_{k=0}^{s-1} \psi_k^2$. On a alors d'après l'égalité de Pythagore (proposition 4.3) :

$$\|(X_t|\mathcal{H}^{\epsilon}_{t,s})\|^2 = \sigma^2 \sum_{k=0}^{s-1} \psi_k^2 = \|X_t\|^2 - \|X_t - (X_t|\mathcal{H}^{\epsilon}_{t,s})\|^2 \leq \|X_t\|^2$$

ce qui conclut la preuve. ■

La suite $(\psi_s)_{s\geq 0}$ étant de carré sommable, la suite $s \to X_{t,s} = \sum_{k=0}^{s} \psi_k \epsilon_{t-k}$ est, pour t fixé, une suite de Cauchy dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Elle admet donc, quand $s \to \infty$, une limite que nous notons :

$$U_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k}$$

et qui est un processus stationnaire au second-ordre. On a, en effet :

$$\mathbb{E}[U_t] = (U_t, 1) = \lim_{s\to\infty} \sum_{k=0}^{s} \psi_k (\epsilon_{t-k}, 1) = 0$$

et

$$\mathbb{E}[U_{t+\tau} U_t] = (U_{t+\tau}, U_t) = \lim_{s\to\infty} \left(\sum_{k=0}^{s} \psi_k \epsilon_{t+\tau-k}, \sum_{k=0}^{s} \psi_k \epsilon_{t-k} \right) = \lim_{s\to\infty} \left(\sum_{k=0}^{s} \psi_k \epsilon_{\tau-k}, \sum_{k=0}^{s} \psi_k \epsilon_{-k} \right)$$

qui est indépendant de t.
Le théorème suivant, connu sous le nom de *décomposition de Wold*, est vraisemblablement le résultat le plus important de la théorie des processus stationnaires au second-ordre.

Théorème 4.5 (Décomposition de Wold)**.** *Soit X_t un processus stationnaire au second ordre et ϵ_t son processus d'innovation. On suppose que X_t est un processus régulier ($\sigma^2 = \|\epsilon_t\|^2 \neq 0$). On note $U_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k}$ où $\psi_k = (X_t, \epsilon_{t-k})/\sigma^2$. Alors il existe un processus V_t tel que :*

$$X_t = U_t + V_t, \tag{4.37}$$

et tel que :

(i). pour tout (t, s), $(V_t, \epsilon_s) = 0$, qui implique que $(V_t, U_s) = 0$,

(ii). $V_t = (X_t|\mathcal{H}^X_{-\infty})$ est la projection orthogonale de X_t sur $\mathcal{H}^X_{-\infty} = \bigcap_{t=-\infty}^{\infty} \mathcal{H}^X_t$,

(iii). U_t est un processus régulier et $\epsilon_t = U_t - (U_t|\mathcal{H}^U_{t-1})$ est l'innovation de U_t. De plus, $\mathcal{H}^{\epsilon}_t = \mathcal{H}^U_t$.

(iv). V_t est un processus déterministe et $\mathcal{H}^V_t = \mathcal{H}^X_{-\infty}$.

Démonstration. Elle est donnée en fin de chapitre. ■

Un processus $\{X_t\}$ tel que $\mathcal{H}^X_{-\infty} = \{0\}$ est dit *purement non déterministe*. Pour un tel processus la partie déterministe de la décomposition de Wold est identiquement nulle. Par exemple, le processus régulier U_t de la décomposition de Wold est purement non déterministe. En effet, en appliquant la décomposition de Wold au processus U_t on a, pour tout t, $U_t = U_t + V_t$ avec $V_t = 0$ et donc, d'après le point (iv), $\mathcal{H}^U_{-\infty} = \{0\}$. Le théorème de Wold permet donc de décomposer tout processus stationnaire au second-ordre sous la forme d'une somme de deux processus orthogonaux, le premier étant *purement non déterministe* et le second étant *déterministe*. La partie purement non-déterministe s'exprime comme le filtrage d'un bruit blanc par un filtre linéaire invariant dans le temps de réponse impulsionnelle $\{\psi_k\}$ *causale* ($\psi_k = 0$ pour $k < 0$) et de carré sommable (pas nécessairement de module sommable).

Exemple 4.15 : Processus MA(1)
Soit $\{Z_t\}$ un bruit blanc et soit le processus $X_t = Z_t + \theta_1 Z_{t-1}$. Remarquons que, par construction, $\mathcal{H}^X_t \subseteq \mathcal{H}^Z_t$ mais que l'inclusion réciproque nést pas nécessairement vérifiée. Montrons par contre que, pour $|\theta_1| < 1$, nous avons effectivement $\mathcal{H}^X_t = \mathcal{H}^Z_t$. En effet, en réitérant p fois l'équation $X_t = Z_t + \theta_1 Z_{t-1}$ et en résolvant par rapport à Z_t, nous obtenons :

$$Z_t = X_t - \theta_1 X_{t-1} + \theta_1^2 X_{t-2} + \cdots + (-1)^p \theta_1^p X_{t-p} - (-1)^p \theta_1^{p+1} Z_{t-p}$$

En prenant la limite en p, nous en déduisons que, si $|\theta_1| < 1$, alors :

$$Z_t = \sum_{k=0}^{\infty} (-\theta_1)^k X_{t-k}$$

ce qui montre que $\mathcal{H}^Z_t \subset \mathcal{H}^X_t$ et donc que $\mathcal{H}^X_t = \mathcal{H}^Z_t$. Dans ce cas, nous pouvons écrire :

$$(X_t|\mathcal{H}^X_{t-1}) = (Z_t|\mathcal{H}^X_{t-1}) + \theta_1(Z_{t-1}|\mathcal{H}^X_{t-1}) = (Z_t|\mathcal{H}^Z_{t-1}) + \theta_1(Z_{t-1}|\mathcal{H}^Z_{t-1}) = 0 + \theta_1 Z_{t-1}$$

en remarquant que $(Z_t|\mathcal{H}^Z_{t-1}) = 0$ car Z_t est un bruit blanc. On en déduit que $X_t - (X_t|\mathcal{H}^X_{t-1}) = X_t - \theta_1 Z_{t-1} = Z_t$. Par conséquent, lorsque $|\theta_1| < 1$, le processus Z_t est l'innovation du processus X_t. Notons que X_t est purement non déterministe *et que les coefficients de la décomposition de Wold sont simplement donnés par $\psi_0 = 1$, $\psi_1 = \theta$, et $\psi_k = 0$ pour $k > 1$.*

4.7 Preuves des théorèmes 4.2, 4.4 et 4.5

Théorème 4.2. *Soit $\mathcal{E}$ est un sous-espace fermé d'un espace de Hilbert $\mathcal{H}$ et soit x un élément quelconque de $\mathcal{H}$, alors :*

(i). il existe un unique élément $\hat{x} \in \mathcal{E}$ tel que :

$$\|x - \hat{x}\| = \inf_{w \in \mathcal{E}} \|x - w\|$$

(ii). $\hat{x} \in \mathcal{E}$ et $\|x - \hat{x}\| = \inf_{w\in\mathcal{E}} \|x - w\|$ si et seulement si $\hat{x} \in \mathcal{E}$ et $x - \hat{x} \perp \mathcal{E}$.

Démonstration. (i). Soit $x \in \mathcal{H}$. On note $h = \inf_{w\in\mathcal{E}} \|x-w\| \geq 0$. Alors il existe une suite $w_1, w_2, \cdots$, de vecteurs de $\mathcal{E}$ tels que :

$$\lim_{m\to+\infty} \|x - w_m\|^2 = h^2 \geq 0 \tag{4.38}$$

L'identité du parallélogramme, $\|a-b\|^2+\|a+b\|^2 = 2\|a\|^2+2\|b\|^2$ avec $a = w_m-x$ et $b = w_n-x$, montre que :

$$\|w_m - w_n\|^2 + \|w_m + w_n - 2x\|^2 = 2\|w_m - x\|^2 + 2\|w_n - x\|^2$$

Comme $(w_m + w_n)/2 \in \mathcal{E}$, nous avons $\|w_m + w_n - 2x\|^2 = 4\|(w_m + w_n)/2 - x\|^2 \geq 4h^2$. D'après 4.38, pour tout $\epsilon > 0$,il existe N tel que et $\forall m, n > N$:

$$\|w_m - w_n\|^2 \leq 2(h^2 + \epsilon) + 2(h^2 + \epsilon) - 4h^2 = 4\epsilon.$$

qui montre que w_n est une suite de Cauchy et donc que w_n tend vers une limite dans $\mathcal{E}$, puisque l'espace $\mathcal{E}$ est fermé. On note y cette limite. On en déduit, par continuité de la norme, que $\|y - x\| = h$. Montrons que cet élément est unique. Supposons qu'il existe un autre élément $z \in \mathcal{E}$ tel que $\|x - z\|^2 = \|x - y\|^2 = h^2$. Alors l'identité du parallélogramme donne :

$$0 \leq \|y - z\|^2 = -4\|(y + z)/2 - x\|^2 + 2\|x - y\|^2 + 2\|x - z\|^2 \leq -4h^2 + 2h^2 + 2h^2 = 0$$

où nous avons utilisé que $(y + z)/2 \in \mathcal{E}$ et que $\|(y + z)/2 - x\|^2 \geq h^2$. Il s'en suit que $y = z$. $\hat{x}$ est appelé la projection orthogonale de x sur $\mathcal{E}$.

(ii). Soit $\hat{x}$ la projection orthogonale de x sur $\mathcal{E}$. Alors, si il existe $u \in \mathcal{E}$ tel que $x - u \perp \mathcal{E}$, on peut écrire :

$$\begin{aligned} \|x - \hat{x}\|^2 &= (x - u + u - \hat{x}, x - u + u - \hat{x}) = \|x - u\|^2 + \|u - \hat{x}\|^2 + 2(u - \hat{x}, x - u) \\ &= \|x - u\|^2 + \|u - \hat{x}\|^2 + 0 \geq \|x - u\|^2 \end{aligned}$$

et donc $u = \hat{x}$. Réciproquement supposons que $u \in \mathcal{E}$ et $x - u \not\perp \mathcal{E}$. Alors choisissons $y \in \mathcal{E}$ tel que $\|y\| = 1$ et tel que $c = (x - u, y) \neq 0$ et notons $\tilde{x} = u + cy \in \mathcal{E}$. On a :

$$\begin{aligned} \|x - \tilde{x}\|^2 &= (x - u + u - \tilde{x}, x - u + u - \tilde{x}) = \|x - u\|^2 + \|u - \tilde{x}\|^2 + 2(u - \tilde{x}, x - u) \\ &= \|x - u\|^2 + c^2 - 2c(y, x - u) = \|x - u\|^2 - c^2 < \|x - u\|^2 \end{aligned}$$

Par conséquent $\tilde{x} \in \mathcal{E}$ est strictement plus proche de x que ne l'est u.

■

Théorème 4.4. *Soit le processus $\{X_t\}$ régulier. Alors, pour tout p, $\phi_p(z) \neq 0$ pour $|z| \leq 1$. Tous les zéros des polynômes prédicteurs sont à l'extérieur du cercle unité.*

Démonstration. Nous allons tout d'abord montrer que le prédicteur optimal n'a pas de racines sur le cercle unité. Raisonnons par contradiction. Supposons que le polynôme $\phi_p(z)$ ait deux racines complexes conjuguées, de la forme $\exp(\pm i\pi\theta)$, sur le cercle unité. (on traite de façon similaire le cas de racines réelles, $\theta = 0$ ou π). Nous pouvons écrire :

$$\phi_p(z) = \phi_p^*(z)(1 - 2\cos(\theta)z + z^2)$$

On note $\bar{\nu}_X(d\lambda) = \nu_X(d\lambda)|\phi_p^*(e^{-i\lambda})|^2$. $\bar{\nu}_X$ est une mesure positive sur $[-\pi, \pi]$ de masse finie. On note $\bar{\gamma}(\tau)$ la suite des coefficients de Fourier associés à $\bar{\nu}_X$:

$$\bar{\gamma}(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\tau\lambda}\bar{\nu}_X(d\lambda)$$

Nous avons donc :

$$\sigma_p^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}(1 - 2\cos(\theta)e^{-i\lambda} + e^{-2i\lambda})\bar{\nu}_X(d\lambda) = \inf_{\psi\in\mathcal{P}_2}\frac{1}{2\pi}\int_{-\pi}^{\pi}|1 + \psi_1 e^{-i\lambda} + \psi_2 e^{-2i\lambda}|^2\bar{\nu}_X(d\lambda).$$

Comme on l'a dit (page 67), la minimisation de σ_p^2 par rapport à ψ_1 et ψ_2 est équivalent à la résolution des équations de Yule-Walker *à l'ordre* $p = 2$ pour la suite des covariances $\bar{\gamma}(h)$. Par conséquent la suite des coefficients $\{1, -2\cos(\theta), 1\}$ doit vérifier l'équation :

$$\begin{bmatrix} \bar{\gamma}(0) & \bar{\gamma}(1) & \bar{\gamma}(2) \\ \bar{\gamma}(1) & \bar{\gamma}(0) & \bar{\gamma}(1) \\ \bar{\gamma}(2) & \bar{\gamma}(1) & \bar{\gamma}(0) \end{bmatrix}\begin{bmatrix} 1 \\ -2\cos(\theta) \\ 1 \end{bmatrix} = \begin{bmatrix} \sigma_p^2 \\ 0 \\ 0 \end{bmatrix}$$

De cette équation il s'en suit (les première et troisième lignes sont égales) que $\sigma_p^2 = 0$. Ce qui est contraire à l'hypothèse que le processus est régulier.
Démontrons maintenant que les racines des polynômes prédicteurs sont toutes *strictement à l'extérieur du cercle unité*. Raisonnons encore par l'absurde. Supposons que le polynôme prédicteur à l'ordre p ait m racines $\{a_k, |a_k| < 1, 1 \leq k \leq m\}$ à l'intérieur du cercle unité et $(p-m)$ racines $\{b_\ell, |b_\ell| > 1, 1 \leq \ell \leq p-m\}$ à l'extérieur du cercle unité. Le polynôme prédicteur à l'ordre p s'écrit donc :

$$\phi_p(z) = \prod_{k=1}^{m}(1 - a_k^{-1}z)\prod_{\ell=1}^{p-m}(1 - b_\ell^{-1}z)$$

Considérons alors le polynôme :

$$\bar{\phi}_p(z) = \prod_{k=1}^{m}(1 - a_k^{*}z)\prod_{\ell=1}^{p-m}(1 - b_\ell^{-1}z)$$

Il a d'une part toutes ses racines strictement à l'extérieur du cercle unité et d'autre part il vérifie $|\bar{\phi}_p(e^{-i\lambda})|^2 < |\phi_p(e^{-i\lambda})|^2$. On a en effet $|1-a_k^* e^{-i\lambda}| = |1-a_k e^{i\lambda}| = |a_k||1-a_k^{-1}e^{-i\lambda}|$ et donc $|\bar{\phi}_p(e^{-i\lambda})|^2 =$

$\left(\prod_{k=1}^{m}|a_k|^2\right)|\phi_p(e^{-i\lambda})|^2$, ce qui démontre le résultat annoncé compte tenu du fait que $|a_k|<1$. On en déduit alors que :

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|\bar{\phi}_p(e^{-i\lambda})|^2\nu_X(d\lambda)<\sigma_p^2$$

ce qui contredit que $\phi_p(z)=\inf_{\psi\in\mathcal{P}_p}(2\pi)^{-1}\int_{-\pi}^{\pi}|\psi(e^{-i\lambda})|^2\nu_X(d\lambda)$. ■

Théorème 4.5. *Soit X_t un processus stationnaire au second ordre et ϵ_t son processus d'innovation. On suppose que X_t est un processus régulier ($\sigma^2=\|\epsilon_t\|^2\neq 0$). On note $U_t=\sum_{k=0}^{\infty}\psi_k\epsilon_{t-k}$ où $\psi_k=(X_t,\epsilon_{t-k})/\sigma^2$. Alors il existe un processus V_t tel que :*

$$X_t=U_t+V_t, \tag{4.39}$$

et tel que :

(i). pour tout (t,s), $(V_t,\epsilon_s)=0$, qui implique que $(V_t,U_s)=0$,

(ii). $V_t=(X_t|\mathcal{H}_{-\infty}^X)$ est la projection orthogonale de X_t sur $\mathcal{H}_{-\infty}^X=\bigcap_{t=-\infty}^{\infty}\mathcal{H}_t^X$,

(iii). U_t est un processus régulier et $\epsilon_t=U_t-(U_t|\mathcal{H}_{t-1}^U)$ est l'innovation de U_t. De plus, $\mathcal{H}_t^\epsilon=\mathcal{H}_t^U$.

(iv). V_t est un processus déterministe et $\mathcal{H}_t^V=\mathcal{H}_{-\infty}^X$.

Démonstration. (i). Par définition, $V_t=X_t-\sum_{k=0}^{\infty}\psi_k\epsilon_{t-k}\in\mathcal{H}_t^X$. Pour $s>t$, $\epsilon_s\perp\mathcal{H}_t^X$, et donc $(V_t,\epsilon_s)=0$. Pour $s\leq t$, $(V_t,\epsilon_s)=(X_t,\epsilon_s)-\psi_{t-s}\sigma^2$ qui est égal à 0 par définition de ψ_k.

(ii). Montrons tout d'abord que $V_t\in\mathcal{H}_{-\infty}^X$. La preuve se fait par récurrence. Nous avons $V_t\in\mathcal{H}_t^X$ et $V_t\perp\epsilon_t$ (d'après la propriété précédente). Comme $\mathcal{H}_t^X=\mathcal{H}_{t-1}^X\oplus\text{span}\{\epsilon_t\}$, on en déduit que $V_t\in\mathcal{H}_{t-1}^X$. Supposons à présent que $V_t\in\mathcal{H}_{t-s}^X$, pour $s\geq 0$. Comme $V_t\perp\epsilon_{t-s}$ et que $\mathcal{H}_{t-s}^X=\mathcal{H}_{t-s-1}^X\oplus\text{span}\{\epsilon_{t-s}\}$, nous avons $V_t\in\mathcal{H}_{t-s-1}^X$. On a donc $V_t\in\mathcal{H}_{-\infty}^X=\bigcap_{s=-\infty}^{\infty}\mathcal{H}_s^X$. Il reste à montrer que $(X_t-V_t)=\sum_{k=0}^{\infty}\psi_k\epsilon_{t-k}$ est orthogonal à $\mathcal{H}_{-\infty}^X$. Pour cela considérons $Y\in\mathcal{H}_{-\infty}^X$. Nous avons :

$$(X_t-V_t,Y)=\left(\sum_{k=0}^{\infty}\psi_k\epsilon_{t-k},Y\right)=\lim_{s\to+\infty}\sum_{k=0}^{s}\psi_k(\epsilon_{t-k},Y)$$

Mais, par définition, $Y\in\mathcal{H}_{-\infty}^X$ implique que, pour tout t, $Y\in\mathcal{H}_t^X$. Comme $\epsilon_{t-k}\perp\mathcal{H}_{t-s-1}^X$ pour $0\leq k\leq s$, nous avons $\sum_{k=0}^{s}\psi_k(\epsilon_{t-k},Y)=0$. Et donc, pour tout $Y\in\mathcal{H}_{-\infty}^X$, on a :

$$(X_t-V_t,Y)=(U_t,Y)=0 \tag{4.40}$$

(iii). Notons que (4.40) implique que, pour tout t, $U_t\perp\mathcal{H}_{-\infty}^X$ et donc $\mathcal{H}_t^U=\overline{\text{span}}\{U_s,s\leq t\}\perp\mathcal{H}_{-\infty}^X$. On peut alors poser $\mathcal{L}_t=\mathcal{H}_t^U\oplus\mathcal{H}_{-\infty}^X$. La décomposition $X_t=U_t+V_t$ et la propriété précédente ($V_t=(X_t|\mathcal{H}_{-\infty}^X)$) impliquent que, pour tout t, $\mathcal{H}_t^X\subset\mathcal{L}_t$, et donc $\epsilon_t\in\mathcal{L}_t$. Comme, pour tout t, $\epsilon_t\perp\mathcal{H}_{t-u}$ pour tout $u\geq 0$, $\epsilon_t\perp Y$ pour tout $Y\in\mathcal{H}_{-\infty}^X$, puisque, en particulier, $Y\in\mathcal{H}_{t-u}$. Nous avons $\epsilon_t\perp\mathcal{H}_{-\infty}^X$. Et donc $\epsilon_t\in\mathcal{H}_t^U$. Cela entraîne que $\sum_{k=1}^{\infty}\psi_k\epsilon_{t-k}\in\mathcal{H}_{t-1}^U$. Notons que $\sum_{k=1}^{\infty}\psi_k\epsilon_{t-k}=U_t-\epsilon_t$ ($\psi_0=1$). Par conséquent, pour tout $Y\in\mathcal{H}_{t-1}^U$ on a :

$$\left(U_t-\sum_{k=1}^{\infty}\psi_k\epsilon_{t-k},Y\right)=(\epsilon_t,Y)=0$$

Cela implique que $\sum_{k=1}^{\infty} \psi_k \epsilon_{t-k}$ est la projection orthogonale de U_t sur $\mathcal{H}_{t-1}^U$ et donc que :

$$\epsilon_t = U_t - (U_t | \mathcal{H}_{t-1}^U)$$

Cela signifie que ϵ_t est le processus d'innovation de U_t. Comme, par hypothèse, $\sigma^2 = \|\epsilon_t\|^2 \neq 0$, U_t est donc régulier. Remarquons que, comme $\epsilon_t \in \mathcal{H}_t^U$, nous avons $\mathcal{H}_t^\epsilon \subset \mathcal{H}_t^U$. Comme, par construction, $\mathcal{H}_t^U \subset \mathcal{H}_t^\epsilon$, nous avons $\mathcal{H}_t^U = \mathcal{H}_t^\epsilon$.

(iv). Montrons tout d'abord que, pour tout t, on a :

$$\mathcal{H}_t^V = \overline{\text{span}}\{V_s, s \leq t\} = \mathcal{H}_{-\infty}^X \tag{4.41}$$

Pour tout t, $V_t \in \mathcal{H}_{-\infty}^X$ et donc $\mathcal{H}_t^V \subseteq \mathcal{H}_{-\infty}^X$. D'un autre côté, puisque $X_t = \sum_{k=0}^{+\infty} \psi_k \epsilon_{t-k} + V_t$, $\mathcal{H}_t^X = \mathcal{H}_t^\epsilon \oplus \mathcal{H}_t^V$. Et donc, quel que soit $Y \in \mathcal{H}_{-\infty}^X$, alors $Y \in \mathcal{H}_{s-1}^X$ pour tout s, de telle sorte que $(Y, \epsilon_s) = 0$ et donc $Y \in \mathcal{H}_t^V$, ce qui implique que $\mathcal{H}_{-\infty}^X \subseteq \mathcal{H}_t^V$. Ce qui démontre (4.41). Partant de (4.41), on déduit que $(V_t | \mathcal{H}_{t-1}^V) = (V_t | \mathcal{H}_{-\infty}^X) = (V_t | \mathcal{H}_t^V) = V_t$ et que $\|V_t - (V_t | \mathcal{H}_{t-1}^V)\|^2 = 0$: V_t est donc déterministe.

■

Chapitre 5

Estimation des processus ARMA

Dans ce chapitre nous nous intéressons aux problèmes de l'estimation des paramètres d'un processus ARMA(p, q) à partir d'une suite de n observations. Nous supposons que les données ont été préalablement traitées de façon à supprimer d'éventuelles tendances affine et/ou saisonnière. L'estimation des paramètres d'un processus ARMA(p, q) comprend aussi, en principe, l'estimation des ordres p et q. Ce problème est complexe et ne sera pas traité dans ce chapitre. Nous supposons donc que p et q sont connus et nous nous intéressons uniquement à l'estimation des paramètres $\{\phi_k; 1 \leq k \leq p\}$, $\{\theta_k; 1 \leq k \leq q\}$ et σ^2 intervenant dans l'équation récurrente définissant le processus (voir équation (1.39) chapitre 1). Dans le cas de l'estimation d'un processus AR(p), on verra que, pour obtenir de bons estimateurs de $\{\phi_k; 1 \leq k \leq p\}$ et de σ^2, il suffit de partir des $(p + 1)$ premiers coefficients d'autocovariance empirique et de résoudre les équations de Yule-Walker. Cela signifie que, quel que soit n, les observations n'interviennent, dans l'expression de l'estimateur, que par un nombre fixé, égal à $p + 1$, de valeurs de la covariance empirique :

$$\hat{\gamma}_n(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_{t+h} - \hat{\mu}_n)(X_t - \hat{\mu}_n)$$

où $0 \leq h \leq p$ et $\hat{\mu}_n = n^{-1} \sum_{t=1}^{n} X_t$. Cela n'est plus vrai pour un processus ARMA(p, q) avec $q > 1$ (comme par exemple pour un MA(q)) : la construction de bons estimateurs ne peut se faire avec un nombre fixé (indépendant de n) de valeurs de la suite des covariances empiriques. Cela rend plus complexe l'estimation ARMA. Il s'en suit que, contrairement au cas de l'estimation AR, il existe de nombreuses méthodes. La solution retenue en pratique établit un compromis entre biais, variance et complexité de mise en œuvre.

5.1 Estimation AR

Nous avons établi, chapitre 1, une relation simple (équations (1.36) de Yule-Walker) entre les $(p+1)$ coefficients du modèle et les $(p+1)$ premiers coefficients d'autocovariance d'un processus AR(p) causal défini par l'équation récurrente :

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$$

En posant $\boldsymbol{\phi} = \begin{bmatrix}\phi_1 & \dots & \phi_1\end{bmatrix}^T$, $\boldsymbol{\gamma}_p = \begin{bmatrix}\gamma(1) & \dots & \gamma(p)\end{bmatrix}^T$ et :

$$\Gamma_p = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-1) \\ \vdots & & \ddots & \\ \gamma(p) & \gamma(p-1) & \cdots & \gamma(0) \end{bmatrix}$$

les équations de Yule-Walker ont pour expression matricielle :

$$\begin{aligned} \Gamma_p \boldsymbol{\phi} &= \boldsymbol{\gamma}_p \\ \sigma^2 &= \gamma(0) - \boldsymbol{\phi}^T \boldsymbol{\gamma}_p \end{aligned} \tag{5.1}$$

En substituant, dans ces relations, les covariances $\gamma(h)$ par les covariances empiriques $\hat{\gamma}(h)$, on obtient un système linéaire qui fournit les estimateurs $\hat{\boldsymbol{\phi}}_n$ et $\hat{\sigma}_n^2$ comme solution de :

$$\hat{\Gamma}_p \hat{\boldsymbol{\phi}}_n = \hat{\boldsymbol{\gamma}}_p \tag{5.2}$$

$$\hat{\sigma}_n^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}_n^T \hat{\boldsymbol{\gamma}}_p \tag{5.3}$$

On a vu chapitre 2 que, si $\hat{\gamma}(0) > 0$, alors $\hat{\Gamma}_p$ est de rang plein. En divisant alors les deux membres de $\hat{\Gamma}_p \hat{\boldsymbol{\phi}}_n = \hat{\boldsymbol{\gamma}}_p$ par $\hat{\gamma}(0)$ et en introduisant l'autocorrélation empirique $\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$, on aboutit aux deux équations :

$$\hat{\boldsymbol{\phi}}_n = \hat{C}_p^{-1} \hat{\boldsymbol{\rho}}_p \tag{5.4}$$

$$\hat{\sigma}_n^2 = \hat{\gamma}(0)(1 - \hat{\boldsymbol{\rho}}_p^T \hat{C}_p^{-1} \hat{\boldsymbol{\rho}}_p) \tag{5.5}$$

où $\hat{\boldsymbol{\rho}}_p = \begin{bmatrix}\hat{\rho}(1) & \dots & \hat{\rho}(p)\end{bmatrix}^T$ et :

$$\hat{C}_p = \begin{bmatrix} \hat{\rho}(0) & \hat{\rho}(1) & \cdots & \hat{\rho}(p) \\ \hat{\rho}(1) & \hat{\rho}(0) & \cdots & \hat{\rho}(p-1) \\ \vdots & & \ddots & \\ \hat{\rho}(p) & \hat{\rho}(p-1) & \cdots & \hat{\rho}(0) \end{bmatrix}$$

Le fait que la matrice $\hat{R}_p$ (comme la matrice $\hat{C}_p$) soit, par construction, de Toëplitz et de type défini positif (voir théorème 4.4 chapitre 4) implique que les coefficients estimés $\hat{\phi}_p$ sont tels que le polynôme $\hat{\phi}(z) = 1 - \sum_{k=1}^{p} \hat{\phi}_k z^k$ a toutes ses racines strictement à l'extérieur du cercle unité : cette façon de procéder aboutit donc nécessairement à un processus AR(p) causal. Ses $(p+1)$ premiers coefficients de covariance coïncident alors avec les coefficients de covariance empiriques. La méthode qui consiste pour estimer des paramètres à substituer, dans une relation telle que (5.1), les moments par des estimateurs consistants, porte le nom de *méthode des moments*. En règle générale, elle conduit à des estimateurs des paramètres qui sont moins efficaces que ceux obtenus par la *méthode des moindres carrés* ou encore par la *méthode du maximum de vraisemblance*. Cependant, dans le cas d'un modèle AR(p) *gaussien*, on montre que les estimateurs $\hat{\boldsymbol{\phi}}$ et $\hat{\sigma}^2$, donnés par (5.2) et (5.3), ont le même comportement asymptotique, quand n tend vers l'infini, que ceux du maximum de vraisemblance. Nous avons vu,

chapitre 4 exemple 4.13, que les coefficients de l'équation récurrente d'un AR(p) causal sont directement reliés aux coefficients du meilleur prédicteur linéaire donnant X_t à partir de ses valeurs passées : plus précisément, *pour tout* $m \geq p$, la suite des m coefficients de prédiction $\boldsymbol{\phi}_m = \{\phi_{1,m}, \ldots, \phi_{m,m}\}$ coïncide avec $\{\phi_1, \ldots, \phi_p, 0, \ldots, 0\}$. Par conséquent, pour un AR(p) causal, l'algorithme de Levinson-Durbin fournit une résolution rapide aux équations de Yule-Walker. On voit aussi que, si, ne connaissant pas la vraie valeur de p, on prend un ordre $m > p$, on peut espérer que les $(m-p)$ derniers coefficients de prédiction seront de faibles valeurs.
Les théorèmes suivants précisent le comportement asymptotique de la suite ϕ et permettent alors de construire des intervalles de confiance ou de fournir des tests d'hypothèse.

Théorème 5.1. *Soit X_t un processus AR(p) causal où $Z_t \sim \mathrm{IID}(0, \sigma^2)$ et soit un échantillon $\{X_1, \ldots, X_n\}$ de taille n. On note $\hat{\boldsymbol{\phi}}_n = \hat{C}_p^{-1}\hat{\boldsymbol{\rho}}_p$ et $\hat{\sigma}_n^2 = \hat{\gamma}(0)(1 - \hat{\boldsymbol{\rho}}_p^T \hat{C}_p^{-1}\hat{\boldsymbol{\rho}}_p)$. Alors, quand $n \to \infty$, on a :*

$$\begin{cases} \hat{\sigma}_n^2 \to_{\mathbb{P}} \sigma^2 \\ \sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}) \to_d \mathcal{N}(0, \sigma^2 \Gamma_p^{-1}) \end{cases} \tag{5.6}$$

Théorème 5.2. *Soit X_t un processus AR(p) causal où $Z_t \sim \mathrm{IID}(0, \sigma^2)$ et soit un échantillon $\{X_1, \ldots, X_n\}$ de taille n. On note $\hat{\boldsymbol{\phi}}_n = \hat{C}_m^{-1}\hat{\boldsymbol{\rho}}_m$ où $m > p$. Alors, quand $n \to \infty$, on a :*

$$\sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_m) \to_d \mathcal{N}(0, \sigma^2 \Gamma_m^{-1}) \tag{5.7}$$

où $\boldsymbol{\phi}_m = \{\phi_1, \ldots, \phi_p, 0, \ldots, 0\}$ est la suite du meilleur prédicteur linéaire de X_t en fonction de $\{X_{t-1}, \ldots, X_{t-m}\}$.

En particulier, le m-ème coefficient de corrélation partielle $\hat{k}_n(m) = \hat{\phi}_{m,m}$ vérifie :

$$\sqrt{n}\,\hat{k}_n(m) \to_d \mathcal{N}(0, 1) \tag{5.8}$$

On en déduit le résultat pratique suivant : si un modèle autorégressive est approprié pour une suite d'observations, il doit y avoir une valeur m à partir de laquelle les valeurs observées de $\hat{k}_n(m)$ sont compatibles avec la distribution $\mathcal{N}(0, 1/n)$. En particulier si m est supérieur à l'ordre du modèle, $\hat{k}_n(m)$ doit être compris entre $\pm 1.96/\sqrt{n}$ avec une probabilité proche de 95%. Ce résultat suggère d'utiliser comme estimateur de p la plus petite valeur r au delà de laquelle $|\hat{k}_n(m)| < 1.96/\sqrt{n}$ pour tout $m > r$. Cette valeur peut servir de valeur initiale à des algorithmes plus performants d'estimation de p.

Exemple 5.1 : Suite des coefficients de réflexion d'un processus AR(2)
Le théorème 5.2 montre que le coefficient de réflexion $\phi_{m,m}$ pour $m > 1$ se comporte comme une variable aléatoire gaussienne de moyenne nulle et de variance de l'ordre de $1/n$. Nous avons représenté figure 5.1 les suites, obtenues au cours de 7 simulations, de $\phi_{m,m}$ en fonction de m pour un échantillon AR(2) de longueur $n = 500$. Les valeurs des paramètres sont $\phi_1 = 1.6$, $\phi_2 = -0.9$ et $\sigma^2 = 1$. Le calcul théorique donne $\phi_{1,1} = 0.8$, $\phi_{2,2} = -0.9$ et, pour $m \geq 2$, $\phi_{m,m} = 0$. Nous avons aussi représenté l'intervalle de confiance à 95% pour $m \geq 2$.

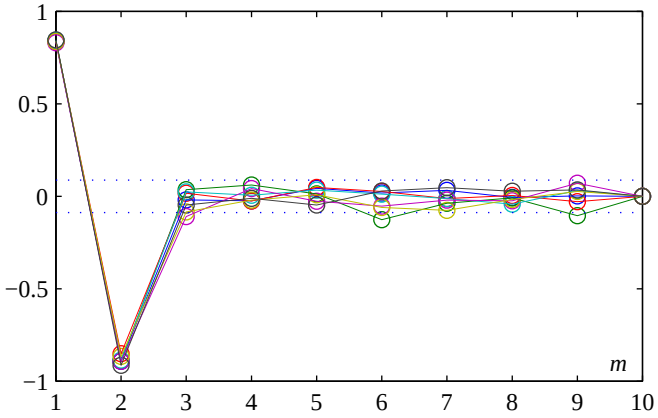

FIG. 5.1 – *Suites, obtenues au cours de 7 simulations, des coefficients de réflexion en fonction de m, pour un échantillon de longueur $n = 500$ d'un processus $AR(2)$ défini par $\phi_1 = 1.6$, $\phi_2 = -0.9$ et $\sigma^2 = 1$.*

Méthode du maximum de vraisemblance

Considérons un AR(p) causal où $Z_t \sim \text{IID}(0, \sigma^2)$ dont la loi de probabilité a pour densité $p_Z(z; \boldsymbol{\eta})$ où $\boldsymbol{\eta}$ désigne un paramètre vectoriel à estimer. Soit $(X_1, \ldots, X_n)$ une observation de taille n. On peut alors écrire :

$$\begin{cases} X_{p+1} = \phi_1 X_p + \cdots + \phi_p X_1 + Z_{p+1} \\ \quad \vdots \\ X_n = \phi_1 X_{n-1} + \cdots + \phi_p X_{n-p} + Z_n \end{cases}$$

Rappelons que, pour un AR(p) causal ($\phi(z) \neq 0$ pour $|z| \leq 1$), les variables aléatoires $\{X_1, \ldots, X_p\}$ appartiennent à $\mathcal{H}_p^Z = \overline{\text{span}}\{Z_s; s \leq p\}$. Par conséquent, le vecteur aléatoire $[X_1, \ldots, X_p]$ est une fonction mesurable de $\{Z_s; s \leq p\}$. Comme les variables aléatoires Z_t sont supposées (conjointement) indépendantes, les variables aléatoires $\{X_1, \ldots, X_p\}$ sont indépendantes des variables aléatoires $\{Z_{p+1}, \ldots, Z_n\}$. On en déduit que la loi conditionnelle de $(X_{p+1}, \ldots, X_n)$ par rapport à $(X_1, \ldots, X_p)$ a pour log-densité :

$$\log p_{X_{p+1},\ldots,X_n|X_1,\ldots,X_p}(x_1, \ldots, x_n; \boldsymbol{\theta}) = \sum_{k=p+1}^{n} \log p_Z\left(x_k - \boldsymbol{\phi}^T \mathbf{x}_k; \boldsymbol{\eta}\right) \tag{5.9}$$

où $\mathbf{x}_k = [x_k \quad \ldots \quad x_{k-p+1}]^T$, $\boldsymbol{\phi} = [\phi_1 \quad \ldots \quad \phi_p]^T$ et $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\eta})$. L'estimateur du maximum de vraisemblance consiste à trouver, pour une suite d'observations $(x_1, \ldots, x_n)$, la valeur de $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\eta})$ qui maximise (5.9). Dans ce contexte, la fonction (5.9) à maximiser s'appelle la *log-vraisemblance*. D'où le nom de l'estimateur obtenu. Dans le cas où la loi de Z_t est gaussienne, $2 \log p_Z(z; \sigma^2) = -\log(2\pi\sigma^2) - z^2/\sigma^2$

et l'expression (5.9) s'écrit :

$$\begin{aligned}\log p_{X_{p+1},\ldots,X_n|X_1,\ldots,X_p}(x_1,\ldots,x_n;\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=p+1}^{n}(x_k - \boldsymbol{\phi}^T\mathbf{x}_k)^2 \\ &= -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{X} - \boldsymbol{\mathcal{X}}\boldsymbol{\phi}\|^2\end{aligned}$$

où $\mathbf{X} = [x_{p+1} \quad \ldots \quad x_n]^T$ et :

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} x_p & \cdots & x_1 \\ x_{p+1} & \cdots & x_2 \\ & \vdots & \\ x_{n-1} & \cdots & x_{n-p} \end{bmatrix}$$

En annulant le gradient de la log-vraisemblance par rapport à $\boldsymbol{\phi}$, il vient $\boldsymbol{\mathcal{X}}^T(\mathbf{X} - \boldsymbol{\mathcal{X}}\hat{\boldsymbol{\phi}}) = 0$ dont on tire $\hat{\boldsymbol{\phi}} = (\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}^T\mathbf{X}$ qui est l'estimateur des moindres carrés. On notera que, contrairement à la méthode de Yule-Walker, la matrice de type positif $\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}$, à inverser, n'a pas une structure de Toëplitz. La conséquence majeure est que la suite des coefficients $\{\hat{\phi}_k\}$ qui en sont déduits ne sont pas nécessairement ceux d'un AR causal. Il peut arriver que les zéros du polynôme $\hat{\phi}(z)$ associé soient à l'intérieur du cercle unité.
Dans le cas où la loi de Z_t n'est pas gaussienne, l'expression d'un estimateur du maximum de vraisemblance ne possède pas de forme simple et on doit, en général, faire appel à des techniques numériques.

5.2 Estimation MA

Nous avons vu que le modèle MA correspond à un filtre linéaire dont la fonction de transfert $\theta(z) = 1 + \sum_{k=1}^{q}\theta_k z^k$ est un polynôme en z. On rencontre cette modélisation pour les canaux de propagation comportant des trajets multiples (en nombre fini), chaque trajet introduisant un retard et/ou une atténuation. C'est, par exemple, le cas des canaux de communication en radio-mobile ou encore de certains canaux de propagation acoustique. Le problème majeur rencontré en modélisation MA est l'impossibilité de retrouver à partir des propriétés du second ordre les paramètres du modèle. En effet la densité spectrale d'un MA a pour expression :

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|1 + \sum_{k=1}^{q}\theta_k e^{-ik\lambda}\right|^2$$

Elle ne définit donc pas, de *manière unique*, un processus MA(q). Tous les processus MA(q) de fonction de transfert :

$$\theta'(z) = \theta(z)\prod_{s=1}^{m}\frac{1 - z_s^* z}{z - z_s}$$

où $\{z_s\}_{1\leq s\leq m\leq q}$ sont une sous-suite quelconque de m zéros de $\theta(z)$, ont même densité spectrale. En effet $(1-z_s^* e^{-i\lambda})/(e^{-i\lambda}-z_s)$ est de module égal à 1. Par conséquent, partant de $f(x)$, on peut construire plusieurs processus MA(q) suivant que l'on place un zéro à l'intérieur où à l'extérieur du cercle unité.

Nous avons vu théorème 1.10 que, parmi toutes ces solutions, celle qui a tous ses zéros à l'extérieur du cercle unité est *inversible* (on dit aussi que le processus est *à phase minimale*). Sous l'hypothèse que le processus MA(q) observé est inversible, le problème de la détermination des paramètres à partir de la suite des covariances a une solution unique. Malheureusement dans certaines situations pratiques, en particulier en communications numériques, l'hypothèse de phase minimale n'est pas vérifiée. Dans ce cas il faut faire appel à des statistiques d'ordre supérieur à 2 pour résoudre le problème. Notons que, dans le cas gaussien, il est donc impossible de résoudre le problème puisque, par définition, les moments de tout ordre d'une variable gaussienne sont fonction des moments d'ordre 2.
Dans la suite nous supposerons que le MA est inversible.

Exemple 5.2 : Estimation MA(1) : méthode des moments
Soit un processus MA(1) défini par $X_t = Z_t + \theta_1 Z_{t-1}$. On suppose que $|\theta_1| \leq 1$ et donc $\theta(z) = 1+\theta_1 z$ s'annule en $z_0 = 1/\theta_1$ qui est à l'extérieur du cercle unité. Le modèle est donc inversible. La fonction d'autocorrélation s'écrit :

$$\rho(h) = \begin{cases} \theta_1/(1+\theta_1^2) & si\ h = \pm 1 \\ 0 & si\ |h| \geq 2 \end{cases}$$

La méthode des moments consiste à substituer à $\rho(1)$ la corrélation empirique $\hat{\rho}_n(1)$ et à résoudre par rapport à θ_1. En supposant que $|\theta_1| < 1$, il vient :

$$\hat{\theta}_1 = \begin{cases} -1 & si\ \hat{\rho}_n(1) < -1/2 \\ (1-(1-4\hat{\rho}_n^2(1))^{1/2})/2\hat{\rho}(1) & si\ |\hat{\rho}_n(1)| \leq 1/2 \\ +1 & si\ \hat{\rho}(1) > 1/2 \end{cases}$$

Une fois $\hat{\theta}_1$ estimé, on obtient une estimation de σ^2 en utilisant, par exemple, l'expression de $\gamma(1)$ qui donne, par la méthode des moments, $\hat{\sigma}^2 = \hat{\theta}_1/\hat{\gamma}(1)$. Malheureusement cet estimateur est de performances inférieures à celles de l'estimateur du maximum de vraisemblance même dans le cas gaussien. De façon plus précise l'estimateur n'est pas même consistant. Le problème est que l'estimateur précédent est construit uniquement à partir du couple de statistiques $\hat{\rho}_n(0)$ et $\hat{\rho}_n(1)$. Or on montre que, quand n tend vers l'infini, il n'y a pas de statistiques de dimension finie qui soit suffisante. On peut alors envisager de trouver un estimateur du maximum de vraisemblance. Dans le cas où Z_t est un bruit blanc gaussien, la log-vraisemblance de l'observation a pour expression :

$$\log p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\sigma^2) = \\ -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log\det(C(\theta_1)) - \frac{1}{2\sigma^2}[x_1 \quad \ldots \quad x_n]C^{-1}(\theta_1)\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

où $\Gamma(\theta_1)$ de dimension $n \times n$ a pour expression :

$$C(\theta_1) = \begin{bmatrix} 1+\theta_1^2 & \theta_1 & 0 & \cdots & 0 \\ \theta_1 & 1+\theta_1^2 & \theta_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \cdots & \theta_1 & 1+\theta_1^2 & \theta_1 \\ 0 & \cdots & 0 & \theta_1 & 1+\theta_1^2 \end{bmatrix}$$

La maximisation par rapport à θ_1 et σ^2 ne conduit pas des expressions analytiques simples. Par contre nous verrons un algorithme récursif qui permet de déterminer cet estimateur.

Méthode de Durbin

La méthode proposée par Durbin s'appuie sur le fait qu'un processus MA(q), défini par $X_t = Z_t + \sum_{k=1}^{q} \theta_k Z_{t-k}$, peut être approché par un AR(p) suffisamment long. Plus précisément supposons que $\theta(z) \neq 0$ pour $|z| \leq 1$. On a vu que $\psi(z) = 1/\theta(z) = 1 - \sum_{k=1}^{\infty} \psi_k z^k$ où $\{\psi_k\}$ est une suite de module sommable et que $Z_t = X_t - \sum_{k=1}^{\infty} \psi_k X_{t-k}$. Mais, puisque $\theta(z)$ est continue, il existe $M > 0$ tel que, pour tout $|z| \leq 1$, on a $|\theta(z)| \leq M$ et donc $|\psi(z)| \geq 1/M = m > 0$. Posons $\psi_p(z) = 1 - \sum_{k=1}^{p} \psi_k z^k$. Alors il existe p suffisamment grand tel que, pour tout $|z| \leq 1$, $|\psi(z) - \psi_p(z)| < m/2$. On en déduit que $m \leq |\psi(z)| = |\psi(z) - \psi_p(z) + \psi_p(z)| \leq |\psi(z) - \psi_p(z)| + |\psi_p(z)| \leq m/2 + |\psi_p(z)|$ qui implique que $|\psi_p(z)| \geq m/2 > 0$. En conclusion, pour tout $|z| \leq 1$, il existe p suffisamment grand tel que $|\psi_p(z)| > 0$. On en déduit que le processus défini par l'équation récurrente $\tilde{X}_t = Z_t + \sum_{k=1}^{p} \psi_k X_{t-k}$ est un AR(p) causal. De plus $X_t - \tilde{X}_t = \sum_{k=p+1}^{\infty} \psi_k X_{t-k}$ et donc $\mathbb{E}\left[|X_t - \tilde{X}_t|^2\right] \leq \gamma(0) \left(\sum_{k=p+1}^{\infty} |\psi_k|\right)^2$ qui tend vers 0 quand p tend vers l'infini.
La méthode de Durbin, qui estime un MA(q) inversible comme un AR(p) causal long, comporte donc une première étape pour estimer les p coefficients $\{\psi_1, \ldots, \psi_p\}$ de prédiction linéaire, obtenus comme solution des équations de Yule-Walker. Il faut ensuite estimer la suite $\{\theta_k\}$. En principe on a $\psi(z)\theta(z) = (1 - \sum_{m=1}^{\infty} \psi_m z^m)(1 + \sum_{k=1}^{q} \theta_k z^k) = 1$. On en déduit que, pour tout $h \geq 1$:

$$\sum_{k=0}^{\min(h,q)} \phi_{h-k}\theta_k = 0$$

où $\theta_0 = \phi_0 = 1$ et $\phi_k = -\psi_k$ pour $k \geq 1$. En adoptant une approche de type moindres carrés, on peut alors minimiser la norme du vecteur $\mathbf{e}$ de composantes $\epsilon_h = \sum_{k=0}^{\min(h,q)} \hat{\phi}_{h-k}\hat{\theta}_k$ où $1 \leq h \leq p+q$. Ce qui s'écrit encore :

$$\begin{bmatrix} -\hat{\psi}_1 \\ -\hat{\psi}_2 \\ \vdots \\ -\hat{\psi}_{p-1} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\hat{\psi}_1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ -\hat{\psi}_p & & \ddots & 1 \\ 0 & \ddots & & -\hat{\psi}_1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\hat{\psi}_p \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_q \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{p+q} \end{bmatrix}$$

Avec des notations matricielles évidentes, cette équation peut encore s'écrire $\hat{\boldsymbol{\psi}} = -\hat{\Psi}\hat{\boldsymbol{\theta}} + \mathbf{e}$. La solution qui minimise $\mathbf{e}^T\mathbf{e}$ a pour expression :

$$\hat{\boldsymbol{\theta}} = -(\hat{\Psi}^T\hat{\Psi})^{-1}\hat{\Psi}^T\hat{\boldsymbol{\psi}} \tag{5.10}$$

On remarque que l'équation (5.10) a la même forme que la solution des équations de Yule-Walker en prenant pour suite des "observations" les $p+1$ quantités $\{\psi_0 = 1, -\hat{\psi}_1, \ldots, -\hat{\psi}_p\}$. L'*algorithme de Durbin*, qui estime un MA(q) à partir de n données, peut alors se résumer de la façon suivante :

- Choisir une valeur de p ($q \ll p \ll n$) et estimer les coefficients de l'AR(p) à partir des n observations.
- Estimer les coefficients de l'AR(q) à partir des p "observations" $\{1, -\hat{\psi}_1, \ldots, -\hat{\psi}_p\}$.

Dans la méthode de Durbin, qui estime un MA(q) comme un AR(p) long, se pose le problème du choix optimal de p. Ce problème ne sera pas traité ici de façon générale. Nous nous limiterons à l'exemple numérique qui suit et qui montre qu'il y a un compromis à trouver entre biais et variance. Remarquons à ce sujet que, plus les zéros de $\theta(z)$ sont proches du cercle unité, plus la valeur de p doit être choisie grande si on veut avoir une bonne précision et donc un biais faible. D'un autre côté, plus p est grand, plus la dispersion de l'estimateur est grande, du fait d'une "mauvaise" estimation de certains coefficients de covariance. Dans tous les cas la suite d'estimateurs n'est pas consistante. La méthode peut cependant fournir une bonne valeur d'initialisation pour des algorithmes plus complexes, comme celui du maximum de vraisemblance.

Exemple 5.3 : Estimation MA(1) : méthode de Durbin
Le tableau 5.1 donne la moyenne, la variance et le risque, estimés empiriquement à partir de 200 *réalisations, de l'estimateur de Durbin pour un processus MA*(1) *(où* $\theta_1 = 0.95$*) et pour différentes valeurs de p. La taille de l'échantillon est* $n = 300$*. On observe que, quand p augmente, la variance augmente, tandis que la moyenne et le risque passent par un minimum.*

p	20	40	70	120	250
biais	-0.1008	-0.0863	-0.0841	-0.0840	-0.0939
variance	0.0007	0.0009	0.0012	0.0016	0.0018
risque	0.0108	0.0083	0.0082	0.0087	0.0106

TAB. *5.1 – Biais, variance et risque empiriques de l'estimateur de Durbin pour un processus MA(1) pour différentes valeurs de p.*

Méthode des innovations partielles

Soit un processus MA(q) inversible. On note $(X_t|H^X_{t,p}) = \sum_{k=1}^{p} \psi_{k,p} X_{t-k}$ la prédiction linéaire optimale de X_t à partir de $\{X_{t-1}, \ldots, X_{t-p}\}$ et $Z_{p,t} = X_t - (X_t|H^X_{t,p})$ le processus d'innovation partielle. Nous avons vu chapitre 4 que, pour un processus stationnaire au second ordre (voir expression (4.33)), le processus d'innovation partielle tendait en moyenne quadratique, quand p tend vers l'infini, vers le processus d'innovation qui est précisément Z_t pour un MA(q) inversible. D'où l'idée de remplacer dans l'équation $X_t = \sum_{k=1}^{q} \theta_k Z_{t-k}$, le processus Z_t par une estimation du processus d'innovation partielle $\hat{Z}_{p,t}$. Cette estimation peut être réalisée par une estimation des coefficients de prédiction suivie d'un filtrage de la suite X_t observée par le filtre à réponse impulsionnelle finie $\{1, -\hat{\psi}_{1,p}, \ldots, -\hat{\psi}_{p,p}\}$. Une autre façon est d'estimer les coefficients de corrélation partielle et d'utiliser la structure de filtrage en treillis donnée figure 4.1. Une fois la suite $\hat{Z}_{p,t}$ estimée, on peut ensuite estimer la suite $\{\theta_k\}$, par une

approche de type moindres carrés, en minimisant $\|\mathbf{x} - \hat{\mathcal{Z}}\hat{\boldsymbol{\theta}}\|^2$ où $\mathbf{x} = [X_{p+q} \quad \dots \quad X_n]^T$ et :

$$\hat{\mathcal{Z}} = \begin{bmatrix} \hat{Z}_{p,p+q} & \dots & \hat{Z}_{p,p} \\ \hat{Z}_{p,p+q+1} & \dots & \hat{Z}_{p,p+1} \\ \vdots & & \\ \hat{Z}_{p,n} & \dots & \hat{Z}_{p,n-q+1} \end{bmatrix}$$

On obtient $\hat{\boldsymbol{\theta}} = (\hat{\mathcal{Z}}^T\hat{\mathcal{Z}})^{-1}\hat{\mathcal{Z}}^T\mathbf{x}$. L'un des avantages de cette méthode est qu'elle peut être appliquée à tout processus ARMA(p, q) causale et inversible.

Méthode du maximum de vraisemblance approchée

On considère le processus $X_t = Z_t + \sum_{k=1}^{q} \theta_k Z_{t-k}$ où Z_t est un bruit blanc, centré, gaussien. Soit $\{X_1, \dots, X_n\}$ une suite de n observations. On peut alors écrire :

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \theta_1 & 1 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & 1 & 0 \\ 0 & \cdots & & \theta_1 & 1 \end{bmatrix} \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} + \Theta_0 \begin{bmatrix} Z_0 \\ \vdots \\ Z_{-(q-1)} \end{bmatrix} = \Theta \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} + \Theta_0 \mathbf{Z}_0$$

où Θ_0 est une matrice, de dimension $n \times q$, dont seul le triangle supérieur, de dimension $q \times q$, est constitué de termes non nuls. Comme Z_t est un processus aléatoire gaussien, X_t est aussi un processus aléatoire gaussien. L'approche adoptée ici consiste à négliger le terme $\Theta_0\mathbf{Z}_0$. En remarquant que $\det\Theta = 1$, la loi de $\mathbf{X}$ a donc pour densité :

$$\log p_{X_1,\dots,X_n}(x_1, \dots, x_n; \boldsymbol{\theta}, \sigma^2) \approx -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} C(\boldsymbol{\theta}) \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

où $C(\boldsymbol{\theta}) = (\Theta^{-1})^T\Theta^{-1}$. On note $c_{km}(\boldsymbol{\theta})$ les éléments de $C(\boldsymbol{\theta})$. La maximisation par rapport à σ^2 donne :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k,m=1}^{n} c_{km}(\boldsymbol{\theta}) X_k X_m$$

En portant cette expression dans la log-vraisemblance, la maximisation à effectuer est équivalente à la minimisation, par rapport à $\boldsymbol{\theta}$, de l'expression :

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{k,m=1}^{n} c_{km}(\boldsymbol{\theta}) X_k X_m$$

5.3 Estimation ARMA

Equations de Yule-Walker pour un ARMA

Considérons un processus ARMA(p,q) causal défini par :

$$X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \sum_{k=1}^{q} \theta_k Z_{t-k} + Z_t$$

où $\phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k \neq 0$ pour $|z| \leq 1$. On note $\gamma(h)$ sa fonction de covariance. Alors en multipliant les deux membres de l'équation récurrente par X_{t-h}, en prenant l'espérance et en utilisant le fait que $\mathbb{E}[Z_t X_{t-h}] = 0$ pour $h \geq q+1$, il vient :

$$\gamma(h) = \sum_{k=1}^{p} \phi_k \gamma(h-k)$$

En regroupant pour $q+1 \leq h \leq p+q$ les p équations sous forme matricielle, on obtient :

$$\begin{bmatrix} \gamma(q) & \gamma(q-1) & \cdots & \cdots & \gamma(q-p+1) \\ \gamma(q+1) & \gamma(q) & \cdots & \cdots & \gamma(q+p-2) \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ \gamma(q+p-1) & \gamma(q+p-2) & \cdots & & \gamma(q) \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \gamma(q+1) \\ \gamma(q+2) \\ \vdots \\ \gamma(q+p) \end{bmatrix} \tag{5.11}$$

Cette expression matricielle a une forme analogue aux équations de Yule-Walker d'un AR(p). On notera cependant que la matrice n'est plus symétrique. En substituant aux covariances les covariances empiriques $\hat{\gamma}(q-p+1), \ldots, \hat{\gamma}(q+p)$ on obtient une estimation de la suite ϕ_k. Contrairement à l'estimation des coefficients d'un AR(p), par la résolution des équations de Yule-Walker, la résolution de (5.11) ne donne pas nécessairement un polynôme $\hat{\phi}(z)$ dont les racines sont toutes strictement à l'extérieur du cercle unité. Une façon de procéder est de déterminer les racines de $\hat{\phi}(z)$ et "d'inverser" celles qui se trouvent à l'intérieur. Du point de vue spectral, cette construction est justifiée puisqu'elle ne change pas la densité spectrale. En fait comme pour un processus MA(q) on peut améliorer l'estimation en partant d'un système sur-dimensionné et en déterminant une solution de norme minimale.
Une fois la suite $\{\hat{\phi}_1, \ldots, \hat{\phi}_p\}$ estimée, il reste à estimer $\{\theta_1, \ldots, \theta_q, \sigma^2\}$. Théoriquement si nous disposions de la "vraie" suite $\{\phi_k\}$, le processus $e_t = X_t - \sum_{k=1}^{p} \phi_k X_{t-k}$ est simplement le processus MA(q) défini par $e_t = Z_t + \sum_{k=1}^{q} \theta_k Z_{t-k}$. Une façon simple de procéder est donc de filtrer la suite $\{X_1, \ldots, X_n\}$ par le filtre de réponse impulsionnelle $\{1, -\hat{\phi}_1, \ldots, -\phi_p\}$ puis d'utiliser, par exemple, la méthode de Durbin pour estimer $\theta_1, \ldots, \theta_q, \sigma^2$. Une autre façon est d'utiliser à nouveau l'idée de Durbin qui est que $\theta(z)/\phi(z)$ peut être approchée par un AR(m) causal suffisamment long. Notons $\psi_{1,m}, \ldots, \psi_{m,m}$ la suite des coefficients, obtenus par prédiction linéaire, de ce processus AR. On peut alors écrire que $(1 - \sum_{k=1}^{m} \psi_{k,m} z^k)(1 + \sum_{k=1}^{q} \theta_k z^k) = 1 - \sum_{k=1}^{p} \hat{\phi}_k z^k$. En notant ϵ_k les coefficients de

z^k pour $p+1 \leq k \leq m+q$ et en adoptant des notations matricielles évidentes, on peut écrire :

$$\begin{bmatrix} -\psi_{p+1,m} \\ -\psi_{p+2,m} \\ \vdots \\ -\psi_{m,m} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} -\psi_{p,m} & & \cdots & -\psi_{p-q+1,m} \\ -\psi_{p+1,m} & & \ddots & \vdots \\ \vdots & & \ddots & \\ -\psi_{m,m} & & \ddots & \\ 0 & \ddots & & \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\psi_{m,m} \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_q \end{bmatrix} = \begin{bmatrix} \epsilon_{p+1} \\ \vdots \\ \epsilon_{m+q} \end{bmatrix}$$

qui peut encore écrire, de façon plus compacte, $\hat{\boldsymbol{\psi}} = -\hat{\Psi}\hat{\boldsymbol{\theta}} + \mathbf{e}$. La solution qui minimise $\mathbf{e}^T\mathbf{e}$ a pour expression :

$$\hat{\boldsymbol{\theta}} = -(\hat{\Psi}^T\hat{\Psi})^{-1}\hat{\Psi}^T\hat{\boldsymbol{\psi}} \tag{5.12}$$

notons ici que, contrairement à l'expression (5.10), la matrice à inverser dans (5.12) n'est pas une matrice de Toëplitz et ne peut donc inverser, de façon rapide, par l'algorithme de Levinson.
Comme dans le cas de l'estimation MA(q), aucune de ces deux méthodes n'est vraiment précise. Toutefois elles fournissent des estimées correctes pour l'initialisation d'algorithmes itératifs.

Méthode du maximum de vraisemblance approchée

Comme dans le cas MA(q), partant de l'équation $X_t = Z_t + \sum_{k=1}^{q} \theta_k Z_{t-k} + \sum_{k=1}^{p} \phi_k X_{t-k}$ où Z_t est un bruit blanc, centré, gaussien, on peut écrire :

$$\begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\phi_1 & 1 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & 1 & 0 \\ 0 & \cdots & & -\phi_1 & 1 \end{bmatrix} \begin{bmatrix} X_p \\ \vdots \\ X_n \end{bmatrix} + \Phi_0 \begin{bmatrix} X_{p-1} \\ \vdots \\ X_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \theta_1 & 1 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & 1 & 0 \\ 0 & \cdots & & \theta_1 & 1 \end{bmatrix} \begin{bmatrix} Z_p \\ \vdots \\ Z_n \end{bmatrix} + \Theta_0 \begin{bmatrix} Z_{p-1} \\ \vdots \\ Z_{p-q} \end{bmatrix}$$

On peut alors déterminer une expression approchée de la log-vraisemblance conditionnelle de $\{X_p, \ldots, X_n\}$ par rapport à $\{X_1, \ldots, X_{p-1}\}$, en négligeant le terme contenant $\{Z_{p-1}, \ldots Z_{p-q}\}$. Il vient :

$$\log p_{X_p,\ldots,X_n|X_1,\ldots,X_{p-1}}(x_1, \ldots, x_n; \boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2) \approx -\frac{n-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} C(\boldsymbol{\theta}, \boldsymbol{\phi}) \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

où $C(\boldsymbol{\theta}, \boldsymbol{\phi}) = (\Theta^{-1}\Phi)^T\Theta^{-1}\Phi$. La maximisation de $\log p_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \boldsymbol{\theta}, \boldsymbol{\phi}, \sigma^2)$ par rapport à $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ et σ^2 peut être faite par des techniques numériques.

Première partie

Annexes

Annexe A

Eléments de probabilité et de statistique

Nous donnons dans cette annexe quelques éléments succincts de la théorie des probabilités et de l'estimation statistique

A.1 Eléments de probabilité

A.1.1 Espace de probabilité

On se donne un espace abstrait Ω, appelé *espace des épreuves*, muni d'une tribu $\mathcal{F}$, c'est à dire d'une ensemble de parties de Ω vérifiant les propriétés suivantes :

1. $\Omega \in \mathcal{F}$,
2. si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$ ("stabilité par passage au complémentaire"),
3. si $(A_n, n \in \mathbb{N})$ est une suite de parties de Ω, alors, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$ ("stabilité par réunion dénombrable")

Un élément ω de Ω est appelé une *épreuve* ou une *réalisation*. L'ensemble Ω est souvent appelé l'ensemble des *épreuves* ou des *réalisations*. Un élément d'une tribu s'appelle un *un événement* (en théorie de la mesure, de tels éléments sont appelés *ensembles mesurables*). Deux événements A et B sont dits *incompatibles*, si $A \cap B = \emptyset$. L'ensemble vide est appelé l'événement impossible. A l'inverse, Ω est l'événement certain. Le couple $(\Omega, \mathcal{F})$ constitué d'un ensemble d'épreuves et d'une tribu d'événements est un *espace probabilisable*. L'ensemble des parties de Ω, $\mathcal{P}(\Omega)$ est une tribu. Toutes les tribus définies sur Ω sont des sous-ensembles de $\mathcal{P}(\Omega)$. L'ensemble $\{\emptyset, \Omega\}$ est aussi une tribu. Cette tribu est contenue dans toutes les tribus définies sur Ω. L'intersection d'une famille quelconque de tribus est encore une tribu.

Définition A.1. *La tribu engendrée par une classe de parties $\mathcal{A}$ de Ω est la plus petite tribu contenant $\mathcal{A}$ (c'est l'intersection de toutes les tribus contenant $\mathcal{A}$)*

Notons que toute classe $\mathcal{A} \subset \mathcal{P}(\Omega)$, et donc qu'il existe toujours au moins une tribu contenant $\mathcal{A}$. On note $\sigma(\mathcal{A})$ la tribu engendrée par $\mathcal{A}$ La notion de *tribu borélienne* est liée à la structure "topologique"

de l'ensemble de base : c'est la tribu engendrée par l'ensemble des ouverts de la topologie. Nous considérerons dans ce chapitre uniquement la tribu borélienne de $\mathbb{R}^d$, en commençant par le cas le plus simple de la droite réelle $\mathbb{R}$.

Définition A.2. *La tribu borélienne ou tribu de Borel de $\mathbb{R}$ est la tribu engendrée par la classe des intervalles ouverts. On la note $\mathcal{B}(\mathbb{R})$. Un élément de cette tribu est appelé une partie borélienne ou un borélien.*

Tout intervalle ouvert, fermé, semi-ouvert, appartient à $\mathcal{B}(\mathbb{R})$. Il en est de même de toute réunion finie ou dénombrable d'intervalles (ouverts, fermés, ou semi-ouverts). La tribu $\mathcal{B}(\mathbb{R})$ est aussi la tribu engendrée par l'une quelconque des quatre classes suivantes d'ensembles :

$$\mathcal{I} = \{]-\infty, x]; x \in \mathbb{R}\} \; \mathcal{I}' = \{]-\infty, x]; x \in \mathbb{Q}\}$$
$$\mathcal{J} = \{]-\infty, x[; x \in \mathbb{R}\} \; \mathcal{J}' = \{]-\infty, x[; x \in \mathbb{Q}\}$$

De façon similaire, la tribu borélienne $\mathcal{B}(\mathbb{R}^d)$ de $\mathbb{R}^d$ est la tribu engendrée par les rectangles ouverts $\prod_{i=1}^d]a_i, b_i[$. Le théorème suivant sera d'un usage constant dans la suite

Théorème A.1 (Classe monotone)**.** *Soient $\mathcal{C} \subset \mathcal{M} \subset \mathcal{P}(\Omega)$. On suppose que*
- *$\mathcal{C}$ est stable par intersection finie,*
- *$\Omega \subset \mathcal{M}$ et pour $A, B \in \mathcal{M}$, $A \subset B$ implique que $B \setminus A \in \mathcal{M}$,*
- *$\mathcal{M}$ est stable par limite croissante*

Alors, $\sigma(\mathcal{C}) \subset \mathcal{M}$.

Probabilité

Définition A.3. *On appelle* **probabilité** *sur $(\Omega, \mathcal{F})$, une application de $P : \mathcal{F} \to [0, 1]$, qui vérifie les propriétés suivantes*

1. $\mathbb{P}(\Omega) = 1$,
2. *("σ-additivité) si $(A_n, n \in \mathbb{N})$ est une suite d'éléments de $\mathcal{F}$ deux à deux disjoints, (*i.e.$A_i \cap A_j = \emptyset$ *pour $i \neq j$)*
$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

On vérifie aisément les propriétés suivantes : A_n,A et B étant des événements

$$A \subset B, \; \mathbb{P}(A) \leq \mathbb{P}(B), \; \mathbb{P}(A^c) = 1 - \mathbb{P}(A),$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$
$$A_n \uparrow B, \; \mathbb{P}(A_n) \uparrow \mathbb{P}(A), \; A_n \downarrow A, \; \mathbb{P}(A_n) \downarrow \mathbb{P}(A), \; \mathbb{P}(\bigcup_n A_n) \leq \sum_n \mathbb{P}(A_n)$$

Définition A.4. *On dit qu'un ensemble $A \subset \Omega$ est $\mathbb{P}$-négligeable (ou plus simplement négligeable, s'il n'y a pas d'ambiguïté sur la mesure de probabilité) si il existe un ensemble $B \in \mathcal{F}$, tel que $A \subset B$ et $\mathbb{P}(B) = 0$.*

Remarquons que les ensembles négligeables ne sont pas nécessairement des éléments de la tribu $\mathcal{F}$. Une propriété est dite $\mathbb{P}$-presque sûre, si la propriété est vérifiée sur un ensemble dont le complémentaire est $\mathbb{P}$-négligeable.

Définition A.5. *Le triplet* $(\Omega, \mathcal{F}, \mathbb{P})$ *définit un* espace de probabilité.

Définition A.6. *On dira que la tribu* $\mathcal{F}$ *est* complète *si tous les ensembles négligeables de* Ω *sont éléments de* $\mathcal{F}$.

Il est facile de construire une tribu $\mathcal{F}'$ qui contient $\mathcal{F}$ et d'étendre $\mathbb{P}$ à $\mathcal{F}'$ de telle sorte que $\mathcal{F}'$ soit complète pour l'extension de $\mathbb{P}$. Pour éviter des complications techniques inutiles, nous supposerons désormais que toutes les tribus que nous manipulerons sont complètes. Rappelons pour conclure ce paragraphe deux résultats techniques d'usage constant.

Définition A.7. *On appelle un* π-système *une famille d'ensembles stable par intersection finie.*

Théorème A.2. *Soient* μ *et* ν *deux mesures sur* $(E, \mathcal{E})$ *et soit* $\mathcal{C} \subset \mathcal{B}$ *un* π*-système. On suppose que pour tout* $C \in \mathcal{C}$, $\mu(C) = \nu(C) < \infty$. *Alors* $\mu(A) = \nu(A)$ *pour tout* $A \in \sigma(\mathcal{C})$.

Soit E un ensemble. Une famille $\mathcal{E}_0$ de sous-ensembles de E est appelé une *algèbre* si **(i)** $E \in \mathcal{E}_0$, **(ii)** $F \in \mathcal{E}_0 \Longrightarrow F^c \in \mathcal{E}_0$ et **(iii)** $F, G \in \mathcal{E}_0 \Longrightarrow F \cup G \in \mathcal{E}_0$. Une fonction d'ensembles μ définie sur $\mathcal{E}$ est dite σ-additive, si pour toute union dénombrables d'éléments $F_i \in \mathcal{E}_0$, $F_i \cap F_j = \emptyset$, telle que $\bigcup_i F_i \in \mathcal{E}_0$, $\mu(\bigcup_i F_i) = \sum_i \mu(F_i)$.

Théorème A.3 (Théorème d'extension de Carathéodory)**.** *Soit* E *un ensemble et* $\mathcal{E}_0$ *une algèbre sur* E. *Soit* μ_0 *une fonction d'ensembles* σ*-additive, telle que* $\mu_0(E) < \infty$. *Il existe une unique mesure* μ *sur* $\mathcal{E} := \sigma(\mathcal{E}_0)$ *telle que* $\mu = \mu_0$ *sur* $\mathcal{E}_0$.

Exemple A.1
Pour illustrer l'utilisation de ce théorème, rappelons la construction de la mesure de Lebesgue (voir chapitre sur l'intégration sur l'intervalle $[0,1]$. *Soit* $\mathcal{C}$ *l'ensemble des parties de* $[0,1]$ *pouvant s'écrire sous la forme d'une union finie d'intervalles semi-ouverts, semi-fermés,* i.e.$F \in \mathcal{C}$ *si*

$$F = (a_1, b_1] \cup \cdots \cup (a_r, b_r].$$

On vérifie facilement que $\mathcal{C}$ *est stable par intersection finie (*$\mathcal{C}$ *est en fait une algèbre). La tribu engendrée par* $\mathcal{C}$, $\sigma(\mathcal{C}) = \mathcal{B}([0,1])$ *est la tribu borélienne sur* $[0,1]$. *Pour* $F \in \mathcal{F}_0$ *considérons*

$$\lambda_0(F) = \sum_i (b_i - a_i).$$

On vérifie que λ_0 *est une fonction positive et additive. On peut démontrer que* λ_0 *est* σ*-additive,* i.e.*pour toute union dénombrable d'ensembles* $F_i \in \mathcal{F}_0$ *disjoints 2 à 2 tels que* $\bigcup_i F_i \in \mathcal{F}_0$, $\lambda_0(F) = \sum_i \lambda_0(F_i)$ *(cette partie de la preuve n'est pas immédiate). Le théorème de Carathéodory permet de montrer que* λ_0 *a une extension unique* λ *sur* $\mathcal{B}([0,1])$, *appelée* mesure de Lebesgue *sur* $[0,1]$.

A.1.2 Variables aléatoires

Définition A.8. *Soit* E *un espace muni d'une tribu* $\mathcal{E}$. *On appelle variable aléatoire (en abrégé v.a.) à valeurs dans* E *toute application mesurable de* $(\Omega, \mathcal{F}) \to (E, \mathcal{E})$.

Soit X une v.a. à valeurs dans $(E, \mathcal{E})$. En vertu de la définition précédente, pour tout $A \in \mathcal{E}$, on a $X^{-1}(A) \in \mathcal{F}$. Si E est dénombrable et $\mathcal{E} = \mathcal{P}(E)$, on dit que X est une v.a. discrète. Si $E = \bar{\mathbb{R}^+}$ et $\mathcal{E} = \mathcal{B}(\bar{\mathbb{R}^+})$, on dit que X est une v.a. positive. Si $E = \mathbb{R}$ et $\mathcal{E} = \mathcal{B}(\mathbb{R})$, on dit que X est une v.a. réelle. Si $E = \mathbb{R}^d$ et $\mathcal{E} = \mathcal{B}(\mathbb{R}^d)$, on dit que X est une variable vectorielle (ou vecteur aléatoire). Soit $(X_i, i \in I)$ une famille de v.a. à valeurs dans $(E, \mathcal{E})$ (I étant un ensemble quelconque, non nécessairement dénombrable).

Définition A.9. *On appelle* tribu engendrée *par $(X_i, i \in I)$ la plus petite tribu $\mathcal{X}$ de Ω qui soit telle que tous les v.a. X_i soit $\mathcal{X}$ mesurable.*

A titre d'illustration, soit $Y : \Omega \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ une v.a. ; $\sigma(Y)$, la tribu engendrée par Y est définie par

$$\sigma(Y) := (\{\omega : Y(\omega) \in B\}, B \in \mathcal{B}(\mathbb{R})).$$

Si $Z : \Omega \to \mathbb{R}$ est $\sigma(Y)$-mesurable, s'il existe une fonction borélienne $f : \mathbb{R} \to \mathbb{R}$ telle que $Z = f(Y)$ De même, si $Y_1, \cdots, Y_n : \Omega \to \mathbb{R}$ sont des v.a.,

$$\sigma(Y_1, \cdots, Y_n) = \sigma(\{Y_k \in B_k\}, B_k \in \mathcal{B}(\mathbb{R}), k = 1, \cdots, n).$$

et $Z : \Omega \to \mathbb{R}$ est $\sigma(Y_1, \cdots, Y_n)$ mesurable s'il existe une fonction borélienne $f : \mathbb{R}^n \to \mathbb{R}$ telle que $Z = f(Y_1, \cdots, Y_n)$.

Espérance d'une variable aléatoire

Nous rappelons dans le paragraphe suivant succinctement des éléments de théorie d'intégration. Le lecteur se reportera avec profit au cours d'intégration. On dit qu'une variable aléatoire X de $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs réelle est étagée si

$$X = \sum_{k=1}^{n} a_k I_{A_k}$$

avec $A_k \in \mathcal{F}$, où I_A est la fonction indicatrice de A. On note dans la suite $e\mathcal{F}$ l'ensemble des variables étagées. Le résultat suivant est à la base de la construction de l'intégrale

Lemme A.1. *Toute v.a. X positive est limite d'une suite croissante de fonctions étagées.*

Il suffit de considérer la suite

$$X_n(\omega) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} I_{\{k/2^n \leq X(\omega) \leq (k+1)/2^n\}} + nI_{X(\omega) \geq n}$$

L'espérance d'une v.a. étagée $X = \sum_{k=1}^{n} a_k I_{A_k}$ est définie par

$$\mathbb{E}[X] := \int X(\omega) d\mathbb{P}(\omega) = \sum_{k=1}^{n} a_k \mathbb{P}(A_k).$$

On remarque facilement que, si $X, Y \in e\mathcal{F}$,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \text{and} X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y].$$

Le résultat technique suivant est la clef de voûte de la construction

Lemme A.2. *Soient $X_n, Y_n \in e\mathcal{F}$ deux suites croissantes telles que* $\lim \uparrow X_n = \lim \uparrow Y_n$. *Alors,* $\lim \uparrow \mathbb{E}[X_n] = \lim \uparrow \mathbb{E}[Y_n]$.

Notons $\mathcal{F}^+$ l'ensemble des v.a. positives. Soit $X \in \mathcal{F}^+$. Le lemme A.1 montre qu'il existe une suite $X_n \in e\mathcal{F}$ telle que $X_n \uparrow X$; la monotonicité de l'espérance assure que $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$. On pose $\mathbb{E}[X] = \lim \uparrow \mathbb{E}[X_n]$. Le lemme A.2 montre que cette limite ne dépend pas du choix de la suite X_n. On a en particulier

$$\mathbb{E}[X] = \lim \uparrow \sum_{k=0}^{n2^n} \frac{k}{2^n} \mathbb{P}(\{\omega : k/2^n \leq X(\omega) < (k+1)/2^n\}) + n\mathbb{P}(\{\omega : X(\omega) \geq n\}).$$

Par passage à la limite, on obtient immédiatement que pour tout $X, Y \in \mathcal{F}^+$, et $a, b \in \mathbb{R}^+$, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ et que, si $X \leq Y$, $\mathbb{E}[X] \leq \mathbb{E}[Y]$. On dira que $X \in \mathcal{F}^+$ est *intégrable* si $\mathbb{E}[X] < \infty$. Notons $f\mathcal{F}$ l'ensemble des v.a. mesurables réelles. On pose

$$\mathcal{L}^1 = \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}) = \{X \in f\mathcal{F}, \mathbb{E}[|] X| < \infty\}$$

Si $f \in \mathcal{L}^1$, nous définissons X^+ et X^- les parties positives et négatives de X,

$$X^+ := X \vee 0 \text{ and } X^- := (-X) \vee 0$$

X^+ et X^- sont des v.a. positives intégrables (car $X^+ \leq |X|$ et $X^- \leq |X|$) , et $X = X^+ - X^-$. L'espérance de X est définie par

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

Il est facile de voir que $\mathcal{L}^1$ est un espace vectoriel (car $|X + Y| \leq |X| + |Y|$, et par monotonicité de l'espérance) et que $X \to \mathbb{E}[X]$ est une forme linéaire positive. De plus, pour $X \in \mathcal{L}^1$, $|\mathbb{E}[X]| \leq \mathbb{E}[|] X|$.

Passages à la limite

Soit X_n une suite de v.a.s. Nous disons que $X_n \to X$ $\mathbb{P}$-p.s., si

$$\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}^c$$

est $\mathbb{P}$-négligeable. Les propriétés suivantes découlent directement des théorèmes classiques de la théorie de la mesure (à savoir, le théorème de convergence monotone, ou théorème de Beppo-Levi, le lemme de Fatou, et le théorème de convergence dominée)

Proposition A.1.
- *("Convergence monotone") si* $0 \leq X_n \uparrow X$, *alors* $\mathbb{E}[X_n] \uparrow \mathbb{E}[X] \leq \infty$
- *("Lemme de Fatou") Si* $X_n \geq 0$, *alors* $\mathbb{E}[\liminf X_n] \leq \liminf \mathbb{E}[X_n]$,
- *("Convergence dominée") Si, pour tout* $n \geq 1$, $|X_n(\omega)| \leq Y(\omega)$, $\mathbb{P}$*-ps, et* $Y \in \mathcal{L}^1$, *alors* $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$

Nous utiliserons de façon très fréquente dans la suite les résultats ci-dessus ; nous donnons toutefois sans attendre quelques exemples d'applications très utiles :

Exemple A.2 – *Soit* (Z_k) *une suite de v.a.s positives. Alors* $\mathbb{E}[\sum Z_k] = \sum \mathbb{E}[Z_k] \leq \infty$ *(application de la convergence monotone et de la linéarité de l'espérance).*

– *Soit (Z_k) une suite de v.a.s positives, telle que $\sum \mathbb{E}[Z_k] < \infty$. Alors $\sum Z_k$ est fini p.s. et donc $Z_k \to 0$ p.s.*

Nous admettrons le résultat suivant (cf. le cours d'intégration)

Théorème A.4. *Soit X une v.a. de $(\Omega, \mathcal{F})$ dans $(E, \mathcal{E})$ et $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{F})$. La formule $\mathbb{P}_X(A) := \mathbb{P}(X^{-1}(A))$ définit une probabilité sur $(E, \mathcal{E})$, appelée probabilité image de $\mathbb{P}$ par X. Cette probabilité vérifie, pour toute fonction f positive mesurable*

$$\int f \circ X(\omega) d\mathbb{P}(\omega) = \int f(x) d\mathbb{P}_X(x)$$

Définition A.10. *On appelle* loi de X *la probabilité image de $\mathbb{P}$ par X.*

La loi d'une variable aléatoire réelle est donc une probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. On définit souvent la loi d'une variable aléatoire en spécifiant une "densité" par rapport à une mesure positive sur $(E, \mathcal{E})$. Plus précisément, soit μ une mesure positive et soit g une fonction mesurable positive, telle que

$$\int_E g(x) d\mu(x) = 1.$$

Pour $A \in \mathcal{E}$, on définit $\mathbb{P}_X : \mathcal{E} \to [0, 1]$

$$\mathbb{P}_X(A) = \int_A g(x) d\mu(x).$$

On vérifie aisément que $\mathbb{P}_X$ défini par la relation précédente spécifie bien une mesure de probabilité sur $(E, \mathcal{E})$. Nous donnons ci-dessous quelques exemples élémentaires

– La mesure de Lebesgue sur $[0, 1]$ est une probabilité, que l'on appelle généralement loi uniforme sur $[0, 1]$. Plus généralement, pour $a < b$, on appelle loi uniforme sur $[a, b]$, la mesure de probabilité $(b - a)^{-1} I_{[a,b]}(x) dx$, où I_A est l'indicatrice de l'ensemble A.
– La mesure sur $\mathbb{R}$ de densité $\pi^{-1}(1 + x^2)^{-1} dx$ est de masse 1, et définit donc bien une mesure de probabilité sur $\mathbb{R}$. On remarque que le moment d'ordre 1 de cette mesure est infini. Cette loi est appelée loi de Cauchy standard.
– La loi de densité $p_X(x)$,
$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$
par rapport à la mesure de Lebesgue est appelée "loi gaussienne". La moyenne de cette loi est μ et sa variance, $\int (x - \mu)^2 p_X(x) \lambda(dx) = \sigma^2$.

Il est souvent pratique de spécifier la loi de probabilité d'une variable aléatoire réelle par la donnée de sa fonction de répartition, $F_X : \mathbb{R} \to [0, 1]$, définie par

$$F_X(x) = \mathbb{P}_X(]-\infty, x]) = \mathbb{P}(X \leq x).$$

La fonction de répartition est une fonction croissante, continue à droite : on remarque en effet que $]-\infty, x] = \bigcap]-\infty, x_n]$, pour toute suite décroissante x_n, telle que $\lim_{n\to\infty} x_n = x$. La σ-additivité impose donc que $F_X(x) = \lim_{n\to\infty} F(x_n)$, et donc plus généralement que $\lim_{h\to 0+} F_X(x+h) = F_X(x)$. Un raisonnement similaire montre que $\lim_{h\to 0-} F_X(x + h) = \mathbb{P}_X(]-\infty, x[) =: F_X(x-)$. La fonction de répartition F_X caractérise la loi $\mathbb{P}_X$, puisque pour tout intervalle $]a, b]$ $(b > a)$, on a $\mathbb{P}_X(]a, b]) = F_X(b) - F_X(a)$ et qu'une mesure borélienne sur $\mathbb{R}$ est déterminée par la donnée des masses qu'elle attribue aux intervalles de ce type (cf. cours d'intégration)

Quelques inégalités utiles

L'inégalité élémentaire suivante, appelée inégalité de Markov, joue un rôle fondamental

Proposition A.2. *Soit Z une v.a et $g : \mathbb{R} \to [0, \infty]$ une fonction borélienne croissante. Alors*

$$\mathbb{E}[g(Z)] \geq \mathbb{E}[g(Z)I(Z \geq c)] \geq g(c)\mathbb{P}[Z \geq c].$$

En prenant pour $g(x) = |x|$, nous avons en particulier, pour $X \in \mathcal{L}^1$, $\mathbb{P}[|X| \geq c] \leq \mathbb{E}[|X|]/c$. Une fonction $c : G \to \mathbb{R}$ où G est un intervalle ouvert de $\mathbb{R}$ est dite *convexe* si, pour tout $x, y \in G$ et tout p, q, $p + q = 1$,

$$c(px + qy) \leq pc(x) + qc(y).$$

A titre d'exemples, les fonctions $|x|$, x^2, $e^{\theta x}$ sont des fonctions convexes. La proposition suivante est souvent utiles

Proposition A.3 (Inégalité de Jensen)**.** *Soit $c : G \to \mathbb{R}$ une fonction convexe sur un sous-intervalle ouvert G de $\mathbb{R}$ et soit X une variable aléatoire vérifiant les propriétés suivantes*

$$\mathbb{E}[|]X| < \infty,\ \mathbb{P}[X \in G] = 1,\ \mathbb{E}[|]c(X)| < \infty$$

Alors, $\mathbb{E}[c(X)] \geq c(\mathbb{E}[X])$.

Variance, covariance, corrélation

Si la variable X admet un moment d'ordre 2, alors X admet un moment d'ordre 1 (par monotonicité des semi-normes, $\mathcal{L}^1 \subset \mathcal{L}^2$). On pose alors,

$$\mathrm{var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

quantité que l'on appelle la *variance* de X. De même, lorsque $X, Y \in \mathcal{L}^2$, nous pouvons définir,

$$\mathrm{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

quantité que l'on appelle la covariance de X et de Y. Les variables aléatoires sont dites *décorrélées*, si le coefficient de covariance $\mathrm{cov}(X, Y) = 0$. Lorsque $X := (X_1, \cdots, X_d)^T$, $d \in \mathbb{N}$ est un vecteur aléatoire, la matrice de covariance $\Gamma(X)$ (ou matrice de variance / covariance) est définie comme la matrice $d \times d$ dont les éléments sont donnés par

$$\Gamma(X)_{i,j} = \mathrm{cov}(X_i, X_j) \qquad 1 \leq i, j \leq d$$

Les éléments diagonaux sont égaux à la variance des variables X_i ; les éléments hors-diagonaux sont les coefficients de covariance. La matrice de covariance est une matrice symétrique ($\Gamma(X) = \Gamma(X)^T$) et semi-définie positive. En effet, pour tout d-uplets de nombre réels ou complexes $(a_1, a_2, \cdots, a_d)$, nous avons

$$\mathbb{E}\left[\left|\sum_{i=1}^{d} a_i(X_i - \mathbb{E}[X_i])\right|^2\right] = \sum_{i,j} a_i a_j^* \Gamma(X)_{i,j} \geq 0$$

Notons que, pour tout vecteur a (déterministe)

$$\Gamma(X + a) = \Gamma(X)$$

et que, pour M une matrice (déterministe) $p \times d$,

$$\Gamma(MX) = M\Gamma(X)M^T.$$

Fonction caractéristique

Dans tout ce paragraphe, X désigne une variable aléatoire à valeurs dans $\mathbb{R}^d$. On note $\mathbb{P}_X$ sa loi. L'application $\Phi_X : \mathbb{R}^d \to \mathbb{C}$ donnée par

$$\Phi_X(\lambda) = \mathbb{E}\left[\exp(i(\lambda, X))\right] = \int_{\mathbb{R}^d} \exp(i(\lambda, x))\mathbb{P}_X(dx).$$

où (u, v) désigne le produit scalaire usuel dans $\mathbb{R}^d$, s'appelle la *fonction caractéristique* de X. La fonction caractéristique est la *transformée de Fourier* de la loi $\mathbb{P}_X$. Nous donnons ci-dessous quelques propriétés élémentaires de la fonction caractéristique

- $\Phi_X(0) = 1$ et $|\Phi_X(\lambda)| \leq 1$.
- La fonction caractéristique est continue sur $\mathbb{R}^d$. Cette propriété est une conséquence immédiate de la continuité de l'application $\lambda \longrightarrow \exp(i(\lambda, X)$ et du théorème de convergence dominé.
- Lorsque la loi $\mathbb{P}_X$ admet une densité g par rapport à la mesure de Lebesgue, alors Φ_X est la transformée de g (au sens usuel). Le théorème de Rieman-Lebesgue implique que $\Phi_X(\lambda)$ tend vers 0 lorsque $\lambda \to \infty$.

Comme son nom l'indique, la fonction caractéristique "caractérise" la loi, dans le sens

Proposition A.4. *Deux variables aléatoires à valeurs dans $\mathbb{R}^d$ ont même loi si et seulement si $\Phi_X = \Phi_Y$.*

Le théorème précédent implique en particulier la proposition suivante

Proposition A.5. *Soient $X = (X_1, \cdots, X_n)$; n variables aléatoires réelles variables aléatoires $(X_1, \cdots, X_n)$ sont indépendantes si et seulement si*

$$\Phi_X(\lambda_1, \cdots, \lambda_n) = \prod_{i=1}^{n} \Phi_{X_i}(\lambda_i)$$

Indépendance. Mesures produits

Soient A et B deux événements. On dit que A et B sont *indépendants* si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Les propriétés élémentaires des probabilités montrent que les événements A et B^c, A^c et B, et A^c et B^c sont aussi indépendants. En effet :

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(\Omega \cap B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = (1 - \mathbb{P}(A))\mathbb{P}(B).$$

Les tribus $\mathcal{A} = \{\emptyset, A, A^c, \Omega\}$ et $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ sont donc indépendantes, au sens de la définition suivante

Définition A.11. *Soit $(\mathcal{B}_i, i \in I)$ une famille de tribu. On dit que cette famille est* indépendante *si, pour tout sous-ensemble J fini de I,*

$$\mathbb{P}\left(\bigcap_{j \in J} B_j\right) = \prod_{j \in J} \mathbb{P}(B_j), \quad B_j \in \mathcal{B}_j$$

Le lemme technique suivant donne un critère plus "pratique" pour vérifier l'indépendance de tribus.

Lemme A.3. *Soient $\mathcal{G}$ et $\mathcal{H}$ deux sous-tribus de $\mathcal{F}$ et soit $\mathcal{I}$ et $\mathcal{J}$ deux π-systèmes tels que $\mathcal{G} := \sigma(\mathcal{I})$ et $\mathcal{H} := \sigma(\mathcal{J})$. Alors, les tribus $\mathcal{G}$ et $\mathcal{H}$ sont indépendantes si et seulement si $\mathcal{I}$ et $\mathcal{J}$ sont indépendantes,* i.e.

$$\mathbb{P}(I \cap J) = \mathbb{P}(I)\mathbb{P}(J), \quad I \in \mathcal{I}, J \in \mathcal{J}.$$

Démonstration. Supposons que les familles $\mathcal{I}$ et $\mathcal{J}$ sont indépendantes. Pour $I \in \mathcal{I}$ donné, considérons les mesures

$$H \to \mathbb{P}(I \cap H) \text{ et } H \to \mathbb{P}(I)\mathbb{P}(H).$$

Ces mesures sont définies $(\Omega, \mathcal{H})$ et coïncident sur $\mathcal{J}$. Le théorème A.2 montre que ces deux mesures coïncident sur $\mathcal{H}$

$$\mathbb{P}(I \cap H) = \mathbb{P}(I)\mathbb{P}(H), \quad I \in \mathcal{I}, H \in \mathcal{H}.$$

Pour H donné dans $\mathcal{H}$, les mesures

$$G \to \mathbb{P}(G \cap H) \text{ et } G \to \mathbb{P}(G)\mathbb{P}(H)$$

sont définies sur $\mathcal{G}$ et coïncident sur $\mathcal{I}$. Par le théorème extension, elles coïncident sur $\mathcal{G}$, et donc $\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H)$, pour tout $G \in \mathcal{G}$ et $H \in \mathcal{H}$. ■

De façon générale, on a

Proposition A.6. *Soient $(\mathcal{C}_i, i \in I)$ une famille de π-systèmes indépendants. Alors les tribus $(\sigma(\mathcal{C}_i), i \in I)$ sont indépendantes.*

Il résulte immédiatement de la définition A.11 que si $\mathcal{B}'_i$ est une sous-tribu de $\mathcal{B}_i$, la famille $(\mathcal{B}'_i, i \in I)$ est une famille indépendante si $(\mathcal{B}_i, i \in I)$ l'est. Nous avons aussi

Proposition A.7. *Si la famille $(\mathcal{B}_i, i \in I)$ est indépendante et si $(I_j, j \in J)$ est une partition de I, la famille $(\sigma(B_i, i \in I_j), j \in J)$ est indépendante.*

De cette définition découle toutes les notions d'indépendance dont nous aurons besoin dans la suite. Si $(A_i, i \in I)$ est une famille d'événements, on dira que cette famille est indépendante si la famille $(\sigma(A_i), i \in I)$ l'est. Si $(X_i, i \in I)$ est une famille de v.a., on dira que cette famille est indépendante si la famille $(\sigma(X_i), i \in I)$ l'est. Si X est une v.a. et $\mathcal{G}$ une tribu, on dira que X et $\mathcal{G}$ sont indépendantes si les tribus $\sigma(X)$ et $\mathcal{G}$ sont indépendantes. Enfin, si $(X_i, i \in I)$ et $(Y_j, j \in J)$ sont indépendantes si les tribus $(\sigma(X_i), i \in I)$ et $(\sigma(Y_j), j \in J)$ le sont.

Exemple A.3
Soient (X_1, X_2, X_3, X_4) quatre v.a. indépendantes. Alors, les couples (X_1, X_2) et (X_3, X_4) sont indépendants, puisque les tribus $\sigma(X_1, X_2)$ et $\sigma(X_3, X_4)$ le sont. Alors $Y_1 := f(X_1, X_2)$ et $Y_2 = g(X_3, X_4)$ (avec f, g boréliennes) sont indépendantes car $\sigma(Y_1) \subset \sigma(X_1, X_2)$ et $\sigma(Y_2) \subset \sigma(X_3, X_4)$.

Avant d'aller plus loin, rappelons quelques résultats sur les mesures produits (on se reportera avec profit au cours d'intégration). Soient $(E_1, \mathcal{B}_1, \nu_1)$ et $(E_2, \mathcal{B}_2, \nu_2)$ deux espaces mesurés et ν_1, ν_2 deux mesures σ-finies. Alors

$$\mathcal{B}_1 \otimes \mathcal{B}_2 := \sigma(A_1 \times A_2, A_1 \in \mathcal{B}_1, A_2 \in \mathcal{B}_2)$$

est une tribu sur $E_1 \times E_2$ appelée *tribu produit* de $\mathcal{B}_1$ et de $\mathcal{B}_2$ et il existe une unique mesure, notée $\nu_1 \otimes \nu_2$ définie sur $\mathcal{B}_1 \otimes \mathcal{B}_2$ telle que

$$\nu_1 \otimes \nu_2(A_1 \times A_2) = \nu_1(A_1)\nu_2(A_2),\ A_1 \in \mathcal{B}_1, A_2 \in \mathcal{B}_2.$$

Pour toute fonction borélienne positive ou bornée f, nous avons (théorème de Fubini)

$$\begin{aligned}\int f d(\nu_1 \otimes \nu_2) &= \int \left(\int f(x_1, x_2) d\nu_1(x_1) \right) d\nu_2(x_2), \\ &= \int \left(\int f(x_1, x_2) d\nu_2(x_2) \right) d\nu_1(x_1)\end{aligned}$$

Ces résultats s'étendent directement pour le produit de n espaces. Il résulte alors de ces rappels et du théorème de classe monotone que

Théorème A.5. *Soient $(X_1, \cdots, X_n)$ des v.a. à valeurs dans $(E_i, \mathcal{E}_i)$, $i \in \{1, \cdots, n\}$. Il y a équivalence entre*

1. *les v.a $X_1, \cdots, X_n$ sont indépendantes,*
2. *Pour tout $A_k \in \mathcal{E}_k$,*
$$\mathbb{P}[X_1 \in A_1, \cdots, X_n \in A_n] = \prod_1^n \mathbb{P}[X_k \in A_k]$$
3. *Pour tout $A_k \in \mathcal{C}_k$, avec $\mathcal{C}_k$ π-système tel que $\sigma(\mathcal{C}_k) = \mathcal{E}_k$,*
$$\mathbb{P}[X_1 \in A_1, \cdots, X_n \in A_n] = \prod_1^n \mathbb{P}[X_k \in A_k]$$
4. *La loi du vecteur aléatoire $(X_1, \ldots, X_n)$, notée $\mathbb{P}_{(X_1, \cdots, X_n)}$ est égale au produit des lois des v.a X_k,*
$$\mathbb{P}_{(X_1, \cdots, X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}.$$
5. *Pour toutes fonctions f_k boréliennes positives (respectivement bornées, respectivement $f_k \in \mathcal{L}^1(E_k, \mathcal{E}_k, \mathbb{P}_k)$),*
$$\mathbb{E}\left[f_1(X_1) \cdots f_n(X_n)\right] = \prod_1^n \mathbb{E}\left[f_k(X_k\right]$$

Exemple A.4
Soient X, Y deux v.a.r. Alors, vu que $\sigma([a, b[, a < b \in \mathbb{R}) = \mathcal{B}(\mathbb{R})$, il résulte du théorème précédent que X et Y sont indépendantes si et seulement si

$$\mathbb{P}(a \leq X < b, c \leq Y < d) = \mathbb{P}(a \leq X < b)\mathbb{P}(c \leq Y < d),$$

pour tout a, b, c, d. Dans ce cas, si $\mathbb{E}\left[|\right] X| < \infty$, $\mathbb{E}\left[|\right] Y| < \infty$, on a $\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$, résultat que l'on utilise sans cesse en probabilité.

A.1.3 Espaces $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ et $L^p(\Omega, \mathcal{F}, \mathbb{P})$

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Pour $p > 0$, on dit que X admet un moment d'ordre p

$$\mathbb{E}\left[|X|^p\right] = \int |X(\omega)|^p P(d\omega) < \infty.$$

Nous notons $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ l'ensemble des variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ admettant un moment d'ordre p. Nous notons, pour $X \in \mathcal{L}^p$, $\|X\|_p = \mathbb{E}\left[|X|^p\right]^{1/p}$. Il est facile de voir que la fonction $\| \bullet \|_p : (\Omega, \mathcal{F}, \mathbb{P}) \mapsto \mathbb{R}$ est positive. Cette fonction vérifie aussi l'inégalité triangulaire, appelée dans ce contexte, *inégalité de Minkovski*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

L'inégalité de Minkovski montre que, pour tout $X, Y \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ et tout $\alpha, \beta \in \mathbb{R}$, nous avons

$$\|\alpha X + \beta Y\|_p \leq |\alpha| \|X\|^p + |\beta| \|Y\|_p$$

et donc que $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ est un espace vectoriel sur $\mathbb{R}$. On omettra la dépendance en $(\Omega, \mathcal{F}, \mathbb{P})$ lorsqu'il n'y a pas d'ambiguïté sur l'espace de probabilité sous-jacent. La fonction $x \mapsto \|x\|_p$ est positive et vérifie l'inégalité triangulaire. Ce n'est toutefois pas une norme, car la relation $\|X\|_p = 0$ entraîne seulement que $X = 0$ $\mathbb{P}$-p.s ($\mathbb{P}(\omega, X(\omega) = 0) = 1$). On dit que $\| \bullet \|_p$ est une *semi-norme*. Comme nous le verrons ci-dessous, il est possible de "quotienter" l'espace par la relation d'équivalence

$$X \equiv Y \Longleftrightarrow \mathbb{P}[\{\omega \in \Omega, X(\omega) = Y(\omega)\}] = 1$$

On note $L^p(\Omega, \mathcal{F}, \mathbb{P})$ l'espace quotient de $\mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$ par la relation d'équivalence $\equiv$. Les éléments de $L^p(\Omega, \mathcal{F}, \mathbb{P})$ sont des **classes d'équivalence**. Si X et Y sont deux éléments de la même classe d'équivalence, alors $\|X\|_p = \|Y\|_p$. Lorsque l'on choisit un élément d'une classe d'équivalence on dit que l'on choisit une *version* de X : X désigne selon les cas sa classe ou une version de la classe. Les (semi)-normes $\| \bullet \|_p$ sont monotones dans le sens suivant

Proposition A.8. *Soit $1 \leq p \leq r < \infty$ et $Y \in \mathcal{L}^r$. Alors, $Y \in \mathcal{L}^p$ et $\|Y\|_p \leq \|Y\|_r$.*

Cette dernière inégalité découle directement de l'inégalité de Jensen appliquée avec $c(x) = x^{r/p}$. L'inégalité suivante est souvent utile

Proposition A.9. *Soient $p, q \geq 1$ tels que $p^{-1} + q^{-1} = 1$. Nous avons (*inégalité de Hölder*)*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

La proposition suivante (en particulier lorsque $p = 2$) joue un rôle clef.

Proposition A.10. *Soit $p \in [1, \infty)$. Soit (X_n) une suite de Cauchy dans $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$, i.e.,*

$$\lim_{k \to \infty} \sup_{r,s \geq k} \|X_r - X_s\|_p = 0.$$

Il existe une variable aléatoire $X \in \mathcal{L}^p$ telle que $X_r \to X$ dans $\mathcal{L}_p$, i.e. $\|X_r - X\|_p \to 0$. De plus, on peut extraire de X_n une sous-suite $Y_k = X_{n_k}$ qui converge vers X $\mathbb{P}$-p.s.

Démonstration. C'est un résultat classique d'analyse ; nous en donnons toutefois une démonstration de nature "probabiliste" afin d'illustrer les résultats et les techniques introduites précédemment. Soit $k_n \uparrow \infty$ une suite telle que

$$\forall (r,s) \geq k_n, \|X_r - X_s\| \leq 2^{-n}$$

Nous avons, par monotonicité des semi-normes $\| \bullet \|_p$, nous avons pour $p \geq 1$,

$$\mathbb{E}\left[|X_{k_{n+1}} - X_{k_n}|\right] \leq \|X_{k_{n+1}} - X_{k_n}\|_p \leq 2^{-n},$$

ce qui implique, en appliquant le théorème de Fubini, que

$$\mathbb{E}\left[\sum |X_{k_{n+1}} - X_{k_n}|\right] < \infty.$$

Ceci implique que la série de terme général $U_n := (X_{k_{n+1}} - X_{k_n})$ converge absolument $\mathbb{P}$-p.s., et donc que

$$\sum_{n \geq 1} U_n = \lim_n X_{k_n}$$

existe $\mathbb{P}$-p.s. Définissons, pour tout $\omega \in \Omega$

$$X(\omega) := \limsup X_{k_n}(\omega)$$

X est une v.a. (en tant que limite supérieure d'une suite de v.a.s) et $\lim_n X_{k_n} = X$, $\mathbb{P}$-p.s. Soit $\epsilon > 0$ et soit m tel que $2^{-m} \leq \epsilon$. Pour tout $r \geq k_m$, et tout $n \geq m$, nous avons

$$\|X_r - X_{k_n}\|_p \leq \epsilon$$

et l'application du lemme de Fatou montre que

$$\|X_r - X\|_p \leq \left(\int \liminf_m |X_r(\omega) - X_{k_m}(\omega)|^p \mathbb{P}(d\omega) \right)^{1/p} \leq \liminf_n \|X_r - X_{k_n}\|_p \leq \epsilon.$$

et donc $\lim_{r \to \infty} \|X_r - X\|_p = 0$. L'inégalité de Minkovski montre que

$$\|X\|_p \leq \|X_r - X\|_p + \|X\|_p$$

et donc que $X \in \mathcal{L}^p$. ■

Le résultat précédent permet de montrer que l'espace quotient $L^p(\Omega, \mathcal{F}, \mathbb{P})$ est complet.

A.1.4 Variables aléatoires Gaussiennes

Définition A.12 (v.a. gaussienne standardisée)**.** *On dit qu'une variable X est Gaussienne standardisée (ou standard) si la loi de X admet la densité (par rapport à la mesure de Lebesgue)*

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}). \tag{A.1}$$

Définition A.13. *On dit qu'une variable aléatoire X est gaussienne de moyenne m et de variance σ^2, s'il existe une variable gaussienne standard Z telle que $X = m + \sigma Z$.*

Lorsque$\sigma > 0$, X admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, densité donnée par

$$f_{m,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-m)^2}{2\sigma^2}). \tag{A.2}$$

On note cette densité $\mathcal{N}(m, \sigma^2)$. Par abus de langage, nous identifierons les variables gaussiennes de variance nulle aux mesures de Dirac au point m. Un calcul élémentaire montre que, pour tout $\lambda \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \exp(\lambda x) dx = \exp(\frac{\lambda^2}{2}).$$

Par prolongement analytique, la fonction caractéristique d'une variable gaussienne standard est donc donnée par

$$\Phi_X(\lambda) = \exp(-\lambda^2/2).$$

Notons que si X est une variable aléatoire de fonction caractéristique $\Phi_X(\lambda)$, la fonction caractéristique de la variable aléatoire $Y = a + bX$ est donnée par

$$\Phi_Y(\lambda) = \exp(i\lambda a) \exp(-b^2\lambda^2/2)$$

Par conséquent, la fonction caractéristique de la loi normale de moyenne m et de variance σ^2 est donnée par

$$\Phi_X(\lambda) = \exp(i\lambda m - \lambda^2\sigma^2/2) \tag{A.3}$$

On en déduit la proposition suivante

Proposition A.11. *Soient X_i, $i \in \{1, \ldots, d\}$, d v.a.r gaussiennes indépendantes de moyenne μ_i et de variance σ_i^2 et soient $a_i \in \mathbb{R}$, $i \in \{1, \cdots, d\}$. La v.a.r $Y = a_1X_1 + \cdots + a_dX_d$ est une v.a.r gaussienne de moyenne $\sum_{i=1}^d a_i m_i$ et de variance $\sum_{i=1}^d a_i^2\sigma_i^2$.*

Démonstration. en utilisant la proposition A.5, la fonction caractéristique de Y est donnée par

$$\phi_Y(t) = \prod_{k=1}^{d} \phi_{X_k}(a_k t), \tag{A.4}$$

$$= \exp\left[it \sum_{k=1}^{d} a_k m_k - \sum_{k=1}^{d} a_k^2 \sigma_k^2 t^2/2 \right], \tag{A.5}$$

et on conclut en utilisant la proposition A.4. ■

Définition A.14 (vecteur gaussien)**.** *Un vecteur aléatoire $X = [X_1, \cdots, X_d]^T$ est dit gaussien, si pour tout vecteur $a \in \mathbb{R}^d$, $a^T X := a_1X_1 + \cdots + a_dX_d$ est une v.a.r. gaussienne*

Cette définition implique en particulier que chaque composante X_k est une v.a.r gaussienne. A l'inverse, le fait que toutes les variables X_k soient gaussiennes ne suffit pas pour assurer que le vecteur X est gaussien. Par construction, la famille de lois gaussiennes est stable par transformation linéaire. Plus précisément

Lemme A.4. *Soit X un vecteur gaussien à valeurs dans $\mathbb{R}^d$ de moyenne m et de matrice de covariance K. Pour tout $b \in \mathbb{R}^r$, et toute matrice M de dimension $(r \times d)$, le vecteur aléatoire $Y = b + MX$ est un vecteur gaussien à valeurs dans $\mathbb{R}^r$, de moyenne $b + Mm$ et de covariance MKM^T.*

En effet, pour tout vecteur $a \in \mathbb{R}^r$, $a^T Y = a^T b + (a^T M)X$ est une v.a. gaussienne. On a $\mathbb{E}[Y] = m + M\mathbb{E}[X]$ et $K(Y) = MKM^T$. Le théorème de caractérisation suivant joue un rôle central

Théorème A.6. *Soit X un vecteur aléatoire de moyenne m et de matrice de covariance K. Le vecteur X est gaussien si et seulement si sa fonction caractéristique est donnée par*

$$\phi_X(\lambda) = \exp[i\lambda^T m - \frac{1}{2}\lambda^T K \lambda]$$

Ce théorème montre que toute loi gaussienne est déterminée par la donnée de sa moyenne et de sa matrice de covariance. Lorsque la matrice de covariance K est inversible, la loi d'un vecteur aléatoire gaussien de moyenne m et de covariance K a une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ et cette densité est donnée par

$$p(x; m, K) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det(K)}} \exp\left(-\frac{1}{2}(x-m)^T K^{-1}(x-m)\right)$$

La loi d'un vecteur gaussien étant entièrement spécifiée par la donnée de sa moyenne et de sa matrice de covariance, les notions d'indépendance et de décorrélation sont confondues (propriété qui n'est pas vérifiée de façon générale).

Théorème A.7. *Soit $Y = [Y_1, \cdots, Y_n^T]^T$ un vecteur gaussien ($(d_1 + \cdots + d_n) \times 1$). Les vecteurs Y_i ($d_i \times 1$, $i \in \{1, \cdots, n\}$) sont indépendants si et seulement si, pour toute suite de vecteurs a_i ($d_i \times 1$, $i \in \{1, \cdots, n\}$) $\mathrm{cov}[a_i^T Y_i, a_j^T Y_j] = 0$, $i \neq j \in \{1, \cdots, n\}$.*

A.1.5 Modes de convergence et Théorèmes limites

Les théorèmes limites sont au coeur même de la théorie des probabilités. Nous ne donnons ici que quelques définitions et énoncés essentiels, en nous limitant aux notions que nous utiliserons dans la suite. Le lecteur se reportera à Resnick ou Williams pour une introduction. Introduisons tout d'abord les différents "modes" de convergence. Soit $(X_n, n \in \mathbb{N})$ une famille de v.a. définies sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. On note $|x| = \left(\sum_{k=1}^d x_k^2\right)^{1/2}$ la norme euclidienne. Soit finalement X une v.a. définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Définition A.15 (Convergence p.s.)**.** *On dit que X_n converge presque-sûrement vers X (on note : $X_n \to_{\mathbb{P}-\text{p.s.}} X$) si et seulement si*

$$\mathbb{P}\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = 1.$$

De façon équivalente, $X_n \to_{\mathbb{P}-\text{p.s.}} X$ si et seulement si , pour tout $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{\bigcup_{k\geq n} \{|X_k - X| \geq \delta\right\} = 0.$$

Définition A.16 (Convergence dans $\mathcal{L}^r$)**.** *On dit que X_n converge dans $\mathcal{L}^r$ vers X (on note : $X_n \to_{\mathcal{L}_r} X$) si et seulement si*

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^r\right] = 0.$$

Définition A.17 (Convergence en probabilité)**.** *Soit $\{X_n\}$ une suite de variables aléatoires et X une autre variable aléatoire, toutes définies sur le même espace de probabilité $\{\Omega, \mathcal{F}, P\}$, à valeurs dans $\mathbb{R}^k$. On dit que X_n converge en $\mathbb{P}$-probabilité vers X et l'on note $X_n \to_{\mathbb{P}} X$, si et seulement si, pour tout $\delta > 0$, $\lim_{n\to\infty} \mathbb{P}\left[\|X_n - X\| > \delta\right] = 0$ où $\|\cdot\|$ désigne la norme euclidienne dans $\mathbb{R}^k$.*

Définition A.18 (Convergence en loi)**.** *On dit que X_n converge en loi (ou en distribution) vers X et l'on note $X_n \to_d X$, si et seulement si l'une des trois conditions équivalentes est satisfaite :*

1. *pour toute fonction f continue bornée $\mathbb{R}^d \to \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right].$$

2. *pour tout $u := (u_1, \cdots, u_d)$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\exp(iu^T X_n)\right] = \mathbb{E}\left[\exp(iu^T X)\right],$$

3. *Pour tout pavé $A = [a_1, b_1] \times \cdots \times [a_d, b_d]$ tel que $\mathbb{P}(X \in \partial A) = 0$ (où ∂A désigne la frontière de A),*

$$\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A).$$

Le théorème suivant permet de hiérarchiser les différents modes de convergence.

Théorème A.8. *1. Si $X_n \to_{\mathbb{P}-\text{p.s.}} X$, alors $X_n \to_{\mathbb{P}} X$.*

2. Si $X_n \to_{\mathcal{L}_r} X$, alors $X_n \to_{\mathbb{P}} X$.

3. Si $X_n \to_{\mathbb{P}} X$, alors $X_n \to_d X$.

4. Si $X_n \to_{\mathbb{P}} X$, alors on peut extraire une sous-suite $(X_{n_k}, k \in \mathbb{N})$, telle que $X_{n_k} \to_{\mathbb{P}-\text{p.s.}} X$.

Théorème de Helley et preuve du Théorème d'Herglotz

Théorème A.9. *Soit μ_n une suite de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, telle que, pour tout $\epsilon > 0$, il existe un ensemble compact $\mathcal{K}_\epsilon$, tel que $\mu_n(\mathcal{K}_\epsilon) \geq (1 - \epsilon)$. Alors, pour toute sous-suite $\{\mu_{n_k}\}$, il existe une sous-suite $\{\mu_{n_{k(j)}}\}$ extraite de $\{\mu_{n_k}\}$ et une probabilité μ telle que $\mu_{n_{k(j)}} \to_d \mu$ faiblement.*

La suite γ étant de type positif, $g_N(t) \geq 0$. Notons μ_N la mesure (positive) de densité g_N par rapport à la mesure de Lebesgue sur $\mathbb{T}$. On a

$$\hat{\mu}_N(p) = \int_{-\pi}^{\pi} g_N(t) e^{-ipt} dt = \left(1 - \frac{|p|}{N}\right) \gamma(-p).$$

pour $|p| \leq N$. En particulier on a $\mu_N(\mathbb{T}) = \gamma(0)$. De toute sous-suite $\{\nu_k = \mu_{N_k}\}$ de la suite $\{\mu_N\}$, on peut extraire une sous-suite $\{\nu_{k}\}$ qui converge étroitement vers une mesure positive μ (dépendant a

priori du choix de la sous suite) de masse totale $c(0)$ (théorème de Prohorov). On a, pour tout p pour tout $p \in \mathbb{Z}$

$$\hat{\mu}(p) = \lim_k \hat{\mu}_k(p) = \gamma(-p)$$

La limite $\hat{\nu}(p)$ ne dépend pas du choix de la sous-suite, et donc de toute sous-suite de la suite $\{\mu_N\}$, on peut extraire une sous-suite qui converge vers la **même** mesure limite μ. On en déduit que la suite μ_N converge étroitement vers μ. Lorsque $\sum_k |\gamma(k)| < \infty$, alors $g_N(t)$ converge vers $f(t)$ par application du théorème de convergence dominé. Les théorèmes suivants sont à la base des statistiques.

Théorème A.10 (Loi forte des grands nombres)**.** *Soient* $(X_n, n \in \mathbb{N})$ *une suite de v.a. indépendantes et identiquement distribuées (i.i.d) telles que* $\mathbb{E}[|] X_1| < \infty$*. Alors,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{\mathbb{P}-\text{p.s.}} \mu =: \mathbb{E}[X_1].$$

Ce théorème montre que la moyenne empirique $n^{-1}\sum_{i=1}^n X_i$ d'une suite de v.a i.i.d intégrables converge p.s. vers la moyenne de ces variables.

Théorème A.11 (Théorème de la limite centrale)**.** *Soient* $(X_n, n \in \mathbb{N})$ *une suite de v.a. i.i.d. telles que* $\mathbb{E}[X_i] = \mu$ *et* $\mathbb{E}\left[(X_i - \mu)^2\right] = \sigma^2 < \infty$*. Alors,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \to_d \mathcal{N}(0, \sigma^2).$$

Ce théorème permet d'évaluer la "vitesse" à laquelle la moyenne empirique $n^{-1}\sum_1^n X_i$ converge vers la moyenne $\mathbb{E}[X_1] =: \mu$. Ceci permet en particulier de déterminer, en statistique, des *intervalles de confiance.*

A.1.6 Espérance conditionnelle

Nous allons voir que, dans le cadre des variables aléatoires de carré intégrable, l'espérance conditionnelle par rapport à une sous-tribu $\mathcal{B}$ est la projection orthogonale sur l'ensemble des variables aléatoires de carré intégrable qui sont $\mathcal{B}$-mesurables. Ainsi $\mathbb{E}[X|Y]$ peut être vue comme la fonction de Y qui fournit la meilleure prédiction quadratique de X. En général cette fonction n'est pas *linéaire* de Y sauf dans le modèle gaussien. Nous allons tout d'abord donner une définition élémentaire de l'espérance conditionnelle à partir d'événements simples, puis nous étendrons cette définition aux variables aléatoires de carré intégrable. Enfin nous donnerons une définition plus générale pour les variables aléatoires posotives ou intégrables.

Construction élémentaire

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Soit $B \in \mathcal{F}$ un événement tel que $\mathbb{P}(B) > 0$ et $A \in \mathcal{F}$ un autre événement. On appelle *probabilité conditionnelle* de A sachant B la quantité :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

En notant $\mathbf{I}_A$ la variable aléatoire qui vaut 1 si $\omega \in A$ et 0 sinon (indicatrice de A), on peut alors écrire que :

$$\mathbb{P}(A|B) = \frac{1}{\mathbb{P}(B)} \int_B \mathbf{I}_A d\mathbb{P}$$

que l'on note $\mathbb{E}[\mathbf{I}_A|B]$. En généralisant cette expression à toute variable aléatoire X intégrable, on définit l'*espérance conditionnelle* de X sachant B par la quantité :

$$\mathbb{E}[X|B] = \frac{1}{\mathbb{P}(B)} \int_B X d\mathbb{P}$$

L'espérance conditionnelle $\mathbb{E}[X|B]$ représente l'espérance de la variable aléatoire X sachant que l'événement B s'est réalisé.

Exemple A.5
Soit X une variable aléatoire à valeurs dans l'ensemble des entiers naturels $\mathbb{N}$. La loi de X est spécifiée par la donnée des probabilités $p_k = \mathbb{P}(X = k)$, pour $k \in \mathbb{N}$. La moyenne de X est donnée par $\mathbb{E}[X] = \sum_{k\in\mathbb{N}} kp_k$. Considérons l'événement $B = \{X \geq k_0\}$. Nous avons $\mathbb{P}(B) = \sum_{k\geq k_0} p_k$ que nous supposerons non nul par hypothèse. L'espérance conditionnelle de X sachant B est donnée par :

$$\mathbb{E}[X|B] = \frac{1}{\sum_{k\geq k_0} p_k} \sum_{k\geq k_0} kp_k$$

Considérons maintenant la tribu $\mathcal{B} = \{\emptyset, \Omega, B, B^c\}$ (c'est-à-dire la plus petite tribu contenant B). On appelle l'*espérance conditionnelle de X sachant la tribu $\mathcal{B}$ la variable aléatoire*, notée $\mathbb{E}[X|\mathcal{B}]$ et définie par :

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\,\mathbf{I}_B + \mathbb{E}[X|B^c]\,\mathbf{I}_{B^c}$$

Cette variable aléatoire prend, suivant le résultat de l'expérience, soit la valeur $\mathbb{E}[X|B]$ soit la valeur $\mathbb{E}[X|B^c]$. De façon plus générale, si $\{B_k, k \geq 0\}$ désigne une famille d'événements formant une partition de Ω et telle que $\mathbb{P}(B_k) > 0$ et si $\mathcal{B}$ est la plus petite tribu engendrée par ces événements, on définit l'*espérance conditionnelle* de X sachant $\mathcal{B}$ par la *variable aléatoire* :

$$\mathbb{E}[X|\mathcal{B}] = \sum_{k\geq 0} \mathbb{E}[X|B_k]\,\mathbf{I}_{B_k} \tag{A.6}$$

On remarque que la variable aléatoire $\mathbb{E}[X|\mathcal{B}]$ est $\mathcal{B}$-mesurable et que, pour tout $B \in \mathcal{B}$, $\int_B \mathbb{E}[X|\mathcal{B}]\,dP = \int_B X dP$. On a donc la caractérisation suivante :

Proposition A.12. *L'espérance conditionnelle de la variable aléatoire X sachant la tribu $\mathcal{B}$ est l'unique variable aléatoire $\mathbb{E}[X|\mathcal{B}]$ qui soit $\mathcal{B}$-mesurable et telle que, pour tout $B \in \mathcal{B}$, on ait :*

$$\int_B \mathbb{E}[X|\mathcal{B}]\,dP = \int_B X dP \tag{A.7}$$

D'après l'équation (A.7), on remarque que, pour tout $B \in \mathcal{B}$, on a :

$$\int_\Omega (\mathbb{E}[X|\mathcal{B}] - X)\mathbf{I}_B d\mathbb{P} = 0$$

et donc que toute variable aléatoire $\mathcal{B}$-mesurable de la forme $Y = \sum_{k\geq 0} y_k \mathbf{I}_{B_k}$ (où y_k est une suite de réels), $\mathbb{E}[(\mathbb{E}[X|\mathcal{B}] - X)Y)] = 0$.

Espérance conditionnelle pour les variables aléatoires de carré intégrable

Le théorème A.12, qui suit, généralise la notion précédente d'espérance conditionnelle aux variables aléatoires de carré intégrable. Ce théorème est la conséquence directe de la structure Hilbertienne de l'ensemble $L^2(\Omega, \mathcal{B}, \mathbb{P})$ des variables aléatoires de carré intégrable et du théorème 4.2 de projection.

Théorème A.12. *Soit $\{\Omega, \mathcal{F}, \mathbb{P}\}$ un espace de probabilité et $\mathcal{B} \subset \mathcal{F}$ une sous-tribu de $\mathcal{F}$. On note $L^2(\Omega, \mathcal{F}, \mathbb{P})$ (resp. $L^2(\Omega, \mathcal{F}, \mathbb{P})$) l'espace des variables aléatoires $\mathcal{F}$-mesurables (resp. $\mathcal{B}$-mesurables) de carré intégrable. Soit X une variable aléatoire de $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Alors il existe une unique variable aléatoire appartenant à $L^2(\Omega, \mathcal{B}, \mathbb{P})$, notée $\mathbb{E}[X|\mathcal{B}]$ et qui vérifie simultanément, pour tout $Y \in L^2(\Omega, \mathcal{B}, \mathbb{P})$, les deux relations suivantes :*

$$\|X - \mathbb{E}[X|\mathcal{B}]\|^2 \leq \|X - Y\|^2 \tag{A.8}$$

$$(X - \mathbb{E}[X|\mathcal{B}], Y) = 0 \tag{A.9}$$

Remarquons, que, si $\mathcal{B}$ est une sous-tribu de $\mathcal{F}$, l'espace $L^2(\Omega, \mathcal{B}, \mathbb{P})$ est un sous-espace linéaire de $L^2(\Omega, \mathcal{F}, \mathbb{P})$, fermé par application de la proposition A.10. Nous pouvons donc appliquer le théorème de projection. Le théorème A.12 donne un sens à l'espérance conditionnelle pour des variables aléatoires de carré intégrable. Pour étendre cette définition aux variables aléatoires positives et/ou intégrables, nous avons besoin du lemme élémentaire d'unicité suivant :

Lemme A.5. *Soient X et Y deux variables aléatoires $\mathcal{B}$-mesurables toutes deux positives ou toutes deux intégrables vérifiant, pour tout $B \in \mathcal{B}$:*

$$\int_B X d\mathbb{P} \geq \int_B Y d\mathbb{P} \ (resp. \ =)$$

Alors, $X \geq Y$ (resp. $=$) $\mathbb{P}$-p.s.

Théorème A.13. *Soit X une variable aléatoire positive (resp. intégrable). Il existe une variable aléatoire Y positive (resp. intégrable) $\mathcal{B}$-mesurable, telle que, pour tout $B \in \mathcal{B}$, on ait :*

$$\int_B X d\mathbb{P} = \int_B Y d\mathbb{P}$$

Cette variable est unique à une équivalence près.

Démonstration. L'unicité découle du lemme A.5. Montrons l'existence. On suppose tout d'abord que $X \geq 0$. Pour $n \in \mathbb{N}$, définissons $X_n = X \wedge n := \min(X, n)$. $X_n \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, et il existe donc une v.a. $Y_n \geq 0$, $\mathcal{B}$-mesurable, unique à une équivalence près, telle que, pour tout $B \in \mathcal{B}$, on ait :

$$\int_B X_n d\mathbb{P} = \int_B Y_n d\mathbb{P}$$

Par application de A.5, Y_n est $\mathbb{P}$-p.s. une suite positive et croissante. En effet, pour tout $B \in \mathcal{B}$, on a :

$$\int_B Y_{n+1} d\mathbb{P} = \int_B X_{n+1} d\mathbb{P} \geq \int_B X_n d\mathbb{P} = \int_B Y_n d\mathbb{P}$$

Définissons $Y = \lim \uparrow Y_n$. Y est $\mathcal{B}$-mesurable, et par application du théorème de Beppo-Levi, pour tout $B \in \mathcal{B}$, on a :

$$\int_B Y d\mathbb{P} = \lim \uparrow \int_B Y_n d\mathbb{P} = \lim \uparrow \int_B X_n d\mathbb{P} = \int_B X dP$$

Notons que, si X est intégrable, alors Y l'est aussi (prendre $B = \Omega$). Pour étendre le résultat au cas intégrable, nous allons prouver que, pour X, Y deux v.a. positives intégrables, et pour $a, b \in \mathbb{R}$, nous avons (linéarité de l'espérance conditionnelle) :

$$\mathbb{E}[aX + bY|\mathcal{F}] = a\mathbb{E}[X|\mathcal{F}] + b\mathbb{E}[Y|\mathcal{F}]$$

Il suffit en effet de remarquer que, pour tout $B \in \mathcal{B}$, on a :

$$\begin{aligned}\int_B \mathbb{E}[aX + bY|\mathcal{F}]\, dP &= \int_B (aX + bY) dP = a\int_B X dP + b\int Y dP \\ &= a\int_B \mathbb{E}[X|\mathcal{B}]\, dP + b\int_B \mathbb{E}[Y|\mathcal{B}]\, dP = \int_B (a\mathbb{E}[X|\mathcal{B}] + b\mathbb{E}[Y|\mathcal{B}]) dP\end{aligned}$$

et on conclut en utilisant A.5. Pour $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, on pose $X = X^+ - X^-$, où $X^+ = \max(X, 0)$ et $X^- = \max(-X, 0)$ (on rappelle que, par définition, si $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, on a $\mathbb{E}[|X|] < +\infty$ et donc on a aussi $\mathbb{E}[X^+] < +\infty$ et $\mathbb{E}[X^-] < +\infty$) et nous concluons en utilisant l'existence de l'espérance conditionnelle pour les variables aléatoires positives et la linéarité de l'espérance conditionnelle. ■

Proposition A.13. *On note $L^1(\Omega, \mathcal{F}, \mathbb{P})$ l'ensemble des variables aléatoires intégrables définies sur l'espace de probabilité $\{\Omega, \mathcal{F}, \mathbb{P}\}$. On note $\mathcal{B}$ une sous-tribu de $\mathcal{F}$.*

1. *Pour tout couple de variables aléatoires $X, Y \geq 0$ (resp. $\in L^1(\Omega, \mathcal{F}, \mathbb{P})$) et pour tout couple de constantes $a, b \geq 0$ (resp. réelles), on a $\mathbb{E}[aX + bY|\mathcal{B}] = a\mathbb{E}[X|\mathcal{B}] + b\mathbb{E}[Y|\mathcal{B}]$.*
2. *Pour tout couple de variables aléatoires $X, Y \geq 0$ (ou $\in L^1(\Omega, \mathcal{F}, \mathbb{P})$), l'inégalité $X \leq Y$ $\mathbb{P}$-p.s. implique $\mathbb{E}[X|\mathcal{B}] \leq \mathbb{E}[Y|\mathcal{B}]$ $\mathbb{P}$-p.s.*
3. *Pour tout couple de variables aléatoires $X, Y \geq 0$ (ou $\in L^1(\Omega, \mathcal{F}, \mathbb{P})$) où Y est $\mathcal{B}$-mesurable, on a $\mathbb{E}[(X - \mathbb{E}[X|\mathcal{B}])Y] = 0$.*
4. *Pour toute variable aléatoire $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ et toute variable aléatoire Y bornée et $\mathcal{B}$-mesurable, on a $\mathbb{E}[(X - \mathbb{E}[X|\mathcal{B}])Y] = 0$.*

La proposition, qui suit, regroupe des propriétés essentielles de l'espérance conditionnelle.

Proposition A.14. *On note $L^1(\Omega, \mathcal{F}, \mathbb{P})$ l'ensemble des variables aléatoires intégrables définies sur $\{\Omega, \mathcal{F}, \mathbb{P}\}$.*

1. *Soit $\mathcal{G}$ la tribu grossière : $\mathcal{G} = \{\Omega, \emptyset\}$. Alors, pour tout $X \geq 0$ (ou $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$), on a $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.*
2. *Soit $\mathcal{A} \subset \mathcal{B}$ deux sous-tribus de $\mathcal{F}$. Alors, pour toute variable aléatoire $X \geq 0$ (ou $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$), on a :*
$$\mathbb{E}[\mathbb{E}[X|\mathcal{B}]\,|\mathcal{A}] = \mathbb{E}[X|\mathcal{A}]$$
3. *Soit $X \geq 0$ (ou $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$) une variable aléatoire indépendante de $\mathcal{B}$ alors on a $\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X]$.*

4. *Soit $X \geq 0$ (ou $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$) et $Y \geq 0$ (ou $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$) une variable aléatoire $\mathcal{B}$-mesurable, alors on a $\mathbb{E}[XY|\mathcal{B}] = Y\mathbb{E}[X|\mathcal{B}]$.*

Démonstration. Les fonctions mesurables par rapport à la tribu grossière sont les fonctions constantes. Or, pour tout $B \in \mathcal{G}$ ($B = \emptyset$ ou $B = \Omega$), on a :

$$\int_B \mathbb{E}[X]\, d\mathbb{P} = \int_B X d\mathbb{P}$$

et donc la fonction constante $\mathbb{E}[X]$ vérifie (A.7), ce qui prouve le point (1). Prouvons maintenant (2). Soit Y une variable aléatoire $\mathcal{A}$-mesurable bornée. Notons que $\mathcal{A} \subset \mathcal{B}$ implique que Y est aussi $\mathcal{B}$-mesurable. Par conséquent, par définition de l'espérance conditionnelle appliquée à la variable aléatoire $Z = \mathbb{E}[X|\mathcal{B}]$, on a successivement :

$$\mathbb{E}[\mathbb{E}[Z|\mathcal{A}]\, Y] = \mathbb{E}[ZY] = \mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[X|\mathcal{A}]\, Y]$$

et donc, pour toute variable aléatoire Y qui est $\mathcal{A}$-mesurable bornée, on a $\mathbb{E}[\mathbb{E}[Z|\mathcal{A}]\, Y] = \mathbb{E}[\mathbb{E}[X|\mathcal{A}]\, Y]$. Ce qui entraîne que les deux variables aléatoires $\mathcal{A}$-mesurables $\mathbb{E}[Z|\mathcal{A}]$ et $\mathbb{E}[X|\mathcal{A}]$ coïncident, ce qui prouve (2). Soit maintenant X une variable aléatoire indépendante de $\mathcal{B}$. Alors, par définition de l'indépendance, pour toute variable aléatoire Y qui est $\mathcal{B}$-mesurable bornée, on a $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. On en déduit que :

$$\mathbb{E}[\mathbb{E}[X|\mathcal{B}]\, Y] = \mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[X]\, Y]$$

ce qui prouve (3). Considérons finalement (4). On a, pour toute variable aléatoire Z bornée $\mathcal{B}$-mesurable :

$$\mathbb{E}[\mathbb{E}[XY|\mathcal{B}]\, Z] = \mathbb{E}[YXZ] = \mathbb{E}[(\mathbb{E}[Y|\mathcal{B}]\, X)Z]$$

la dernière égalité est justifiée puisque XZ est $\mathcal{B}$-mesurable. Comme la variable aléatoire $\mathbb{E}[Y|\mathcal{B}]\, X$ est elle-même $\mathcal{B}$-mesurable, elle s'identifie à $\mathbb{E}[XY|\mathcal{B}]$. Ce qui prouve (4). ■

Proposition A.15. *Les propriétés suivantes sont l'extension à l'espérance conditionnelle de propriétés fondamentales de l'espérance.*

1. *(Convergence monotone conditionnelle) Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires telles que $0 \leq X_n \uparrow X$. Alors $\mathbb{E}[X_n|\mathcal{B}] \uparrow \mathbb{E}[X|\mathcal{B}]$.*
2. *(Lemme de Fatou conditionnel) Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires positives. Alors $\mathbb{E}[\liminf X_n|\mathcal{B}] \leq \liminf \mathbb{E}[X_n|\mathcal{B}]$.*
3. *(Convergence dominée conditionnelle) Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires telle que $|X_n| \leq V$ $\mathbb{P}$-p.s., avec $\mathbb{E}[V] < \infty$ et $X_n \to X$ $\mathbb{P}$-p.s. Alors, $\mathbb{E}[X_n|\mathcal{B}] \to \mathbb{E}[X|\mathcal{B}]$ $\mathbb{P}$-p.s.*
4. *(Inégalité de Jensen conditionnelle) Soit $c : \mathbb{R} \to \mathbb{R}$ convexe telle que $\mathbb{E}[|c(X)|] < \infty$. Alors, $\mathbb{E}[c(X)|\mathcal{B}] \leq c(\mathbb{E}[X|\mathcal{B}])$.*
5. *(Contraction des normes) Pour $p \geq 1$, $\|\mathbb{E}[X|\mathcal{B}]\|_p \leq \|X\|_p$, où $\|Y\|_p := (\mathbb{E}[|Y|^p])^{1/p}$.*

Définition A.19. *Soit deux variables aléatoires définies sur le même espace de probabilité $\{\Omega, \mathcal{F}, \mathbb{P}\}$. On appelle espérance conditionnelle de X par rapport à Y :*

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$$

où $\sigma(Y)$ désigne la tribu engendré par Y (la plus petite tribu rendant Y mesurable).

A.2 Estimation statistique

Lors d'une expérience aléatoire, l'observation est modélisée comme un point d'un espace mesurable $\{H, \mathcal{H}\}$ dont la loi de probabilité nous est inconnue. Le but de l'estimation ponctuelle est de fournir, à partir d'une suite d'observations d'une expérience aléatoire, la valeur d'un paramètre relié à la loi de probabilité inconnue. Dans la suite, le plus souvent, ce paramètre est un scalaire ou un vecteur de dimension fini. Un estimateur est alors défini comme une fonction mesurable, arbitraire, de l'observation à valeurs dans l'espace du paramètre. D'où le problème de définir, au moyen de critères raisonnables, ce que l'on entend par "un estimateur est *bon*" et comment, à partir d'un critère, construire, si possible, le meilleur d'entre eux. Dans ce paragraphe nous donnons les définitions du biais et de la dispersion quadratique ainsi que des propriétés asymptotiques. Toutes ces notions sont à la base de la comparaison des estimateurs entre eux.

A.2.1 Biais, dispersion d'un estimateur

Définition A.20 (Modèle statistique)**.** *Un modèle statistique est un triplet $\{H, \mathcal{H}, \mathcal{P}\}$ où $\{H, \mathcal{H}\}$ est un espace mesurable et $\mathcal{P}$ est une famille de mesures de probabilité définies sur $\{H, \mathcal{H}\}$.*

Dans la suite, le plus souvent, les observations sont réelles : on aura alors, dans le cas des échantillons de taille n finie, $H = \mathbb{R}^n$ et, dans le cas de l'étude des propriétés asymptotiques, $H = \mathbb{R}^{\mathbb{N}}$. En estimation statistique, il est d'usage de distinguer deux approches : l'approche liée aux modèles paramétriques et celle liée aux modèles non-paramétriques. Dans le premier cas, la famille $\mathcal{P}$ possède une structure dépendant d'un paramètre d'intérêt de dimension finie : si on connaît alors la vraie valeur du paramètre, on dispose très exactement de la loi de probabilité de l'observation. Dans le second cas, on fait très peu d'hypothèses sur la famille $\mathcal{P}$ et la connaissance du paramètre d'intérêt ne permet plus de reconstruire la loi de probabilité de l'observation. Dans ce dernier cas, il est même possible que le paramètre d'intérêt ne soit plus de dimension finie.

Exemple A.6 : MA(1) gaussien
On observe la suite $(X_1, \ldots, X_n)$ d'un processus MA(1) défini par $X_t = Z_t + \theta_1 Z_{t-1}$ où Z_t est un bruit gaussien, centré, blanc (fort) de variance σ^2. Le modèle est paramétrique. La loi de l'observation ne dépend, en effet, que de $\theta = (\theta_1, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Sa densité a pour expression :

$$p_X(x_1, \ldots, x_n; \theta) = \frac{1}{(2\pi)^{n/2}\sigma^n \sqrt{\det(C(\theta_1))}} \exp\left\{-\frac{1}{2\sigma^2}(x_1, \ldots, x_n)C^{-1}(\theta_1)(x_1, \ldots, x_n)^T\right\}$$

où

$$C(\theta_1) = \begin{bmatrix} 1+\theta_1^2 & \theta_1 & 0 & \cdots & 0 \\ \theta_1 & 1+\theta_1^2 & \theta_1 & \cdots & 0 \\ \vdots & & & & \\ 0 & & & 1+\theta_1^2 & \theta_1 \\ 0 & \cdots & 0 & \theta_1 & 1+\theta_1^2 \end{bmatrix}$$

Si on omet l'hypothèse gaussienne, on ne peut plus, connaissant uniquement θ, écrire la loi de l'observation. Dans ce cas, le modèle est dit semi-paramétrique. Si, à présent, on omet aussi l'hypothèse que le processus est un processus MA(1) et que l'on suppose uniquement que l'observation provient d'un processus stationnaire au second ordre, il n'y a plus, à proprement parler, de paramètres d'intérêt de dimension finie. On dit alors que le modèle est non-paramétrique.

Définition A.21 (Estimateur)**.** *Soit le modèle statistique $\{H, \mathcal{H}, \mathcal{P}\}$. On suppose que $\mathbb{P} \in \mathcal{P}$ dépend d'un paramètre θ élément d'un espace mesurable $\{\Theta, \mathcal{B}(\Theta)\}$. On appelle estimateur de $\theta \in \Theta$ toute fonction mesurable de $\{H, \mathcal{H}\}$ dans $\{\Theta, \mathcal{B}(\Theta)\}$.*

Définition A.22 (Biais d'un estimateur)**.** *Soit $\{\mathbb{R}^n, \mathcal{B}^n, \mathcal{P}\}$ un modèle statistique, soit $\theta \in \Theta \subset \mathbb{R}^k$ un paramètre à estimer et soit $\hat{\theta} : \{\mathbb{R}^n, \mathcal{B}^n\} \mapsto \{\Theta, \mathcal{B}(\Theta)\}$ un estimateur de θ. On appelle biais de $\hat{\theta}$ le vecteur de $\mathbb{R}^k$ défini par :*

$$b(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\hat{\theta}(X_1, \ldots, X_n) \right] - \theta \tag{A.10}$$

Un estimateur est dit sans biais si $b(\theta, \hat{\theta}) = 0$ pour tout $\theta \in \Theta$.

Définition A.23 (Dispersion et risque quadratique)**.** *Soit $\{\mathbb{R}^n, \mathcal{B}^n, \mathcal{P}\}$ un modèle statistique, soit $\theta \in \Theta \subset \mathbb{R}^k$ un paramètre à estimer et soit $\hat{\theta} : \{\mathbb{R}^n, \mathcal{B}^n\} \mapsto \{\Theta, \mathcal{B}(\Theta)\}$ un estimateur de θ. On appelle matrice de dispersion de l'estimateur $\hat{\theta}$ la matrice, de dimension $k \times k$, définie par :*

$$D(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[(\hat{\theta}(X_1, \ldots, X_n) - \theta)(\hat{\theta}(X_1, \ldots, X_n) - \theta)^T \right] \tag{A.11}$$

On dit que $\hat{\theta}^{(1)}(X_1, \ldots, X_n)$ est meilleur que $\hat{\theta}^{(2)}(X_1, \ldots, X_n)$, si, pour tout $\theta \in \Theta$, on a :

$$D(\theta, \hat{\theta}^{(1)}) \leq D(\theta, \hat{\theta}^{(2)}) \tag{A.12}$$

On appelle risque quadratique de $\hat{\theta}$:

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[(\hat{\theta}(X_1, \ldots, X_n) - \theta)^T (\hat{\theta}(X_1, \ldots, X_n) - \theta) \right] = \text{Trace}(D(\theta, \hat{\theta}))$$

La notation $\mathbb{E}_\theta$ indique que l'espérance doit être calculée avec la loi de l'observation lorsque la valeur du paramètre inconnu est précisément θ :

$$\mathbb{E}_\theta = \int_{\mathbb{R}^n} \hat{\theta}(X_1, \ldots, X_n) \mathbb{P}_\theta(dx)$$

Il s'en suit qu'en règle générale, le biais et la dispersion quadratique dépendent du paramètre inconnu θ. Il est important de noter que la relation (A.12) ne permet pas d'ordonner totalement les estimateurs, dans le sens où deux estimateurs ne sont pas nécessairement comparables. Il est donc vain de vouloir trouver un estimateur qui soit meilleur que tous les autres pour toute valeur de θ. Ajoutons par ailleurs que, dans les situations rencontrées en pratique, le calcul explicite du biais et de la dispersion est souvent impossible. On peut alors, pour juger des performances, soit calculer des bornes, la plus utilisée étant la borne inférieure de Cramer-Rao, soit déterminer les performances lorsque la taille de l'échantillon tend vers l'infini.

Théorème A.14 (Borne de Cramer-Rao)**.** *Soit une modèle statistique $\{H, \mathcal{H}, \mathcal{P}\}$ dominé par la mesure μ et soit Θ une partie ouverte de $\mathbb{R}^k$. On note $p(x; \theta)$ la densité de $P_\theta \in \mathcal{P}$ par rapport à μ. On suppose :*

- *que θ, $p(x; \theta)$ est, μ-presque partout, continûment dérivable,*
- *et que la matrice d'information de Fisher, de dimension $k \times k$, :*

$$F(\theta) = \int_H \frac{\partial \log p(x; \theta)}{\partial \theta} \frac{\partial \log p(x; \theta)}{\partial \theta}^T p(x; \theta) \mu(dx)$$

est définie positive pour toute valeur du paramètre θ et continue par rapport à θ.

Soit $\hat{\theta}(X_1, \ldots, X_n)$ un estimateur de θ. On note :

$$b(\theta, \hat{\theta}) = [b_1(\theta, \hat{\theta}) \quad \ldots \quad b_k(\theta, \hat{\theta})]^T = \mathbb{E}_\theta \left[\hat{\theta}(X_1, \ldots, X_n)\right] - \theta$$

le biais de cet estimateur. Alors le risque quadratique vérifie :

$$R(\theta, \hat{\theta}) \geq (I_k + \partial_\theta b(\theta, \hat{\theta}))F^{-1}(\theta)(I_k + \partial_\theta b(\theta, \hat{\theta}))^T + b(\theta, \hat{\theta})b(\theta, \hat{\theta})^T \tag{A.13}$$

$\partial_\theta b(\theta, \hat{\theta})$ désigne la matrice de dimension $k \times k$ dont l'élément général est $\partial b_m(\theta, \hat{\theta})/\partial_{\theta_j}$. On montre que :

$$F(\theta) = -\int_H \partial^2_{\theta^2} \log p(x; \theta) p(x; \theta) \mu(dx) \tag{A.14}$$

où $\partial^2_{\theta^2} \log p(x; \theta)$ désigne la matrice Hessien d'élément général $\partial^2 \log p(x; \theta))/\partial_{\theta_j}\partial_{\theta_m}$.

Dans la classe des estimateurs sans biais, la borne de Cramer-Rao a pour expression :

$$R(\theta, \hat{\theta}) \geq F^{-1}(\theta)$$

A.2.2 Comportement asymptotique d'un estimateur

Voyons à présent quelques résultats concernant les propriétés asymptotiques.

Définition A.24 (Consistance)**.** *Soit un modèle statistique dépendant du paramètre $\theta \in \Theta \subset \mathbb{R}^k$ et soit $\hat{\theta}_n(X_1, \cdots, X_n)$ une suite d'estimateurs de θ. On dit que la suite $\hat{\theta}_n(X_1, \cdots, X_n)$ est consistante si, pour tout $\theta \in \Theta$, la suite de vecteurs aléatoires $\hat{\theta}_n(X_1, \cdots, X_n)$ converge en P_θ-probabilité vers θ.*

Définition A.25 (Normalité asymptotique)**.** *Soit un modèle statistique dépendant du paramètre $\theta \in \Theta \subset \mathbb{R}^k$ et soit $\hat{\theta}_n(X_1, \cdots, X_n)$ une suite d'estimateurs de θ. On dit que la suite $\hat{\theta}_n(X_1, \cdots, X_n)$ est asymptotiquement normale si, il existe une constante $\alpha > 0$ et une $\Gamma(\theta)$ définie positive telle que, pour tout $\theta \in \Theta$:*

$$n^\alpha(\hat{\theta}_n(X_1, \cdots, X_n) - \theta) \to_d \mathcal{N}(0, \Gamma(\theta)) \tag{A.15}$$

où $\mathcal{N}(0, \Gamma)$ désigne la loi gaussienne centrée, de matrice de covariance Γ.

Dans le cas des suites i.i.d., la consistance et la normalité asymptotique sont, le plus souvent, la conséquence directe, d'une part, de la loi des grands nombres et du théorème de la limite centrale et, d'autre part, de théorèmes de continuité.

Théorème A.15 (Loi faible des grands nombres)**.** *Soit $\{X_n\}_{n\geq 1}$ une suite de vecteurs aléatoires de dimension k, indépendants et identiquement distribués, de moyenne $\mathbb{E}[X_1]$ et de variances finies. Alors :*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \to_{\mathbb{P}} \mathbb{E}[X_1]$$

Théorème A.16 (Théorème de la limite centrale)**.** *Soit $\{X_n\}_{n\geq 1}$ une suite de vecteurs aléatoires de dimension k, indépendants et identiquement distribués, de moyenne $\mathbb{E}[X_1]$ et de matrice de covariance $\mathrm{cov}(X_1)$ supposée définie positive. Alors :*

$$n^{1/2}\left(\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbb{E}[X_1]\right) \to_d \mathcal{N}(0, \mathrm{cov}(X_1))$$

Théorème A.17. *Soit $\{X_n\}_{n\geq 0}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^k$. Supposons que $X_n \to_{\mathbb{P}} X$, et soit $\mathcal{X}$ un sous-ensemble borelien de $\mathbb{R}^k$ tel que $\mathbb{P}[X \in \mathcal{X}] = 1$. Si $g : \mathbb{R}^k \to \mathbb{R}^m$ est continue sur $\mathcal{X}$ alors $g(X_n) \to_{\mathbb{P}} g(X)$,*

Théorème A.18. *Soit $\{X_n\}$ une suite de vecteurs aléatoires de dimension k telle que :*

$$n^{\alpha}(X_n - \mu) \to_d \mathcal{N}(0, \Gamma)$$

où α est une constante positive et Γ une matrice de covariance définie positive. Soit $g = (g_1, \cdots, g_m) : \mathbb{R}^k \to \mathbb{R}^m$ une fonction différentiable au point μ, de matrice différentielle D, de dimension $m \times k$, au point μ :

$$D = \left[\frac{\partial g_\ell(\mu)}{\partial x_j}\right]$$

telle que la matrice $\Phi = D\Gamma$, de dimension $m \times m$, soit définie positive. Alors :

$$n^{\alpha}(g(X_n) - g(\mu)) \to_d \mathcal{N}(0, \Phi)$$

Définition A.26 (Quantité pivotale)**.** *Pour des observations $X_1, \ldots, X_n$ issues d'un modèle paramétrique de paramètre θ, une quantité T_n fonction de $X_1, \ldots, X_n$ et de θ est dite pivotale si sa distribution ne dépend pas du paramètre θ. Dans les cas où cette propriété n'est pas vérifiée à n fini mais ou néanmoins T_n converge en distribution vers une loi ne dépendant pas de θ, la quantité1 T_n est dite asymptotiquement pivotale.*

Un exemple simple de cette situation est le cas d'un paramètre de centrage où les observations sont supposées iid de loi $f(x-\mu)$ pour une loi $f(x)$ connue, μ étant le paramètre. Dans ce cas, on vérifie directement que pour l'estimateur de la moyenne empirique $\hat{\mu}_n = n^{-1}\sum_{t=1}^{n} X_t$, la quantité $\hat{\mu}_n - \mu$ est pivotale. A n fini, cette propriété peut néanmoins être difficile à exploiter dans la mesure où la loi de $\hat{\mu}_n - \mu$ n'a pas forcément une expression simple (sa fonction caractéristique par contre vaut $\Phi_f(\frac{\lambda}{n})^n$ où $\Phi_f(\lambda)$ est la fonction caractéristique associée à f). On note cependant que $\sqrt{n}(\hat{\mu}_n - \mu)$ est une quantité asymptotiquement pivotale dans la mesure où le théorème de la *limite centrale* A.16 indique que dès que $\mathbb{E}\left[(X_i - \mu)^2\right] = \sigma^2 < \infty$,

$$\sqrt{n}(\hat{\mu}_n - \mu) \to_d \mathcal{N}(0, \sigma^2) \tag{A.16}$$

En pratique, même dans le modèle de centrage, il est fréquent que la variance σ^2 soit également un paramètre inconnu à estimer. Il est néanmoins possible d'obtenir une quantité asymptotiquement pivotale en remplaçant σ^2 par une estimation consistante :

Propriété A.1. *Si μ_n est une séquence asymptotiquement normale telle que*

$$\sqrt{n}(\hat{\mu}_n - \mu) \to_d \mathcal{N}(0, \sigma^2)$$

et σ_n est un estimateur consistant de σ, on a

$$\sqrt{n}\sigma_n^{-1}(\hat{\mu}_n - \mu) \to_d \mathcal{N}(0, 1)$$

ce qui implique que $\sqrt{n}\sigma_n^{-1}(\hat{\mu}_n - \mu)$ est une quantité asymptotiquement pivotale.

Cette propriété montre que dès qu'un estimateur est asymptotiquement normal, il est général au moins possible de trouver des quantités asymptotiquement pivotales. Cette propriété est capitale pour la construction d'intervalles de confiance qui mesurent la fiabilité du résultat d'estimation ainsi que pour le test, c'est à dire la validation d'hypothèses concernant certains paramètres du modèle.

Définition A.27 (Intervalle de confiance asymptotique)**.** *Un intervalle de confiance asymptotique de niveau α pour le paramètre scalaire inconnu θ est une suite d'intervalles, de la forme $J_n = [T_{1,n}, T_{2,n}]$ où $T_{1,n} = T_1(X_1, X_2, \cdots, X_n)$ et $T_{2,n} = T_2(X_1, X_2, \cdots, X_n)$ sont des variables aléatoires, telle que :*

$$\lim_{n\to\infty} \mathbb{P}(\theta \in J_n) = \alpha \tag{A.17}$$

Dans le cas du paramètre de centrage en supposant que la variance σ est *connue*, nous avons, d'après (A.16)

$$\lim_{n\to\infty} \mathbb{P}(\mu \in [T_{1,n}, T_{2,n}]) = \lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}(\hat{\mu}_n - \mu) \in [-c, c]\right) = 2\int_0^c \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

où nous avons posé $T_{1,n} = \hat{\mu}_n - c\sigma/\sqrt{n}$ et $T_{2,n} = \hat{\mu}_n + c\sigma/\sqrt{n}$. Et donc, si nous choisissons c de telle sorte que l'intégrale soit égale à α, nous obtenons un intervalle $J_n = [T_{1,n}, T_{2,n}]$ qui vérifie l'expression (A.17). Ainsi, par exemple, pour $\alpha = 0.95$ on $c = 1.96$ et :

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\mu}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) = 95\% \tag{A.18}$$

Dans le cas où σ est inconnu, il est possible de le remplacer par un estimateur consistant σ_n en vertu de la propriété A.1.

Définition A.28 (Test asymptotique)**.** *Un test asymptotique pour l'hypothèse $\theta = \theta_0$ est une fonction T_n des observations $X_1, \ldots, X_n$ et de θ_0 à valeur dans $\{0, 1\}$ (1 pour l'acceptation de l'hypothèse, 0 pour son rejet) telle que*

$$\lim_n \mathbb{P}_\theta(T_n = 1) = 0 \quad \textit{quand } \theta \neq \theta_0$$

et

$$\lim_n \mathbb{P}_{\theta_0}(T_n = 1) = \alpha$$

où $1 - \alpha$ est dite probabilité d'erreur de première espèce (ou de rejet à tort de l'hypothèse $\theta = \theta_0$).

L'existence de quantités pivotales est également un point clé pour le test puisque pour l'exemple du paramètre de centrage, l'expression (A.18), donnant l'intervalle de confiance asymptotique avec un niveau de confiance de 95% pour μ, peut être encore écrite sous la forme

$$\lim_{n\to\infty} \mathbb{P}\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \hat{\mu}_n \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 95\%$$

Ainsi la fonction T_n permettant de tester que $\mu = \mu_0$ dans ce modèle est donnée par $\mathbb{I}_{]}\mu_0 - 1.96\frac{\sigma}{\sqrt{n}}, \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}](\hat{\mu}_n)$ où $\mathbb{I}$ désigne la fonction indicatrice. Par construction, $\lim_n \mathbb{P}_{\mu_0}(T_n = 1) = 95\%$. De plus dès que $\hat{\mu}_n$ est un estimateur consistant du paramètre inconnu μ et dans la mesure où les bornes de l'intervalle $[\mu_0 - 1.96\frac{\sigma}{\sqrt{n}}, \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}]$ se rapprochent (à la vitesse $1/\sqrt{n}$), il est immédiat que $\lim_n \mathbb{P}_{\mu}(T_n = 1) = 0$ lorsque $\mu \neq \mu_0$. Comme dans le cas de l'intervalle de confiance, la propriété A.1 permet également de traiter le cas où la variance limite σ^2 est inconnue (du moment que l'on dispose d'un estimateur consistant de cette dernière).

Annexe B

Rappels sur la transformée de Fourier

Dans toute la suite, I désigne l'intervalle $I = [-\pi, \pi]$ et $\mathcal{B}(I)$ la tribu de Borel de I construite sur les ouverts de I.

Propriété B.1 (Transformée de Fourier discrète d'une suite sommable)**.** *Soit $R(n)$ une suite complexes de module sommable. Alors :*

$$R(n) = \int_I e^{in\lambda} f(\lambda) d\lambda \quad où \quad f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} R(n) e^{-in\lambda}$$

D'après l'absolue sommabilité de $R(n)$, $f(\lambda)$ existe. Du fait que $\int_I \sum_n |R(n)| d\lambda < +\infty$, l'application directe du théorème de Fubini donne :

$$\int_I e^{in\lambda} f(\lambda) d\lambda = \int_I e^{in\lambda} \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} R(k) e^{-ik\lambda} d\lambda = \sum_{k=-\infty}^{\infty} R(k) \frac{1}{2\pi} \int_I e^{i(n-k)\lambda} d\lambda = R(n)$$

Propriété B.2 (Coefficients de Fourier d'une mesure finie)**.** *Soit ν une mesure non-négative, définie sur $\{I, \mathcal{B}(I)\}$, finie (i.e. telle que $\int_I \nu(d\lambda) < +\infty$) et soit $n \in \mathbb{Z}$. On appelle n-ième coefficient de Fourier de ν :*

$$\hat{\nu}(n) = \int_I e^{i\lambda n} \nu(d\lambda)$$

Du fait que la mesure est finie $|\hat{\nu}(n)|$ est fini.

1. *L'application $\nu \to \hat{\nu}$ est injective.*
2. *La suite $\{\hat{\nu}\}$ est de type non-négatif.*
3. *Soit $\{\nu_n\}_{n\geq 0}$ et ν des mesures finies. La suite de mesures $\{\nu_n\}$ converge étroitement vers la mesure ν (quand n tend vers l'infini), si et seulement si, pour tout $k \in \mathbb{Z}$, $\hat{\nu}_n(k)$ converge vers $\hat{\nu}(k)$ (quand n tend vers l'infini).*

1. $\mathcal{C}_b(I)$ désigne l'ensemble des fonctions complexes, continues et bornées, définies sur $I = [-\pi, \pi]$, muni de la topologie associée à la norme uniforme $\|f\|_\infty = \sup_{\lambda \in [-\pi,\pi]} |f(\lambda)|$. Précisons que l'égalité $\nu_1 = \nu_2$ doit être comprise dans le sens où $\int_I f(\lambda)\nu_1(d\lambda) = \int_I f(\lambda)\nu_2(d\lambda)$ pour toute fonction $f \in \mathcal{C}_b(I)$. Le point 1 est alors une conséquence directe du fait que les combinaisons

linéaires d'exponentielles complexes, de la forme $e^{in\lambda}$, sont denses dans $\mathcal{C}_b(I)$. L'application qui, à tout $f \in \mathcal{C}_b(I)$ fait correspondre le nombre complexe $c_\nu(f) = \int f(\lambda)\nu(d\lambda) \in \mathbb{C}$ est une forme linéaire continue sur $\mathcal{C}_b(I)$, qui associe aux exponentielles complexes de la forme $e^{in\lambda}$ les coefficients de Fourier $c_\nu(e^{in\bullet}) = \hat{\nu}(n)$. Par conséquent, si pour deux mesures ν et ν, les formes linéaires associées, c_ν et c_ν, coïncident pour les exponentielles complexes (i.e. $\hat{\nu}(n) = \hat{\nu}(n)$), alors elles coïncident pour toute fonction de $\mathcal{C}_b(I)$. Ce qui démontre le point 1.

2. Soit $(z_1, z_2, \cdots, z_n)$ des nombres complexes. On a :

$$\sum_{r,s=1}^{d} z_r z_s^* \hat{\nu}(r-s) = \int_I \sum_{r,s=1}^{d} z_s z_r^* e^{i(r-s)\lambda} \nu(d\lambda) = \int_I \left| \sum_{r=1}^{d} z_r e^{-ir\lambda} \right|^2 \nu(d\lambda) \geq 0$$

3. Par définition, la suite de mesure ν_n converge étroitement vers ν si pour toute fonction $f \in \mathcal{C}_b(I)$, $\lim_n c_{\nu_n}(f) = c_\nu(f)$. En particulier, si on prend $f = e^{-ik\bullet}$ (qui est continue et bornée), nous avons $c_{\nu_n}(e^{ik\bullet}) = \hat{\nu}_n(k) \to \hat{\nu}(k)$. Réciproquement, soit $\{\nu_n\}$ une suite de mesures finies sur I telles que, pour tout $k \in \mathbb{Z}$, $\lim_n \hat{\nu}_n(k) = \nu(k)$. Cette propriété implique en particulier que la suite $\hat{\nu}_n(0) = \nu_n(I)$ est convergente, et est donc bornée, $\sup_{n\geq 0} \hat{\nu}_n(0) < \infty$. Remarquons aussi que $|\hat{\nu}_n(k)| \leq \nu_n(0)$. Pour $f \in L^2(I, d\lambda)$ (où $d\lambda$ désigne la mesure de Lebesgue), définissons :

$$\hat{f}(k) = \int_I f(t) e^{-ikt} dt$$

Considérons la classe $\mathcal{F}$ de fonctions f vérifiant $\sum_{k\in\mathbb{Z}} |\hat{f}(k)| < \infty$. La classe $\mathcal{F}$ est dense dans $\mathcal{C}_b(I)$. Notons que, pour toute fonction $f \in \mathcal{F}$, nous avons :

$$f(\lambda) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \hat{f}(k) e^{-ik\lambda}$$

Par conséquent, en appliquant le théorème de Fubini, on a :

$$c_{\nu_n}(f) = \int_I f(\lambda)\nu_n(d\lambda) = \frac{1}{2\pi} \int_I \sum_{k\in\mathbb{Z}} \hat{f}(k) e^{-ik\lambda} \nu_n(d\lambda), = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \hat{f}(k)\hat{\nu}_n(k)$$

Comme $\sup_k \sup_n |\hat{\nu}_n(k)| < \infty$, le théorème de convergence dominée et le théorème de Fubini impliquent que :

$$\lim_n c_{\nu_n}(f) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \hat{f}(k) \lim_{n\to+\infty} \hat{\nu}_n(k) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \hat{f}(k)\hat{\nu}(k) = c_\nu(f)$$

Soit maintenant f une fonction continue. Pour tout $\epsilon > 0$, il existe $f_\epsilon \in \mathcal{F}$ tel que $\|f - f_\epsilon\|_\infty \leq \epsilon$ et nous avons :

$$\begin{aligned} |\nu_n(f) - \nu(f)| &\leq |\nu_n(f_\epsilon) - \nu(f_\epsilon)| + |\nu(f_\epsilon) - \nu(f)| \\ &\leq |\nu_n(f_\epsilon) - \nu(f_\epsilon)| + \|f - f_\epsilon\|_\infty (|\hat{\nu}_n(0)| + |\hat{\nu}(0)|) \end{aligned}$$

et donc puisque $f_\epsilon \in \mathcal{F}$ la limite du premier terme est 0 et on a :

$$\limsup_n |\nu_n(f) - \nu(f)| \leq 2\epsilon |\hat{\nu}(0)|$$

Comme ϵ est arbitraire, nous avons donc $\lim_n \nu_n(f) = \nu(f)$, ce qui conclut la preuve.

Annexe C

Compléments sur les espaces de Hilbert

Théorème C.1. *Si $\mathcal{E}$ est un sous-ensemble d'un espace de Hilbert $\mathcal{H}$, alors $\mathcal{E}^\perp$ est un sous-espace fermé.*

Démonstration. Soit $(x_n)_{n\geq 0}$ une suite convergente d'éléments de $\mathcal{E}^\perp$. Notons x la limite de cette suite. Par continuité du produit scalaire nous avons, pour tout $y \in \mathcal{E}$,

$$(x, y) = \lim_{n\to\infty} (x_n, y) = 0$$

et donc $x \in \mathcal{E}^\perp$. ■

Définition C.1 (Famille orthonormale)**.** *Soit $E = \{e_j; j \in T\}$ un sous ensemble de $\mathcal{H}$. On dit que E est une famille orthonormale ssi $(e_i, e_j) = \delta(i - j)$.*

Exemple C.1

Propriété C.1 (Inégalité de Bessel)**.** *Si x est un vecteur d'un espace de Hilbert $\mathcal{H}$ et si $E = \{e_1, \cdots, e_k\}$ est une famille orthonormale finie, alors :*

$$\sum_{i=}^{k} |(x, e_i)|^2 \leq \|x\|^2$$

Démonstration. Notons $\mathbf{E} = \overline{\text{span}}(E)$ le sous-espace engendre par les vecteurs $\{e_1, \cdots, e_k\}$. Nous avons $\|(x|\mathbf{E})\| \leq \|x\|$. On vérifie aisément que $(x|\mathbf{E}) = \sum_{i=1}^k (x, e_i)e_i$ et que $\|(x|\mathbf{E})\|^2 = \sum_{i=1}^k |(x, e_i)|^2$. Remarquons en effet, pour tout $j \in \{1, \ldots, k\}$,

$$(x - \sum_{i=1}^{k} (x, e_i)e_i, e_j) = (x, e_j) - (x, e_j) = 0$$

■

Définition C.2 (Famille orthonormale complète)**.** *Soit $E = \{e_j; j \in T\}$ une famille orthonormale de $\mathcal{H}$. On dit que E est une famille orthonormale complète ssi $\overline{\text{esp}}(E) = \mathcal{H}$.*

Lemme C.1. *(i). Soit $(\mathcal{M}_n)$ une suite croissante de sous-espaces vectoriels (s.e.v.) fermés d'un espace de Hilbert $\mathcal{H}$ et notons $\mathcal{M}_{-\infty} = \bigcap_n \mathcal{M}_n$. Alors, pour tout $h \in \mathcal{H}$, nous avons*

$$(h|\mathcal{M}_{-\infty}) = \lim_{n\to-\infty}(h|\mathcal{M}_n)$$

(ii). Soit $\mathcal{M}_\infty = \overline{\bigcup}_{n\in\mathbb{Z}}\mathcal{M}_n$. Alors, pour tout $h \in \mathcal{H}$,

$$(h|\mathcal{M}_\infty) = \lim_{n\to\infty}(h|\mathcal{M}_n).$$

(iii). Soit $\{e_k, k \in \mathbb{N}\}$ une famille orthonormale de h, $e_j \perp e_k$, for $j \neq k$, $\|e_j\| = 1$. Soit $E_n = \overline{\text{span}}\{e_l, 0 \leq l \leq n\}$ et $E_\infty = \overline{\bigcup}_{n\geq 0}E_n$. Alors

$$(h|\mathcal{E}_\infty) = \sum_{k=0}^{\infty} a_k e_k.$$

Démonstration. (a) Comme $\mathcal{M}_n$ est un s.e.v. fermé de $\mathcal{H}$ et donc $\mathcal{M}_{-\infty}$ est un s.e.v. fermé de $\mathcal{H}$. Le théorème de projection 4.2 prouve que $(h|\mathcal{M}_{-\infty})$ existe. Pour $m < n$, définissons $\mathcal{M}_n \ominus \mathcal{M}_m$ le complément orthogonal de $\mathcal{M}_m$ dans $\mathcal{M}_n$, c'est à dire l'ensemble des vecteurs $x \in \mathcal{M}_n$ tel que $x \perp \mathcal{M}_m$. $\mathcal{M}_n \ominus \mathcal{M}_m$ est un s.e.v fermé de $\mathcal{H}$. Notons que

$$(h|\mathcal{M}_n \ominus \mathcal{M}_m) = (h|\mathcal{M}_n) - (h|\mathcal{M}_m).$$

On a, pour tout $m \geq 0$,

$$\sum_{n=-m}^{\infty} \|(h|M_n \ominus M_{n-1})\|^2 = \|(h|M_0 \ominus M_{-m})\|^2 \leq \|h\|^2 < \infty$$

et donc la suite $\{(h|M_n), n = 0, -1, -2, \ldots\}$ est une suite de Cauchy. Comme $\mathcal{H}$ est complet, $(h|M_n)$ converge dans $\mathcal{H}$. Notons $z := \lim_{m\to-\infty}(h|M_n)$. Il reste à prouver que $z = (h|M_{-\infty})$. En appliquant le théorème de projection 4.2, nous devons donc démontrer que (i) $z \in \mathcal{M}_{-\infty}$ et (ii) $h - z \perp M_{-\infty}$. Comme $(h|M_n) \in M_p$ pour tout $n \leq p$, nous avons donc $\lim_{n\to-\infty}(h|M_n) \in M_p$ pour tout p et donc $z \in \mathcal{M}_{-\infty}$, ce qui établit (i). Pour prouver (ii), prenons $p \in \mathcal{M}_{-\infty}$. Nous avons $p \in \mathcal{M}_n$ pour tout $n \in \mathbb{Z}$, et donc, pour tout $n \in \mathbb{Z}$, $(h - (h|\mathcal{M}_n), p) = 0$ et (ii) découle de la continuité du produit scalaire. La preuve du point [(b)] est similaire et est laissée au lecteur à titre d'exercice Nous prouvons finalement le point [(c)]. En appliquant [(b)], nous avons

$$(h|\mathcal{E}_\infty) = \lim_{n\to\infty}(h|E_n).$$

On vérifie aisément que

$$(h|E_n) = \sum_{k=1}^{n}(h, e_k)e_k.$$

Notons en effet que $(h|E_n) \in E_n$ et, pour tout $k \in \{1, \cdots, n\}$,

$$(h - (h|E_n), e_k) = (h, e_k) - (h, e_k) = 0.$$

On conclut la preuve en combinant les deux résultats précédents. ■

Dans les espaces de Hilbert le fait qu'il existe une famille orthonormale complète *dénombrable* joue un rôle important. Ce qui conduit à la définition suivante.

Définition C.3 (Espace de Hilbert séparable)**.** *On dit qu'un espace de Hilbert est séparable ssi il existe une famille orthonormale complète dénombrable.*

La plupart des espaces de Hilbert que nous rencontrerons seront séparables. En particulier le sous-espace fermé engendré à partir d'une famille dénombrable d'un espace de Hilbert, que celui-ci soit séparable ou non séparable, est séparable.

Théorème C.2. *Soit $\mathcal{H}$ un espace de Hilbert séparable et soit $\{e_i; i \in \mathbb{N}\}$ une famille orthonormale complète dénombrable. Alors :*

1. $\forall \epsilon > 0$*, il existe un entier* k *et une suite* $c_0, \cdots, c_k$ *t.q.* $\|x - \sum_{i=0}^{k} c_i e_i\| < \epsilon$.
2. $x = \sum_{i=0}^{+\infty} (e_i, x_i) e_i$ *(série de Fourier),*
3. $\|x\|^2 = \sum_{i=0}^{+\infty} |(e_i, x_i)|^2$ *(égalité de Parseval),*
4. $(x, y) = \sum_{i=0}^{+\infty} (x_i, e_i)(e_i, y)$*,*
5. $x = 0$ *ssi* $(e_i, x) = 0$ *pour tout* i*.*

Annexe D

Compléments sur les matrices

Toutes les matrices et tous les vecteurs (colonne) considérés sont de dimensions finies à éléments complexes. On suppose connue la définition du déterminant.

Notations

L'exposant T désigne la transposition, l'exposant H désigne la transposition-conjugaison. I désigne une matrice identité de dimension adéquate. La matrice $\mathrm{diag}(a_1, \ldots, a_N)$ désigne la matrice carrée diagonale de dimension N, dont les éléments diagonaux sont $a_1, \ldots, a_N$. Une matrice carrée U est dite unitaire si $UU^H = U^H U = I$. Une matrice carrée P est un projecteur si $P^2 = P = P^H$. Par exemple, si v désigne un vecteur, la matrice $vv^H/v^H v$ est un projecteur. La trace d'une matrice est la somme de ses éléments diagonaux. La trace vérifie $\mathrm{Trace}(A+B) = \mathrm{Trace}(A) + \mathrm{Trace}(B)$ et $\mathrm{Trace}(AB) = \mathrm{Trace}(BA)$.

Matrice-bloc, déterminant et trace

Pour des matrices carrées ayant des dimensions appropriées, on a les formules suivantes :

- $(AB)^H = B^H A^H$
- $(A^H)^{-1} = (A^{-1})^H$
- $\det(A) = \det(A^T)$
- $\det(AB) = \det(A)\det(B)$
- $\det(I - AB) = \det(I_M - BA)$
- $\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(A)\det(D - CA^{-1}B)$
- $\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B\Delta^{-1}CA^{-1} & -A^{-1}B\Delta^{-1} \\ -\Delta^{-1}CA^{-1} & \Delta^{-1} \end{bmatrix}$
  où $\Delta = D - CA^{-1}B$

Lemme d'inversion matricielle : si A et B sont deux matrices carrées inversibles, alors pour toutes matrices G et H de dimensions appropriées :

$$(A + GBH)^{-1} = A^{-1} - A^{-1}G\left(HA^{-1}G + B^{-1}\right)^{-1} HA^{-1}$$

Valeurs propres

Pour une matrice carrée A de dimension $N \times N$, les vecteurs propres représentent les directions de l'espace $\mathbb{C}^N$ qui sont invariantes. Ce sont par conséquent les vecteurs w définis par l'équation $Aw = \lambda w$. La trace est égale à la somme des valeurs propres et le déterminant à leur produit. Cela s'écrit :

$$\text{Trace}(A) = \sum_{i=1}^{N} \lambda_i \quad \text{et} \quad \det(A) = \prod_{i=1}^{N} \lambda_i$$

Image de A

Soit A une matrice de dimension $M \times N$. On appelle *image* de A le sous-espace de $\mathbb{C}^M$ noté $\mathcal{I}(A)$, qui est engendré par les vecteurs-colonnes de A. On appelle *noyau* de A le sous-espace de $\mathbb{C}^N$ noté $\mathcal{N}(A)$, qui est solution de $Ax = 0$. On appelle *rang-colonne* de A la dimension de son espace image $rang(A) = \dim \mathcal{I}(A)$. C'est aussi le nombre de vecteurs-colonnes de A qui sont indépendants. On montre que :

$$\dim \mathcal{N}(A) + \dim \mathcal{I}(A) = N$$

Si A est de rang-colonne plein, cad $rang(A) = N$, alors soit $A^H A$ est inversible. On définit de la même manière un rang-ligne. Le rang de A est le minimum de son rang-colonne et de son rang-ligne. Dans tous les cas le rang d'une matrice est inférieur à $\min(M, N)$.

Valeurs singulières

Soit A une matrice de dimension $M \times N$ et de rang r. Alors il existe deux matrices carrées unitaires l'une notée U de taille $M \times M$ et l'autre notée V de taille $N \times N$, telles que :

$$A = U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^H$$

où $\Sigma_r = \text{diag}(\sigma_1, \ldots, \sigma_r)$ avec $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Les valeurs σ_i sont dites valeurs singulières de A.

- Les vecteurs colonnes de U de dimension M sont les vecteurs propres de AA^H. Les r premiers vecteurs colonnes de U forment une base orthonormée de l'image de A.
- Les vecteurs colonnes de V de dimension N sont les vecteurs propres de $A^H A$. Les $(N - r)$ derniers vecteurs colonnes de V forment une base orthonormée du noyau de A.

On appelle pseudo-inverse de A la matrice de dimension $N \times M$:

$$A^+ = V \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^H$$

Dans $\mathbb{C}^M$, la matrice carrée AA^+ est le projecteur sur $\mathcal{I}(A)$. Dans $\mathbb{C}^N$, la matrice carrée $(I - A^+A)$ est le projecteur sur $\mathcal{N}(A)$. Si A est de rang plein, alors :

- pour $M = N$, $A^+ = A^{-1}$,
- pour $M > N$, $A^+ = (A^H A)^{-1} A^H$
- et pour $M < N$, $A^+ = A^H (AA^H)^{-1}$

Le rapport entre la plus grande et la plus petite valeur singulière d'une matrice s'appelle son *nombre de conditionnement*. Il mesure la difficulté numérique à calculer sa pseudo-inverse.

Matrice carrée positive

Une matrice carrée R est dite *hermitienne* si elle vérifie $R = R^H$. Une matrice carrée hermitienne R est dite *non-négative*, respectivement *positive* si pour tout vecteur a, on a $a^H Ra \geq 0$ (resp. > 0). Pour les matrices non négatives, la décomposition en valeurs propres et la décomposition en valeurs singulières coïncident. Si R est positive, alors R^{-1} existe et est positive. Si R est non négative, toutes ses valeurs propres sont réelles, non négatives et leur ordre de multiplicité est égal à la dimension du sous-espace propre associé. Si R est une matrice non négative et si ses valeurs propres λ_i sont distinctes, alors les vecteurs propres w_i associés sont deux à deux orthogonaux et on a :

$$R = \sum_{i=1}^{N} \lambda_i w_i w_i^H$$

où tous les λ_i sont non négatifs. On en déduit que :

$$R^n = \sum_{i=1}^{N} \lambda_i^n w_i w_i^H$$

Il est facile d'étendre cette écriture à une fonction polynomiale quelconque. En particulier on en déduit que R vérifie son équation caractéristique ($\det(A - \lambda I) = 0$). Par extension, pour toute fonction f développable en série entière, on peut définir la fonction de matrice :

$$f(R) = \sum_{i=1}^{N} f(\lambda_i) w_i w_i^H$$