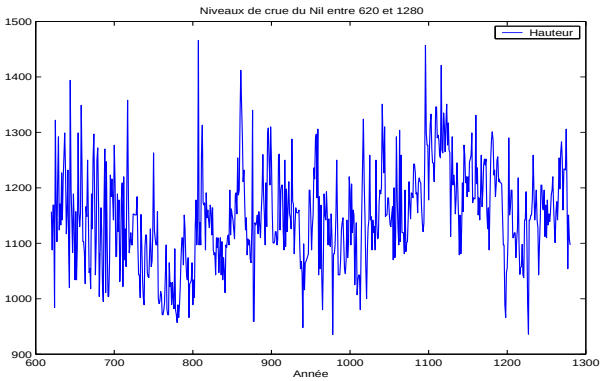
Université Paris I, Panthéon - Sorbonne

MASTER M.A.E.F.

Cours de Statistiques II

JEAN-MARC BARDET (UNIVERSITÉ PARIS 1, SAMM)



U.F.R. 27 et SAMM (Statistique, Analyse et Modélisation Multidisciplinaire)
Université Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris.

Plan du cours

1. Processus aléatoires: premières définitions et propriétés.
2. Estimation de la tendance et de la saisonnalité.
3. Exemples de processus stationnaires à temps discret.
4. Identification d'un processus stationnaire à temps discret.
5. Prédiction pour un processus à temps discret.

References

[1] Amemiya, T. (1985). *Advanced Econometrics.* Cambridge , MA: Harvard University Press.

[2] Azencott, R. et Dacunha-Castelle, D. (1984) *Séries d'observation irrégulières.* Masson, Paris.

[3] Barbe P. et Ledoux M. (1998) *Probabilité.* EDP Sciences.

[4] Brockwell P.J. et Davis R.A. (1991) *Time Series: Theory and Methods.* Wiley.

[5] Brockwell P.J. et Davis R.A. (2002) *Introduction to Time-Series and Forecasting.* SpringerVerlag.

[6] Dacunha-Castelle, D. et Duflo, M. (1983) *Probabilités et statistiques. Tome 1: Problèmes à temps fixe* et *Tome 2: Problèmes à temps mobile.* Collection Mathématiques Appliquées pour la Maîtrise, Masson.

[7] Gourieroux, C. et Montfort, A. *Séries temporelles et modèles dynamiques.* Economica.

[8] Hamilton, J.D. (1994). *Time series analysis.* Princeton University Press, Princeton.

Documents accessibles librement sur internet

- Cours de Paul Doukhan à l'ENSAE: `http://samos.univ-paris1.fr/Teaching`.
- Cours de Xavier Guyon pour STAFAV: `http://www.stafav.org`.
- Aide-mémoire en économétrie de A. Trognon et J.M. Fournier à l'ENSAE: `http://www.ensae.fr/ParisTech/SEC02/ENSAEEconometrieCursusintegre2006.pdf`.
- Cours de R. Bourdonnais: `http://www.dauphine.fr/eurisco/eur-wp/CoursSeriesTemp-Chap*.pdf` où on peut remplacer $*$ par $1, 2, 3$ ou 4.

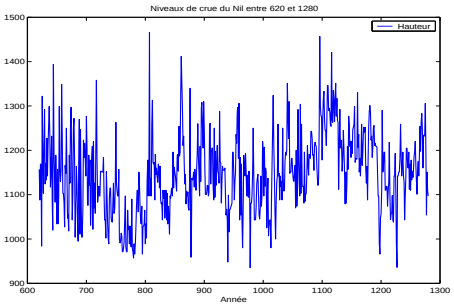
Quelques sites internet intéressants

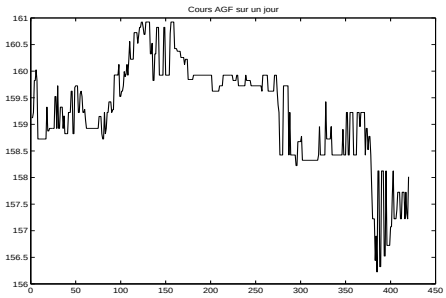
- Le site de Toulouse III: `http://wwww.lsp.ups-tlse.fr`. Regarder les documents pédagogiques.
- Le site de Paris V: `http://wwww.math-info.univ-paris5.fr`. Regarder les documents pédagogiques.
- Le site de Paris VI: `http://wwww.proba.jussieu.fr`. Regarder les documents pédagogiques.
- Le site de la S.M.A.I.: `http://smai.emath.fr`. Regarder la rubrique `Logiciels` dans laquelle de nombreux logiciels de mathématiques peuvent être téléchargés (en particulier, `Scilab` et `Mupad`).
- Le site français d'où l'on peut télécharger le logiciel R: `http://cran.cict.fr`.

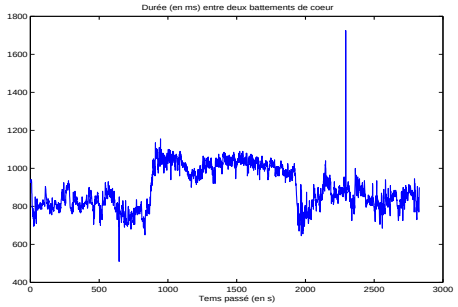
Introduction

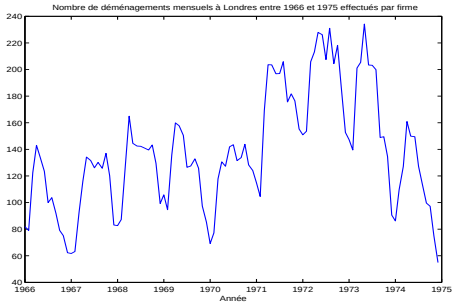
Exemple.

Processus ou séries chronologiques climatiques, biologiques, hydrauliques, financières,...









Objectifs. *Les objectifs de ce cours sont:*

1. *décrire une série chronologique.*
2. *modéliser un processus et tester les modèles proposés.*
3. *comprendre et expliquer des phénomènes attenant à un processus.*
4. *prévoir le comportement futur d'un processus.*

1 Processus aléatoires: premières définitions et propriétés

Nous allons voir en premier lieu qu'un processus diffère de ce que l'on a jusqu'alors essentiellement rencontré en probabilités et statistiques, c'est-à-dire des suites de v.a.i.i.d. C'est surtout sur l'hypothèse d'indépendance que nous allons revenir, en proposant des formes de dépendances que l'on pourra caractériser par les covariances, les probabilités conditionnelles,...

1.1 Processus aléatoire

Définition. *Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité.*

- *On dit que $X = (X_t,\, t \in T)$ est un processus aléatoire (ou encore stochastique) sur T à valeurs dans $\mathbb{R}^k$ lorsque pour tout $t \in T$, X_t est une variable aléatoire sur $(\Omega, \mathcal{A})$ à valeurs dans $\mathbb{R}^k$.*
- *Pour $\omega \in \Omega$, $(X_t(\omega),\, t \in T)$ est appelé une trajectoire du processus X.*
- *On dit que $X = (X_t,\, t \in T)$ est un processus aléatoire du second ordre lorsque pour tout $t \in T$, X_t est une variable aléatoire appartenant à $\mathbb{L}^2(\Omega, \mathcal{A}, P)$.*
- *On appelle fonction espérance, variance, covariance et corrélation d'un processus aléatoire du second ordre à valeurs réelles, pour $(s,t) \in T^2$, les fonctions $m(t) = \mathbb{E}X_t$, $\sigma^2(t) = \mathbb{E}X_t^2 - m^2(t)$, $\gamma(s,t) = \mathbb{E}(X_s - \mathbb{E}X_s)(X_t - \mathbb{E}X_t)$ et $r(s,t) = \gamma(s,t)/(\sigma(s)\sigma(t))$.*

Exemple.

Fonction réelle; Suite de variables indépendantes; Marche aléatoire; Chaîne de Markov; Mouvement brownien.

Définition. *Une série chronologique (temporelle) est un processus aléatoire réel indicé par* $\mathbb{Z}$ *ou* $\mathbb{N}$ *(on dit encore processus à temps discret; quand ce n'est pas le cas, notamment lorsque* $T = \mathbb{R}$*, on parle de processus à temps continu).*

Définition. *On appelle processus aléatoire (ou une série chronologique)* $X = (X_t, t \in T)$ *gaussien, un processus aléatoire tel que* $\forall k \in \mathbb{N}^*$, $\forall (t_1, \cdots, t_n) \in T^n$, $(X_{t_1}, \cdots, X_{t_n})$ *est un vecteur gaussien.*

Remarque.

Rappelons que $(X_{t_1}, \cdots, X_{t_n})$ est un vecteur gaussien si $\forall (u_1, \cdots, u_n) \in \mathbb{R}^n$, $u_1 X_1 + \cdots + u_n X_n$ est une variable gaussienne.

Propriété.

- *Si* X *et* Y *sont des vecteurs gaussiens, alors (Indépendance de* X *et* $Y \iff cov(X, Y) = 0$*).*
- *Un processus gaussien est entièrement défini par ses fonctions espérance* $m(t) = \mathbb{E}X(t)$ *et covariance* $\gamma(s,t) = \mathbb{E}X_s X_t$*. Réciproquement, la connaissance d'une fonction* $\gamma(s,t)$ *définie positive et d'une fonction* $m(t)$*, définit un unique processus gaussien.*

Définition. *On dit qu'un processus aléatoire* $X = (X_t, t \in T)$ *est (strictement) stationnaire lorsque* X *est invariant en distribution par toute translation du temps, c'est-à-dire que* $\forall n \in \mathbb{N}^*$, $\forall (t_1, \cdots, t_n) \in T^n$, $\forall c \in T$, $(X_{t_1}, \cdots, X_{t_n})$ *à la même distribution que* $(X_{t_1+c}, \cdots, X_{t_n+c})$*.*

Remarque.

Une caractérisation de l'égalité en loi est celle obtenue par la fonction caractéristique. On montrera ainsi que pour tout $(u_1, \cdots, u_n) \in \mathbb{R}^n$,

$$\phi_{(X_{t_1}, \cdots, X_{t_n})}(u_1, \cdots, u_n) = \mathbb{E}\big(e^{i \sum_{j=1}^n u_j X_{t_j}}\big) = \phi_{(X_{t_1+c}, \cdots, X_{t_n+c})}(u_1, \cdots, u_n).$$

On pourra ainsi utiliser le fait que la fonction caractéristique de la somme de 2 v.a. indépendantes vaut le produit des fonctions caractéristiques...

Exemple.

Suite de v.a.i.i.d., chaîne de Markov homogène.

Propriété. *Conséquences de la stationnarité sur les fonctions espérance, variance, covariance d'un processus à temps discret* X*:*

- *L'espérance* $m(t) = \mathbb{E}(X_t)$ *est constante.*
- *La variance* $\sigma^2(t) = var(X_t)$ *est constante.*
- *La covariance* $\gamma(s,t) = cov(X_s, X_t)$ *est une fonction ne dépendant que de* $|t - s|$*.*

Définition. *Soit un processus à temps discret* $X = (X_t, t \in T)$*. On dit que* X *est:*

- *Un processus stationnaire d'ordre 2 lorsque: 1/ son espérance* $m(t)$ *est constante, 2/ sa covariance* $cov(X_s, X_t)$ *est une fonction de* $|t - s|$*.*
- *Un processus à accroissements stationnaires lorsque le processus* $Y = \{Y_t, t \in T\}$ *telle que* $Y_t = X_{t+1} - X_t$ *pour* $t \in T$*, est stationnaire.*

Propriété.

- *(Stationnarité stricte* $\Longrightarrow$ *Stationnarité d'ordre 2), mais la réciproque est fausse.*
- *Si* X *est un processus gaussien, alors (Stationnarité stricte* $\iff$ *Stationnarité d'ordre 2).*

Remarque.

On pourrait penser qu'il est plus difficile d'être stationnaire strict que stationnaire d'ordre 2. Cependant la stationnarité d'ordre 2 requiert d'avoir des moments d'ordre 2 ce qui n'est pas demander par la stationnarité stricte. On peut ainsi montrer que pour un ARCH(p) (on verra plus loin la définition d'un tel processus), les conditions de stationnarité d'ordre 2 sont plus fortes que celles de stationnarité stricte.

Définition.
- *Un bruit blanc (fort) est une suite de variables aléatoires identiquement distribuées indépendantes (v.a.i.i.d.) centrées.*
- *Un bruit blanc faible est une suite de variables aléatoires identiquement distribuées centrées non corrélées.*
- *Un bruit blanc gaussien est une suite de v.a.i.i.d. gaussiennes centrées.*

Remarque.

La terminologie de bruit blanc est surtout employée par les spécialistes de traitements du signal. Ceux-ci évoquent aussi parfois des bruits roses pour certaines formes de dépendances entre les données...

1.2 Tendance et composante saisonnière

Les séries de données réelles que l'on peut rencontrer ne sont que très rarement des séries que l'on peut modéliser par un processus à temps discret stationnaire. En effet, très souvent leurs moyennes varient dans le temps, parfois de manière suffisamment régulière pour être facilement modélisées (typiquement une tendance linéaire), parfois non... Voici pour commencer le type principal de "tendances" dites tendances additives:

Définition. *Tout processus à temps discret $X = (X_t)_{t\in T}$ avec $T \subset \mathbb{Z}$ peut s'écrire sous la forme*

$$X_t = a(t) + S(t) + \varepsilon_t, \quad \textit{pour } t \in T,$$

où $t \mapsto a(t)$ et $t \mapsto S(t)$ sont deux fonctions déterministes telles que $\mathbb{E}X_t = a(t) + S(t)$, avec

1. *si $t \mapsto S(t)$ est une fonction périodique non nulle de période $r > 0$ telle que $\sum_{i=1}^{r} S(i) = 0$, alors S est la composante saisonnière, saisonnalité de X,*
2. *si $t \mapsto a(t)$ est non nulle, a est la tendance de X,*
3. *si $\varepsilon = (\varepsilon_t)_{t\in T}$ est non nulle, ε est une série chronologique centrée appelée souvent le bruit de X.*

Exemple.

Tendances polynômiales, saisonnalité annuelle,..

Remarque.

Attention, il n'y a pas unicité de la décomposition précédente. Il y a en revanche unicité du bruit de X puisque $u_t = X_t - \mathbb{E}X_t$ pour tout t.

Définition. *Pour $X = (X_t)_t$ un processus aléatoire ayant pour tendance a et pour saisonnalité S. On appelle:*

- *Série détendancialisée la série $(X_t - a(t))_t$. Si la fonction $a(\cdot)$ n'est pas connue explicitement, ce qui est le plus souvent le cas, $(X_t - \widehat{a}(t))_t$, où $\widehat{a}(t)$ est un estimateur de $a(t)$ sera la série détendancialisée (même dénomination).*
- *Série désaisonnalisé (ou série corrigée des variations saisonnières) la série $(X_t - S(t))_t$. Si la fonction $S(\cdot)$ n'est pas connue explicitement, ce qui est le plus souvent le cas, $(X_t - \widehat{S}(t))_t$, où $\widehat{S}(t)$ est un estimateur de $S(t)$ sera la série désaisonnalisée (même dénomination).*

Il est aussi possible que le processus ait une variance qui varie de manière déterministe en fonction du temps. En ce cas on évoquera une tendance multiplicative:

Définition. *On dira que X possède une tendance multiplicative si $X_t - m(t) = \sigma(t)\, u(t)$ pour tout $t \in T$, où $u = (u_t)_{t\in T}$ est une suite de variables aléatoires centrées de variance constante et $\sigma(\cdot)$ est une fonction positive.*

1.3 Cas particulier des processus stationnaires

Dans toute la suite, on suppose que $X = (X_k)_{k\in\mathbb{Z}}$ est un processus à temps discret **stationnaire** (donc un processus sans tendance additive ou multiplicative non constantes). Ceci induit en particulier que les X_n sont des variables identiquement distribuées. On considérera également que les processus sont centrés, ce qui s'adapte par simple translation au cas non centré.

Définition. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus du second ordre à temps discret centré stationnaire,*

- *on appelle $r(k) = \mathbb{E}X_0X_k = \mathbb{E}X_iX_{i+k}$ pour $k \in \mathbb{Z}$ et $i \in \mathbb{Z}$, la covariance de X. Ainsi, $r(0)$ est la variance de la série.*
- *on appelle $\rho(k) = r(k)/r(0)$ la corrélation de X. On a $-1 \leq \rho(k) \leq 1$ pour $k \in \mathbb{Z}$.*

Définition. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus du second ordre à temps discret centré stationnaire. S'il existe une fonction $f : [-\pi, \pi[\to \mathbb{C}$ telle que $\forall k \in \mathbb{Z}$, $r(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda$, alors on dit que X admet une densité spectrale f.*

Exemple.

Montrer que la densité spectrale d'un bruit blanc faible stationnaire à variance finie existe et la calculer.

Propriété. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus du second ordre à temps discret centré stationnaire. Si la densité spectrale f existe sur $[-\pi, \pi[$ alors f est paire et $f(\lambda) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} r(k) e^{-ik\lambda}$ pour tout $\lambda \in [-\pi, \pi[$.*

Propriété. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus du second ordre à temps discret centré stationnaire.*

1. *(Les covariances $r(k)$ vérifient $\sum |r(k)|^2 < \infty$) $\iff$ (f existe et est de carré intégrable sur $[-\pi, \pi[$).*
2. *(Les covariances $r(k)$ sont telles que $\sum |r(k)| < \infty$) $\implies$ (f existe et f continue sur $[-\pi, \pi[$).*

Exemple.

Soit $(\varepsilon_k)_{k\in\mathbb{N}}$ une suite de v.a.i.i.d. à variance finie et soit $X_k = \varepsilon_k - \varepsilon_{k-1}$ pour $k \in \mathbb{Z}$. Déterminer la densité spectrale de $(X_k)_{k\in\mathbb{Z}}$. Grâce à la notion de stationarité il est possible d'obtenir un résultat théorique très puissant, qui n'a cependant que peu d'intérêt en pratique:

Théorème (Décomposition de Cramèr-Wald)**.** *Soit $X = (X_t)_{t\in\mathbb{Z}}$ un processus à temps discret stationnaire d'ordre 2 tel que $\int_{-\pi}^{\pi} \log(f(\lambda)) d\lambda > -\infty$. Alors il existe un unique bruit blanc faible $(\varepsilon_t)_{t\in\mathbb{Z}}$ (donc $cov(\varepsilon_t, \varepsilon_s) = 0$ pour $t \neq s$) et une famille de réels $(a_i)_{i\in\mathbb{N}}$ telle que $a_0 \geq 0$ et $\sum_{k\in\mathbb{N}} |a_k|^2 < \infty$, vérifiant*

$$X_t = \mathbb{E}X_0 + \sum_{i=0}^{N} a_i \varepsilon_{t-i} \quad \textit{for all } t \in \mathbb{Z}. \tag{1}$$

Ceci permet d'avoir une forme un peu générale d'un processus stationnaire. Cependant nous verrons que l'on peut surtout traiter le cas des processus linéaires (quand le bruit blanc est fort, donc avec l'hypothèse d'indépendance en plus); dans le cas général, la décomposition de Cramèr-Wald n'apporte pas beaucoup de renseignements (voir le cas des processus GARCH).

2 Estimation de la tendance et de la saisonnalité

Cette partie est souvent omise ou vite traitée dans les livres consacrés aux processus ou aux séries chronologiques. Pourtant l'estimation de la tendance et de la saisonnalité est essentielle dans la plupart des travaux concrets portant sur les séries chronologiques, en particulier parce qu'elle apporte une information souvent bien plus importante que la partie bruit en vue de prédictions. Nous ne traitons pour commencer que des processus sans tendance multiplicative, ce dernier cas étant traité à la fin.

2.1 Estimation semi-paramétrique par régression

On voudrait connaître la tendance et la saisonnalité du processus en supposant connue une trajectoire $(X_1, \cdots, X_N)$.
On suppose que la tendance et la saisonnalité s'écrivent sous une forme connue a priori (les f_i, fonctions quelconques, et r sont supposés connues), soit:

$$a(t) = \sum_{i=1}^{k} a_i f_i(t) \quad \text{et} \quad S(t) = \sum_{i=1}^{r-1} s_i (g_i(t) - g_r(t)) \quad \text{pour } t \in T,$$

avec $g_i(t) = \mathbb{I}_{\{t=i,\ [r]\}}$ sont r-périodiques (on a ainsi $\sum_{i=1}^{r} S(i) = 0$).

Notation.
- $X = {}^t(X_1, .., X_N)$.
- $F_i = {}^t(f_i(1), .., f_i(N))$ *pour* $i = 1, .., k$ *et* $G_i = {}^t(g_i(1) - g_r(1), .., g_i(N) - g_r(N))$ *pour* $i = 1, .., r-1$.
- $U = (\varepsilon(1), .., \varepsilon(N))$.

Le modèle s'écrit alors vectoriellement:

$$X = \sum_{i=1}^{k} a_i f_i + \sum_{i=1}^{r} s_i (g_i - g_r) + U.$$

Proposition. *On peut estimer les coefficients* (a_i) *et* (s_i) *par une régression par moindres carrés en minimisant une distance dans* $\mathbb{R}^n$*:*

$$\| X - (a_1 F_1 + \cdots + s_{r-1} G_{r-1}) \|^2, \quad \textit{et on notant } Z \textit{ la matrice } Z = (F_1 \cdots F_k G_1 \cdots G_{r-1}),$$

1. *si* U *est un bruit dont on ne connaît pas la variance, on utilise une estimation par moindres carrés ordinaires et:*
$${}^t(\widehat{a_1}, \cdots, \widehat{a_k}, \widehat{s_1}, \cdots, \widehat{s_{r-1}}) = ({}^t Z.Z)^{-1\,t} Z X,$$
2. *si* U *est un bruit tel que* $\mathbb{E} U {}^t U = \Sigma$ *est une matrice connue, on utilise une estimation par moindres carrés généralisés, et:*
$${}^t(\widehat{a_1}, \cdots, \widehat{a_k}, \widehat{s_1}, \cdots, \widehat{s_{r-1}}) = ({}^t Z \Sigma^{-1} Z)^{-1\,t} Z \Sigma^{-1} X.$$

On déduit aisément de ces expressions un premier résultat de convergence pour les estimateurs de la tendance et de la saisonnalité:

Propriété. *Dans le cadre de régression précédent, si* (ε_n) *est un bruit blanc de variance finie, si la matrice* $({}^t Z\, Z)^{-1}$ *tend vers* 0 *en norme, alors les estimateurs des paramètres sont non biaisés et convergents en probabilité quand* $N \to \infty$.

Sous les mêmes hypothèses, une condition nécessaire et suffisante de convergence presque sûre est $\max_{1 \le i \le k+r-1} \left| \left(Z({}^t Z\, Z)^{-1\,t} Z \right)_{ii} \right| \xrightarrow[N \to +\infty]{\mathcal{L}} 0$. La normalité asymptotique (théorème de la limite centrale vérifié par les estimateurs des paramètres) est également impliquée par cette condition. Cependant, quand on n'est plus dans le cas d'un bruit blanc, les résultats de convergence peuvent être plus complexes, le comportement avec N de la matrice Σ ayant un rôle important...

Exemple.

Soit le cas où la tendance est une constante, où il n'y a pas de saisonnalité et le bruit est stationnaire et admet une matrice de covariance Σ. A-t-on toujours convergence? Et si $\lim_{n \to \infty} \mathbb{E}(u_0 u_n) = 0$?

Remarque.

- On peut noter que ces estimateurs sont des combinaisons linéaires des X_i.

- Une telle régression nécessite la connaissance a priori des fonctions f_i. Pour ce faire on peut faire différentes hypothèses: on peut considérer que $a(\cdot)$ est un polynôme (typiquement $f_i = t^i$), ou bien que $a(\cdot)$ est un polynôme trigonométrique (typiquement $f_i(t) = \sin(i\,t)$ ou bien $f_i(t) = \cos(i\,t)$). Concrètement, on utilisera plutôt une régression polynomiale lorsque les données semblent prendre une certaine direction, alors que la modélisation par un polynôme trigonométrique permet d'avoir des prédictions qui restent dans le même ordre de grandeur que les données connues. Pour aller un peu plus loin, on peut même essayer de décomposer la tendance $a(t)$ dans une certaine base (ce qui peut être fait en Fourier ou avec une base d'ondelettes par exemple).
  La question qui se pose ensuite est celle de savoir comment choisir le nombre k de fonctions f_i considérées. Cela revient dans le cas polynomial à choisir le degré maximal du polynôme que l'on utilise dans la régression. Il est clair que ce nombre k doit être inférieur à $N - r$, sinon les différents paramètres a_i ne pourront pas être estimés. Une technique efficace pour obtenir "mathématiquement" un choix "optimal" de k est d'utiliser un critère de sélection de modèle. Pour ses propriétés de convergence même dans des cas non gaussiens, le **critère BIC** (Bayesian Information Criterium) est un choix souvent intéressant. Pour k fixé, on calcule

$$BIC(k) = -2\log(\text{Vraisemblance maximisée du modèle à (k+r-1) paramètres}) + \frac{\log N}{N}\,(k + r - 1)$$

  et on cherche k qui minimise ce critère. Cela revient, après des approximations, à chercher

$$\widehat{k} = \text{Argmax}_{k=0,1,\ldots,k_{\max}}\Big(\log\left(\widehat{\sigma}^2_{k+r-1}\right) + \frac{\log N}{N}\,k\Big), \quad \text{où} \quad \widehat{\sigma}^2_{k+r-1} = \frac{1}{n}\|X - (\widehat{a}_1 F_1 + \ldots + \widehat{s}_{r-1}G_{r-1})\|^2.$$

  où $k_{\max}$ est un entier suffisamment grand mais plus petit que $N - r$.

Cas particulier de l'estimation du saisonnier

L'estimation par moindres carrés ordinaires permet d'estimer le saisonnier d'une série chronologique détendancialisée. On suppose connue $(X_1, \cdots, X_{rN})$, où r est la période (connue) du saisonnier. On écrit $X_t = S(t) + Y_t$, où $S(t) = \sum_{i=1}^{r-1} s_i(g_i(t) - g_r(t))$ pour $t \in T$, avec $g_i(t) = \mathbb{I}_{\{t=i,\ [r]\}}$ (voir plus haut). Dans ce cas, on a:

$$Z = \begin{pmatrix} 1 & 0 & 0 & . & 0 \\ 0 & 1 & 0 & . & 0 \\ 0 & 0 & 1 & . & 0 \\ . & . & . & . & . \\ -1 & -1 & -1 & . & -1 \\ 1 & 0 & 0 & . & 0 \\ 0 & 1 & 0 & . & 0 \\ . & . & . & . & . \\ -1 & -1 & -1 & . & -1 \\ 1 & 0 & 0 & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ -1 & -1 & -1 & . & -1 \end{pmatrix}, \qquad {}^tZ.Z = N\begin{pmatrix} 2 & 1 & . & 1 \\ 1 & 2 & . & 1 \\ . & . & . & . \\ 1 & 1 & . & 2 \end{pmatrix},$$

et

$$({}^tZ.Z)^{-1} = \tfrac{1}{rN}\begin{pmatrix} r-1 & -1 & . & -1 \\ -1 & r-1 & . & -1 \\ . & . & . & . \\ -1 & -1 & . & r-1 \end{pmatrix}.$$

On montre alors que dans le cadre d'une estimation par moindres carrés ordinaires:

$$\widehat{s}_i = \frac{1}{N}\sum_{k=1}^{N} X_{i+r(k-1)} - \frac{1}{rN}\sum_{k=1}^{rN} X_k, \quad \text{pour } i = 1, \cdots, r-1.$$

Exercice.

Montrer que si $\lim_{n\to\infty} \mathbb{E}u_0u_n = 0$, où $u = (u_k)_k$ est un bruit gaussien stationnaire, il y a convergence en probabilité de cet estimateur.

2.2 Estimation non-paramétrique de la tendance

On supposera donc ici que la série n'admet pas de composante saisonnière, et que l'on a $X_t = a(t) + u_t$ pour $t \in \{t_1, \cdots, t_N\}$. Plutôt que de poser a priori un modèle pour la fonction a comme cela a été fait avec la

régression, on peut estimer directement cette fonction avec une méthode non-paramétrique. On présente ici deux types de méthodes:

• Méthode d'estimation par noyau.

Définition. *On appelle noyau* $K : \mathbb{R} \to \mathbb{R}$ *une fonction mesurable telle que* $\int_{\mathbb{R}} K(t)dt = 1$ *et* $K(0) > 0$.

L'idée de la méthode par noyau est que pour $x_0 \in R$ et $h > 0$ dit taille de fenêtre, $\frac{1}{h} K(\frac{x - x_0}{h})$ converge vers une masse de Dirac lorsque $h \to 0$, dans le sens où: $\int_R \frac{1}{h} K(\frac{x - x_0}{h}) f(x)dx \to f(x_0)$ quand $h \to 0$, pour toute fonction f. Aussi peut-on définir un estimateur de la tendance a par:

$$\widehat{a}_{N,h}(t) = \frac{\frac{1}{Nh} \sum_{j=1}^{N} X_j K(\frac{t-t_j}{h})}{\frac{1}{Nh} \sum_{j=1}^{N} K(\frac{t-t_j}{h})}.$$

Sous certaines hypothèses sur la manière dont h converge vers 0 en fonction de N, on peut montrer que $\widehat{a}_{N,h}(t) \to a(t)$ pour tout t lorsque a est suffisamment régulière.
On peut également estimer de façon automatique une taille de fenêtre $\widehat{h}_N$ adaptée aux données et optimale en un certain sens: pour ce faire on utilise par exemple le principe de la validation croisée, c'est-à-dire que pour $h > 0$ fixé et pour chaque $i = 1, \cdots, N$, l'on calcule $\widehat{a}^{(i)}_{N,h}(i)$ à partir de l'échantillon $(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_N)$ et l'on compare $\widehat{a}^{(i)}_{N,h}(i)$ à X_i qui est la valeur obtenue en $t = i$. Ainsi on pourra choisir

$$\widehat{h}_N = \text{Argmin}_{h>0} \sum_{i=1}^{N} (\widehat{a}^{(i)}_{N,h}(i) - X_i)^2 \quad \Longrightarrow \widehat{a}_{N,\widehat{h}_N}(t) \text{ estimateur de } a.$$

• Régressions localisées

Pour estimer la tendance, il est aussi possible d'utiliser des **régressions localisées** de type Spline, Loess ou Lowess. Celles-ci s'obtiennent en fixant une taille de fenêtre et on fait une régression polynomiale (pour les Splines, de degré 3) ou linéaire mais pondérées (pour Loess ou Lowess) de X dans la fenêtre que l'on fait glisser. En fait ce sont également des estimations non-paramétriques, au sens où elles ne supposent pas connues les fonctions pouvant composer la tendance (comme dans le cas de la régression), mais juste une certaine régularité de la fonction tendance. Plus précisément, considérons l'approche par moyenne mobile: en X_t, une idée pour approcher $a(t)$ serait de supposer que a ne change pas trop autour de t et que l'on peut moyenner sur les valeurs de temps proche de t. Par exemple, on pourra considérer une moyenne mobile $\widehat{a}_\ell(t) = \frac{1}{2\ell+1} \sum_{i=-\ell}^{\ell} X_{t+i}$. L'intérêt réside dans le fait de faire ainsi baisser la variance du bruit u_t en moyennant, ce qui se montrer facilement dans le cas où le bruit u est un bruit blanc. Mais il faut choisir ℓ: trop grand, on lissera trop jusqu'à n'obtenir qu'une constante pour tout t, trop petit, on ne fera guère mieux que d'estimer $a(t)$ par X_t, et la fonction a sera donc très irrégulière. Une idée pour dépasser cela sera de considérer une moyenne mobile pondérée dans laquelle on n'accorde pas le même poids à tous les X_i autour de X_t, mais plutôt un poids décroissant en fonction de la distance entre i et t. Aussi considérera-t-on des pondérations exponentielles: pour $\beta \in [0, 1]$, on définit:

$$\widehat{a}_\beta(t) = \sum_{i=-t+1}^{N-t} \beta^{|t-i|} X_{t+i}.$$

On appelera un tel estimateur un estimateur LOWESS (LOcally WEighted Scatterplot Smoothing). Le choix de β peut aussi se faire par validation croisée, comme cela est présenté ci-dessus pour obtenir $\widehat{\beta}_N = \text{Argmin}_\beta \sum_{i=1}^{N} (\widehat{a}^{(i)}_\beta(i) - X_i)^2$.

Notons que d'une manière générale, la généralité des estimateurs non-paramétriques se traduit souvent pas une moins bonne vitesse de convergence de $\widehat{a}_N$ vers a que les estimateurs paramétriques dans le cas où le vrai modèle est paramétrique.

2.3 Description à l'aide d'un filtre linéaire

Une autre technique de traitement de la tendance et de la composante saisonnière d'une série chronologique peut consister à ne pas les estimer mais à plutôt les éliminer... Pour cela on peut utiliser certains filtres linéaires.

Définition. *Soit $X = (X_k, k \in \mathbb{Z})$ un processus à temps discret. On appelle filtre linéaire une famille de réels $a = (a_i)_{n\in I}$, où $I \subset \mathbb{Z}$, et la série filtrée est $Y = (Y_k, k \in \mathbb{Z})$ telle que $Y_k = \sum_{i\in I} a_i\, X_{k-i}$.*

Exemple.

Filtres des accroissements, moyenne empirique sur des fenêtres mobiles,...

Proposition. *Avec les notations de la définition précédente,*

1. *Si I est fini, alors Y dans $\mathbb{L}^p$ lorsque X existe dans $\mathbb{L}^p$ (avec $p > 0$). De plus si X est stationnaire strict (respectivement du second ordre) alors Y est stationnaire strict (respectivement du second ordre).*
2. *Si I est infini, alors la condition $\sum_{i\in I} |a_i| < +\infty$ et $\sup_{t\in\mathbb{Z}} \mathbb{E}|X_t| < \infty$ garantit que Y existe p.s. De plus si X est stationnaire strict alors Y est stationnaire strict. Sous cette condition, si X est gaussien alors Y est gaussien, si X est centré alors Y est centré.*
3. *Si I est infini, alors la condition $\sum_{i\in I} |a_i| < +\infty$ et $\sup_{t\in\mathbb{Z}} \mathbb{E}|X_t|^2 < \infty$ garantit que Y existe dans $\mathbb{L}^2$. De plus si X est stationnaire strict (respectivement du second ordre) alors Y est stationnaire strict (respectivement du second ordre).*
4. *Si I est infini et X est un bruit blanc fort (respectivement faible) et $\sum_{i\in I} |a_i|^2 < +\infty$ alors Y existe p.s. et dans $\mathbb{L}^2$, est appelé un processus linéaire et Y est stationnaire strict (respectivement du second ordre).*

Proof. (1) L'existence ne pose pas de problème. Pour la stationarité, en supposant $I = -m, \cdots, m$, comme X est stationnaire, on sait que pour tout $n \in \mathbb{N}^*$, pour tout $t_1, \cdots, t_n$ dans $\mathbb{Z}$, alors pour tout $c \in \mathbb{Z}$, $(X_{t_1-m}, X_{t_1-m+1}, \cdots, X_{t_1+m}, X_{t_2-m}, \cdots, X_{t_n+m})$ a même loi que $(X_{t_1-m+c}, X_{t_1-m+1+c}, \cdots, X_{t_1+m+c}, X_{t_2-m+c}, \cdots, X_{t_n+m+c})$. Maintenant, en considérant la fonction $g : \mathbb{R}^{(2m+1)k} \to \mathbb{R}^k$ telle que $g(X_{t_1-m}, X_{t_1-m+1}, \cdots, X_{t_1+m}, X_{t_2-m}, \cdots, X_{t_n+m}) = \left(\sum_{i=-m}^m a_i\, X_{t_1-i}, \cdots, \sum_{i=-m}^m a_i\, X_{t_n-i}\right)$, g étant continue donc mesurable, on montre bien que $g(X_{t_1-m}, X_{t_1-m+1}, \cdots, X_{t_1+m}, X_{t_2-m}, \cdots, X_{t_n+m})$ à la même loi que $g(X_{t_1-m+c}, X_{t_1-m+1+c}, \cdots, X_{t_1+m+c}, X_{t_2-m+c}, \cdots, X_{t_n+m+c})$ donc $(Y_{t_1}, \cdots, Y_{t_n})$ a la même loi que $(Y_{t_1+c}, \cdots, Y_{t_n+c})$: (Y_t) est bien stationnaire.
Si X est stationnaire d'ordre 2, il est facile voir que $\mathbb{E}Y_t$ est constante. On a $\mathrm{cov}(Y_s, Y_t) = \sum_{i\in I}\sum_{i'\in I} a_i a_{i'} \mathrm{cov}(X_{t-i}, X_{s-i'}) = \sum_{i\in I}\sum_{i'\in I} a_i a_{i'} \gamma(t-s-i+i')$ en notant $\gamma(k) = \mathrm{cov}(X_0, X_k)$. Donc $\mathrm{cov}(Y_s, Y_t)$ est bien une fonction de $t-s$. De plus, on peut intervertir i et i' et du fait de la parité de γ on voit bien que $\mathrm{cov}(Y_s, Y_t)$ est une fonction de $|t-s|$.
(2) On a $\mathbb{E}|Y_t| \le \mathbb{E}\left(\sum_{i\in I} |a_i|\,|X_{t-i}|\right) \le \left(\sum_{i\in I} |a_i|\right) \sup_{j\in\mathbb{Z}} \mathbb{E}|X_j|$ d'après le Théorème de Lebesgue, donc $\mathbb{E}|Y_t| < \infty$: (Y_t) est bien finie avec une probabilité 1.
Pour la stationarité stricte, on procède comme précédemment en considérant des restrictions $(Y_t^{(m)})_t$ qui sont bien stationnaires. Du fait de l'existence d'une limite, on a $(Y_{t_1}^{(m)}, \cdots, Y_{t_n}^{(m)})$ qui tend dans $\mathbb{L}^1$ vers $(Y_{t_1}, \cdots, Y_{t_n})$ lorsque m tend vers l'infini, donc on a également convergence en loi. Aussi comme $(Y_{t_1}^{(m)}, \cdots, Y_{t_n}^{(m)})$ a la même loi que $(Y_{t_1+c}^{(m)}, \cdots, Y_{t_n+c}^{(m)})$, cette égalité a également lieu à la limite: (Y_t) est bien stationnaire.
(3) On a $\mathbb{E}(Y_t - Y_t^{(m)})^2 = \mathbb{E}\left(\sum_{|i|>m}\sum_{|i'|>m} a_i\, a_{i'}\, \mathbb{E}(X_{t-i}X_{t-i'}\right) \le \left(\sum_{|i|>m} |a_i|\right)^2 \sup_{j\in\mathbb{Z}} \mathbb{E}|X_j^2|$ d'après le Théorème de Lebesgue, donc $\mathbb{E}(Y_t - Y_t^{(m)})^2 \to 0$ $(m \to \infty)$ car $\sum_{i\in I} |a_i| < \infty$: (Y_t) existe dans $\mathbb{L}^2$.
La stationarité stricte s'obtient comme dans le cas précédent et la stationarité d'ordre 2 utilise le point (1): on a $(Y_t^{(m)})_t$ qui est stationnaire d'ordre 2 et comme on a convergence dans $\mathbb{L}^2$ donc en loi, on a bien convergence de l'espérance et de la covariance des $(Y_t^{(m)})$: (Y_t) est bien aussi stationnaire d'ordre 2.
(4) Dans le cas d'un processus linéaire, on peut donc obtenir l'existence et la stationarité sous la condition $\sum_{i\in I} a_i^2 < \infty$ qui est plus faible que la condition $\sum_{i\in I} |a_i| < \infty$. En effet, on a $\mathbb{E}(Y_t - Y_t^{(m)})^2 = \mathbb{E}\left(\sum_{|i|>m}\sum_{|i'|>m} a_i\, a_{i'}\, \mathbb{E}(X_{t-i}X_{t-i'}\right) = \sum_{|i|>m} a_i^2\, \mathbb{E}(X_0^2) \to 0$ $(m \to \infty)$ dès que (X_t) est un bruit blanc. La preuve de la stationarité est immédiate (voir (3)), et pour la stationarité d'ordre 2, on a clairement $\mathrm{cov}(Y_s, Y_t) = \mathbb{E}X_0^2 \sum_{i\in I} a_i a_{i+s-t}$ qui est une fonction dépendant de $|t-s|$ et qui existe (d'après Cauchy-Schwarz). □

Propriété. *Avec les notations de la définition précédente, la composée de 2 filtres linéaires est un filtre linéaire.*

Propriété. *On considère les filtres linéaires particuliers suivants:*

1. $a_i = 1/k$ *pour* $i = 1, \cdots, k$ *(à une translation près). On dit alors que la série* Y *est une moyenne mobile. Alors si* X *a une composante saisonnière de période* r*, elle est annulée dès que* k *est un multiple de* r*. De plus, si la partie bruit de* X *est un bruit blanc, alors le processus à temps discret* Y *a un bruit de variance plus petite que celle de* X*. Enfin, seuls les polynômes de degré* 0 *et* 1 *sont invariants par ce filtre.*

2. $a_i = (-1)^i C_i^k$ *pour* $i = 0, 1, \cdots, k$ *(à une translation près). Alors si la tendance de* X *est une tendance polynômiale de degré* $k-1$*, elle est annulée.*

3. $a_0 = -1$*,* $a_r = 1$ *et* $a_i = 0$ *pour* $i = 1, \cdots, r-1$*. Alors si* X *a une composante saisonnière de période* r*, elle est annulée.*

Propriété. *Quelque soit le filtre linéaire utilisé, si* Y *est le processus à temps discret filtré, on peut reconstruire le processus original* X *à partir de* Y *(+ les premières valeurs de* X*).*

Remarque.

Attention, il ne faut pas croire qu'utiliser des filtres linéaires est une solution "magique" aux problèmes de tendance et de stationnarité. Il faut bien avoir en tête que la structure de la partie bruit change après le passage d'un filtre et généralement le nouveau bruit est plus compliqué. Par exemple, si le bruit initial est blanc, après l'application d'un filtre, le nouveau bruit à la structure d'un processus MA (voir ci-dessous): il est devenu "dépendant" (non blanc).
Cependant la technique suivante peut être intéressante pour estimer la tendance $a(\cdot)$ et la saisonnalité $S(\cdot)$ (de période r connue) d'un processus X:

1. On utilise d'abord le filtre $a_i = 1/2r$ pour $i = -r+1, -r+2, \cdots, r-1, r$. Ce filtre va donc permettre d'annuler la saisonnalité et en même temps de "moyenner" autour de chaque point, donc d'une certaine manière d'approcher la tendance. Soit Y la série filtrée.

2. On considère ensuite la série $X - Y$, qui est une approximation de la série détendancialisée. On peut alors estimer la saisonnalité, soit $\widehat{S}(\cdot)$ sur cette série, à l'aide par exemple de la méthode par régression présentée plus haut (donc avec des moyennes pour chaque $1 \leq t \leq r-1$).

3. On peut maintenant considérer $X_t - \widehat{S}(t)$, série désaisonnalisée, et estimer la tendance (par exemple avec une des méthodes paramétriques ou non-paramétriques vues plus haut ou bien avec une moyenne mobile).

Une telle méthode, finalement assez simple, peut concurrencer la méthode de régression vue plus haut pour estimer conjointement la tendance et la saisonnalité.

2.4 Présence d'une tendance multiplicative

Tout comme dans le cas du modèle linéaire avec résidus hétéroscedastiques, certaines transformations peuvent permettre d'éliminer la tendance multiplicative tout en l'estimant. On suppose donc que l'on a su estimer la partie $m(t) = \mathbb{E}X_t$ (voir ci-dessous différentes techniques) et que l'on dispose des $\widehat{Y}_t = X_t - \widehat{m}(t)$. On supposera ici que $\sigma(t) = \gamma(m(t))$ ce qui est assez réaliste en pratique. Il est donc possible de représenter en fonction de $\widehat{m}(t)$, les $\widehat{Y}_t$, ce qui donne une idée du graphe de $m \to \gamma(m)$. Si ce graphe est de type $\gamma(m) \simeq C\, m^a$ avec $a > 0$ et $a \neq 1$, alors on transformera les $\widehat{Y}_t$ en $\widehat{Z}_t = (\widehat{Y}_t)^{1-a}$. Si on a plutôt $\gamma(m) \simeq C\, m$, on transformera les $\widehat{Y}_t$ en $\widehat{Z}_t = \log \widehat{Y}_t$. On peut s'attendre alors à ce que les $\widehat{Z}_t$ aient une variance constante, donc plus de tendance multiplicative.

3 Exemples de processus stationnaires à temps discret

Nous allons donner ici plusieurs exemples de processus à temps discret. Tout d'abord nous présenterons les processus ARMA et succinctement les processus ARCH qui seront des exemples de processus permettant de modéliser des données de type "continues". Les chaînes de Markov seront ensuite évoquées comme exemple de modèle pour des données de type "discrètes".

3.1 Un premier exemple de processus à temps discret: les processus ARMA

Les processus ARMA et ARIMA sont sans doute l'exemple le plus simple de processus à temps discret et à valeurs dans $\mathbb{R}$. Certains résultats généralisant à l'infini des résultats déjà vus pour les filtres linéaires finis doivent d'abord être établis.

Définition. *Pour $(X_k)_{k\in\mathbb{Z}}$ un processus à temps discret, on appelle:*

- *opérateur retard B l'application linéaire qui à X associe $Y=(Y_n)_{n\in\mathbb{Z}}=B\cdot X$ tel que $Y_n=X_{n-1}$ pour $n\in\mathbb{Z}$ (par abus de notation $B\cdot X_n=X_{n-1}$).*
- *puissance B^k la k-ième composée de B par elle même qui est donc telle $B^k\cdot X_n=X_{n-k}$.*
- *avec le polynôme $P(B)=\sum_{j=0}^p a_jB^j$ où $(a_i)_{0\leq p}\in\mathbb{R}^{p+1}$, $P(B)\cdot X=(\sum_{j=0}^p a_jX_{n-j})_{n\in\mathbb{Z}}$.*
- *B^{-1} vérifiant $B^{-1}\cdot X_n=Y_{n+1}$ est tel que $B^{-1}B=BB^{-1}=I_d$ d'où l'extension à B^k pour $k\in\mathbb{Z}$.*
- *si $\sum_{i=-\infty}^{\infty}|a_i|<\infty$, alors $f(B)=\sum_{i=-\infty}^{\infty}a_iB^i$ transforme X en Y tel que $Y_n=\sum_{i=-\infty}^{\infty}a_iX_{n-i}$.*

Du fait de la structure linéaire des applications B, $P(B)$ ou $f(B)$, on a les propriétés suivantes:

Propriété.

- *Additivité des fonctions de B: $f(B)\cdot(\lambda X+\lambda X')=\lambda f(B)\cdot X+\lambda' f(B)\cdot X'$.*
- *Commutativité du produit de fonctions de B: $f_1(B)f_2(B)=f_2(B)f_1(B)$.*

Proposition. *Soit $X=(X_n)_{n\in\mathbb{Z}}$ et $Y=(Y_n)_{n\in\mathbb{Z}}=(1-\lambda B)\cdot X$. Alors,*

- *si $|\lambda|<1$, $X_n=(1-\lambda B)^{-1}\cdot Y_n=\sum_{i=0}^{\infty}\lambda^iY_{n-i}$;*
- *si $|\lambda|>1$, $X_n=-\lambda^{-1}B^{-1}(1-\lambda^{-1}B^{-1})^{-1}\cdot Y_n=-\sum_{i=1}^{\infty}\lambda^{-i}Y_{n+i}$;*
- *si $|\lambda|=1$, $1-\lambda B$ n'est pas inversible.*

Conséquence.

1. Si on suppose que $P(B)$ est un polynôme dont les racines ne sont pas situées sur le cercle trigonométrique (donc les racines sont telles que $|z|\neq 1$), alors il existe une unique série de la forme $(P(B))^{-1}=\sum_{i=-\infty}^{\infty}a_iB^i$ avec $\sum_{i=-\infty}^{\infty}|a_i|<\infty$.
2. Si les racines de P sont hors du disque unité (donc les racines sont telles que $|z|>1$) alors alors il existe une unique série de la forme $(P(B))^{-1}=\sum_{i=0}^{\infty}a_iB^i$ avec $\sum_{i=0}^{\infty}|a_i|<\infty$. On dira alors que la relation entre les deux processus est causale.

Propriété. *Soit $(\varepsilon_n)_{n\in\mathbb{Z}}$ un bruit blanc et soit $(a_i)_{i\in\mathbb{Z}}$ une famille de réels tels que $\sum_{i=-\infty}^{\infty}|a_i|<\infty$. Alors le processus à temps discret $X=(X_n)_{n\in\mathbb{Z}}$ tel que $X_n=\sum_{i=-\infty}^{\infty}a_i\varepsilon_{n-i}$ pour $n\in\mathbb{Z}$ est stationnaire et appelée processus linéaire (ou processus à moyennes mobiles infinies). Si $a_i=0$ pour $i<0$ alors on dira que X est un processus linéaire causal.*

Propriété. *Soit le processus à temps discret* $X = (X_n)_{n\in\mathbb{Z}}$ *tel que* $X_n = \sum_{i=-\infty}^{\infty} a_i \varepsilon_{n-i}$ *pour* $n \in \mathbb{Z}$, *avec* $(\varepsilon_n)_{n\in\mathbb{Z}}$ *un bruit blanc et* $\sum_{i=-\infty}^{\infty} |a_i| < \infty$. *Alors* X *admet une densité spectrale* f *telle que*

$$f(\lambda) = \frac{1}{2\pi}\Big| \sum_{j=-\infty}^{\infty} a_j e^{-ij\lambda}\Big|^2, \quad \text{pour } \lambda \in [-\pi, \pi[.$$

Propriété. *Soit le processus à temps discret* $Y = (Y_n)_{n\in\mathbb{Z}}$ *tel que* $Y_n = \sum_{i=-\infty}^{\infty} a_i X_{n-i}$ *pour* $n \in \mathbb{Z}$, *avec* $X = (X_n)_{n\in\mathbb{Z}}$ *un processus à temps discret stationnaire de densité spectrale* f_X *et* $\sum_{i=-\infty}^{\infty} |a_i| < \infty$. *Alors* Y *est un processus à temps discret stationnaire de densité:*

$$f_Y(\lambda) = f_X(\lambda)\Big| \sum_{j=-\infty}^{\infty} a_j e^{-ij\lambda}\Big|^2, \quad \text{pour } \lambda \in [-\pi, \pi[.$$

Définition. *Soit* P *et* Q *deux polynômes de degrés respectifs* p *et* q *tels que* P *et* Q *soient premiers entre eux dans* $\mathbb{C}[X]$. *On suppose que les racines de* P *ne sont pas sur le cercle unité. Soit également* $\varepsilon = (\varepsilon_n)_{n\in\mathbb{Z}}$ *un bruit blanc (fort).*

- *Un processus* $X = (X_n)_{n\in\mathbb{Z}}$ *tel que* $P(B)\cdot X = \varepsilon$ *est appelé un processus AR(p). Ceci revient à écrire qu'il existe* $(b_1, \cdots, b_p) \in \mathbb{R}^p$ *tel que:*

$$X_n + b_1 X_{n-1} + \cdots + b_p X_{n-p} = \varepsilon_n \quad \text{pour tout } n \in \mathbb{Z}.$$

- *Un processus* $X = (X_n)_{n\in\mathbb{Z}}$ *tel que* $X = Q(B)\cdot \varepsilon$ *est appelé un processus MA(q). Ceci revient à écrire qu'il existe* $(c_1, \cdots, c_q) \in \mathbb{R}^q$ *tel que:*

$$X_n = \varepsilon_n + c_1\varepsilon_{n-1} + \cdots + c_q\varepsilon_{n-q} \quad \text{pour tout } n \in \mathbb{Z}.$$

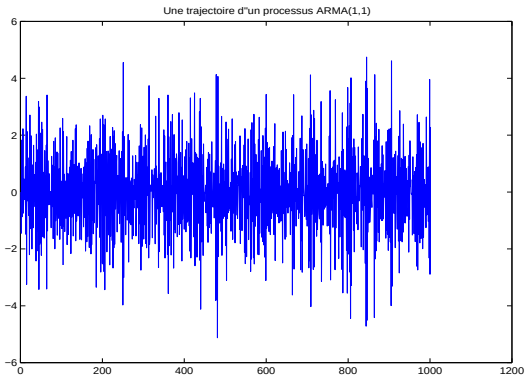
- *Un processus* $X = (X_n)_{n\in\mathbb{Z}}$ *tel que* $P(B)\cdot X = Q(B)\cdot \varepsilon$ *est appelé un processus ARMA(p, q). Ceci revient à écrire qu'il existe* $(b_1, \cdots, b_p) \in \mathbb{R}^p$ *et* $(c_1, \cdots, c_q) \in \mathbb{R}^q$ *tels que:*

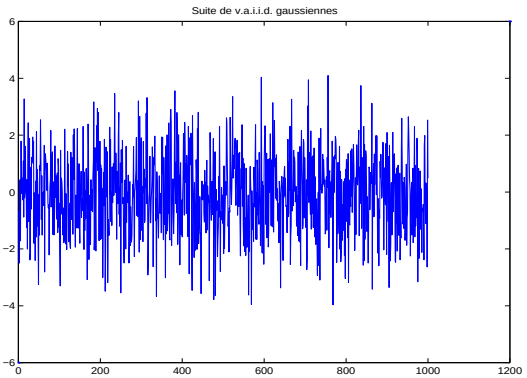
$$X_n + b_1 X_{n-1} + \cdots + b_p X_{n-p} = \varepsilon_n + c_1\varepsilon_{n-1} + \cdots + c_q\varepsilon_{n-q} \quad \text{pour tout } n \in \mathbb{Z}.$$

Exemple.

Exemples d'AR(1), de MA(1), d'ARMA(1, 1).

Voici un exemple (à gauche) d'une trajectoire du processus ARMA(1, 1) avec $b_1 = 0.6$ et $c_1 = -0.5$, à comparer avec une trajectoire (à droite) de variables gaussiennes indépendantes et identiquement distribuées:





Propriété. *Si les racines de P sont à l'extérieur (strictement) du disque trigonométrique, alors un processus ARMA est stationnaire et causal. Si les racines de P sont à l'intérieur (strictement) du disque trigonométrique, le processus ARMA est stationnaire mais non causal.*

Exemple.

Exemples d'ARMA stationnaires et non-stationnaires.

Propriété. *Lorsque les racines de P sont à l'extérieur du disque trigonométrique, on peut établir des relations de récurrence permettant de calculer la fonction covariance d'un processus ARMA(p,q) tel que*

$$P(B)\cdot X_n = X_{n+p} + b_1 X_{n+p-1} + \cdots + b_p X_n = Q(B)\cdot \varepsilon_n,$$

et on obtient

$$r(n+p) + b_1 r(n+p-1) + \cdots + b_p r(n) = 0 \quad \textit{pour tout} \quad n \geq p-q.$$

Corollaire. *En particulier, on obtient les équations de Yule-Walker, vérifiant le système:*

$$\begin{pmatrix} r(q) & r(q-1) & \cdots & r(q-p+1) \\ r(q+1) & r(q) & \cdots & r(q-p+2) \\ \vdots & \vdots & \vdots & \vdots \\ r(q+p-1) & r(q+p-2) & \cdots & r(q) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = - \begin{pmatrix} r(q+1) \\ r(q+2) \\ \vdots \\ r(q+p) \end{pmatrix}.$$

Conséquence.

- La covariance d'un processus ARMA décroît asymptotiquement exponentiellement vite.
- S'il existe un estimateur convergent de la covariance (ou de la corrélation), l'inversion des équations de Yule-Walker permet d'obtenir des estimateurs des paramètres a_i quand une trajectoire $(X_1, \cdots, X_N)$ est connue.

Propriété. *Lorsque les racines de P sont à l'extérieur du disque trigonométrique, un processus ARMA(p,q) (tel que $P(B)\cdot X = Q(B)\cdot \varepsilon$ avec (ε_n) un bruit blanc de variance σ^2) a pour densité spectrale:*

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left| \frac{Q(e^{i\lambda})}{P(e^{i\lambda})} \right|^2, \quad \textit{pour } \lambda \in [-\pi, \pi[.$$

3.2 Une brève présentation des processus ARCH

Les processus ARCH(p) ont été introduits par Engle (1982) et Bollersev (1986) dans le cadre de données financières où la "volatilité" (la variance) dépend conditionnellement du passé.

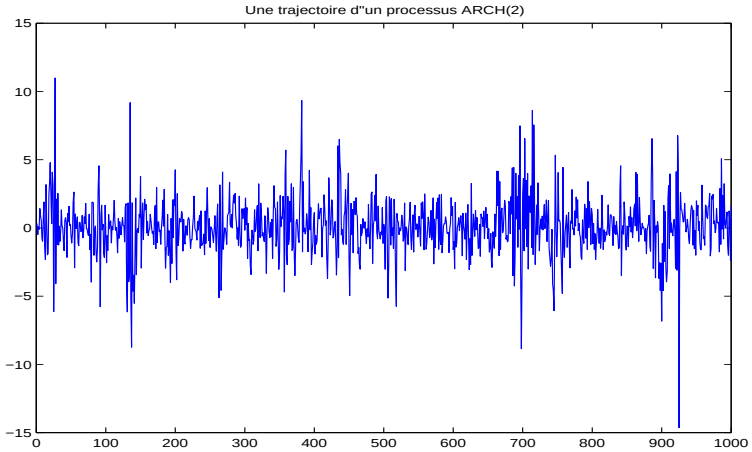
Définition. *Soit $(\varepsilon_k)_{k\in\mathbb{Z}}$ un bruit blanc. Alors on dit que $(X_k)_{k\in\mathbb{Z}}$ est un processus ARCH(p), où $p \in \mathbb{N}^*$ s'il existe $(a_0, a_1, \cdots, a_p)$ une famille de constantes réelles positives telles que*

$$X_k = \varepsilon_k \sqrt{a_0 + \sum_{j=1}^{p} a_j X_{k-j}^2} \quad \textit{pour tout } k \in \mathbb{Z}. \tag{2}$$

On peut se demander s'il est possible de trouver des processus stationnaires qui peuvent vérifier une telle équation de récurrence. Nous contenterons ici de donner une réponse de la cas de la stationnarité d'ordre 2:

Propriété. *Un processus ARCH(p) vérifiant l'équation (2) est stationnaire d'ordre 2 si et seulement si* $\mathbb{E}\left[\varepsilon_0^2\right] \sum_{k=1}^{p} a_k < 1$.

Ce résultat est très récent et nous n'en donnerons pas la preuve (en particulier la condition nécessaire n'a été montrée qu'en 2000...). Voici un exemple d'une trajectoire du processus ARCH(2) avec $a_0 = 1$, $a_1 = 0.3$ et $a_2 = 0.5$:



Propriété. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus ARCH(p) stationnaire d'ordre 2. Alors X_k n'est pas indépendant de X_0, mais $r(k) = cov(X_0, X_k) = 0$ pour tout $k \neq 0$. De plus la densité spectrale f de X est la même que celle d'un bruit blanc: $f(\lambda) = C$ pour tout $\lambda \in [-\pi, \pi[$, où $C > 0$.*

On s'aperçoit donc qu'un processus ARCH(p) ne pourra pas être identifié à partir de sa densité spectrale. Une conséquence de ceci est également qu'un ARCH(p) ne peut pas être gaussien (car sinon ce serait un bruit blanc, ce qu'il n'est pas). On peut cependant montrer qu'en considérant $(X_k^2)_k$ au lieu de $(X_k)_k$ on se ramène à un processus ARMA:

Propriété. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus ARCH(p) stationnaire d'ordre 2. Alors $Y = (X_k^2)_{k\in\mathbb{Z}}$ est un processus ARMA(p).*

3.3 Chaînes de Markov à espace d'états fini

On revient pour commencer sur quelques définitions relatives aux chaînes de Markov:

Définition. *Soit $E = (e_1, \cdots, e_p)$ un espace d'état finis. On dit que $X = (X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov sur E si pour tout $n \in N$, pour tout $x_0, x_1, \cdots, x_{n+1} \in E$,*

$$\Pr(X_{n+1} = x_{n+1} \,|\, X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_0 = x_0) = \Pr(X_{n+1} = x_{n+1} \,|\, X_n = x_n).$$

Cette propriété qui fait que le "coup" suivant ne dépende que du présent et non du passé est appelée propriété de Markov. On peut ainsi associée à une telle chaîne une matrice de transition:

Définition. *Soit $X = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov sur $E = (e_1, \cdots, e_p)$. On appelle matrice de transition de la chaîne la matrice $P^{(n)} = (p_{ij}^{(n)})_{1\leq i,j\leq p}$ telle que*

$$p_{ij}^{(n)} = \Pr(X_{n+1} = e_j \,|\, X_n = e_i) \quad \textit{pour tout } 1 \leq i, j \leq p.$$

Si $P^{(n)}$ ne dépend pas de n, on dit que X est une chaîne de Markov homogène.

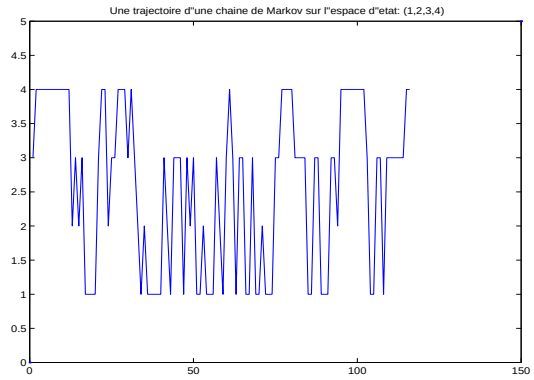
Exemple.

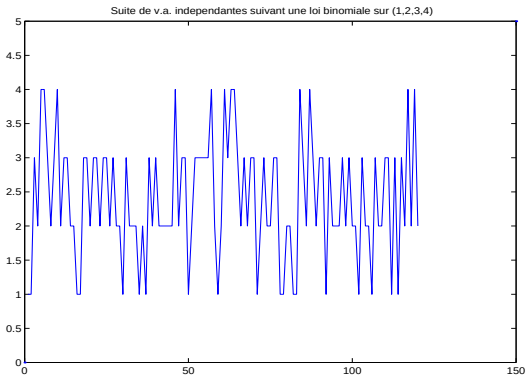
Chaîne de Markov à deux états.

Propriété. *La matrice de transition P d'une chaîne de Markov homogène est dite stochastique, c'est-à-dire que la somme des termes sur une même ligne vaut 1, ou encore $\sum_{j=1}^{p} p_{ij} = 1$.*

Proposition. *Soit $X = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène sur $E = (e_1, \cdots, e_p)$ de matrice de transition P. On suppose que X_0 suit une loi P_0 sur E. Alors la loi de X_n est $P_0 \cdot P^n$.*

Ci-dessous on peut observer une trajectoire (à droite) d'une suite de v.a.i.i.d. binomiales sur $\{1, 2, 3, 4\}$ avec celle (à gauche) d'une chaîne de Markov sur $\{1, 2, 3, 4\}$ dont la matrice de transition est $P = \begin{pmatrix} 0.5 & 0.25 & 0.25 & 0 \\ 0.5 & 0 & 0.4 & 0.1 \\ 0.25 & 0.25 & 0.30 & 0.2 \\ 0 & 0.1 & 0.2 & 0.7 \end{pmatrix}$





La différence la plus marquante entre la chaîne et la suite de v.a.i.i.d. réside dans le fait que la chaîne peut rester parfois longtemps dans le même état. On s'aperçoit aussi que l'état 2 est nettement le moins visité par la chaîne, les états 1, 3 et 4 étant à peu près visités autant. Si on laissait la chaîne "tourner" plus longtemps ce comportement aurait tendance à se reproduire, comme si les X_n pour n grand finissait par avoir tous la même loi, comme si donc X_n convergeait en loi vers une certaine mesure...

Définition. *Soit $X = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène sur $E = (e_1, \cdots, e_p)$ de matrice de transition P. Soit $\mu = (\mu_1, \cdots, \mu_p)$ un vecteur correspondant à une mesure de probabilité sur E (la mesure que l'on nommera aussi μ est donc $\sum_{i=1}^{p} \mu_i \delta_{e_i}$). Alors on dit que μ est invariante pour la chaîne de Markov de matrice de transition P si:*

$$\mu \cdot P = \mu.$$

On voit donc que si μ est invariante alors ${}^t\mu$ est une valeur propre de ${}^tP$. D'où la propriété suivante:

Propriété. *Soit $X = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène sur $E = (e_1, \cdots, e_p)$ de matrice de transition P. Alors il existe au moins une mesure de probabilité invariante pour cette chaîne.*

D'autres propriétés peuvent servir à caractériser les chaînes de Markov.

Définition. *Soit $X = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène sur $E = (e_1, \cdots, e_p)$ de matrice de transition P. On dit que X est irréductible si*

$$\forall i, j \in \{1, \cdots, p\}, \quad \exists k \in \mathbb{N}^* \quad \textit{tel que} \quad (P^k)_{ij} = \Pr(X_k = e_j \,|\, X_0 = e_i) > 0.$$

En particulier si la matrice P n'a aucun 0 alors la chaîne est irréductible.

4 Identification d'un processus stationnaire à temps discret

On s'intéresse ici à l'identification de la partie bruit d'un processus à temps discret. Lorsque l'on considère une suite de v.a.i.i.d., on connaît un certain nombre de techniques (estimation, tests) et des résultats asymptotiques qui permettent d'en savoir plus sur cette suite (par exemple estimer sa variance, la probabilité d'une occurrence, tester si les variables sont gaussiennes,...). Pour un processus à temps discret, il faut compter avec la structure de dépendance du bruit, ce qui nous oblige à donner des extensions aux résultats asymptotiques classiques (loi des grands nombres, théorème de la limite centrale,...).

4.1 Comportement asymptotique des processus stationnaires à temps discret

Proposition. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus du second ordre à temps discret stationnaire d'espérance m et de densité spectrale f continue en 0. Alors:*

$$\begin{aligned} \overline{X_n} &\xrightarrow[n\to\infty]{p.s.} m \quad et \\ \lim_{n\to\infty} \mathbb{E}\big(\sqrt{n}(\overline{X_n} - m)\big)^2 &= 2\pi f(0). \end{aligned}$$

Proposition. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus gaussien stationnaire d'espérance m et de densité spectrale f. Alors:*

1. *Si* $\lim_{k\to\infty} r(k) = 0$ *alors* $\overline{X_n} \xrightarrow[n\to\infty]{p.s.} m$.

2. *Si la densité spectrale* f *existe et est continue en* 0, *alors* $\sqrt{n}(\overline{X_n} - m) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 2\pi f(0))$.

On peut maintenant donner une loi des grands nombres et un théorème de la limite centrale vérifiés par les chaînes de Markov.

Théorème (Théorème ergodique)**.** *Soit* $X = (X_n)_{n\in\mathbb{N}}$ *une chaîne de Markov homogène irréductible sur* $E = (e_1, \cdots, e_p)$ *ayant pour unique mesure de probabilité invariante* μ. *Alors pour toute fonction* $h : E \to \mathbb{R}$ *mesurable, et quelque soit la loi de* X_0,

$$\frac{1}{n+1}\sum_{j=0}^{n} h(X_i) \xrightarrow[n\to\infty]{p.s.} \int h(x)d\mu(x) = \sum_{i=1}^{p} h(e_i)\mu(e_i).$$

Théorème (Théorème de la limite centrale pour les chaînes de Markov à espace d'états)**.** *Soit* $X = (X_n)_{n\in\mathbb{N}}$ *une chaîne de Markov homogène irréductible sur* $E = (e_1, \cdots, e_p)$ *ayant pour unique mesure de probabilité invariante* μ. *Alors pour toute fonction* $h : E \to \mathbb{R}$ *mesurable, et quelque soit la loi de* X_0,

$$\sqrt{n}\Big(\frac{1}{n+1}\sum_{j=0}^{n} h(X_i) - \int h(x)d\mu(x)\Big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(0, \int h^2(x)d\mu(x) - \big(\int h(x)d\mu(x)\big)^2\Big).$$

Voyons maintenant des utilisations de ces résultats. En premier lieu, on peut définir les moments empiriques d'ordre 2 d'un processus à temps discret:

Définition. *Soit* $X = (X_k)_{k\in\mathbb{Z}}$ *un processus à temps discret. On appelle covariance empirique de* $(X_1, \cdots, X_n)$ *pour* $p = 0, 1, \cdots, n-1$,

$$\widehat{c}_n(p) = \frac{1}{n}\sum_{k=1}^{n-p}(X_{k+p} - \bar{X}_n)(X_k - \bar{X}_n) \quad \textit{et } \widehat{c}_n(-p) = \widehat{c}_n(p).$$

Lorsque l'on sait que le processus est centré alors pour $p = 0, 1, \cdots, n-1$,

$$\widehat{c}_n(p) = \frac{1}{n}\sum_{k=1}^{n-p} X_{k+p}X_k \quad \textit{et } c_n(-p) = c_n(p).$$

On appelle corrélation empirique de $(X_1, \cdots, X_n)$ *pour* $p = 0, 1, \cdots, n-1$,

$$\widehat{\rho}_n(p) = \frac{\widehat{c}_n(p)}{\widehat{c}_n(0)}.$$

Remarque.

Un certain nombre de logiciels propose de tracer le "correlogram" de la série (souvent la fonction `ACF`) soit le graphe de $k \mapsto \widehat{\rho}_n(k)$ pour $k \in \{0, 1, \cdots, m\}$, avec m souvent proche de $\sqrt{N}$. Il est bien clair que l'on a toujours $\widehat{\rho}_n(0) = 1$ et nous allons voir que sous des hypothèses assez générales on peut espérer que ce corrélogramme soit une bonne approximation du graphe de la fonction de corrélation. On définit maintenant l'estimateur "naturel" de la densité spectrale qui consiste à remplacer la covariance par la covariance empirique dans l'expression de la densité spectrale.

Définition. *Soit* $X = (X_k)_{k\in\mathbb{Z}}$ *un processus à temps discret centré. On appelle périodogramme de* $(X_1, \cdots, X_n)$ *la fonction* $I_n(\lambda) : [-\pi, \pi[\to \mathbb{R}_+$ *telle que:*

$$I_n(\lambda) = \frac{1}{2\pi}\sum_{k=1-n}^{n-1} \widehat{c}_n(k)e^{-ik\lambda} = \frac{1}{2\pi n}\Big|\sum_{k=1}^{n} X_k e^{-ik\lambda}\Big|^2.$$

On peut alors montrer la propriété suivante:

Propriété. *Soit* $X = (X_k)_{k\in\mathbb{Z}}$ *un processus à temps discret centré stationnaire ayant des moments d'ordre* 2. *Si les covariances* $r(k)$ *sont telles que* $\sum |r(k)| < \infty$ *alors*

1. $\lim_{n\to\infty} \mathbb{E}(I_n(\lambda)) = f(\lambda)$ *pour tout* $\lambda \in [-\pi, \pi[$.
2. *si* X *est un processus gaussien,* $\lim\limits_{n\to\infty} var(I_n(\lambda)) = \begin{cases} f^2(\lambda) & si\ \lambda \in]-\pi, 0[\cup]0, \pi[\\ 2f^2(\lambda) & si\ \lambda = -\pi, 0. \end{cases}$
3. *si* X *est un processus gaussien,* $\lim\limits_{n\to\infty} cov(I_n(\lambda), I_n(\lambda')) = 0$ *si* $\lambda \neq \lambda'$.

Avec cette propriété on voit que I_n est un estimateur asymptotiquement non biaisé de f, mais un estimateur non convergent dans le cas gaussien. On pourrait généraliser cette propriété dans le cas plus général d'un processus ayant des moments d'ordre 4. Cependant cet estimateur sera intéressant lorsqu'on l'intègre par rapport à une fonction:

Définition. *Soit* $X = (X_k)_{k\in\mathbb{Z}}$ *un processus à temps discret du second ordre. Pour* g *une fonction continue sur* $[-\pi, \pi[$*, on définit le périodogramme intégré par:*

$$J_n(g) = \int_{-\pi}^{\pi} g(\lambda) I_n(\lambda) d\lambda.$$

Proposition. *Soit* $X = (X_k)_{k\in\mathbb{Z}}$ *un processus centré stationnaire ARMA(p, q) ou ARCH(p) tel que* $\mathbb{E}X_0^4 < \infty$ *(ce qui, dans le cas d'un processus ARCH(p) signifie que* $\left(\mathbb{E}\varepsilon_0^4\right)^{1/2} \sum_{j=1}^{p} b_i < 1$*), alors:*

1. *Si Pour* g *une fonction continue sur* $[-\pi, \pi[$, $J_n(g) \xrightarrow[n\to\infty]{p.s.} \int_{-\pi}^{\pi} f(\lambda) g(\lambda) d\lambda$.
2. *Pour* $(g_i)_{1\leq i\leq m}$ *une famille de fonctions continues sur* $[-\pi, \pi[$ *alors:*

$$\sqrt{n}\Big(J_n(g_i) - \int_{-\pi}^{\pi} f(\lambda) g(\lambda) d\lambda\Big)_{1\leq i\leq m} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_m(0, \Sigma)$$

avec $\Sigma = (\sigma_{ij})_{1\leq i,j\leq m}$ *et* $\sigma_{ij} = 4\pi \int_{-\pi}^{\pi} f^2(\lambda) g_i(\lambda) g_j(\lambda) d\lambda$.

Conséquence.

- Rappelons que pour un ARCH(p) stationnaire d'ordre 2, f est une fonction constante.
- Ce double résultat s'applique aussi pour le carré d'un processus ARCH(p) puisque ce carré est un processus ARMA. La condition sera donc que X soit stationnaire et tel que $\mathbb{E}X_0^8 < \infty$, ce qui est possible dès que $\left(\mathbb{E}\varepsilon_0^8\right)^{1/4} \sum_{j=1}^{p} b_i < 1$.
- Ce double résultat s'étend également aux chaînes de Markov homogène à espace d'états fini irréductibles, et ceci quelque soit la loi de X_0.
- Comme $\widehat{c}_n(p) = \int_{-\pi}^{\pi} \cos(p\lambda) I_n(\lambda) d\lambda$, on peut appliquer le théorème de la limite centrale précédent. On obtient ainsi que:

$$\sqrt{n}\Big(\widehat{c}_n(p) - r(p)\Big)_{0\leq p\leq m} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{m+1}\Big(0, \Big(4\pi \int_{-\pi}^{\pi} f^2(\lambda) \cos(\lambda i) \cos(\lambda j) d\lambda\Big)_{0\leq i,j\leq m}\Big).$$

- En utilisant la Delta-méthode pour la fonction $G : (x_0, \cdots, x_m) \in \mathbb{R}^* \times \mathbb{R}^m \mapsto \left(\frac{x_1}{x_0}, \frac{x_2}{x_0}, \cdots, \frac{x_m}{x_0}\right) \in \mathbb{R}^m$, on peut déduire un théorème de la limite centrale pour $(\widehat{\rho}_n(1), \cdots, \widehat{\rho}_n(m))$. Cela permet notamment de tester si X est un bruit blanc car sous une telle hypothèse, $f(\lambda) = \sigma^2/2\pi$ et on a alors

$$\sqrt{n}\big(\widehat{\rho}_n(i)\big)_{1\leq i\leq m} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_m(0, I_m).$$

  Comme dit plus haut ce résultat est également vrai pour un ARCH(p), ce qui limite la puissance d'un tel test.
- En utilisant un noyau comme fonction g que l'on fait se rapprocher d'une impulsion de Dirac en λ_0, on obtient un estimateur de la densité $f(\lambda_0)$.

4.2 Estimation paramétrique et semi-paramétrique pour un processus stationnaire à temps discret

Notation: On suppose que l'on a observé une trajectoire $(X_1, \cdots, X_N)$ d'un processus paramétrique de paramètre $\theta = (\theta_1, \cdots, \theta_\ell)$ où $\ell \in \mathbb{N}^*$ tel que la "vraie" valeur prise par ce paramètre θ^* soit inconnue. La méthode qui suit pourrait être généralisée à d'autres exemples de processus, mais ici nous ne considérerons que les cas paramétriques suivants:

- Les paramètres $b_1, \cdots, b_p, c_1, \cdots, c_q$ d'un processus ARMA(p,q) stationnaire dont on supposera que le bruit blanc $(\varepsilon_k)_k$ admet une densité par rapport à la mesure de Lebesgue;
- Les paramètres $b_1, \cdots, b_p$ d'un processus ARCH(p) stationnaire dont on supposera que le bruit blanc $(\varepsilon_k)_k$ admet une densité par rapport à la mesure de Lebesgue;
- Les probabilités de transition $p_{ij} = \Pr(X_1 = e_j \,|\, X_0 = e_i)$ d'une chaîne de Markov homogène à espace d'états fini irréductibles. Comme la matrice est stochastique, il n'y a que $p(p-1)$ inconnues puisque par exemple $p_{ip} = 1 - \sum_{j=1}^{p-1} p_{ij}$.

On définit $L_N(\theta, \omega)$ une vraisemblance pour $\theta \in \Theta$ (ensemble localement compact de $\mathbb{R}^\ell$). Soit

$$\widehat{\theta}_N = \operatorname*{Argmax}_{\theta\in\Theta} L_N(\theta) = \operatorname*{Argmax}_{\theta\in\Theta} (\log L_N(\theta)).$$

Par "une" vraisemblance on veut dire: soit la densité du vecteur $(X_1, \cdots, X_N)$ (pour les processus ARMA(p,q)), soit la densité conditionnelle par rapport à $X_0, X_{-1}, \cdots$ (pour les processus ARCH(p) ou les chaînes de Markov à espace d'états fini irréductibles). Rappelons que dans ce dernier cas

$$L_N(\theta) = f_{((X_1,\cdots,X_N)\ |X_0,X_{-1},\cdots)} = f_{(X_N\ |X_{N-1},X_{N-2},\cdots)} f_{(X_{N-1}\ |X_{N-2},X_{N-3},\cdots)} \times \cdots \times f_{(X_1\ |X_0,X_{-1},\cdots)}$$

(attention à ne pas confondre cette densité f, qui peut d'ailleurs être une probabilité dans le cas d'une mesure discrète, avec la densité spectrale vue avant). Dans le cas d'une chaîne de Markov homogène irréductible sur $(e_1, \cdots, e_p)$, le modèle est totalement posé et la densité revient à un produit de probabilités et le logarithme de L_N devient:

$$\log L_N(\theta) = \sum_{i=1}^{p}\sum_{j=1}^{p-1} \log(p_{ij}) \sum_{k=1}^{N-1} \mathbb{I}_{(X_k,X_{k+1})=(e_i,e_j)} + \sum_{i=1}^{p} \log(1 - p_{i,1} - \cdots - p_{i,p-1}) \sum_{k=1}^{N-1} \mathbb{I}_{(X_k,X_{k+1})=(e_i,e_p)}.$$

Notons que dans un tel cas la statistique constituée par le vecteur $\big(\sum_{k=1}^{N-1} \mathbb{I}_{(X_k,X_{k+1})=(e_i,e_j)}\big)_{1\le i\le p,\, 1\le j\le p-1}$ est exhaustive (on ne peut pas prendre tous les i et j car la statistique serait liée).

Théorème. *Sous les hypothèses précédentes alors*

1. *si* $\mathbb{E}\big[|\log f_{(X_1\ |X_0,X_{-1},\cdots)}|\big] < \infty$ *on a* $\quad \widehat{\theta}_N \xrightarrow[N\to\infty]{p.s.} \theta^*$.
2. *si* $\mathbb{E}\big[\log^2 f_{(X_1\ |X_0,X_{-1},\cdots)}\big] < \infty$ *on a* $\quad \sqrt{N}(\widehat{\theta}_N - \theta^*) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_m(0, I(\theta^*)^{-1})$, *où*

$$I_{ij}(\theta^*) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \Big(\frac{\partial \log f_\theta(\lambda)}{\partial \theta_i}\Big)_{\theta^*} \cdot \Big(\frac{\partial \log f_\theta(\lambda)}{\partial \theta_j}\Big)_{\theta^*} d\lambda.$$

Remarque.

On retrouve donc ici une vitesse en $\sqrt{N}$ comme dans le cas de variables indépendantes. Cependant une des limitations de la méthode du maximum de vraisemblance réside dans le fait qu'il faille connaître exactement la loi du processus pour en déduire $L_N(\theta)$. L'utilisation de cet estimateur reste malgré tout intéressant pour les processus ARCH(p) (pour lesquels on peut utiliser la vraisemblance conditionnelle gaussienne même si l'innovation n'est pas gaussienne) et les chaînes de Markov. Mais dans le cas d'un processus ARMA(p,q), on préférera utiliser une approximation de l'estimateur $\widehat{\theta}_N$ dite approximation de Whittle (voir ci-dessous) qui offre deux avantages: pas de nécessité de connaître a priori la loi du processus et un grand gain en terme de calcul tout en conservant une vitesse de convergence en $\sqrt{N}$.

Théorème. *Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus ARMA(p,q) stationnaire et de densité spectrale f_θ. On considère le contraste, dit encore contraste de Whittle:*

$$U_N(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(f_\theta(\lambda)) + \frac{I_N(\lambda)}{f_\theta(\lambda)} d\lambda.$$

Soit $\widetilde{\theta_N} = Arg\min_{\theta\in\Theta} U_N(\theta)$, appelé estimateur de θ par minimum de contraste (ou maximum de vraisemblance approché de Whittle). Alors, sous les conditions du Théorème précédent et avec une matrice $J(\theta^*)$:

$$\sqrt{N}(\widetilde{\theta_N} - \theta^*) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_m(0, J(\theta^*)^{-1}).$$

Remarque.

Notons tout d'abord que la matrice $J = I$ lorsque le processus est gaussien. Sinon, elle fait intervenir les moments d'ordre 4 de l'innovation. Ensuite, concrètement, quand on travaille sur des données, les intégrales sont approchées par des sommes de Riemann calculées pour les $\lambda_k = \pi k/n$. On peut montrer que l'erreur d'approximation n'est pas préjudiciable au théorème de la limite centrale ci-dessus.

Conséquence.

On préférera donc nettement l'estimateur par contraste de Whittle pour les processus ARMA(p,q) car il peut s'appliquer facilement à des échantillons de grande taille (alors que par exemple dans le cas gaussien, l'estimateur de maximum de vraisemblance classique requiert d'inverser la matrice de covariance ce qui peut être impossible numériquement dès que $n > 10^4$) et comme dit précédemment, il ne requiert pas la connaissance de la loi de X ou de ε. Pour un processus ARCH(p), cet estimateur peut aussi être utiliser en l'appliquant aux carrés de ce processus.

4.3 Identification et test d'adéquation pour les processus ARMA et ARCH

Estimation de l'ordre d'un processus ARMA: Une autre question apparaît une fois l'estimation des paramètres obtenue: comment estimer l'ordre (p,q) de l'ARMA qui dans les deux techniques précédentes est supposé connu? Pour ce faire on utilisera un critère de sélection de modèle, par exemple le critère BIC vu précédemment. Dans le cas d'un processus ARMA celui-ci s'écrit:

$$BIC(p,q) = \log\big(\frac{1}{N}\sum_{j=1}^{N} \widehat{\varepsilon}_j^2\big) + \frac{\log N}{N}(p+q),$$

où $\widehat{\varepsilon}_j = \dfrac{\widehat{P}_N(B)}{\widehat{Q}_N(B)} \cdot X_j$ (il faut donc que Q soit inversible causal, donc que les racines de Q soient en dehors du disque trigonométrique). Les estimateurs $\widehat{P}_N(B)$ et $\widehat{Q}_N(B)$ sont calculés en remplaçant leurs coefficients a_i et b_j par les estimateurs obtenus par une des méthodes évoquées plus haut. On calcule ce critère pour tous $0 \le p \le p_{\max}$ et $0 \le q \le q_{\max}$ et on choisira

$$(\widehat{p}_N, \widehat{q}_N) = \operatorname*{Argmin}_{0\le p\le p_{\max},\ 0\le q\le q_{\max}} BIC(p,q).$$

Estimation de l'ordre d'un processus ARCH: On peut soit reprendre la technique qui vient d'être présentée pour les ARMA en l'appliquant aux carrés du processus. On peut également reprendre la formule générale du critère BIC et l'appliquer à l'estimation de la log-vraisemblance conditionnelle. Auquel cas, on a:

$$BIC(p,q) = \log\big(\prod_{j=1}^{N} \widehat{f}_{(X_j \mid X_{j-1}, X_{j-2}, \cdots)}\big) + \frac{\log N}{N} p,$$

où $\widehat{f}$ est obtenue en remplaçant les coefficients $(b_0, \cdots, b_p)$ par leurs estimations par maximum de vraisemblance.

Test d'adéquation à un processus ARMA: Pour tester l'adéquation à un processus ARMA, on peut utiliser un test dit de portemanteau (ce qui signifie "fourre-tout" en anglais). Après avoir calculé les résidus

estimés $(\widehat{\varepsilon}_1,\dots,\widehat{\varepsilon}_N)$ obtenus à partir $(X_1,\cdots,X_N)$ et des coefficients estimés par une méthode de type maximum de vraisemblance ou minimum de contraste: comme on a $\varepsilon = \dfrac{Q(B)}{P(B)}X$ on définit

$$\widehat{T}_N(k) = N\sum_{j=1}^{k}\big(\widehat{\rho}_\varepsilon(k)\big)^2 \quad \text{où} \quad \widehat{\rho}_\varepsilon(k) = \frac{\frac{1}{N}\sum_{i=p}^{N-k}\widehat{\varepsilon}_i\widehat{\varepsilon}_{i+k} - \Big(\frac{1}{N}\sum_{i=p}^{N}\widehat{\varepsilon}_i\Big)^2}{\frac{1}{N}\sum_{i=p}^{N}\widehat{\varepsilon}_i\widehat{\varepsilon}_{i+k} - \Big(\frac{1}{N}\sum_{i=p}^{N}\widehat{\varepsilon}_i\Big)^2}.$$

Notons que $\widehat{\rho}_\varepsilon(k)$ est la corrélation empirique des résidus, qui doit tendre vers 0 si le modèle est bien un ARMA(p,q) dès que $k\neq 0$, suivant un Théorème de la Limite Centrale que nous avons vu un peu plus haut (d'où le N devant la statistique de test). Notons également que les $\widehat{\varepsilon}_i$ ne sont calculables que lorsque $i\geq p+1$). On doit choisir k suffisamment grand pour donner plus de pertinence au test. Ainsi, sous l'hypothèse que le processus est bien un processus ARMA(p,q), dont le bruit admet un moment d'ordre 4 on peut alors montrer que:

$$\widehat{T}_N(k) \xrightarrow[N\to+\infty]{\mathcal{L}} \chi^2(k-p-q).$$

Le fait que l'on doivent retirer $p+q$ dans le nombre de degrés de liberté du χ^2 limite est assez classique et correspond au fait que cette statistique est obtenue en estimant $p+q$ paramètres.

Test d'adéquation à un processus ARCH: On peut reprendre la technique présentée pour les ARMA et l'appliquer aux carrés d'un processus ARCH. On peut aussi directement mettre en place un test du portemanteau pour les ARCH en "estimant" l'innovation $(\varepsilon_k)_k$ par

$$\widehat{\varepsilon}_k = X_k\big(\widehat{b}_0 + \sum_{j=1}^{p}\widehat{b}_j\,X_{k-j}^2\big)^{-1/2}.$$

Il n'y a plus alors qu'à reprendre la statistique de test $\widehat{T}_N(k)$ décrite plus haut pour les ARMA.

5 Prédiction pour un processus à temps discret

Pouvoir prévoir est généralement une des possibilités offertes par les statistiques, peut-être même la plus importante.

5.1 Définitions et propriétés générales

On commence à faire quelques rappels sur l'espérance conditionnelle qui est essentielle pour donner des formulations théoriques à la prédiction.

Rappel. *Si X et Y sont deux variables aléatoires continues de densité jointe $f_{(X,Y)}(x,y)$ alors*

$$\mathbb{E}(X \mid Y=y) = \frac{\int x f_{(X,Y)}(x,y)dx}{\int f_{(X,Y)}(x,y)dx}.$$

Propriété.

1. *$\mathbb{E}(Y\,|\,\mathcal{B})$ est une variable aléatoire de $\mathbb{L}^2(\Omega,\mathcal{B},P)$; de plus, $\mathbb{E}(Y\,|\,X)=h(X)$, avec h une fonction borélienne.*
2. *Pour Y_1 et Y_2 deux variables aléatoires sur $(\Omega,\mathcal{A},P)$, et $(a,b,c)\in\mathbb{R}^3$, alors*
$$\mathbb{E}(aY_1+bY_2+c\,|\,\mathcal{B}) = a\mathbb{E}(Y_1\,|\,\mathcal{B}) + b\mathbb{E}(Y_2\,|\,\mathcal{B}) + c.$$
3. *Si $Y_1\leq Y_2$, alors $\mathbb{E}(Y_1\,|\,\mathcal{B})\leq\mathbb{E}(Y_2\,|\,\mathcal{B})$.*
4. *Si $Y\in\mathbb{L}^2(\Omega,\mathcal{B},P)$, alors $\mathbb{E}(Y\,|\,\mathcal{B})=Y$; ainsi $\mathbb{E}(g(X)\,|\,X)=g(X)$ pour g une fonction mesurable réelle.*
5. *On a $\mathbb{E}(\mathbb{E}(Y\,|\,\mathcal{B}))=\mathbb{E}Y$, donc $\mathbb{E}(\mathbb{E}(Y\,|\,X))=\mathbb{E}Y$.*

6. *Si $Y^{-1}(\mathcal{B}(\mathbb{R}))$ et $\mathcal{B}$ sont indépendantes alors $\mathbb{E}(Y \mid \mathcal{B}) = \mathbb{E}Y$; ainsi, si X et Y sont indépendantes, $\mathbb{E}(Y \mid X) = \mathbb{E}Y$.*

Proposition. *Si $(Y, X_1, \cdots, X_n)$ est un vecteur gaussien, alors $\mathbb{E}(Y \mid (X_1, \cdots, X_n)) = a_0 + a_1X_1 + \cdots + a_nX_n$, où les a_i sont des réels.*

Définition. *On suppose que $(X_1, \cdots, X_N)$ est connue. On appelle prédiction $\widehat{X}_{N+p}$ à l'horizon p $\widehat{X}_{N+p} = f_p(X_1, \cdots, X_N)$ où f_p est une fonction mesurable.*

Définition. *Soit X un processus à temps discret du second ordre. On appelle erreur moyenne quadratique (MSE) de la prédiction $\widehat{X}_{N+p}$ le réel*

$$\mathbb{E}(\widehat{X}_{N+p} - X_{N+p})^2.$$

Proposition. *On suppose que X un processus à temps discret du second ordre. La prédiction $\widehat{X}_{N+p}$ optimale au sens de la moyenne quadratique (ou des moindres carrés) est définie par:*

$$\widehat{X}_{N+p} = \widehat{f}(X_1, \cdots, X_N) \text{ avec } \widehat{f} = Arg\min_{f \in \mathbb{L}^2} \left(\mathbb{E}(X_{N+p} - f(X_1, \cdots, X_N))^2\right).$$

Alors,

$$\widehat{X}_{N+p} = \mathbb{E}(X_{N+p} \mid (X_1, \cdots, X_N)).$$

$\widehat{X}_{N+p}$ est donc une variable aléatoire dépendant de $(X_1, \cdots, X_N)$ telle que

$$\mathbb{E}(X_{N+p} \mid (X_1, \cdots, X_N)) = \frac{\int x_{N+p} f_{(X_1, \cdots, X_N, X_{N+p}))}(X_1, \cdots, X_N, x_{N+p}))dx_{N+p}}{\int f_{(X_1, \cdots, X_N, X_{N+p}))}(X_1, \cdots, X_N, x_{N+p}))dx_{N+p}}.$$

Propriété (Cas des processus ARMA). *Soit $(X_k)_{k\in\mathbb{Z}}$ un processus ARMA(p, q) avec $P(B) \cdot X = Q(B) \cdot \varepsilon$ où les racines de P et de Q sont à l'extérieur du disque trigonométrique (donc X est stationnaire, causal et inversible). On suppose également que $(X_1, \cdots, X_N)$ est connu. Soit $\widehat{X}_{N+1}$ le prédicteur optimal au sens des moindres carrés. Alors $\widehat{X}_{N+1} = -\sum_{k=0}^{N-1} c_{k+1}X_{N-k}$ où les (c_k) sont définis par le développement en série entière de*

$$\frac{P}{Q}(z) = \sum_{k\geq 0} c_k z^k.$$

Propriété (Cas des processus ARCH(p)). *Soit $(X_k)_{k\in\mathbb{Z}}$ un processus ARCH(p) défini par*

$$X_k = \xi_k \sqrt{a_0 + a_1X_{k-1}^2 + \cdots + a_pX_{k-p}^2} \quad \text{pour tout } k \in \mathbb{Z},$$

où X est stationnaire d'ordre 2 (donc $(a_1 + \cdots + a_p)\mathbb{E}\xi_0^2 < 1$). On suppose également que $(X_1, \cdots, X_N)$ est connu. Soit $\widehat{X}_{N+1}$ le prédicteur optimal au sens des moindres carrés. Alors $\widehat{X}_{N+1} = 0$.

Propriété (Cas des chaînes de Markov à espace d'états fini). *Soit $(X_k)_{k\in\mathbb{N}}$ une chaîne de Markov homogène à espace d'états fini $(e_1, \cdots, e_p) \in \mathbb{R}^p$, de matrice de transition $P = (p_{ij})_{1\leq i,j\leq p}$. Soit $\widehat{X}_{N+1}$ le prédicteur optimal au sens des moindres carrés. Alors $\widehat{X}_{N+1} = \mathbb{E}[X_{N+1} \mid X_N] = \sum_{j=1}^{p} p_{X_N,j}e_j$.*

Proposition. *On suppose que X est un processus à temps discret du second ordre. Alors la prédiction* **linéaire** *$\widehat{X}_{N+p}$ optimale au sens de la moyenne quadratique (ou des moindres carrés) est définie par:*

$$\widehat{X}_{N+p} = \widehat{a_1}X_1 + \cdots + \widehat{a_N}X_N + \widehat{b} \text{ avec}$$

$$(\widehat{a_i})_{1\leq i\leq N} = Arg\min_{(a_i)_{1\leq i\leq N}\in\mathbb{R}^N} \left\{\mathbb{E}\left(X_{N+p} - (a_1X_1 + \cdots + a_NX_N + b)\right)^2\right\}.$$

Les coefficients estimés sont obtenus par régression (théorique).

Proposition. *Si X est un processus gaussien, alors la prédiction linéaire par moindres carrés est aussi la prédiction par moindres carrés (donc l'espérance conditionnelle).*

5.2 Prédiction par filtrage exponentiel

L'étude théorique précédente présuppose que l'on ait choisi un modèle de bruit pour prédire. On peut cependant prédire sans avoir besoin de connaître le modèle de bruit. On suppose cependant que la tendance et la saisonnalité sont composées de certaines fonctions. Le lissage exponentiel est une démarche possible pour prédire X_{N+p} lorsque $(X_1, \cdots, X_N)$ est connu.

On suppose donc que la tendance et la saisonnalité s'écrivent sous une forme connue a priori (les f_i, fonctions non constantes, et r sont connues), soit:

$$a(t) = \sum_{i=1}^{k} a_i f_i(t) \quad \text{et} \quad S(t) = \sum_{i=1}^{r} s_i g_i(t) \quad \text{pour } t \in T,$$

avec $g_i(t) = \mathbb{I}_{\{t=i,\ [r]\}}$ sont r-périodiques (on ne suppose pas ici que $\sum_{i=1}^{r} S(i) = 0$).

Notation.

- $X = (X_1, \cdots, X_N)$.
- $F_i = (f_i(1), \cdots, f_i(N))$ *pour* $i = 1, \cdots, k$ *et* $G_i = (g_i(1), \cdots, g_i(N))$ *pour* $i = 1, \cdots, r$.
- $U = (\varepsilon(1), \cdots, \varepsilon(N))$.

Le modèle s'écrit alors vectoriellement:

$$X = \sum_{i=1}^{k} a_i f_i + \sum_{i=1}^{r} s_i g_i + U.$$

Proposition. *On peut estimer les coefficients* (a_i) *et* (s_i) *par une régression linéaire par moindres carrés pondérés qui accorde un poids exponentiellement plus important à* X_N *qu'à* X_M *si* $M < N$. *Pour cela, on minimise la distance dans* $\mathbb{R}^N$ *dépendant d'un paramètre* β*:*

$$\| X - (a_1 F_1 + \cdots + s_r G_r) \|_\beta^2 = {}^t(X - (a_1 F_1 + \cdots + s_r G_r)\Omega^{-1}(\beta)(X - (a_1 F_1 + \cdots + s_r G_r),$$

en notant $\Omega(\beta) = \begin{pmatrix} \beta^N & 0 & 0 & \cdots & 0 \\ 0 & \beta^{N-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$ *et avec* Z *la matrice* $Z = (F_1 \cdots F_k G_1 \cdots G_r)$,

$$\textit{Alors} \quad \widehat{X_{N+p}} = (f_1(N+p), \cdots, g_r(N+p))\,({}^t Z\Omega(\beta)Z)^{-1\,t} Z\Omega(\beta)X.$$

Remarque.

Il faut avoir en tête que si β est proche de 0 alors on prend seulement en compte les toutes dernières valeurs de la série, alors que β proche de 1 fait que l'on considère toutes les valeurs. En pratique, on peut arbitrairement choisir β entre 0.5 et 0.9. Mais le mieux est d'estimer une valeur de β adaptée au jeu de données. Pour ce faire, on utilise les données précédentes $(X_{10}, \cdots, X_{N-p})$ (le 10 est pris ici de manière arbitraire; on pourrait aussi bien choisir $X_{\sqrt{N}}$ ou $X_{N/2}$). Pour chacune de ces valeurs on connait la valeur réellement obtenue à l'horizon p; il suffit donc pour une grille de valeur de β dans $[0, 1]$, de calculer la somme des carrés des écarts entre ces valeurs obtenues et les prédictions faites avec le filtre exponentielle de paramètre β. On choisira $\widehat{\beta}_N$ qui minimise cette somme de distance au carré. Cette technique peut être étendue à des prédictions linéaires: c'est que l'on appelle une prédiction par le filtre de