Séries temporelles 2A

16 décembre 2015

Table des matières

Chapitre 1

Méthodes de base pour l'analyse des séries temporelles

1.1 L'étude des séries temporelles et de leurs composantes

On peut voir une série temporelle comme une suite d'observations répétées d'un même phénomène à des dates différentes (par exemple la température moyenne journalière en un lieu donné, la consommation moyenne en électricité chaque mois en France, le prix du baril de pétrole chaque jour...). Les dates sont souvent équidistantes (séries journalière, mensuelles, trimestrielles ou annuelles) sauf dans quelques cas (par exemple les données journalières en économie ne sont pas toujours disponibles les jours non ouvrables). On représente habituellement une série temporelle $(x_t)_{1\leq t\leq T}$ (t désigne le numéro de l'observation) à l'aide d'un graphique avec en abscisse les dates et en ordonnée les valeurs observées.
Les figures 1.1 et 1.2 représentent deux séries temporelles qui correspondent respectivement au trafic voyageur SNCF et à la consommation mensuelle d'électricité.

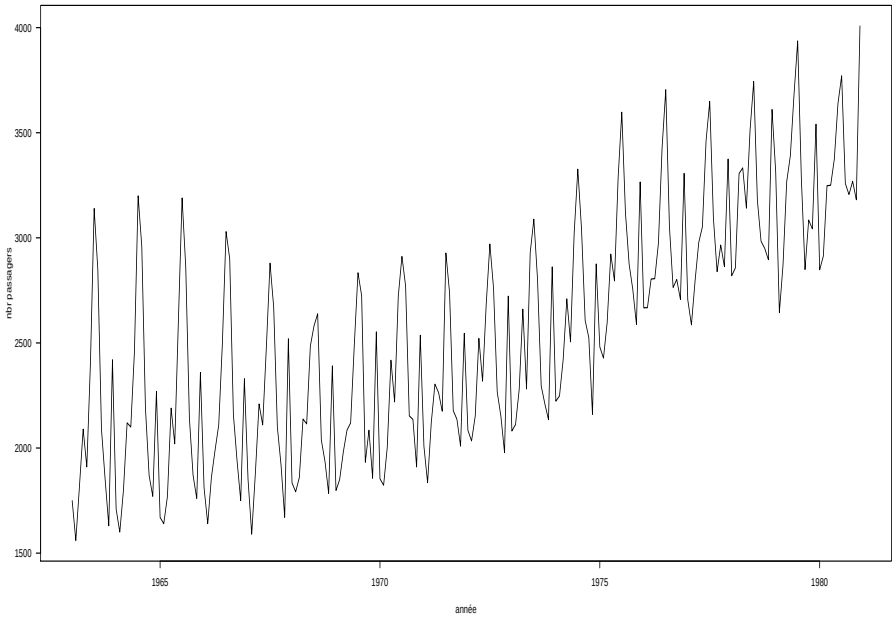


FIGURE 1.1 – Trafic voyageur SNCF en millions de voyageurs kilomètres (observations mensuelles entre 1963 et 1980)

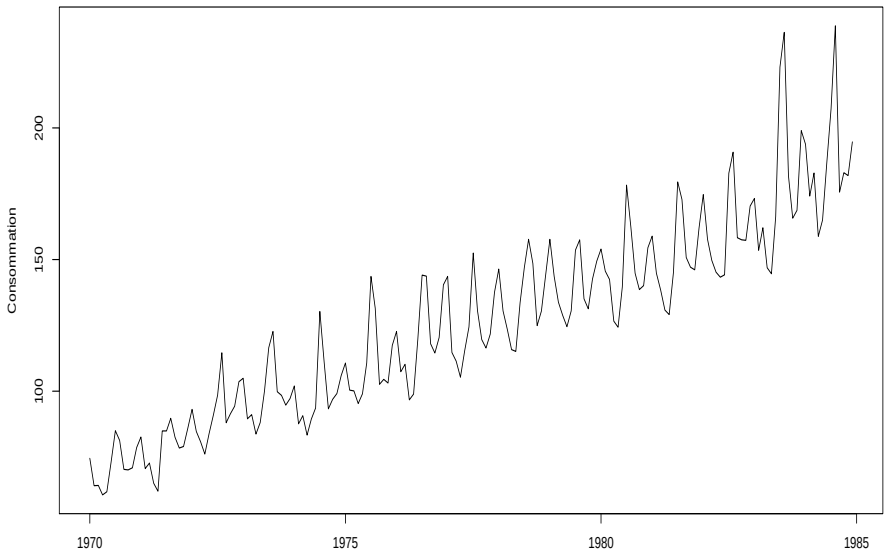


FIGURE 1.2 – Consommation mensuelle d'électricité (en kwh) entre 1970 et 1985

Quelques problèmes posés par les séries temporelles

- Un des problèmes majeurs est celui de la prévision. Peut-on à partir des valeur $x_1, \ldots, x_T$ avoir une idée des valeurs futures $x_{T+1}, x_{T+2} \ldots$? Evidemment on ne peut pas connaitre en pratique la dynamique exacte de la série ; dynamique à travers laquelle les valeurs passées influencent la valeur présente. On pourra toujours tenir compte des valeurs passées ou d'une information auxilliaire (par exemple la consommation d'électricité pourrait être expliquée en tenant compte des températures) mais il existera toujours une composante aléatoire dont il faudra tenir compte d'autant plus que cette composante est en général autocorrélée. Ainsi les valeurs observées $x_1, \ldots, x_T$ seront supposées être des réalisations de variables aléatoires $X_1, \ldots, X_T$ dont il faudra spécifier la dynamique. La prévision pourra alors se faire en estimant la projection de X_{T+1} sur un ensemble de fonctionnelles de $X_1, \ldots, X_T$ (projection linéaire, espérance conditionnelle...). La modélisation doit aussi permettre d'obtenir des intervalles de confiance pour ce type de prévision.
- Résumer la dynamique des valeurs considérées en enlevant les détails de court terme ou les fluctuations saisonnières. On est intéressé ici par l'estimation d'une tendance (on constate à l'oeil une augmentation linéaire pour la consommation d'électricité). Les fluctuations saisonnières captent un comportement qui se répète avec une certaine périodicité (périodicité annuelle très nette pour l'exemple du trafic voyageur).
- L'interprétation du lien entre plusieurs variables (par exemple des variables économiques) ou de l'influence des valeurs passées d'une variable sur sa valeur présente demande de retrancher les composantes tendancielles et saisonnières (sinon on trouverait des corrélations importantes alors qu'un caractère explicatif est peu pertinent).
- Les séries temporelles multivariées (qui correspondent à l'observation simultanée de plusieurs séries temporelles à valeurs réelles) mettent en évidence les effets de corrélation et de causalité entre différentes variables. Il peut alors être intéressant de savoir si les valeurs prises par une variable $x^{(1)}$ sont la conséquence des valeurs prises par la variable $x^{(2)}$ ou le contraire et de regarder les phénomènes d'anticipation entre ces deux variables.
- D'autres problèmes plus spécifiques peuvent aussi se poser : détection de rupture de tendance qui permettent de repérer des changements profonds en macoréconomie, prévision des valeurs extrêmes en climatologie ou en finance, comprendre si les prévisions faites par les entreprises sont en accord avec la conjoncture...

1.2 Modélisations de base pour les séries temporelles

La décomposition additive

Une des décompositions de base est la suivante

$$X_t = m_t + s_t + U_t, \quad 1 \leq t \leq T$$

où

- $(m_t)_t$ est une composante tendancielle déterministe qui donne le comportement de la variable observée sur le long terme (croissance ou décroissance linéaire, quadratique...). Cette composante peut aussi avoir une expression différente pour différentes périodes (affine par mor-

ceaux par exemple). Par exemple, la consommation en électricité représentée Figure 1.2 fait apparaître une tendance affine $m_t = at + b$. Plus généralement, on peut voir cette composante comme une fonction lisse du temps t.

– $(s_t)_t$ est une suite péridiodique qui correspond à une composante saisonnière (par exemple de période 12 pour les séries du trafic voyageur et de la consommation d'électricité, on peut avoir une période 4 pour les séries trimestrielles, 24 pour des séries horaires en météorologie...). Une somme de plusieurs suites de ce type peuvent être pertinentes (par exemple une série de températures horaires observées sur plusieurs années nécessite la prise en compte d'une périodicité quotidienne et annuelle).
– $(U_t)_t$ représente une composante irrégulière et aléatoire, le plus souvent de faible amplitude par rapport à la composante saisonnière mais importante en pratique puisque ce terme d'erreur sera le plus souvent autocorrélé (c'est à dire que la covariance entre U_t et U_{t+h} sera non nulle). Nous verrons quels types de modèles peuvent être utilisés pour l'étude de cette composante.

La décomposition multiplicative

On décompose la série temporelle sous la forme $X_t = m_t s_t U_t, \quad 1 \leq t \leq T$. Les composantes $(m_t)_t$ et $(s_t)_t$ sont de la même forme que pour le modèle additif et la composante irrégulière $(U_t)_t$ a pour moyenne 1. Par une transformation logarithmique, on se ramène à une décomposition additive. Cette décomposition multiplicative est intéressante lorsqu'on observe une variation linéaire des effets saisonniers comme le montre la Figure 1.2. On peut aussi combiner l' approche additive et l'approche multiplicative en décomposant $X_t = m_t s_t + U_t$.

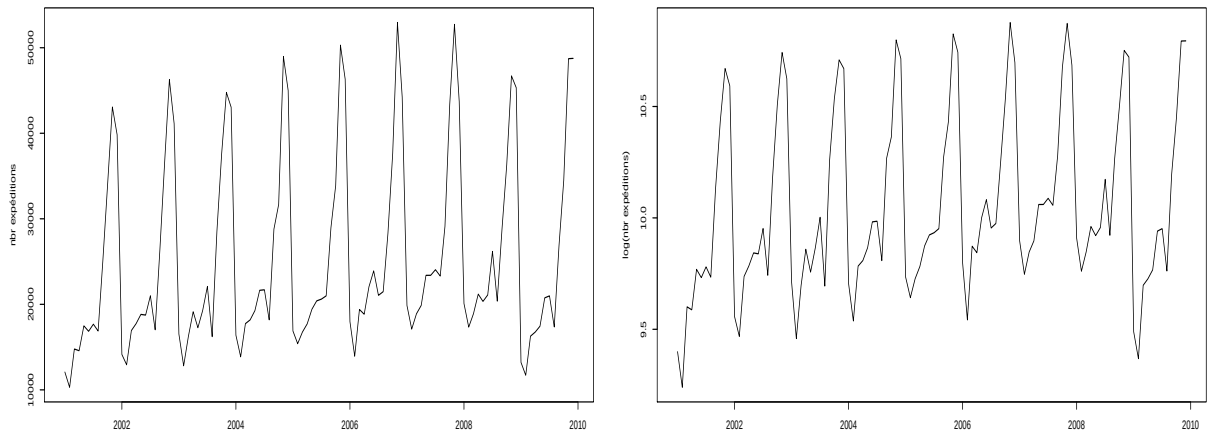


FIGURE 1.3 – Expédition mensuelle de champagne en milliers de bouteilles entre 2001 et 2010 (série intiale à gauche et son logarithme à droite)

Dynamique autorégressive

La dynamique (aléatoire) de la série temporelle est basée sur des équations récursives du type

$$X_t = f\left(X_{t-1}, \ldots, X_{t-p}, \varepsilon_t\right), \quad p+1 \leq t \leq T \tag{1.1}$$

où f est une fonction mesurable qui dépend d'un paramètre inconnu et $(\varepsilon_t)_t$ est une perturbation aléatoire non observée. Comme nous le verrons dans ce cours, l'utilisation de fonctions f linéaires est souvent pertinente pour modéliser la dynamique de la composante irrégulière $(U_t)_t$ des décompositions additives et multiplicatives. Cependant, certaines situations sortiront de ce cadre.

- En pratique, les modélisations de type (1.1) font parfois intervenir d'autres variables observées au temps t (information auxilliaire) $Z_t^{(1)}, \ldots, Z_t^{(k)}$ dans la fonction f.
- Des modèles de type (1.1) avec f non linéaire sont aussi souvent utilisés pour étudier la dynamique des rendements des séries financières (voir la série de l'indice du CAC40, Figure 1.4). L'étude de ce type de série ne sera pas abordée dans ce cours.

FIGURE 1.4 – Indice boursier du CAC40 du 01/01/2006 au 10/10/2014

1.3 Désaisonnaliser par la méthode de la régression linéaire

On supposera dans cette partie que

$$X_t = a + bt + s_t + U_t, \quad t = 1, 2, \ldots, T,$$

où s est un suite périodique de période k connue (saisonnalité) et $(U_t)_{1 \leq t \leq T}$ correspond à un bruit. Pour simplifier, on supposera que $U_1, \ldots, U_T$ sont i.i.d bien que cette hypothèse sera relâchée par

la suite. Il est possible de généraliser l'approche décrite dans cette section à des tendances $(m_t)_t$ plus complexes (par exemple polynomiales avec $m_t = \sum_{j=0}^{\ell} a_j t^j$). Nous allons écrire un modèle linéaire qui permet d'estimer conjointement les deux coefficients de tendance (a, b) ainsi que la saisonnalité (c'est à dire les valeurs $s_1, s_2, \ldots, s_k$). Pour simplifier, nous supposerons que $T = Nk$ pour un entier N (la taille de l'échantillon est proportionnelle à la période). Pour $1 \leq j \leq k$, soit e^j le vecteur de $\mathbb{R}^T$ composé de 0 sauf pour les coordonnées n°$j + \ell k$ qui valent 1. En posant $\alpha_j = s_j$, on a

$$s_t = \sum_{j=1}^{k} \alpha_j e_t^j.$$

En posant $Y = (X_1, X_2, \ldots, X_T)'$, $z = (1, 2, \ldots, T)$ et en notant $\mathbb{1}$ le vecteur formé uniquement de 1, on a l'écriture vectorielle

$$Y = a\mathbb{1} + bz + \sum_{j=1}^{k} \alpha_j e^j + U.$$

Mais il y a trop de paramètres, la somme des vecteurs e^j coïncide avec le vecteur $\mathbb{1}$. Les régresseurs sont donc linéairement dépendants. Pour que le modèle soit identifiable, on supposera que $\sum_{j=1}^{k} \alpha_j = 0$ ce qui revient à imposer que

$$s_{t+1} + s_{t+1} + \ldots + s_{t+k} = 0, \quad t \in \mathbb{Z}, \tag{1.2}$$

au niveau du facteur saisonnier. En remplaçant $\mathbb{1}$ par $\sum_{j=1}^{k} e^j$ et en posant $\beta_j = a + \alpha_j$, on obtient

$$Y = bz + \sum_{j=1}^{k} \beta_j e^j + \varepsilon = bz + E\beta + \varepsilon,$$

où E est la matrice dont les vecteurs colonnes sont $e^1, \ldots, e^k$ et $\beta = (\beta_1, \ldots, \beta_k)'$. Les relations entre les différents coefficients sont

$$a = \frac{1}{k} \sum_{j=1}^{k} \beta_j, \quad \alpha_j = \beta_j - a.$$

Les vecteurs $z, e^1, \ldots, e^k$ sont libres et on peut estimer b, β par moindres carrés sans contrainte. On pose

$$\left(\hat{b}, \hat{\beta}\right) = \arg \min_{b \in \mathbb{R}, \beta \in \mathbb{R}^k} \|Y - bz - E\beta\|^2,$$

où $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^T$.
On définit alors

$$\hat{a} = \frac{1}{k} \sum_{j=1}^{k} \hat{\beta}_j, \quad \hat{\alpha}_j = \hat{\beta}_j - \hat{a}.$$

Proposition 1 *Posons pour $n = 1, \ldots, N$ et $j = 1, \ldots, k$,*

$$\widetilde{x_n} = \frac{1}{k} \sum_{j=1}^{k} X_{(n-1)k+j}, \quad \bar{x}_j = \frac{1}{N} \sum_{n=1}^{N} X_{(n-1)k+j}.$$

Si $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} X_t$*, on a les formules*

$$\hat{b} = \frac{12}{k}\frac{\sum_{n=1}^{N} n\widetilde{x_n} - \frac{N(N+1)}{2}\bar{x}}{N(N^2-1)},$$

$$\hat{a} = \bar{x} - \frac{Nk+1}{2}\hat{b},$$

$$\hat{\alpha}_j = \bar{x}_j - \bar{x} + \hat{b}\left(\frac{k+1}{2} - j\right).$$

Preuve. Le couple $\left(\hat{b}, \hat{\beta}\right)$ vérifie les deux équations

$$z'z\hat{b} + z'E\hat{\beta} = z'Y, \quad E'z\hat{b} + E'E\hat{\beta} = E'Y.$$

On tire de la deuxième relation l'égalité

$$\hat{\beta} = (E'E)^{-1}\left(E'Y - E'z\hat{b}\right). \tag{1.3}$$

En reportant dans la première équation, on obtient

$$\hat{b} = \left(z'z - z'E(E'E)^{-1}E'z\right)^{-1}\left(z'Y - z'E(E'E)^{-1}E'Y\right). \tag{1.4}$$

On a

$$\begin{aligned} z'Y &= \sum_{t=1}^{T} tX_t \\ &= \sum_{n=1}^{N}\sum_{j=1}^{k}((n-1)k+j)\,X_{(n-1)k+j} \\ &= N\sum_{j=1}^{k} j\bar{x}_j + k^2\sum_{n=1}^{N} n\widetilde{x_n} - kT\bar{x}. \end{aligned}$$

A partir des formules

$$z'z = \frac{1}{6}T(T+1)(2T+1), \quad E'Y = \begin{pmatrix} N\bar{x}_1 \\ \vdots \\ N\bar{x}_k \end{pmatrix}, \quad E'z = \left(Nj + \frac{k}{2}N(N-1)\right)_{1\le j\le k}, \quad E'E = NI_k$$

et de quelques calculs, l'expression (1.4) devient

$$\hat{b} = \frac{12}{k}\frac{\sum_{n=1}^{N} n\widetilde{x_n} - \frac{N(N+1)}{2}\bar{x}}{N(N^2-1)}.$$

En reportant cette expression dans (1.3), les expressions annoncées pour les $\hat{\beta}_j$, $\hat{a}$ et $\hat{\alpha}_j$ se déduisent aisément.□

Remarque. Nous verrons dans la section suivante une autre façon d'écrire la saisonnalité $(s_t)_t$ sous la contrainte (1.2), à l'aide des fonctions trigonométriques.

1.3.1 Application au trafic SNCF

Nous considérons ici une application à l'étude de la série du trafic voyageur de la SNCF. En effectuant une régression linéaire à partir d'une saisonnalité mensuelle (donc $k = 12$) et d'une tendance affine, nous obtenons

$$\begin{aligned}\hat{y}_t &= 6.44 \times t + 1531.43 \times S_t^1 + 1431.27 \times S_t^2 + 1633.16 \times S_t^3 + 1872.55 \times S_t^4 + 1799.89 \times S_t^5 \\ &+ 2192.22 \times S_t^6 + 2547.17 \times S_t^7 + 2219.68 \times S_t^8 + 1711.07 \times S_t^9 + 1637.35 \times S_t^10 \\ &+ 1477.19 \times S_t^{11} + 2126.63 \times S_t^{12}.\end{aligned}$$

Le coefficient de détermination vaut $R^2 \approx 0.99$. Toutefois les résidus présentent des aspects atypiques (résidus négatifs au milieu de la série, convexité du graphe). Une tendance quadratique pourrait être considérée pour la série initiale.

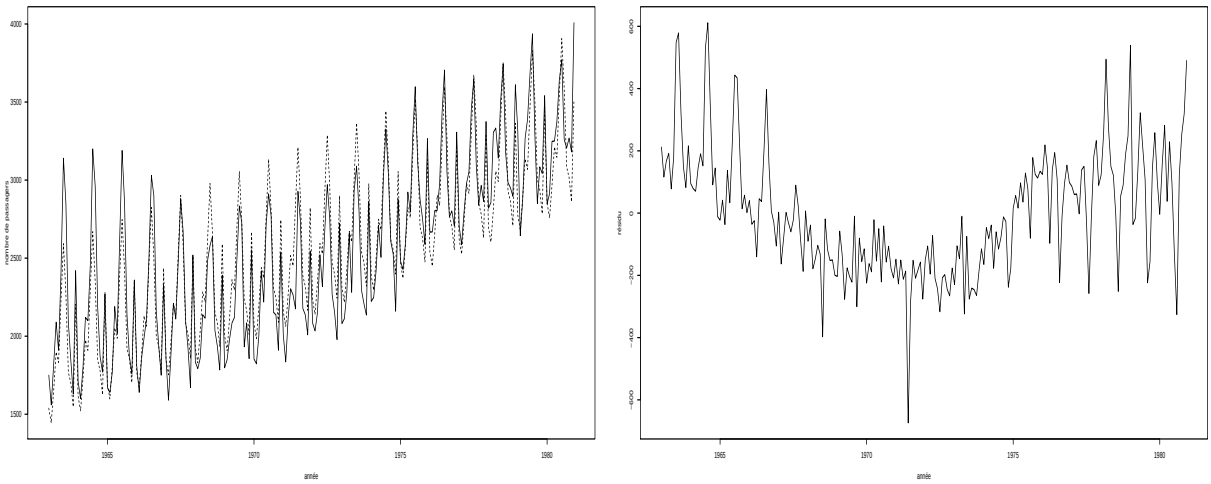


FIGURE 1.5 – Comparaison de la série avec la série des valeurs ajustées (à gauche, les valeurs ajustées sont représentées en pointillés) et graphe des résidus (à droite)

L'inclusion de la variable t^2 dans la régression permet de corriger ces problèmes. Mais le résultat est loin d'être parfait. Au vu des graphes des résidus (et même des données initiales), on pourrait appliquer une transformation logarithmique (utile aussi pour stabiliser la variance). Mais la stabilité des coefficients de saisonnalité est aussi mise en doute (les résidus ne montrent pas une absorption complète de la composante saisonnière). On pourrait alors considérer uniquement la deuxième partie de la série (à partir de 1971, le comportement de la série semble plus homogène). Lorsque la fonction $(s_t)_t$ n'est pas tout à fait périodique, la méthode des moyennes mobiles (présentée dans la section suivante) est une alternative intéressante. Un exemple pour lequel la

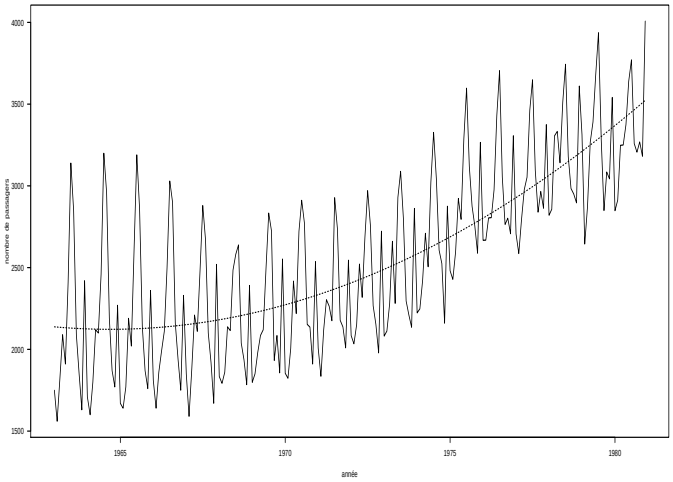


FIGURE 1.7 – Tendance estimée pour la deuxième régression (en pointillé)

désaisonnalisation par régression donne de meilleurs résultats que pour le trafic SNCF sera étudié en TD.

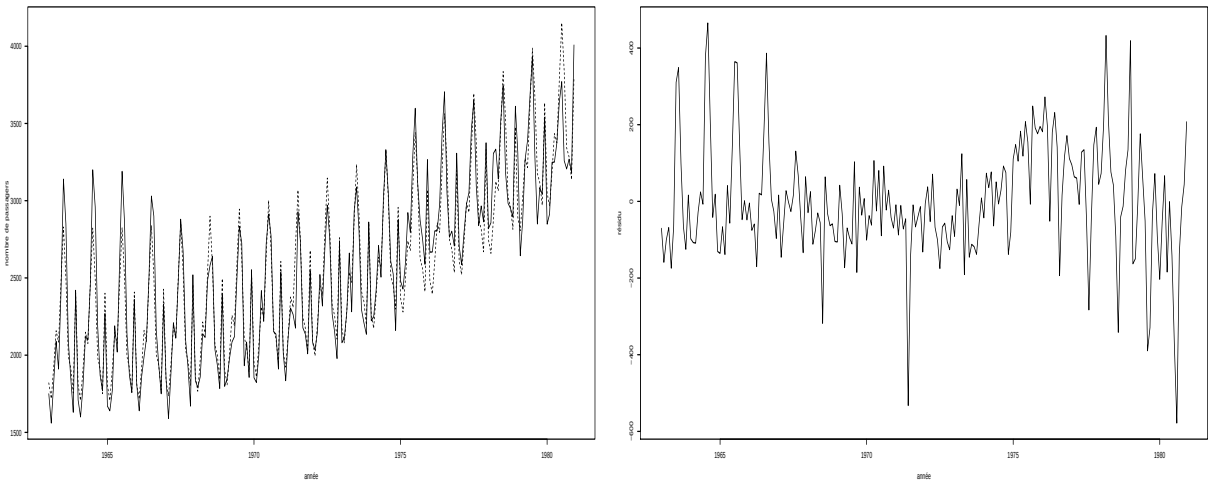


FIGURE 1.6 – Incorporation d'une tendance quadratique. Comparaison de la série avec la série des valeurs ajustées (à gauche, les valeurs ajustées sont représentées en pointillés) et résidus (à droite)

1.4 Désaisonnalisation à partir de moyennes mobiles

Dans cette section, on considère une décomposition additive de la série temporelle,

$$X_t = m_t + s_t + u_t, \quad 1 \leq t \leq T.$$

On supposera dorénavant que la composante saisonnière $(s_t)_t$ vérifie la condition (1.2) c'est à dire que sa moyenne est nulle sur une période. Supposons que l'on veuille estimer une des composantes de la série, par exemple la tendance. Un moyen est d'appliquer une transformation linéaire f à la série qui conserve la tendance T, élimine la saisonnalité s et ce sans trop amplifier la partie stochastique u. Les transformations utilisées sont les moyennes mobiles et l'utilisation de ces transformations peut s'interpréter comme des régressions locales sur des constantes. Un des intérêts majeurs de cette méthode est d'être plus robuste aux changements de régime (e.g des ruptures de tendances).

1.4.1 Moyennes mobiles

Définition 1 *Une moyenne mobile M est un opérateur linéaire du type $M = \sum_{i=-p_1}^{p_2} \theta_i B^{-i}$ où*
- *p_1 et p_2 sont deux entiers positifs,*
- *$\theta_{-p_1}, \ldots, \theta_1, \theta_0, \theta_1, \ldots, \theta_{p_2}$ sont des nombres réels (les coefficients de M).*
- *B est l'opérateur retard (backward) qui a une suite de nombres réels $(x_t)_{t\in\mathbb{Z}}$ associe la suite $(x_{t-1})_{t\in\mathbb{Z}}$ et B^{-1} désigne son inverse.*

L'ordre de la moyenne mobile M est le nombre de coefficients $p_1 + p_2 + 1$ qui la compose.

On a donc pour une suite $(x_t)_{t\in\mathbb{Z}}$,

$$M(x)_t = \sum_{i=-p_1}^{p_2} \theta_i x_{t+i}.$$

Pour la série temporelle $(X_t)_{1\leq t\leq T}$, on ne peut définir $M(X)_t$ que si $p_1 + 1 \leq t \leq T - p_2$. Dans un premier temps nous occulterons les problèmes de bord qui empêchent de définir $M(X)_t$ pour $1 \leq t \leq p_1$ et $T - p_2 + 1 \leq t \leq T$ (en supposant que la série initiale est définie sur $\mathbb{Z}$ par exemple).

Définition 2 *Soit $M = \sum_{i=-p_1}^{p_2} \theta_i B^{-i}$ une moyenne mobile. On dit que M est centrée si $p_2 = p_1 = p$. De plus, si M est centrée, on dira que M est symétrique si $\theta_i = \theta_{-i}$.*

Par exemple la moyenne mobile M telle que $M(X)_t = 2X_{t-1} + X_t + 2X_{t+1}$ est centrée symétrique. Notons qu'une moyenne mobile centrée est automatiquement d'ordre impair.

Exercice 1 Si M_1 et M_2 sont des moyennes mobiles, la moyenne mobile composée, notée M_2M_1, est définie par $M_2M_1(X) = M_2(M_1(X))$. Vérifier que $M_2M_1 = M_1M_2$ (la composition est commutative). Vérifier également que l'ensemble des moyennes mobiles, des moyennes mobiles centrées ou des moyennes mobiles symétriques est stable par composition.

Définition 3 *Soit M une moyenne mobile.*

1. *On dit M conserve $(X_t)_t$ (ou que $(X_t)_t$ est invariante par M) si $M(X) = X$.*
2. *On dit que qu'une série (X_t) est absorbée par M si $M(X) = 0$.*

Exercice 2 Montrer qu'une moyenne mobile M centrée conserve les suites constantes ssi la somme de ses coefficients valent 1. Si M est symétrique et conserve les constantes, montrer que M conserve les polynômes de degré 1 (i.e $X_t = at + b$).

Exercice 3 Si I est l'application identité (i.e $I(X) = X$) soit $M = (I-B)^p$ (on itère p fois l'opérateur $I - B$). Décrire $M(X)_t$ pour $p = 1, 2, 3$. Montrer que M transforme un polynôme de degré p en une constante. On peut ainsi éliminer les tendances de bas degré.

Exercice 4 Montrer qu'une moyenne mobile centrée conserve tout polynôme de degré d si et seulement si ses coefficients vérifient les équations

$$\sum_{i=-p}^{p} \theta_i = 1, \quad \sum_{i=-p}^{p} i^\ell \theta_i = 0, \quad 1 \leq \ell \leq d.$$

Elimination de la tendance et de la saisonnalité

Si la série se décompose sous la forme $X_t = m_t + s_t + U_t$ avec $(m_t)_t$ polynomial et $(s_t)_t$ de période k, on peut toujours éliminer la tendance et la saisonnalité en appliquant une moyenne mobile du type $M = \left(I - B^k\right)(I - B)^d$. C'est le moyen le plus simple de se ramener à une composante résiduelle qui ne présente ni périodicité, ni tendance.

Lien avec la régression

Si M est une moyenne mobile centrée et à coefficients positifs et dont la somme vaut 1, alors $M(X)_t$ minimise la fonction $a \mapsto \sum_{i=-p}^{p} \theta_i (X_{t+i} - a)^2$. On parle de régression locale (au point t) sur une constante puisqu'on n'utilise que quelques valeurs de la série (dont les indices sont proches de t) dans le critère des moindres carrés. En général, l'application d'une moyenne mobile a un effet "lissant" puisqu'elle gomme les irrégularités de la série. Pour une série sans saisonnalité $X_t = m_t + u_t$, $M(X)_t$ peut s'interprêter comme une estimation locale de la tendance (on fait alors en sorte que M préserve la forme de la tendance, par exemple en imposant que M laisse invariant tout polynôme de degré 1). Par rapport à la méthode de régression (non locale) vue à la section précédente, l'utilisation des moyennes mobiles rend les estimations plus robustes à des changements de régime tels que par exemple des ruptures de tendance. Une comparaison des deux méthodes est illustrée Figure 1.8 à partir de la série $(X_t)_{1 \leq t \leq 50}$ définie par

$$X_t = \begin{cases} t + U_t \text{ si } 1 \leq t \leq 25 \\ 2t - 25 + U_t \text{ si } 26 \leq t \leq 50 \end{cases}$$

où les U_t sont i.i.d de loi gaussienne $\mathcal{N}(0, 4)$.

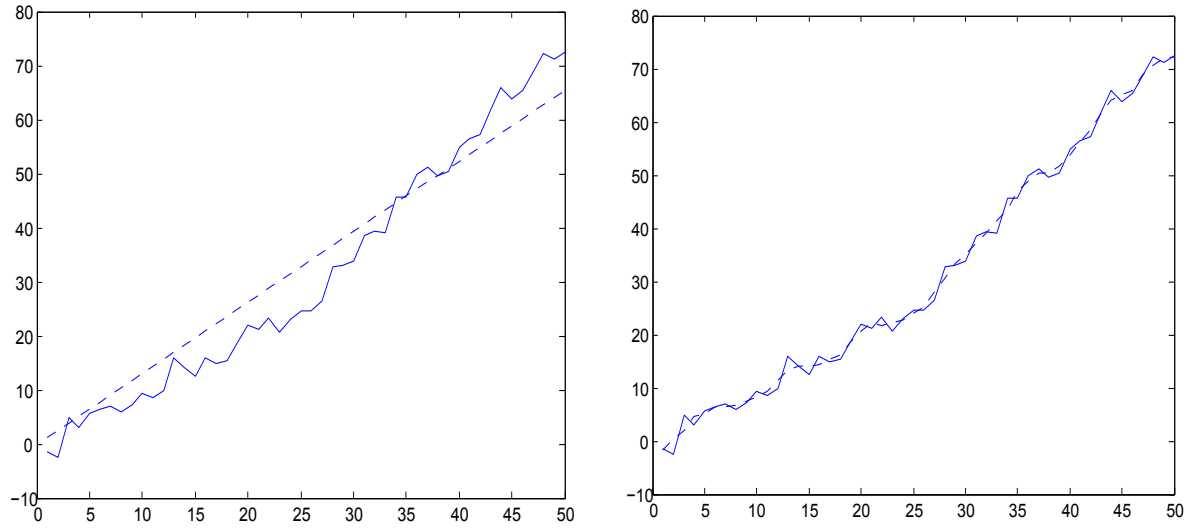


FIGURE 1.8 – Estimation de la tendance par régression (en ignorant la rupture, graphe de gauche) puis par l'application de $M(X)_t = \frac{X_{t-1}+X_t+X_{t+1}}{3}$ qui conserve les droites (la tendance estimée figure en pointillés)

Typiquement, on essaie d'utiliser les moyennes mobiles de la façon suivante. On applique d'abord une moyenne mobile qui laisse la tendance invariante et qui absorbe la saisonnalité. Ceci permet d'avoir une première estimation de la tendance. Ensuite, on retranche cette estimation à la série initiale pour pouvoir estimer la saisonnalité à l'aide d'une autre moyenne mobile (qui conserve la saisonnalité et atténue le bruit). On retranche l'estimation de la saisonnalité à la série initiale pour estimer plus précisément la tendance et ainsi de suite. Notons que l'application d'une moyenne mobile transforme le bruit u_t en $\sum_{i=-p_1}^{p_2} \theta_i u_{t+i}$. Il est alors impératif d'utiliser des moyennes mobiles qui atténue le bruit. Un des critères importants est alors le rapport $\frac{\text{Var}\left(\sum_{i=-p_1}^{p_2} \theta_i u_i\right)}{\text{Var}\,(u_1)}$. Lorsque les variables u_t sont supposées décorrélées deux à deux et de même variance, ce rapport vaut $\sum_{i=-p_1}^{p_2} \theta_i^2$. Pour choisir une moyenne mobile ayant des propriétés données, on cherche souvent à minimiser ce rapport.

Définition 4 *Le rapport de variance d'un moyenne mobile M est la quantité $\sum_{i=-p_1}^{p_2} \theta_i^2$.*

1.4.2 Suites récurrentes linéaires et séries absorbées par une moyenne mobile

Une suite récurrente linéaire est définie par des équations du type

$$y_t = \sum_{j=1}^{p} \alpha_j y_{t-j}, \quad t \in \mathbb{Z}, \tag{1.5}$$

où p est un entier et $\alpha_1, \ldots, \alpha_p$ sont des nombres réels. On supposera $\alpha_p \neq 0$. L'ensemble des solutions de l'équation 1.5 constitue un espace vectoriel de dimension p. Si on cherche des so-

lutions de la forme $y_t = \lambda^t$ où λ est un nombre complexe non nul. On obtient une solution si et seulement si λ est une racine du polynôme $\mathcal{P}(\lambda) = \lambda^p - \sum_{j=1}^{p} \alpha_j \lambda^{p-j}$. Ce polynôme admet k racines complexes. Si $\lambda = r\exp(i\phi)$ est racine simple, on obtient deux solutions réelles $y_t = r^t \cos(\phi t)$ et $y_t = r^t \sin(\phi t)$. Si la racine a une multiplicité $m \geq 2$, on obtient des solutions complexes $y_t = t^\ell \lambda^t$ avec $\ell = 0, 1, \ldots, m-1$. Les solutions réelles associées sont alors $y_t = t^\ell r^t \cos(\phi t)$ et $y_t = t^\ell r^t \sin(\phi t)$ pour $\ell = 0, 1, \ldots, m-1$. La solution générale de l'équation (1.5) s'écrit comme combinaison linéaire de ces solutions de base.

Exemple : l'expression de la saisonnalité Une saisonnalité de période k et de moyenne nulle sur une période vérifie l'équation

$$\sum_{j=1}^{k} s_{t+j} = 0, \quad t \in \mathbb{Z}.$$

On a alors

$$\mathcal{P}(\lambda) = \lambda^{k-1} + \lambda^{k-2} + \cdots + \lambda + 1.$$

Comme $\lambda = 1$ n'est pas racine et que $\mathcal{P}(\lambda) = \frac{1-\lambda^k}{1-\lambda}$ pour $\lambda \neq 1$, les racines de $\mathcal{P}$ vérifie $\lambda^k = 1$. On a donc k racines de l'unité $\lambda = \exp\left(i\frac{2\pi\ell}{k}\right)$ pour $\ell = 1, \ldots, k$.

- Lorsque k est impair, $k = 2k' + 1$, une base de solutions est constituée de $y_t = \cos\left(\frac{2\pi\ell}{k}t\right)$ et $y_t = \sin\left(\frac{2\pi\ell}{k}t\right)$ pour $\ell = 1, \ldots, k'$.
- Lorsque k est pair, $k = 2k'$, une base de solutions est constituée de $y_t = \cos\left(\frac{2\pi\ell}{k}t\right)$ pour $\ell = 1, \ldots, k'$ et $y_t = \sin\left(\frac{2\pi\ell}{k}t\right)$ pour $\ell = 1, \ldots, k' - 1$.

Application aux moyennes mobiles

Définition 5 *On appelle polynôme cararactéristique de la moyenne mobile $M = \sum_{i=-p_1}^{p_2} \theta_i B^{-i}$, le polynôme*

$$\mathcal{P}(\lambda) = \theta_{-p_1} + \theta_{-p_1+1}\lambda + \cdots + \theta_{p_2}\lambda^{p_1+p_2}.$$

On a donc $M = B^{p_1}\mathcal{P}\left(B^{-1}\right)$.

Soit M une moyenne mobile centrée. Alors une série $(y_t)_t$ est absorbée par M ssi $(y_t)_t$ vérifie l'équation de récurrence

$$\sum_{i=-p}^{p} \theta_i y_{t+i} = 0.$$

Lorsque tous les θ_i sont non nuls, l'ensemble des suites solutions est un espace vectoriel de dimension $2p$. Le polynôme caractéristique associé à M est

$$\mathcal{P}(\lambda) = \theta_p \lambda^{2p} + \theta_{p-1}\lambda^{2p-1} + \cdots + \theta_{-p+1}\lambda + \theta_{-p}.$$

Toute série aborbée par M s'écrit donc comme combinaison linéaire de séries du type $y_t = t^\ell \rho^t \cos(\phi t)$ ou $y_t = t^\ell \rho^t \sin(\phi t)$, $0 \leq \ell \leq m(\lambda) - 1$, oú $\lambda = r\exp(i\phi)$ est racine de $\mathcal{P}$ avec multiplicité $m(\lambda)$.

Par exemple, toute moyenne mobile telle que $\theta_i = \theta$ pour $-p \leq i \leq p$ aborbe les suites $(y_t)_t$ telles que

$$\sum_{i=1}^{2p+1} y_{t+i} = 0.$$

Ceci signifie que M absorbe toutes les saisonnalités de période impaire $k = 2p + 1$ et de moyenne nulle sur une période. On a vu que les solutions s'écrivait à partir de cosinus et de sinus.

Autre exemple : une moyenne mobile M absorbe les polynômes de degré d ssi elle absorbe les séries $\left(t^\ell\right)_t$ pour $\ell = 0, \ldots, d$. D'après la caractérisation des solutions, ceci signifie que 1 est racine de $\mathcal{P}$ avec une multiplicité $\geq d + 1$ et donc que $\mathcal{P}$ est divisible par $(1 - \lambda)^{d+1}$.

1.4.3 Les moyennes mobiles arithmétiques

Si on s'intéresse aux moyennes mobiles centrées qui conservent les constantes (i.e $\sum_{i=-p}^{p} \theta_i = 1$) et qui minimisent le rapport de réduction de la variance $\sum_{i=-p}^{p} \theta_i^2$, on trouve la solution $\theta_i = \frac{1}{2p+1}$ et on a

$$M(X)_t = \frac{1}{2p+1}\left(X_{t-p} + \cdots + X_{t+p}\right).$$

D'après la sous-section précédente, on a le résultat suivant.

Proposition 2 *Une moyenne arithmétique d'ordre* $2p + 1$ *absorbe les saisonnalités de périodes* $k = 2p + 1$ *qui sont nulles en moyenne et préserve les polynômes de degré* 1.

Toutefois, en pratique, il est souvent nécessaire de considérer des saisonnalités de période paire (données trimestrielles ou mensuelles). Pour cela, si la période est $k = 2p$, on utilise la moyenne mobile $M = \frac{1}{2}M_1 + \frac{1}{2}M_2$ avec

$$M_1(X)_t = \frac{1}{2p}\left(X_{t-p} + X_{t-p+1} + \cdots + X_{t+p-1}\right), \quad M_2(X)_t = \frac{1}{2p}\left(X_{t-p+1} + X_{t-p+1} + \cdots + X_{t+p}\right).$$

Les moyennes mobiles M_1 et M_2 absorbent les saisonnalités de période $2p$ et de moyenne nulle mais elles ne sont pas symétriques. En revanche

$$M(X)_t = \frac{1}{2p}\left(\frac{1}{2}X_{t-p} + X_{t-p+1} + \cdots + X_{t+p-1} + \frac{1}{2}X_{t+p}\right)$$

et M est symétrique, annule les saisonnalités de période $2p$ et de moyenne nulle et préserve les polynômes de degré 1. On peut montrer que cette moyenne mobile minimise le rapport de variance parmi les moyennes mobiles centrées d'ordre $2p + 1$ qui absorbent les saisonnalités de période $2p$ et dont la somme des coefficients vaut 1. Les séries trimestrielles (resp. mensuelles) correspondent au cas $p = 2$ (resp. $p = 6$) et la moyenne mobile correspondante est notée $M_{2,4}$ (resp. $M_{2,12}$).

Estimation de la tendance et de la saisonnalité

Une façon simple d'estimer une tendance linéaire (ou même linéaire par morceaux) est d'appliquer une moyenne mobile arithmétique qui annule la saisonnalité. On obtient une estimation $\hat{m}_t$ de la tendance et $X_t - \hat{m}_t$ est une estimation de $s_t + U_t$. On définit ensuite pour $1 \leq \ell \leq k$, w_ℓ comme la moyenne sur j des $X_{\ell+kj} - \hat{m}_{\ell+kj}$ et on pose $\hat{s}_\ell = w_\ell - \frac{1}{k}\sum_{i=1}^{k} w_i$ (pour que la moyenne sur une période soit bien nulle). On peut alors appliquer ensuite une moyenne mobile à $X_t - \hat{s}_t$ pour estimer plus précisément la tendance puis de nouveau la saisonnalité. C'est ce type d'idées qui est utilisée dans la méthode $X11$ étudiée plus loin.

1.4.4 Moyenne mobile d'Henderson

Une des propriétés souhaitées pour une moyenne mobile est la réduction des oscillations de la série initiale. Soit par exemple la série $X_t = \mathbb{1}_{t=0}$. Si M est une moyenne mobile centrée d'ordre $2p+1$, on voit que $M(X)_t = \theta_{-t}\mathbb{1}_{-p\leq t\leq p}$. La régularité de $t \mapsto M(X)_t$ est équivalente à la régularité des coefficients dans ce cas.

Henderson a proposé de trouver des moyennes mobiles M d'ordre $2p+1$, qui conservent les polynômes de degré 2 et qui minimise la quantité

$$C(M) = \sum_{i=-p+3}^{p} \left((I-B)^3(\theta)_i\right)^2.$$

L'annulation de $(I-B)^3(\theta)$ a lieu lorsque les points (i, θ_i) sont situés sur une parabole. On peut alors interpréter $C(M)$ comme un écart des coefficients de M avec la forme parabolique et donc comme un indicateur de régularité pour $i \mapsto \theta_i$. Il faut alors minimiser $C(M)$ sous les contraintes

$$\sum_{i=-p}^{p} \theta_i = 1, \quad \sum_{i=-p}^{p} i\theta_i = 0, \quad \sum_{i=-p}^{p} i^2\theta_i = 0.$$

On peut montrer que ce problème se résout explicitement.

1.4.5 Autres moyennes mobiles

Pour estimer la saisonnalité lorsque une estimation de la tendance a été retranchée à la série, on utilise des moyennes symétriques du type 3×3 ou 3×5. Les moyennes mobiles du type 3×3 sont composées des coefficients $\{1,2,3,2,1\}/9$ mais entre lesquels on intercale des coefficients nuls (11 dans le cas de série mensuelles et 3 dans pour les séries trimestrielles). L'idée est calculer une moyenne pour chaque mois ou chaque trimestre de l'année. Les moyennes mobiles de type 3×5 sont définies de la même façon, mais à partir des coefficients $\{1,2,3,3,3,2,1\}/15$. L'application de ces moyennes mobiles est suivie par l'application de la moyenne mobile $I - M$ où M est la moyenne arithmétique associée à période k, ce qui permet de rencentrer les coefficients de saisonnalité estimés (et donc de respecter la contrainte (1.2)).

1.4.6 Procédures disponibles et traitement des extrémités de la série

Les procédures $X11$ et $X12$ (implémentées notamment sous SAS) sont basées sur des applications successives de moyennes mobiles avec au final une estimation des trois termes m_t, s_t et U_t (ou des trois facteurs puisqu'il existe une version multiplicative). Les moyennes mobiles utilisées correspondent à celles vues aux sous-sections précédentes (arithmétique, de Henderson et 3×3 ou 3×5). On pourra consulter [4] pour un exposé détaillé de ces procédures. Le principe de base de la méthode $X11$ pour les séries mensuelles est la suivante.

1. On applique la moyenne mobile $M_{2,12}$ à la série initiale, ce qui permet d'estimer la tendance par $\hat{m}_t^{(1)} = M_{2,12}(X)_t$.

2. La saisonnalité est alors estimée par $\hat{s}_t^{(1)} = X_t - \hat{m}_t^{(1)}$.

3. Ensuite, on affine l'estimation de la saisonnalité à l'aide de la moyenne mobile 3×3 (notée $M_{3,3}$) et on recentre les coefficients estimés. On obtient alors $\hat{s}_t^{(2)} = (I - M_{2,12})\, M_{3,3}\left(\hat{s}^{(1)}\right)_t$.

4. Une première estimation de la série corrigée des valeurs saisonnières est donnée par $\hat{X}_t^{(1)} = X_t - \hat{s}_t^{(2)}$.

5. On affine l'estimation de la tendance en appliquant la moyenne mobile de Henderson sur 13 termes (notée M_{13} dans la suite). On a donc $\hat{m}_t^{(2)} = M_{13}\left(\hat{X}^{(1)}\right)_t$.

6. La composante saisonnière est de nouveau estimée : $\hat{s}_t^{(3)} = X_t - \hat{m}_t^{(2)}$.

7. L'application de la moyenne mobile 3×5 et le recentrage des coefficients permettent enfin de fournir un nouvelle estimation de la saisonnalité $\hat{s}_t^{(3)}$.

8. L'estimation finale de la tendance est donnée par $\hat{m}_t^{(3)} = X_t - \hat{s}_t^{(3)}$.

On voit alors que l'estimation de la tendance et de la saisonnalité correspond à l'application d'une moyenne mobile (par composition de moyennes mobiles successives) d'ordre 169. Mais l'application d'une telle moyenne mobile ne peut se faire que si $85 \leq t \leq T - 84$ ce qui pose un gros problème pour connaitre la série corrigée des variations saisonnières au début et à la fin. On peut considérer alors deux solutions pour y remédier. La première est d'utiliser des moyennes mobiles asymétriques au bord (c'est ce que fait la méthode $X11$ initiale). La seconde est de prolonger la série sur les bords à l'aide d'une méthode de prévision puis d'augmenter la série avec ces prévisions. C'est cette deuxième solution qui est utilisée dans les méthodes $X11$ ARIMA et $X12$ ARIMA. Les modèles ARIMA seront étudiés plus loin dans ce cours. Notons aussi que ces procédures ont des options permettant la correction des valeurs aberrantes ou encore la prise en compte des jours fériés.

1.4.7 Illustration de la méthode $X-11$ sur les données SNCF

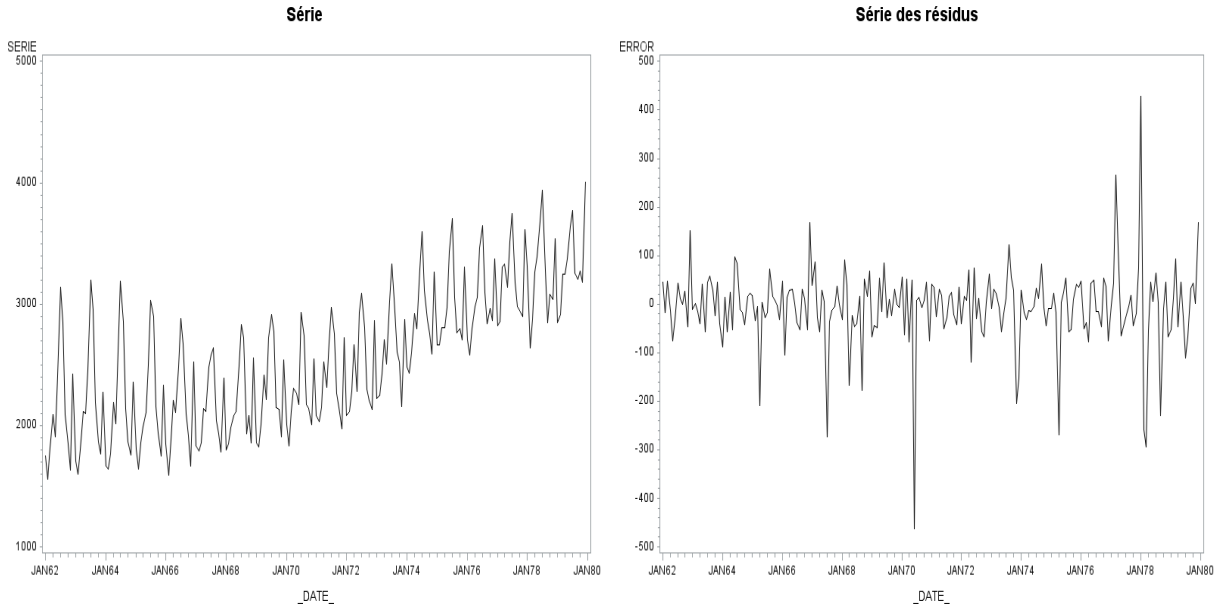


FIGURE 1.9 – La série et les résidus après application de la procédure $x11$

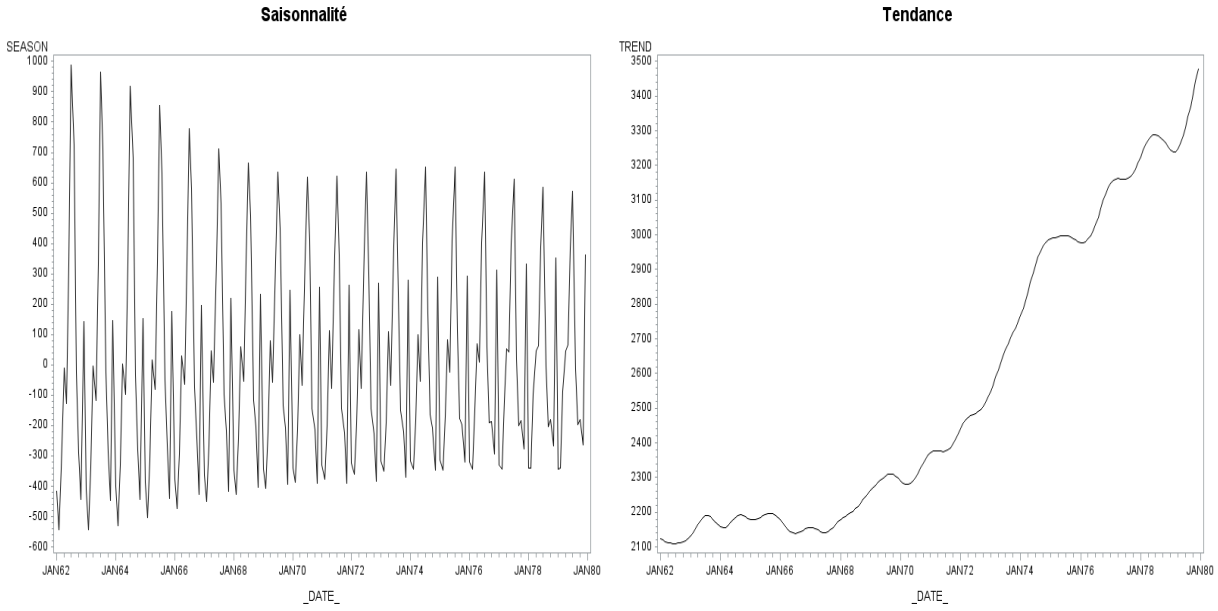


FIGURE 1.10 – Estimations de la tendance et de la saisonnalité obtenues à partir de la procédure $x11$

1.5 Lissage exponentiel

On dispose d'une série temporelle $X_1, X_2, \ldots, X_T$ et on souhaite prévoir les valeurs futures : X_{T+1},X_{T+2},.... On note $\hat{X}_T(h)$ la prévision de X_{T+h} pour $h \in \mathbb{N}^*$. L'entier h est appelé horizon de la prévision. Pour ce type de méthodes, on ne suppose pas de modèle statistique particulier.

1.5.1 Lissage exponentiel simple

Pour une constante β appelée constante de lissage ($0 < \beta < 1$), on définit

$$\hat{X}_T(h) = (1-\beta)\sum_{j=0}^{T-1}\beta^j X_{T-j}. \tag{1.6}$$

Cette prévision ne dépend de h qu'à travers β. Si β ne dépend pas de h, la prévision à l'horizon h sera égale à la prévision à l'horizon 1. Lorsque β est proche de 1, la prévision tient compte d'un grand nombre de de valeurs passées. Lorsque β est proche de zéro, seules les valeurs récentes de la série ont une importance. Dans la suite, on oublie l'entier h dans la notation et on note $\hat{X}_T$ au lieu de $\hat{X}_T(h)$. Une autre écriture de (1.6) est

$$\hat{X}_T = \hat{X}_{T-1} + (1-\beta)\left(X_T - \hat{X}_{T-1}\right). \tag{1.7}$$

Lorsque $h = 1$, la formule (1.7) montre que la prévision $\hat{X}_T$ s'interprète comme la prévision faite à l'instant précédent corrigée par un terme proportionnel à l'erreur de prévision correspondante. On obtient aussi une formule de mise à jour, que l'on peut initialiser par exemple par $\hat{X}_1 = X_1$ (noter que comme $0 < \beta < 1$, le valeur initiale aura peu d'influence lorsque T est grand).
L'interprétation suivante conduira à des généralisations de cette méthode. Si on souhaite minimiser en $C \mapsto \sum_{j=0}^{T-1}\beta^j\left(X_{T-j} - C\right)^2$, on trouve l'expression

$$\hat{C} = \frac{1-\beta}{1-\beta^T}\sum_{j=0}^{T-1}\beta^j X_{T-j}$$

qui correpond à peu près à $\hat{X}_T$ lorsque T est grand. La valeur $\hat{X}_T(\omega)$ s'interprète alors comme la constante qui approxime le mieux la série au voisinage de T (les β^j pondèrent l'importance de la $T-j$-ième observation). **Il est donc préférable de ne pas appliquer cette méthode lorsque la série présente une tendance non constante ou d'importantes fluctuations (de sorte qu'une approximation localement constante soit peu réaliste).**

Choix de la constante de lissage β

Une méthode naturelle et utilisée en pratique est de minimiser la somme des carrés des erreurs de prévision :

$$\hat{\beta} = \arg\min_{\beta}\left\{\sum_{t=1}^{T-1}\left[X_{t+1} - (1-\beta)\sum_{j=0}^{t-1}\beta^j X_{t-j}\right]^2\right\}.$$

Le β optimal tient compte de la qualité de la prévision à l'ordre 1. Si $h \geq 2$, on pourrait aussi obtenir un coefficient de lissage plus adapté en minimisant

$$\beta \mapsto \sum_{t=1}^{T-h} \left(X_{t+h} - (1-\beta) \sum_{j=0}^{t-1} \beta^j X_{t-j} \right)^2 .$$

Lien avec la notion de modèle

Les équations de prévision s'appliquent indépendamment du modèle considéré pour $(X_1, \ldots, X_T)$. Le désavantage est de ne pouvoir bénéficier d'intervalle de confiance pour la prévision. Considérons le cas de la prévision à l'horizon 1 et considérons le processus des prévisions et des erreurs. En posant $\varepsilon_t = X_t - \hat{X}_{t-1}$, on a les deux équations suivantes.

$$\hat{X}_t = \hat{X}_{t-1} + (1-\beta)\varepsilon_t, \quad X_t = \hat{X}_{t-1} + \varepsilon_t.$$

Ces deux équations peuvent s'écrire en une seule,

$$X_t - X_{t-1} = \varepsilon_t - \beta\varepsilon_{t-1}. \tag{1.8}$$

Nous verrons que l'équation (1.8) correspond à un cas particulier des modèles ARIMA que nous étudierons plus loin dans ce cours. Des intervalles de confiance peuvent être obtenus en supposant les ε_t i.i.d de loi gaussienne centrée réduite ce qui permet d'estimer le paramètre β dans (1.8) par maximum de vraisemblance.

Exercice 5 On suppose la suite $(X_t)_t$ centrée, de variance 1 et qu'il existe $-1 < \alpha < 1$ tel que $\mathrm{Cov}\,(X_t, X_{t+h}) = \alpha^h$ pour $h \in \mathbb{N}$ (on verra plus loin dans ce cours l'existence d'une telle suite). Montrer que

$$\mathbb{E}\left[X_{T+1} - \hat{X}_T(1)\right]^2 = \frac{2(1-\alpha)}{(1+\beta)(1-\alpha\beta)}.$$

Déterminer la valeur optimale du paramètre β suivant la valeur de α. Pour quelles valeurs de α la méthode du lissage exponentiel simple est-elle inadaptée ?

1.5.2 Lissage exponentiel double

Une façon de généraliser le lissage exponentiel simple est de supposer que la série s'approche localement par une fonction affine du temps (au lieu d'une fonction constante). On suppose alors que $X_{T+h} \approx b + ah$ pour h petit. Pour une constante de lissage β, il s'agit alors de minimiser la fonction

$$(a, b) \mapsto \sum_{j=0}^{T-1} \beta^j \left(X_{T-j} - b + aj \right)^2 .$$

En annulant les dérivées partielles, on a le système

$$\begin{cases} \sum_{j=0}^{T-1} \beta^j X_{T-j} - b \sum_{j=0}^{T-1} \beta^j + a \sum_{j=0}^{T-1} j\beta^j = 0 \\ \sum_{j=0}^{T-1} j\beta^j X_{T-j} - b \sum_{j=0}^{T-1} j\beta^j + a \sum_{j=0}^{T-1} j^2\beta^j = 0 \end{cases}$$

Pour simplifier, on fait les approximations

$$\sum_{j=0}^{T-1}\beta^j \approx \frac{1}{1-\beta}, \quad \sum_{j=0}^{T-1} j\beta^j \approx \frac{\beta}{(1-\beta)^2}, \quad \text{`} \sum_{j=0}^{T-1} j^2\beta^j \approx \frac{\beta(1+\beta)}{(1-\beta)^3}.$$

De plus, pour obtenir des formules de mise à jour, on note $S_1(T) = (1-\beta)\sum_{j=0}^{T-1}\beta^j X_{T-j}$ et on réécrit

$$\begin{aligned}(1-\beta)\sum_{j=0}^{T-1} j\beta^j X_{T-j} &= (1-\beta)\sum_{k=0}^{T-2}\sum_{j=k+1}^{T-1}\beta^j X_{t-j} \\ &= \sum_{k=0}^{T-2}\beta^{k+1} S_1(T-k-1) \\ &= \sum_{k=0}^{T-1}\beta^k S_1(T-k) - S_1(T).\end{aligned}$$

En posant $S_2(T) = (1-\beta)\sum_{k=0}^{T-1}\beta^k S_1(T-k)$, le système d'équations se réécrit

$$\begin{cases} S_1(T) - b + \frac{a\beta}{1-\beta} = 0 \\ S_2(T) - (1-\beta)S_1(T) - b\beta + \frac{a\beta(1+\beta)}{1-\beta} = 0 \end{cases}$$

On trouve les solutions

$$\hat{b}(T) = 2S_1(T) - S_2(T), \quad \hat{a}(T) = \frac{1-\beta}{\beta}\left(S_1(T) - S_2(T)\right),$$

ces deux égalités étant équivalentes à

$$S_1(T) = \hat{b}(T) - \frac{\beta}{1-\beta}\hat{a}(T), \quad S_2(T) = \hat{b}(T) - \frac{2\beta}{1-\beta}\hat{a}(T).$$

La prévision $\hat{X}_T$ est donnée à l'horizon h par $\hat{X}_T(h) = \hat{b}(T) + h\hat{a}(T)$. Pour choisir β, on peut, comme pour le lissage exponentiel simple, minimiser $\sum_{t=1}^{T-1}\left(X_{t+1} - \hat{X}_t(1)\right)^2$. Il existe aussi des formules de mise à jour :

$$\begin{cases} \hat{b}(T) = \hat{b}(T-1) + \hat{a}(T-1) + (1-\beta^2)\left[X_T - \hat{X}_{T-1}(1)\right] \\ \hat{a}(T) = \hat{a}(T-1) + (1-\beta)^2\left[X_T - \hat{X}_{T-1}(1)\right] \end{cases} \tag{1.9}$$

Pour interpréter ces formules de mise à jour, on pourra noter que si la prévision de X_T était exacte (i.e $\hat{X}_{T-1}(1) = X_T$) alors on conserve la même droite à l'étape suivante. Pour intialiser ces formules de mise à jour, on remarque que les solutions du problème de minimisation sont données pour $T = 2$ par $\hat{b}(2) = X_2$ et $\hat{a}(2) = X_2 - X_1$.

1.5.3 Le lissage de Holt-Winters

Méthode non saisonnière

Reprenons les formules de mise à jour (1.9) du lissage exponentiel double. Dans la première égalité, on remplace $\hat{X}_{T-1}(1)$ par $\hat{b}(T-1) + \hat{a}(T-1)$ et dans la deuxième, on utilise l'expression

$$X_T - \hat{X}_{T-1}(1) = \frac{\hat{b}(T) - \hat{a}(T-1) - \hat{b}(T-1)}{1-\beta^2},$$

déduite de la première égalité. On obtient alors

$$\begin{cases} \hat{b}(T) = \beta^2 \left(\hat{a}(T-1) + \hat{b}(T-1)\right) + \left(1-\beta^2\right) X_T \\ \hat{a}(T) = \hat{a}(T-1)\left(1 - \frac{(1-\beta)^2}{1-\beta^2}\right) + \frac{(1-\beta)^2}{1-\beta^2}\left(\hat{b}(T) - \hat{b}(T-1)\right) \end{cases}$$

On peut alors généraliser ces formules en supposant que pour deux paramètres $0 < \alpha, \delta < 1$ à calibrer,

$$\begin{cases} \hat{b}(T) = \alpha \left(\hat{a}(T-1) + \hat{b}(T-1)\right) + (1-\alpha) X_T \\ \hat{a}(T) = \gamma \hat{a}(T-1) + (1-\gamma)\left(\hat{b}(T) - \hat{b}(T-1)\right) \end{cases}$$

Les deux paramètres α et γ peuvent être choisis en minimisant la somme des carrés des erreurs de prévision, comme pour le lissage exponentiel simple ou double. L'initialisation des formules récursives de mise à jour est identique à celle pour le lissage exponentiel double.

Méthode saisonnière additive

Ici on suppose que $X_{t+h} \approx b + ah + s_{t+h}$ où s_t est un facteur saisonnier de période k. Cette hypothèse est adaptée au cas où $X_t = c + at + s_t + U_t$ (on a alors $b = c + aT$). Les formules de mise à jour sont les suivantes et dépendent de trois paramètres $0 < \alpha, \gamma, \delta < 1$.

$$\begin{cases} \hat{b}(T) = \alpha \left(\hat{a}(T-1) + \hat{b}(T-1)\right) + (1-\alpha)\left(X_T - \hat{s}_{T-k}\right) \\ \hat{a}(T) = \gamma \hat{a}(T-1) + (1-\gamma)\left(\hat{b}(T) - \hat{b}(T-1)\right) \\ \hat{s}_T = (1-\delta)\left[X_T - \hat{b}(T)\right] + \delta \hat{s}_{T-k} \end{cases}$$

Les deux premières équations sont naturelles et correspondent à celles de la méthode non-saisonnière. La dernière équation montre que la prévision de s_T est une combinaison linéaire de la prévision précédente $\hat{s}_{T-k}$ et de l'écart entre X_T et l'ordonnée à l'origine de la droite de prévision utilisée au temps T (terme de correction). La prévision est donnée par

$$\hat{X}_T(h) = \hat{b}(T) + h\hat{a}(T) + \hat{s}_{T+h-k}, \quad 1 \le h \le k.$$

Le choix des trois paramètres α, γ, δ se fait en général en minimisant $\sum_{t=1}^{T-1}\left[X_{t+1} - \hat{X}_t(1)\right]^2$. Les valeurs d'initialisation doivent être choisies. On peut par exemple définir $\hat{b}(k)$, $\hat{b}(k-1)$ par $M(X)_k$, $M(X)_{k-1}$ où M est une moyenne mobile qui annule la saisonnalité, puis $\hat{a}(k) = \hat{a}(k-1) = \hat{b}(k) - \hat{b}(k-1)$, $\hat{s}_k = X_k - \hat{b}(k)$, $\hat{s}_{k-1} = X_{k-1} - \hat{b}(k-1)$ et enfin $\hat{s}_{k-i} = X_{k-i} - \hat{b}(k-1) + \hat{a}(k)(i-1)$ pour $2 \le i \le k-1$.

Méthode saisonnière multiplicative

On suppose l'approximation

$$X_{t+h} \approx [b + ah]\, s_t$$

au voisinage de t. Les formules de mise à jour sont alors pour trois paramètres α, γ, δ à calibrer

$$\begin{cases} \hat{b}(T) = \alpha\left(\hat{a}(T-1) + \hat{b}(T-1)\right) + (1-\alpha)\frac{X_T}{\hat{S}_{T-k}} \\ \hat{a}(T) = \gamma\hat{a}(T-1) + (1-\gamma)\left(\hat{b}(T) - \hat{b}(T-1)\right) \\ \hat{S}_T = (1-\delta)\frac{X_T}{\hat{b}(T)} + \delta\hat{s}_{T-k} \end{cases}$$

Les initialisations sont analogues à celles du paragraphe précédent (on calcule la tendance linéaire à partir de $t = k, k-1$ puis on estime s_i en divisant X_i par la tendance linéaire évaluée en $t = i$).

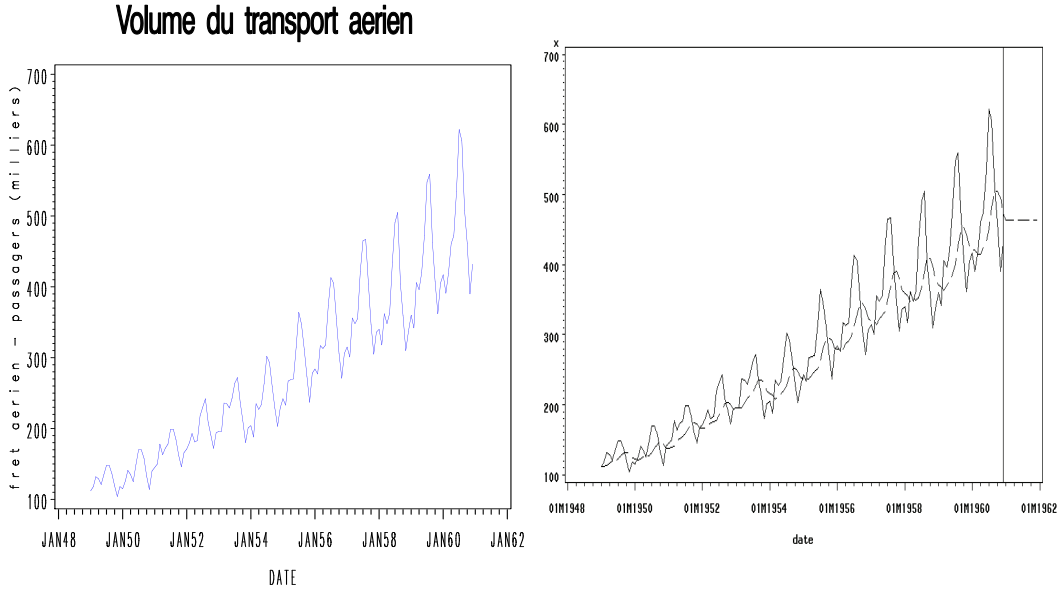


FIGURE 1.11 – Données de trafic aérien (à gauche) et lissage exponentiel simple (à droite en pointillés)

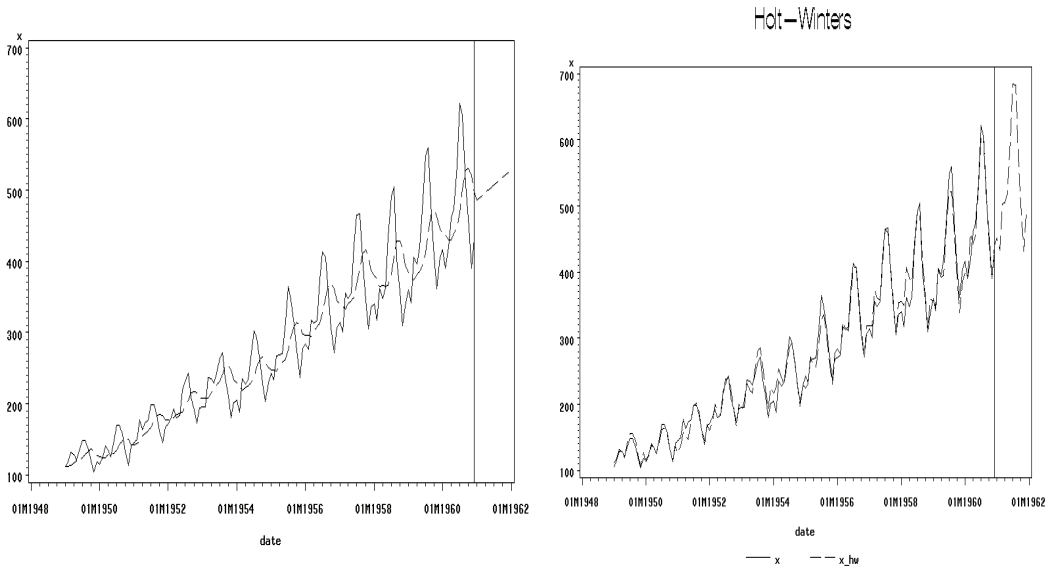


FIGURE 1.12 – Lissage exponentiel double (à gauche) et méthode de Holt-Winters saisonnière multiplicative (à droite)

Chapitre 2

Introduction à la théorie des processus stationnaires à temps discret

Lorsque les composantes tendancielles et saisonnières auront été retirées, ou bien lorsque une opération de différenciation aura permis de les absorber, la série temporelle obtenue doit être modélisée mathématiquement. Il serait trop simpliste de modéliser ces séries à l'aide de suites de variables aléatoires indépendantes et identiquement distribuées. Par exemple, au niveau empirique, on observe le plus souvent de la corrélation entre les vecteurs $(x_1, \ldots, x_{T-1})$ et $(x_2, \ldots, x_{T-1})$ (voir par exemple la Figure 2.1). Tenir compte de cette corrélation est important pour prévoir plus précisément les valeurs futures. L'objectif de ce chapitre est d'introduire un formalisme mathématique qui permettra de modéliser ces composantes aléatoires.

2.1 Quelques généralités sur les processus stochastiques

Comme pour la statistique des données indépendantes, il est très utile du point de vue mathématique de considérer une infinité de variables aléatoires toutes définies sur le même espace probabilisé. Dans la suite, on note I un ensemble qui représentera $\mathbb{N}$ ou $\mathbb{Z}$.

Définition 6 *Un processus stochastique est une famille de variables aléatoires* $\{X_t : t \in I\}$ *toutes définies sur* $(\Omega, \mathcal{A}, \mathbb{P})$. *Les applications* $t \mapsto X_t(\omega)$, $\omega \in \Omega$, *sont appelées trajectoires du processus.*

Notons que la possibilité de définir une infinité de variables aléatoires sur un même espace probabilisé est un problème non trivial en général. Par exemple, pour pouvoir définir k variables aléatoires indépendantes et toutes de loi à densité $f : \mathbb{R} \to \mathbb{R}$, on peut se contenter de poser $\Omega = \mathbb{R}^k$, $\mathcal{A} = \otimes_{i=1}^{k} \mathcal{B}(\mathbb{R})$ et $\mathbb{P} = \mu_k$ où

$$\mu_k(B) = \int_B f(x_1) \cdots f(x_k) dx_1 \cdots dx_k, \quad B \in \mathcal{A}. \tag{2.1}$$

Il suffit alors de poser $X_i(\omega) = \omega_i$ pour $\omega \in \Omega$ et $1 \leq i \leq k$. Mais cela ne permet pas de définir toute une suite de variables aléatoires. Le théorème suivant, appelé théorème de Kolmogorov, guarantit la possibilité de définir un processus lorsqu'on dispose d'une famille de lois dites

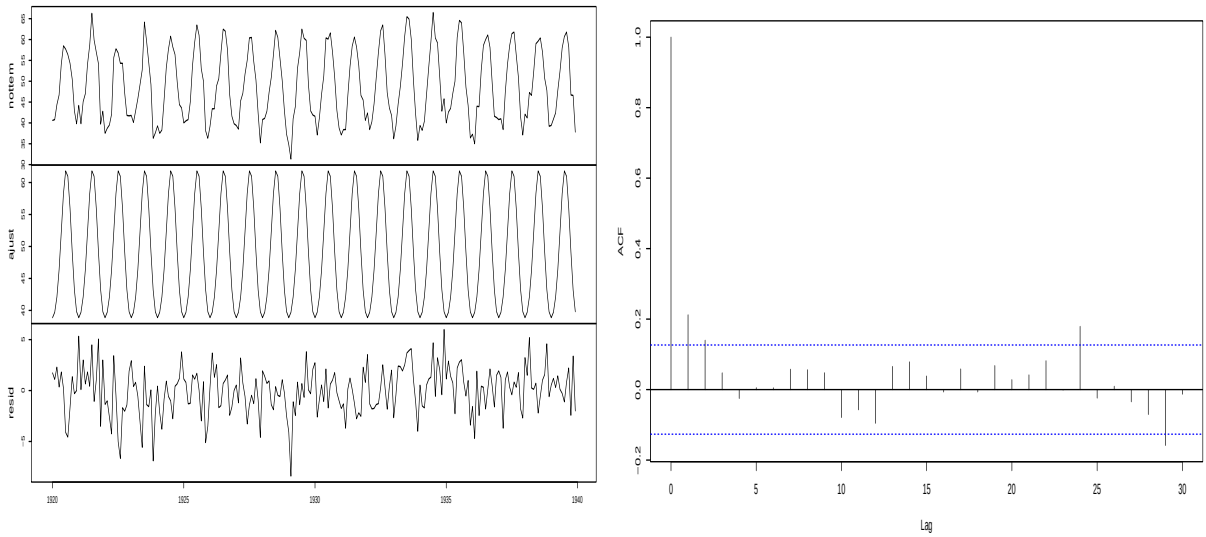


FIGURE 2.1 – Série temporelle désaisonnalisée par régression et autocorrélation des résidus obtenus (sur la figure de droite la barre n°i représente la valeur du coefficient de corrélation empirique entre $(x_1, \ldots, x_{T-i})$ et $(x_{i+1}, \ldots, x_T)$).

fini-dimensionnelles (telles que μ_k) qui satisfait une condition de compatibilité assez naturelle. Ci-dessous, nous l'énonçons pour $I = \mathbb{N}$ uniquement.

Théorème 1 *Soit* $\left\{\mu_n : \mathcal{B}\left(\mathbb{R}^{n+1}\right) \to [0,1]; n \in \mathbb{N}\right\}$ *une famille de mesures de probabilité qui vérifie la condition suivante : pour tout* $n \in \mathbb{N}$*, et* $A_0, \ldots, A_n \in \mathcal{B}(\mathbb{R})$*,*

$$\mu_n\left(A_0 \times \cdots \times A_n\right) = \mu_{n+1}\left(A_0 \times \cdots \times A_n \times \mathbb{R}\right).$$

Alors il existe un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ *et un processus stochastique* $\{X_t : t \in \mathbb{N}\}$ *tel que pour tout entier n, la loi de* $(X_0, \ldots, X_n)$ *sous* $\mathbb{P}$ *coïncide avec* μ_n*.*

Le théorème précédent s'applique en particulier pour construire une suite $(X_t)_{t\in\mathbb{N}}$ de variables aléatoires i.i.d et toutes de loi μ. Il suffit de poser $\mu_n = \mu^{\otimes n}$. On peut également, grâce à ce théorème, définir une suite de variables aléatoires indépendantes mais non identiquement distribuées.

Hormis le cas des variables aléatoires indépendantes, une classe importante de processus stochastiques est celle des processus gaussiens.

Définition 7 *On dit qu'un processus* $\{X_t : t \in \mathbb{N}\}$ *est gaussien si pour tout entier n, le vecteur* $(X_0, \ldots, X_n)$ *suit une loi gaussienne sur* $\mathbb{R}^{n+1}$*.*

La loi d'un processus gaussien ne dépend donc que des moyennes $m(t) = \mathbb{E}(X_t)$ et des covariances $\Gamma(s,t) = \mathrm{Cov}(X_s, X_t)$ pour $s, t \in \mathbb{N}$. La fonction Γ (dite fonction de covariance) vérifie les propriétés suivantes.

Proposition 3 *Toute fonction de covariance* Γ *vérifie les trois propriétés suivantes.*

1. $\Gamma(s,t) = \Gamma(t,s)$ *pour tous* $s,t \in \mathbb{N}$.
2. $|\Gamma(s,t)| \leq \sqrt{\Gamma(s,s)} \cdot \sqrt{\Gamma(t,t)}$ *pour tous* $s,t \in \mathbb{N}$.
3. Γ *est semi-définie positive, i.e pour tous* $n \in \mathbb{N}$ *et* $a_0, \ldots, a_n \in \mathbb{R}$, *on a* $\sum_{i,j=0}^{n} a_i a_j \Gamma(i,j) \geq 0$.

Inversement, toute fonction $\Gamma : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ *qui satisfait les propriétés* 1 *et* 3 *est la fonction de covariance d'un processus gaussien.*

Preuve

Si Γ est une fonction de covariance, les deux premières propriétés sont évidentes (la deuxième résulte de l'inégalité de Cauchy-Schwarz). Pour la dernière propriété, il suffit de remarquer que

$$\sum_{i,j=0}^{n} a_i a_j \Gamma(i,j) = \mathrm{Var}\left(\sum_{i=0}^{n} a_i X_i\right).$$

Inversement, si Γ vérifie les propriétés 1 à 3 alors on peut appliquer le Théorème 1 en définissant pour $n \in \mathbb{N}$, μ_n comme étant la loi gaussienne de moyenne 0 et de matrice de variance-covariance $(\Gamma(i,j))_{0 \leq i,j \leq n}$.□

Exemple

Soit $\rho \in (-1,1)$. Alors il est possible de construire un processus gaussien de covariance $\Gamma(s,t) = \rho^{|t-s|}$ (la corrélation de dépend que de l'écart $t-s$ et décroît à vitesse exponentielle. Pour vérifier la condition de positivité, on peut remarquer qu'il suffit de prouver que pour tous $n \in \mathbb{N}$, la matrice $A_n = (\Gamma(i-j))_{0 \leq i,j \leq n}$ est définie positive pour tout $n \in \mathbb{N}^*$. Pour cela, on pourra vérifier en utilisant des opérations sur les colonnes que les mineurs principaux de A_n valent $(1-\rho^2)^j$, pour $j = 0, 1, \ldots, n$.

2.2 Stationnarité stricte et stationnarité faible

Définition 8

1. *On dit que* X *est strictement stationnaire si pour tout* $h \in I$, *le processus* $(X_{t+h})_{t \in I}$ *a la même loi que* X.
2. *On dit que* X *est faiblement stationnaire (ou stationnaire au second ordre, ou stationnaire en covariance) s'il vérifie les trois proriétés suivantes.*
   (a) *Pour tout* $t \in I$, $\mathbb{E}\left(X_t^2\right) < +\infty$.
   (b) *Il existe* $m \in \mathbb{R}$ *tel que pour tout* $t \in I$, $\mathbb{E}(X_t) = m$.
   (c) *Il existe une fonction* $\gamma : I \to \mathbb{R}$ *telle que pour tous* $s,t \in I$, $Cov(X_s, X_t) = \gamma(t-s)$.

Dire que X est strictement stationnaire signifie que le vecteur aléatoire $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h})$ a la même loi que $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ et ce pour tous $h \in I$, $n \in \mathbb{N}^*$ et $t_1 < t_2 < \cdots < t_n$ dans I. Evidemment, tout processus strictement stationnaire et de carré intégrable est faiblement stationnaire.

La fonction γ associée à un processus stationnaire est appelée fonction d'autocovariance du processus. On pourra remarquer que c'est une fonction paire (symétrique) : $\gamma(-h) = \gamma(h)$ pour tout $h \in \mathbb{Z}$.

A partir de maintenant, nous ne considérerons plus que des processus stationnaires indexés par $I = \mathbb{Z}$. Ce choix, qui sera plus approprié à la façon dont nous représenterons les processus stationnaires, n'est nullement restrictif. On peut toujours étendre un processus stationnaire indexé par $\mathbb{N}$ en un processus stationnaire indexé par $\mathbb{Z}$ et qui a les mêmes caractéristiques.

Exemples

1. Une suite de variables aléatoires i.i.d $(X_t)_{t\in\mathbb{Z}}$ est strictement stationnaire.
2. Le processus gaussien $(X_t)_{t\in\mathbb{Z}}$ construit à la section précédente (avec $\mathbb{E}(X_t) = 0$ et $\mathrm{Cov}\,(X_s, X_t) = \rho^{|t-s|}$) est faiblement stationnaire. On pourra remarquer que tout processus gaussien faiblement stationnaire est automatiquement strictement stationnaire.
3. Un exemple de processus stationnaire très utilisé dans la suite est le bruit blanc. On dit que $(\varepsilon_t)_{t\in I}$ est un bruit blanc (faible) si $\mathbb{E}\,(\varepsilon_t) = 0$, $\mathrm{Var}\,(\varepsilon_t) = \sigma^2$ et $\mathrm{Cov}\,(\varepsilon_s, \varepsilon_t) = 0$ si $s \neq t$. Lorsque les variables sont en plus i.i.d, on parle de bruit blanc fort. Un bruit blanc fort est un bruit blanc faible mais le contraire n'est pas nécessairement vrai. Pour construire un contre-exemple, on peut par exemple considérer une suite $(\varepsilon_t)_{t\in\mathbb{Z}}$ de variables aléatoires indépendantes et telle que
$$\mathbb{P}\left(\varepsilon_t = \sqrt{|t|+1}\right) = \mathbb{P}\left(\varepsilon_t = -\sqrt{|t|+1}\right) = \frac{1}{2\,(|t|+1)}, \quad \mathbb{P}\,(\varepsilon_t = 0) = 1 - \frac{1}{|t|+1}.$$
4. Lorsque $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc, on peut construire des processus stationnaires sous la forme de moyennes mobiles en posant
$$X_t = \sum_{i=-p_1}^{p_2} \theta_i \varepsilon_{t+i}.$$
En effet, X_t est de carré intégrable car somme de variables aléatoires de carré intégrable. De plus $\mathbb{E}(X_t) = 0$ et si $h \geq 0$, $s \in \mathbb{Z}$,
$$\mathrm{Cov}\,(X_s, X_{s+h}) = \sigma^2 \sum_{i,j=-p_1}^{p_2} \theta_i\theta_j \mathbb{1}_{i=h+j}.$$
On voit alors que $(X_t)_{t\in\mathbb{Z}}$ est stationnaire en posant
$$\gamma(h) = \begin{cases} \sigma^2 \sum_{j=-p_1}^{p_2-h} \theta_j\theta_{j+h} \text{ si } 0 \leq h \leq p_2 + p_1 \\ 0 \text{ si } h > p_2 + p_1 \end{cases}$$
Il n'est pas utile de refaire le calcul pour $h < 0$. En effet, on a pour $h < 0$,
$$\mathrm{Cov}\,(X_s, X_{s+h}) = \mathrm{Cov}\,(X_{s+h}, X_{s+h-h}) = \gamma(-h),$$
d'après les calculs précédents. Sans tenir compte du signe de h, on posera
$$\gamma(h) = \begin{cases} \sigma^2 \sum_{j=-p_1}^{p_2-|h|} \theta_j\theta_{j+|h|} \text{ si } |h| \leq p_2 + p_1 \\ 0 \text{ sinon} \end{cases}$$

5. On peut aussi construire des moyennes mobiles d'ordre infini à partir d'un bruit blanc $(\varepsilon_t)_{t\in\mathbb{Z}}$ et d'une suite $(\theta_i)_{i\in\mathbb{N}}$ de nombres réels tels que $\sum_{i\in\mathbb{N}}\theta_i^2 < +\infty$. Pour $t \in \mathbb{Z}$, posons $X_t = \sum_{i=0}^{+\infty}\theta_i\varepsilon_{t-i}$. La somme infinie considérée a bien un sens. On la définit comme la limite dans $\mathbb{L}^2$ de la variable $X_t^{(N)} = \sum_{i=0}^{N}\theta_i\varepsilon_{t-i}$ (formellement, on peut vérifier que la suite de $\left(X_t^{(N)}\right)_{t\in\mathbb{Z}}$ est une suite de Cauchy dans $\mathbb{L}^2$, espace des variables aléatoires de carré intégrable). Il alors clair que

$$\mathbb{E}(X_t) = \lim_{N\to+\infty}\sum_{i=0}^{N}\theta_i\mathbb{E}(\varepsilon_{t+i}) = 0.$$

De plus, pour $h \in \mathbb{N}$ donné, on a

$$\begin{aligned}\mathrm{Cov}(X_t, X_{t+h}) &= \lim_{N\to+\infty}\mathrm{Cov}\left(X_t^{(N)}, X_{t+h}^{(N)}\right)\\ &= \lim_{N\to+\infty}\sigma^2\sum_{i=0}^{N-h}\theta_i\theta_{i+h}\\ &= \sigma^2\sum_{i\in\mathbb{N}}\theta_i\theta_{i+h}.\end{aligned}$$

Tous les processus stationnaires que nous rencontrerons dans ce cours s'écriront sous cette forme. Comme nous le verrons plus loin dans ce chapitre, on obtient presque tous les processus faiblement stationnaires à partir de ces moyennes mobiles d'ordre infini. On peut aussi construire des moyennes mobiles bilatérales d'ordre infini en posant $X_t = \sum_{i=-\infty}^{+\infty}\theta_i\varepsilon_{t-i}$ en convenant que

$$\sum_{i=-\infty}^{+\infty}\theta_i^2 = \sum_{i=0}^{+\infty}\theta_i^2 + \sum_{i=1}^{+\infty}\theta_{-i}^2 < +\infty.$$

La fonction d'autocovariance est alors donnée par $\gamma(h) = \sum_{i=-\infty}^{+\infty}\theta_i\theta_{i+|h|}$ (on peut même supprimer la valeur absolue car $\sum_{i\in\mathbb{Z}}\theta_i\theta_{i+h} = \sum_{i\in\mathbb{Z}}\theta_i\theta_{i-h}$).

6. Soient $(X_t)_{t\in\mathbb{Z}}$ un processus faiblement stationnaire de moyenne m_X et de fonction d'autocovariance γ_X et $(\theta_i)_{i\in\mathbb{Z}}$ une famille de nombres réels absolument sommable (c'est à dire que la valeur absolue est sommable au sens de la définition donnée dans l'exemple précédent). Alors, en posant

$$Y_t = \sum_{i\in\mathbb{Z}}\theta_i X_{t+i}, \quad t \in \mathbb{Z}, \tag{2.2}$$

on définit un nouveau processus stationnaire. Notons $\|X\|_{\mathbb{L}^2} = \sqrt{\mathbb{E}(X^2)}$ la norme $\mathbb{L}^2$ d'une variable aléatoire de carré intégrable. On peut remarquer $\mathbb{E}\left(X_t^2\right) = \gamma_X(0) + m_X^2$ et donc que $\|X_t\|_{\mathbb{L}^2}$ ne dépend pas de t. Pour vérifier que la somme dans (2.2) a bien un sens, il suffit de voir que la série est normalement convergente dans $\mathbb{L}^2$,

$$\sum_{i=0}^{+\infty}\|\theta_i X_{t+i}\|_{\mathbb{L}^2} \le \sup_{t\in\mathbb{Z}}\|X_t\|_{\mathbb{L}^2}\sum_{i=0}^{+\infty}|\theta_i| < +\infty$$

et de conclure en utilisant la sommabilité des θ_i (on montre de la même manière que la série $\sum_{i=1}^{+\infty} \theta_{-i} X_{t-i}$ est normalement convergente dans $\mathbb{L}^2$). On voit alors facilement que

$$\mathbb{E}(Y_t) = \sum_{i\in\mathbb{Z}} \theta_i \mathbb{E}(X_{t+i}) = \sum_{i\in\mathbb{Z}} \theta_i m_X.$$

De plus, on peut montrer que

$$\mathrm{Cov}(Y_t, Y_{t+h}) = \sum_{i\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \theta_i \theta_j \gamma_X(i-j+h).$$

7. On appelle marche aléatoire tout processus de la forme $X_t = \sum_{i=1}^{t} \varepsilon_i$ pour $t \in \mathbb{N}^*$ où $(\varepsilon_i)_{i\in\mathbb{N}^*}$ désigne un bruit blanc fort. Alors $\mathbb{E}(X_t)$ ne dépend pas de t uniquement si $\mathbb{E}(\varepsilon_t) = 0$. Mais $\mathrm{Var}(X_t) = t\mathrm{Var}(\varepsilon_1)$, donc le processus n'est pas stationnaire.
8. Un processus construit à partir d'une série temporelle qui présente une tendance ou une saisonnalité n'est pas stationnaire (la moyenne dépend du temps t).

Définition 9 *La fonction d'autocorrélation ρ d'un processus stationnaire X est défini par $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$ où γ désigne la fonction d'autocovariance du processus X.*

On pourra remarquer que pour un processus stationnaire X, on a $\rho_X = \gamma_Y$ où $Y_t = \frac{X_t}{\sqrt{\gamma_X(0)}}$ correspond au processus renormalisé (de variance 1). La fonction ρ_X qui prend ses valeurs dans $[0, 1]$ et vérifie $\rho_X(-h) = \rho_X(h)$

Exemple

Considérons le cas d'une moyenne mobile infinie $X_t = \sum_{i\in\mathbb{Z}} \theta_i \varepsilon_{t-i}$ où $(\varepsilon_i)_{i\in\mathbb{Z}}$ est un bruit blanc faible de variance σ^2. Alors

$$\gamma(h) = \sigma^2 \sum_{i\in\mathbb{Z}} \theta_i \theta_{i+|h|}$$

et $\mathrm{Var}(X_0) = \sigma^2 \sum_{i\in\mathbb{Z}} \theta_i^2$. Ainsi si $\theta_i = 0$ lorsque $|i| > p$ (c'est la cas pour une moyenne mobile d'ordre $p+1$), on $\gamma(h) = \rho(h) = 0$ lorsque $|h| > 2p$.

Le graphe de la fonction d'autocorrélation est appelée autocorrélogramme. C'est un graphe très important en pratique, il permet souvent à déterminer le type de modèles à ajuster.

Exercice 6 Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc et $X_t = (-1)^t \varepsilon_t$.

1. Montrer que $(X_t)_{t\in\mathbb{Z}}$ est faiblement stationnaire. Lorsque $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc gaussien (i.e un bruit blanc qui soit aussi un processus gaussien), montrer que $(X_t)_{t\in\mathbb{Z}}$ est strictement stationnaire.
2. Donner une exemple de bruit blanc $(\varepsilon_t)_{t\in\mathbb{Z}}$ pour lequel $(X_t)_{t\in\mathbb{Z}}$ n'est pas strictement stationnaire.
3. A partir de cet exemple, montrer que la somme de processus stationnaires n'est pas nécessairement stationnnaire et que la stationnarité de $(|X_t|)_{t\in\mathbb{Z}}$ n'entraine pas la stationnairité de $(X_t)_{t\in\mathbb{Z}}$.

2.3 Extension de la loi des grands nombres

On aimerait que pour tout processus strictement stationnaire $(X_t)_{t\in\mathbb{Z}}$ la moyenne empirique $\frac{1}{T}\sum_{t=1}^{T} X_t$ converge p.s vers $\mathbb{E}(X_1)$ lorsque $T \to +\infty$ si X_1 est intégrable. Malheureusement, une telle extension de la loi des grands nombres n'a pas toujours lieu. Par exemple si $(Y_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires i.i.d et intégrables, indépendante d'une variable aléatoire intégrable Z non dégénérée, alors en posant $X_t = Y_t + Z$, on a

$$\frac{1}{T}\sum_{t=1}^{T} X_t = \frac{1}{T}\sum_{t=1}^{T} Y_t + Z \to \mathbb{E}(Y_1) + Z \neq \mathbb{E}(X_1).$$

Pour guarantir la convergence des moyennes empiriques vers les moyennes théoriques (espérance mathématique), il faut exclure des cas de dépendance trop forte entre les variables. Nous admettrons le théorème suivant qui sera suffisant pour justifier la consistance des estimateurs que nous utiliserons.

Théorème 2 *Soient $H : \left(\mathbb{R}^d\right)^{\mathbb{Z}} \to \mathbb{R}$ est une fonction mesurable et $(\varepsilon_i)_{i\in\mathbb{Z}}$ une suite i.i.d de variables aléatoires à valeurs dans $\mathbb{R}^d$. Posons $X_t = H\left((\varepsilon_{t-i})_{i\in\mathbb{Z}}\right)$. Alors la suite $(X_t)_{t\in\mathbb{Z}}$ est strictement stationnaire. De plus, si X_0 est intégrable, alors*

$$\lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T} X_t = \mathbb{E}(X_0),\ p.s.$$

Par exemple, considérons une moyenne mobile infinie $X_t = \sum_{i\in\mathbb{Z}} \theta_i \varepsilon_{t-i}$ où $\sum_{i\in\mathbb{Z}} \theta_i^2 < +\infty$ et $(\varepsilon_i)_{i\in\mathbb{Z}}$ une suite de v.a.r i.i.d centrées et de carré intégrable. Alors, le théorème précédent guarantit que $\lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T} X_t = 0$ p.s.
Toujours avec les mêmes hypothèses, si h est un nombre entier positif, on a aussi

$$\lim_{T\to+\infty} \frac{1}{T}\sum_{t=h+1}^{T} X_t X_{t-h} = \mathbb{E}(X_0 X_{-h}) = \mathbb{E}(X_h X_0),\ \text{p.s.}$$

En effet, il suffit d'appliquer le théorème précédent avec la fonction $H : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ définie par

$$H(x) = \sum_{i\in\mathbb{Z}} \theta_i x_i \cdot \sum_{i\in\mathbb{Z}} \theta_i x_{i+h}.$$

2.4 Quelques rappels sur les projections

2.4.1 Rappels sur l'espace $\mathbb{L}^2$. Théorème de projection

L'ensemble des variables aléatoires X telles que $\mathbb{E}\left(X^2\right) < +\infty$ peut être muni d'une structure d'espace vectoriel normé. La norme d'une variable aléatoire X est définie par $\|X\|_{\mathbb{L}^2} = \sqrt{\mathbb{E}(X^2)}$. On note $\mathbb{L}^2$ cet espace. La norme utilisée est associée à un produit scalaire, ce qui permet de définir une notion orthogonalité dans $\mathbb{L}^2$. Le produit scalaire en question est l'application $(X, Y) \mapsto \mathbb{E}(XY)$ et on dit que X et Y sont orthogonaux lorsque $\mathbb{E}(XY) = 0$. Un des résultats fondamentaux est le théorème de projection.

Projection sur un sous-espace vectoriel fermé

Supposons que F soit un sous-espace vectoriel fermé de $\mathbb{L}^2$ (la fermeture est automatique lorsque F est engendré par un nombre fini de variables aléatoires $Y_1, Y_2, \ldots, Y_p$ de $\mathbb{L}^2$). Alors il existe une unique variable aléatoire Z_X de F telle que

$$\|X - Z_X\|_{\mathbb{L}^2} = \inf_{Z \in F} \|X - Z\|_{\mathbb{L}^2}.$$

Cette variable aléatoire est caractérisée par les relations d'orthogonalité (ou égalités de produits scalaires)

$$\mathbb{E}(XZ) = \mathbb{E}(Z_X Z), \quad \forall Z \in F.$$

On dit que Z_X est la projection orthogonale de X sur F. Lorsque le sous-espace vectoriel F est engendré par des variables aléatoires $Y_1, Y_2, \ldots$, Z_X est caractérisé par les équations

$$\mathbb{E}(XY_i) = \mathbb{E}(Z_X Y_i), \quad i \in \mathbb{N}^*.$$

Z_X s'interprète alors comme la meilleure fonction linéaire des Y_i qui approche X. Dans le cas où les Y_i sont orthogonaux deux à deux, on a l'expression

$$Z_X = \sum_{i=1}^{+\infty} \frac{\mathbb{E}(XY_i)}{\mathbb{E}\left(Y_i^2\right)} Y_i.$$

De plus l'application $p_F : X \mapsto Z_X$ est linéaire et on a l'égalité (théorème de Pythagore)

$$\|X\|_{\mathbb{L}^2}^2 = \|p_F(X)\|_{\mathbb{L}^2}^2 + \|X - p_F(X)\|_{\mathbb{L}^2}^2.$$

On a $p_F(X) = X$ lorsque $X \in F$. Enfin, si G est une sous-espace vectoriel de F alors $p_G = p_G \circ p_F$.

Projection et espérance conditionnelle

Supposons toujours que F soit engendré par des variables aléatoires $Y_1, Y_2, \ldots$. Notons $\mathcal{B}$ la tribu engendrée par les Y_i. Alors l'espérance conditionnelle $\mathbb{E}(X|\mathcal{B})$ est une variable aléatoire de $\mathbb{L}^2$, pouvant s'écrire comme une fonction de $Y_1, Y_2, \ldots$ et qui permet d'approcher la variable aléatoire X dans $\mathbb{L}^2$ avec moins d'erreur que $p_F(X)$ en général (lorsque l'espérance conditionnelle est une fonction non linéaire des Y_i). Toutefois, en pratique, la projection linéaire p_F est plus facile à calculer que l'espérance conditionnelle, sauf si on fait des hypothèses fortes sur les variables aléatoires considérées. Il est important de noter que dans le cas de variables gaussiennes(i.e $(Y_1, \ldots, Y_d, X)$ est un vecteur gaussien) les deux notions coïncident puisqu'on peut montrer que l'espérance conditionnelle est linéaire en $Y_1, \ldots, Y_d$.

2.4.2 Projection linéaire sur un sous-espace vectoriel de dimension fini de $\mathbb{L}^2$

Dans le reste de cette section, $X = (X_t)_{t \in \mathbb{Z}}$ désignera un processus stationnaire au second ordre. Nous noterons γ sa fonction d'autocovariance. On supposera que sa moyenne m est nulle (si ce

n'est pas le cas, il suffit de considérer $Y_t = X_t - m$). Si on veut projeter X_t sur le sous-espace vectoriel $F_{t,k}$ de $\mathbb{L}^2$ engendré par les variables $X_{t-1}, \ldots, X_{t-k}$ (projection sur le passé), on a une expression du type

$$p_{F_{t,k}}(X_t) = \sum_{j=1}^{k} b_{k,j} X_{t-j}.$$

Pour trouver une expression des coefficients $b_{k,j}$, il suffit d'appliquer les relations d'orthogonalité

$$\mathbb{E}\left(X_t X_{t-i}\right) = \sum_{j=1}^{k} b_{k,j} \mathbb{E}\left(X_{t-j} X_{t-i}\right), \quad i = 1, \ldots, k.$$

Ces relations se réécrivent

$$\gamma(i) = \sum_{j=1}^{k} b_{k,j} \gamma(i-j), \quad i = 1, \ldots, k.$$

En notant Γ_k la matrice $[\gamma(i-j)]_{1 \leq i,j \leq k}$, il vient

$$\begin{pmatrix} b_{k,1} \\ \vdots \\ b_{k,k} \end{pmatrix} = \Gamma_k^{-1} \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(k) \end{pmatrix}.$$

De plus, on pourra remarquer, en effectuant les mêmes calculs, que la projection de X_t sur $F_{t+k+1,k} = \text{Vect}\left(X_{t+1}, \ldots, X_{t+k}\right)$ (projection sur le futur) s'écrit

$$p_{F_{t+k+1,k}}(X_t) = \sum_{j=1}^{k} b_{k,j} X_{t+j}.$$

De plus, on remarquera que dans les deux cas, la variance de l'erreur de prévision est la même et est donnée par

$$\begin{aligned} \text{Var}\left(X_t - p_{F_{t,k}}(X_t)\right) &= \text{Var}\left(X_t - p_{F_{t+k+1,k}}(X_t)\right) \\ &= \text{Var}\,(X_t) - \text{Var}\left(p_{F_{t+k+1,k}}(X_t)\right) \\ &= \gamma(0) - \left(\gamma(1), \ldots, \gamma(k)\right) \Gamma_k^{-1} \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(k) \end{pmatrix}. \end{aligned}$$

2.5 Fonction d'autocorrélation partielle

Conservons les notations de la sous-section précédente.

Définition 10 *On appelle fonction d'autocorrélation partielle la fonction $r : \mathbb{N} \to \mathbb{R}$ définie par $r(0) = 1$, $r(1) = \rho(1)$ et pour tout entier $k \geq 2$, $r(k)$ désigne la corrélation entre $X_t - P_{F_{t,k-1}}(X_t)$ et $X_{t-k} - P_{F_{t,k-1}}(X_{t-k})$*

La fonction d'autocorrélation partielle mesure donc la liaison linéaire entre X_t et X_{t-k} lorsqu'on retranche la projection sur les variables intermédiaires. Une expression différente de cette fonction est donnée dans la proposition suivante.

Proposition 4 *Supposons $k \geq 1$. Alors si $b_{k,1}X_{t-1} + \cdots + b_{k,k}X_{t-k}$ désigne la projection de X_t sur $F_{t,k} = Vect_{\mathbb{L}^2}(X_{t-1}, \ldots, X_{t-k})$, on a $r(k) = b_{k,k}$.*

Preuve

C'est clair si $k = 1$ (exercice). Si $k \geq 2$, on a, en reprenant les notations de la définition précédente,

$$p_{F_{t,k}}(X_t) = \sum_{j=1}^{k} b_{k,j}X_{t-j} = \sum_{j=1}^{k-1} b_{k,j}X_{t-j} + b_{k,k}\left(X_{t-k} - P_{F_{t,k-1}}(X_{t-k})\right) + b_{k,k}P_{F_{t,k-1}}(X_{t-k}).$$

En projetant sur $F_{t,k-1}$, on en déduit que $p_{F_{t,k-1}}(X_t) = \sum_{j=1}^{k-1} b_{k,j}X_{t-j} + b_{k,k}P_{F_{t,k-1}}(X_{t-k})$ (puisque $X_{t-k} - P_{F_{t,k-1}}(X_{t-k})$ est orthogonal à $X_{t-1}, \ldots, X_{t-k+1}$). On obtient

$$\begin{aligned} & \mathrm{Cov}\left(X_t - p_{F_{t,k-1}}(X_t), X_{t-k} - P_{F_{t,k-1}}(X_{t-k})\right) \\ = \; & \mathrm{Cov}\left(X_t - p_{F_{t,k}}(X_t), X_{t-k} - P_{F_{t,k-1}}(X_{t-k})\right) + b_{k,k}\mathrm{Var}\left(X_{t-k} - P_{F_{t,k-1}}(X_{t-k})\right) \\ = \; & b_{k,k}\mathrm{Var}\left(X_{t-k} - P_{F_{t,k-1}}(X_{t-k})\right). \end{aligned}$$

Dans la sous-section précédente, on a vu que $\mathrm{Var}\left(X_{t-k} - P_{F_{t,k-1}}(X_{t-k})\right) = \mathrm{Var}\left(X_t - P_{F_{t,k-1}}(X_t)\right)$. Le résultat annoncé est alors immédiat.□

Algorithme de Durbin

La proposition précédente montre que $r(k)$ correspond au dernier coefficient du vecteur $\Gamma_k^{-1}\begin{pmatrix}\gamma(1)\\ \vdots \\ \gamma(k)\end{pmatrix}$.

On peut aussi obtenir des formules récursives qui évitent de calculer les matrices inverses de Γ_k. Conservons les notations précédentes. On a vu que

$$P_{F_{t,k-1}}(X_t) = \sum_{j=1}^{k-1} b_{k,j}X_{t-j} + b_{k,k}P_{F_{t,k-1}}(X_{t-k}).$$

Comme

$$P_{F_{t,k-1}}(X_t) = \sum_{j=1}^{k-1} b_{k-1,j}X_{t-j}, \quad P_{F_{t,k-1}}(X_{t-k}) = \sum_{j=1}^{k-1} b_{k-1,k-j}X_{t-j},$$

on trouve les relations

$$b_{k,j} = b_{k-1,j} - b_{k,k} b_{k-1,k-j}, \quad j = 1, \dots, k-1.$$

On obtient ensuite une deuxième relation à partir de l'égalité

$$\Gamma_k \begin{pmatrix} b_{k,1} \\ \vdots \\ b_{k,k} \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(k) \end{pmatrix}.$$

La dernière ligne de ce système entraine que

$$\sum_{j=1}^{k} b_{k,j} \gamma(k-j) = \gamma(k).$$

On obtient au final

$$b_{k,j} = b_{k-1,j} - b_{k,k} b_{k-1,k-j}, \quad j = 1, \dots, k-1, \text{ et } b_{k,k} = \frac{\gamma(k) - \sum_{j=1}^{k-1} \gamma(k-j) b_{k-1,j}}{\gamma(0) - \sum_{j=1}^{k-1} \gamma(j) b_{k-1,j}}. \tag{2.3}$$

On peut aussi remplacer les autocovariances par les autocorrélations dans la deuxième formule.

Exemple

Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2 et $X_t = \varepsilon_t - \theta\varepsilon_{t-1}$ (moyenne mobile d'ordre 1) Alors la fonction d'autocorrélation ρ est donnée par $\rho(0) = 1$, $\rho(1) = \frac{-\theta}{1+\theta^2}$ et $\rho(h) = 0$ si $h \geq 2$. Pour le calcul de r, on a pour $k \geq 2$,

$$b_{k,1} = b_{k-1,1} - b_{k,k} b_{k-1,k-1}, \quad b_{k,k} = \frac{-\rho(1) b_{k-1,k-1}}{1 - \rho(1) b_{k-1,1}}.$$

On peut en fait calculer explicitement $b_{k,k}$ sans utiliser cet algorithme (voir TD).

2.6 Autocorrélations et autocorrélations partielles empiriques

On note $\bar{X}_T = \frac{1}{T} \sum_{t=1}^{T} X_t$ la moyenne empirique. La fonction d'autocovariance empirique $\hat{\gamma}$ est définie par

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=h+1}^{T} \left(X_t - \bar{X}_T\right) \cdot \left(X_{t-h} - \bar{X}_T\right)$$

lorsque h est positif. Lorsque le processus X est de carré intégrable et vérifie les hypothèses sur Théorème 2 alors $\lim_{T\to+\infty} \hat{\gamma}(h) = \gamma(h)$ p.s. De même la fonction d'autocorrélation empirique est définie par $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$. Les autocorrélations partielles empiriques sont obtenues en remplaçant les

autocorrélations par les autocorrélations empiriques dans l'algorithme de Durbin ou en considérant la dernière coordonnée du vecteur

$$\hat{\Gamma}_k^{-1} \begin{pmatrix} \hat{\gamma}(1) \\ \vdots \\ \hat{\gamma}(k) \end{pmatrix}.$$

Le graphe de la fonction d'autocorrélation (ACF) empirique (on parle d'autocorrélogramme) et de la fonction des autocorrélations partielles (PACF) empiriques sont fondamentaux. Ils permettent en pratique de voir quel(s) type(s) de processus stationnaire il convient d'ajuster, ce qui ne peut se déduire à partir du graphe de la série.

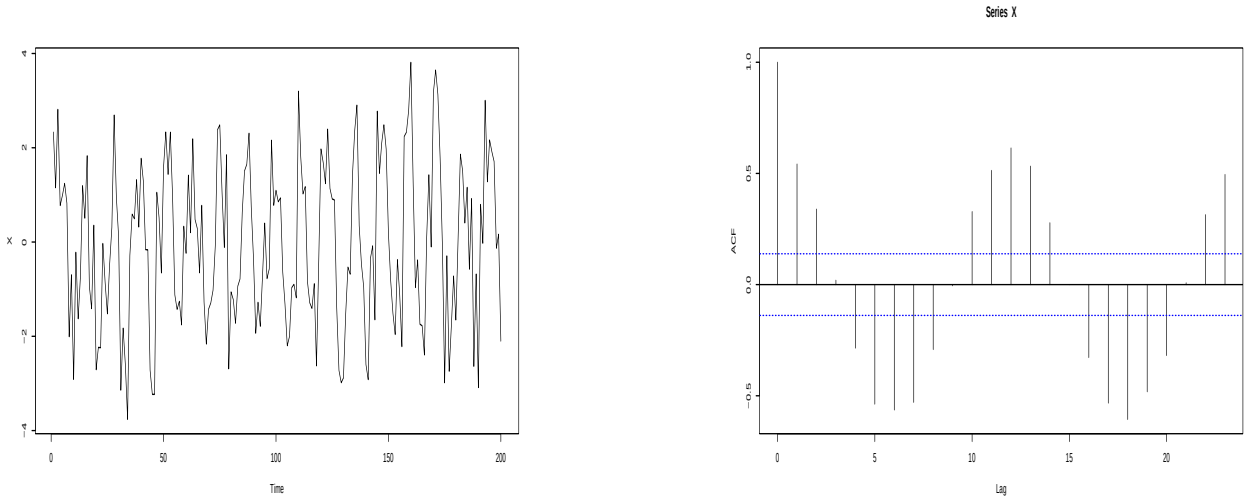


FIGURE 2.2 – Autocorrélogramme d'une série somme d'un processus i.i.d $\mathcal{N}(0,1)$ et d'une composante déterministe de période 12. La série n'est pas stationnaire et on obtient un graphe typique.

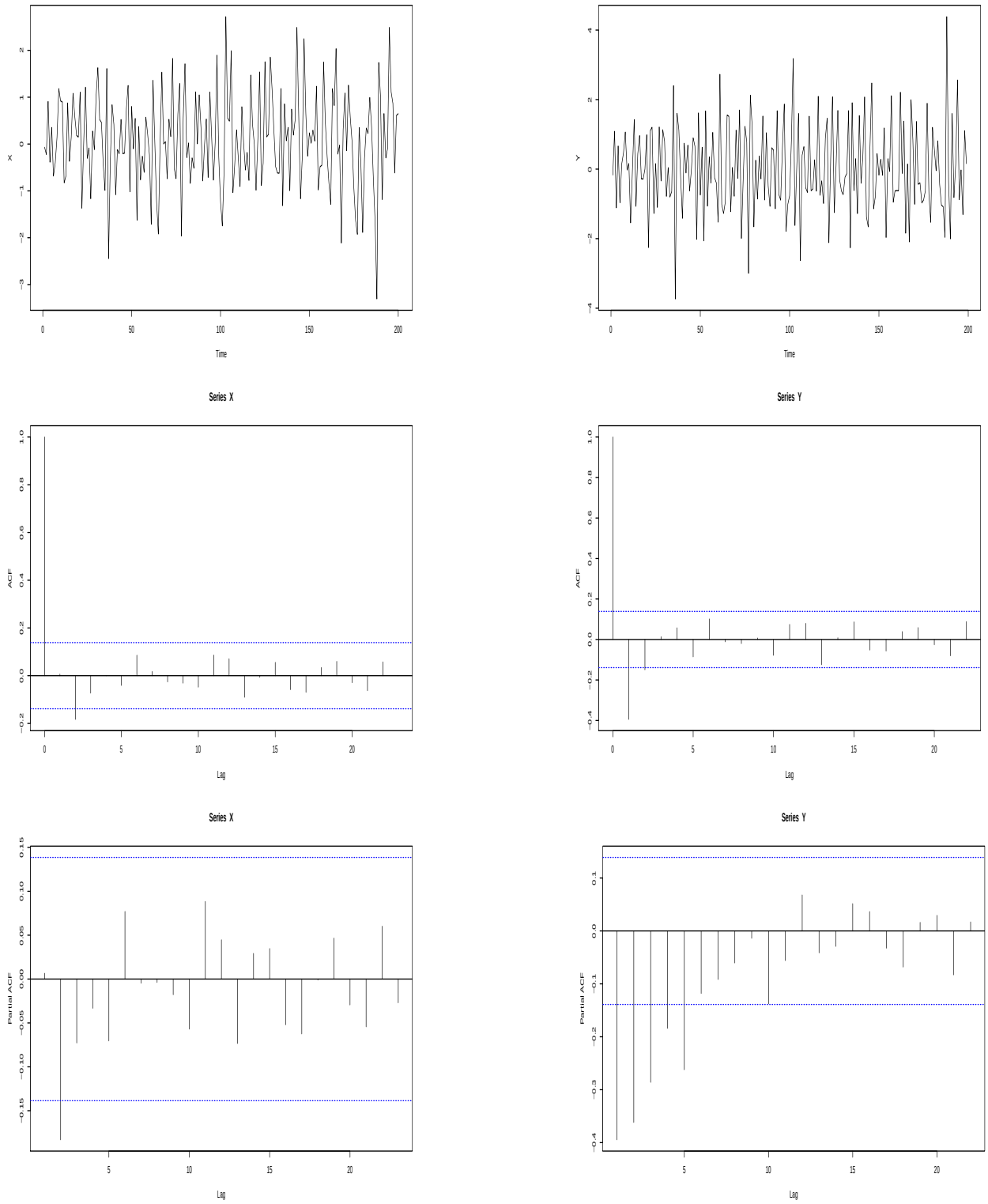


FIGURE 2.3 – Différence entre la série $X_t = \varepsilon_t$ et la série $X_t = \varepsilon_t - \theta\varepsilon_{t-1}$ lorsque $(\varepsilon_t)_t$ est une suite i.i.d $\mathcal{N}(0,1)$

2.7 Théorème de représentation de Wold

Soit $(X_t)_{t\in\mathbb{Z}}$ une série temporelle telle que $\mathbb{E}(X_t) = 0$ et $\text{Var}(X_t) < +\infty$ pour tout $t \in \mathbb{Z}$. Pour tout $t \in \mathbb{Z}$, on pose

$$F_t = \overline{\text{Vect}}(X_{t-1}, X_{t-2}, \ldots).$$

En théorie, la meilleure prévision linéaire de X_t à partir des variables $X_{t-1}, X_{t-2}, \ldots$ est donnée par $P_{F_t}(X_t)$ (projection linéaire de X_t sur F_t). On appelle alors processus d'innovation le processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ formé des erreurs de prévision, i.e $\varepsilon_t = X_t - P_{F_t}(X_t)$. A titre d'exercice, on pourra vérifier que le processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc. Lorsque le processus d'innovation est nul ($\varepsilon_t = 0$ p.s pour tout $t \in \mathbb{Z}$), on dit que le processus $(X_t)_{t\in\mathbb{Z}}$ est déterministe. Tout processus déterministe est donc parfaitement prévisible à l'aide de ces valeurs passées. C'est le cas par exemple, s'il existe une variable aléatoire Y telle que $X_t = Y$ pour tout $t \in \mathbb{Z}$ ou encore si $X_t = (-1)^t Y$ pour tout $t \in \mathbb{Z}$.

Théorème 3 *Si $(X_t)_{t\in\mathbb{Z}}$ est un processus centré et stationnaire au second ordre, alors on a la décomposition*

$$X_t = \sum_{j=0}^{+\infty} \psi_j \varepsilon_{t-j} + V_t, \quad t \in \mathbb{Z}$$

où

- $\psi_0 = 1$ *et* $\sum_{j=0}^{+\infty} \psi_j^2 < +\infty$.
- $(\varepsilon_t)_{t\in\mathbb{Z}}$ *est le processus d'innovation de* $(X_t)_{t\in\mathbb{Z}}$.
- $(V_t)_{t\in\mathbb{Z}}$ *est un processus déterministe.*
- $Cov(\varepsilon_t, V_s) = 0, \forall\ s, t \in \mathbb{Z}$.

Le théorème précédent montre que les moyennes mobiles d'ordre infini sont intéressantes puisque très représentatives des processus stationnaires (la décomposition fait apparaître un terme déterministe que nous ne rencontrerons pas dans la suite). Mais ce résultat informatif est peu utile en pratique puisque les coefficients ψ_j sont inconnus. On peut approcher la moyenne mobile d'ordre infini par une moyenne mobile d'ordre fini. Ce sera parfois pertinent mais nous verrons que certaines modélisations font intervenir une infinité de termes. Quoi qu'il en soit, il est indispensable d'utiliser en pratique des représentations parcimonieuses de ces moyennes mobiles (pour avoir un nombre raisonnable de coefficients à estimer). Tout les processus stationnaires que nous étudierons par la suite admettrons des représentations sous la forme de moyenne mobile d'ordre infini.

2.8 Représentation spectrale

Il y a au moins deux façons de voir un processus stationnaire au second ordre. La première est de travailler dans le domaine temporel en considérant les corrélations entre X_t et X_{t+h}. La deuxième approche est celle des fréquences : on considère les composantes périodiques (aléatoires) qui composent le processus (traitement d'un "signal"). Ces deux approches permettent d'obtenir les mêmes résultats mais il est parfois plus commode d'utiliser l'une ou l'autre. Dans cette section, nous n'introduirons que quelques notions de base qui permettent de travailler dans le domaine des fréquences.

Commençons par un exemple. Soient A et B deux variables aléatoires centrées, de même variance σ^2 et décorrélées. Posons

$$X_t = A\cos(t\alpha) + B\sin(t\alpha), \quad t \in \mathbb{Z}.$$

On pourra vérifier que $(X_t)_t$ est stationnaire au second ordre et que $\mathrm{Cov}(X_t, X_{t+h}) = \sigma^2\cos(\alpha h)$. Les trajectoire du processus $(X_t)_{t\in\mathbb{Z}}$ sont périodiques. On peut généraliser cet exemple. Si $\left\{A_j, B_j : 1 \le j \le k\right\}$ est une famille de variables aléatoires centrées, décorrélées et telles que $\mathrm{Var}\left(A_j\right) = \mathrm{Var}\left(B_j\right) = \sigma_j^2$ pour $1 \le j \le k$ et $-\pi \le \alpha_1 < \alpha_2 < \cdots < \alpha_k \le \pi$ sont des fréquences données, alors

$$X_t = \sum_{j=1}^{k}\left[A_j\cos\left(t\alpha_j\right) + B_j\sin\left(t\alpha_j\right)\right], \quad t \in \mathbb{Z},$$

définit un processus faiblement stationnaire. De plus en posant $Z_1(\alpha) = \sum_{j=1}^{k} A_j \mathbb{1}_{\alpha\le\alpha_j}$, et $Z_2(\alpha) = \sum_{j=1}^{k} B_j \mathbb{1}_{\alpha\le\alpha_j}$, on voit que

$$X_t = \int_{-\pi}^{\pi}\cos(t\alpha)dZ_1(\alpha) + \int_{-\pi}^{\pi}\sin(t\alpha)dZ_2(\alpha).$$

On peut montrer qu'un processus stationnaire arbitraire peut être représenté comme une moyenne de signaux périodiques de différentes fréquences α. Dans la suite, nous ne considérons que le cas des moyennes mobile d'ordre infini,

$$X_t = \sum_{i\in\mathbb{Z}}\theta_i\varepsilon_{t-i}, \quad \sum_{i\in\mathbb{Z}}|\theta_i| < +\infty. \tag{2.4}$$

Les résultats seront énoncés pour des processus stationnaires centrés.

Définition 11 *Un processus stochastique centré $\{Z(\alpha) : -\pi \le \alpha \le \pi\}$ est dit à accroissements orthogonaux si pour tous $-\pi \le t_1 < t_2 \le t_3 < t_4 \le \pi$, les variables aléatoires $Z(t_4) - Z(t_3)$ et $Z(t_2) - Z(t_1)$ sont décorrélées.*

Théorème 4 *On a la représentation*

$$X_t = \int_{-\pi}^{\pi}\cos(t\alpha)\,dZ_1(\alpha) + \int_{-\pi}^{\pi}\sin(t\alpha)\,dZ_2(\alpha),$$

où les processus Z_1 et Z_2 sont à accroissements orthogonaux et décorrélées entre eux. De plus, on a pour $i = 1, 2$ et $\alpha < \beta$,

$$Var\,(Z_i(\beta) - Z_i(\alpha)) = \int_{\alpha}^{\beta} f(u)du,$$

oú $f : [-\pi, \pi] \to \mathbb{R}_+$ est la fonction donnée par

$$\begin{aligned} f(\alpha) &= \frac{1}{2\pi}\sum_{h\in\mathbb{Z}}\gamma(h)\exp(i\alpha h) \\ &= \frac{1}{2\pi}\left\{\gamma(0) + 2\sum_{h=1}^{+\infty}\gamma(h)\cos(\alpha h)\right\}. \end{aligned}$$

Enfin, nous avons la représentation

$$\gamma(h) = \int_{-\pi}^{\pi} f(\alpha) \exp(ih\alpha)\, d\alpha = 2\int_{0}^{\pi} f(\alpha)\cos(h\alpha)\, d\alpha.$$

Il faut retenir l'idée de ce théorème difficile. Un processus stationnaire s'obtient en moyennant des composantes périodiques de différentes fréquences α. Le poids (aléatoire) accordé à une fréquence α a une variance donnée par $f(\alpha)$. La fonction f mesure donc l'importance des composantes périodiques de fréquence donnée dans la décomposition du signal. Plus $f(\alpha)$ est important plus la composante de fréquence α a un rôle important dans la décomposition de X_t. Enfin, la fonction d'autocovariance s'écrit comme la transformée de Fourier de la fonction f. La fonction f est appelée densité spectrale du processus (c'est une fonction paire). Elle peut être estimée directement à partir des données en utilisant les autocovariances empiriques. En pratique, l'estimation de la densité spectrale sert à identifier des cycles (périodicités dans les séries) et aussi à construire des tests (par exemple pour tester si la série est un bruit blanc car dans ce cas la densité spectrale est constante).
Donnons une autre expression de la densité spectrale d'une moyenne mobile d'ordre infinie $X_t = \sum_{j\in\mathbb{Z}} \theta_j \varepsilon_{t-j}$ où $(\varepsilon_i)_{i\in\mathbb{Z}}$ est un bruit blanc de variance σ^2. On a déjà vu que $\gamma(h) = \sigma^2 \sum_{j\in\mathbb{Z}} \theta_j\theta_{j+h}$. Nous obtenons alors

$$f(\alpha) = \frac{\sigma^2}{2\pi}|\Theta\left(e^{i\alpha}\right)|^2,$$

où $\Theta(z) = \sum_{j\in\mathbb{Z}} \theta_j z^j$, $z \in \mathbb{C}$.

Exemples

- Pour un bruit blanc, on a $f \equiv \frac{\sigma^2}{2\pi}$. Réciproquement, on pourra montrer que si $f \equiv \frac{\sigma^2}{2\pi}$ alors le processus est un bruit blanc de variance σ^2 (utiliser l'expression de la covariance en terme de transformée de Fourier).
- Supposons que $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$ où $|\theta| < 1$. D'après l'expression générale de la densité spectrale, il suffit de calculer

  $$|1 + \theta e^{i\alpha}|^2 = 1 + \theta^2 + 2\theta\cos(\alpha).$$

  On a donc

  $$f(\alpha) = \frac{\sigma^2}{2\pi}\left(1 + \theta^2 + 2\theta\cos(\alpha)\right).$$

- Considérons la moyenne mobile infinie

  $$X_t = \sum_{j=0}^{+\infty} \phi^j \varepsilon_{t-j},$$

  où $|\phi| < 1$. On obtient

  $$f(\alpha) = \frac{\sigma^2}{2\pi}\frac{1}{|1-\phi e^{i\alpha}|^2} = \frac{\sigma^2}{2\pi}\frac{1}{1+\phi^2-2\phi\cos(\alpha)}.$$

  Remarquons l'égalité $X_t = \phi X_{t-1} + \varepsilon_t$.

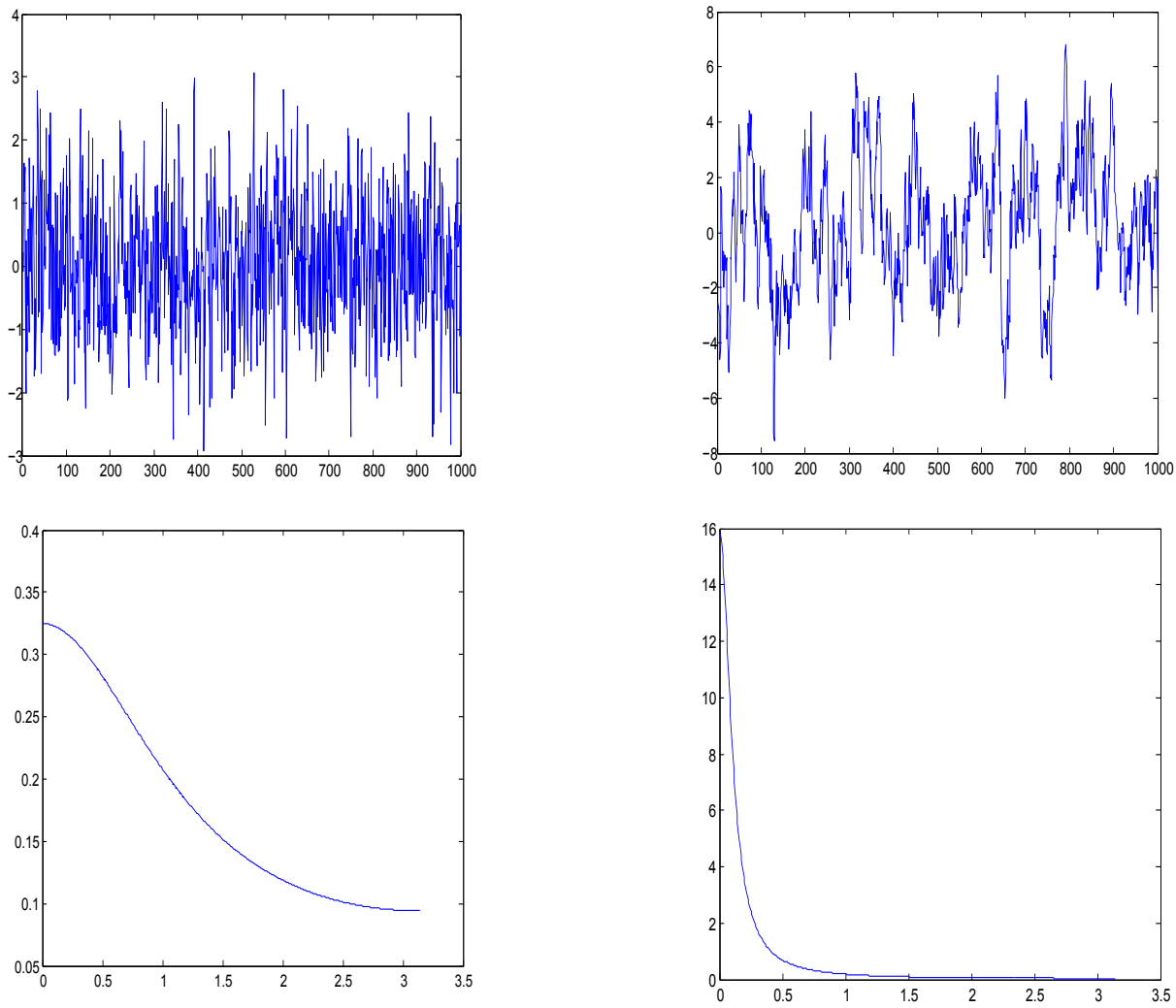


FIGURE 2.4 – Trajectoire d'une moyenne mobile infinie lorsque le bruit blanc est gaussien et $\phi = 0.3$ (à gauche) ou $\phi = 0.9$ (à droite). Les graphes du bas représentent la densité spectrale correspondante (entre $\alpha = 0$ et $\alpha = \pi$). Le cas $\phi = 0.9$ laisse apparaitre une prédominance des basses fréquences.

La proposition suivante, que l'on pourra démontrer en exercice, est très utile pour le calcul de la densité spectrale d'une transformation linéaire d'un processus stationnaire.

Proposition 5 *Soit $(X_t)_{t\in\mathbb{Z}}$ stationnaire au second ordre pouvant s'écrire sous la forme (2.4) et de densité spectrale f_X. Si $Y_t = \sum_{j\in\mathbb{Z}} a_j X_{t-j}$ avec $\sum_{j\in\mathbb{Z}} |a_j| < +\infty$. La densité spectrale f_Y du processus $(Y_t)_{t\in\mathbb{Z}}$ est donnée par*

$$f_Y(\alpha) = f_X(\alpha)\left|\sum_{j\in\mathbb{Z}} a_j e^{ij\alpha}\right|^2 .$$

2.9 Les processus ARMA

A partir de maintenant, on notera 1 pour l'application identité I définie sur l'ensemble des suites indexées par $\mathbb{Z}$ et à valeurs réelles. On rappelle que B est l'opérateur retard défini par $B(x)_t = x_{t-1}$. Dans la suite la notation Bx_t sera utilisée pour désigner $B(x)_t$.

2.9.1 Inversion d'opérateurs

Considérons l'opérateur

$$\Phi(B) = \sum_{i=0}^{p} \phi_i B^i = 1 + \phi_1 B + \ldots + \phi_p B^p,$$

où les ϕ_i sont des nombres réels avec $\phi_0 = 1$. Soit $\mathcal{B} = \{(x_t)_{t\in\mathbb{Z}} : \sup_{t\in\mathbb{Z}} |x_t| < +\infty\}$. Alors $\Phi(B)$ envoie $\mathcal{B}$ dans lui-même. Nous aurons besoin de définir l'inverse d'un tel opérateur. On a la décomposition

$$\Phi(B) = \prod_{i=1}^{p} (1 - \lambda_i B),$$

où $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$ sont les racines (complexes) du polynôme Φ. Un opérateur du type $1 - \lambda B$ est inversible si et seulement si $|\lambda| \neq 1$. On a plus précisément

$$(1 - \lambda B)^{-1} = \begin{cases} \sum_{i=0}^{+\infty} \lambda^i B^i \text{ si } |\lambda| < 1 \\ -\sum_{i=1}^{+\infty} \lambda^{-i} B^{-i} \text{ si } |\lambda| > 1 \end{cases}$$

Lorsque toutes les racines de Φ sont de module différent de 1, on a

$$\Phi(B)^{-1} = \prod_{i=1}^{p} (1 - \lambda_i B)^{-1}$$

en remarquant que les différentes opérateurs commutent deux à deux (l'ordre n'a donc pas d'importance). Ces considérations permettent de prouver la proposition suivante.

Proposition 6 *L'opérateur $\Phi(B)$ est inversible si et seulement si toutes les racines de Φ ont un module différent de 1. Dans ce cas*

$$\Phi(B)^{-1} = \sum_{j=-\infty}^{+\infty} \psi_j B^j,$$

où $\sum_{j=-\infty}^{+\infty} |\psi_j| < +\infty$. On a $\psi_j = 0$ pour tout $j < 0$ si et seulement si toutes les racines de Φ sont de module strictement plus grand que 1 (on dit aussi toutes les racines à l'extérieur du disque unité).

Remarque. On peut aussi voir Φ comme une application polynomiale définie sur $\mathbb{C}$ le corps des nombres complexes. Dans ce cas, la relation $\Phi(B) \cdot \sum_{j\in\mathbb{Z}} \psi_j B^j = 1$ (composition des opérateurs) se traduit par

$$\Phi(z) \times \sum_{j\in\mathbb{Z}} \psi_j z^j = 1, \quad |z| = 1.$$

On a donc $\sum_{j\in\mathbb{Z}} \psi_j z^j = \frac{1}{\Phi(z)}$ lorsque z est sur le cercle unité.

2.9.2 Les moyennes mobiles

Définition 12 *Soient $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc et $q \in \mathbb{N}^*$. Tout processus $(X_t)_{t\in\mathbb{Z}}$ de la forme*

$$X_t = m + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \quad t \in \mathbb{Z},$$

avec $m, \theta_i \in \mathbb{R}$, sera appelé processus moyenne mobile d'ordre q et est noté $MA(q)$.

Définition 13 *On dit qu'une moyenne mobile d'ordre q est inversible si le polynôme Θ tel que $\Theta(B) = 1 - \sum_{i=1}^{q} \theta_i B^i$ a toutes ses racines à l'extérieur du disque unité.*

Remarques

1. Lorsqu'une moyenne mobile est inversible, on peut écrire $\varepsilon_t = \Theta(B)^{-1}(X_t - m)$. Il est donc possible de reconstruire ε à partir d'une combinaison linéaire (infinie) des valeurs passées de X (recentrées si $m \neq 0$). Les coefficients de

$$\Theta(B)^{-1} = \sum_{i=0}^{+\infty} \eta_i B^i$$

peuvent être déterminés en résolvant $\Theta(B) \sum_{i=0}^{+\infty} \eta_i B^i = 1$ ce qui donne lieu à un système d'équations pouvant se résoudre de façon récursive. En se rappelant du théorème de Wold, le processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ est le processus d'innovation du processus $(X_t)_{t\in\mathbb{Z}}$.

2. Si le polynôme Θ a certaines de ses racines à l'intérieur du disque unité, on peut montrer que quitte à changer le bruit blanc $(\varepsilon_t)_{t\in\mathbb{Z}}$ et les coefficients θ_j, on peut se ramener à une moyenne mobile inversible. Ceci est fondamental car dans la formulation intiale ε_t ne correspond pas à l'innovation. Par exemple si $X_t = \Theta(B)\varepsilon_t$ où

$$\Theta(B) = (1 - z_1 B) \cdots \left(1 - z_q B\right)$$

avec $z_1^{-1}, \ldots, z_q^{-1}$ les racines de Θ. Supposons que $|z_k| > 1$ pour $1 \leq k \leq r < q$. La densité

spectrale de X est donnée par la formule

$$\begin{aligned} f_X(\alpha) &= \frac{\sigma^2}{2\pi}\prod_{j=1}^{q}|1-z_j e^{i\alpha}|^2 \\ &= \frac{\sigma^2\prod_{j=1}^{r}|z_j|^2}{2\pi}\prod_{j=1}^{r}|1-z_j^{-1}e^{i\alpha}|^2\prod_{j=r+1}^{q}|1-z_j e^{i\alpha}|^2 \\ &= \frac{\sigma^2\prod_{j=1}^{r}|z_j|^2}{2\pi}\left|\Theta_*\left(e^{i\alpha}\right)\right|^2. \end{aligned}$$

Le polynôme Θ_* a toutes ses racines a l'extérieur du disque unité. En posant $\eta_t = \Theta_*(B)^{-1}X_t$ on obtient que

$$f_\eta(\alpha) = \frac{f_X(\alpha)}{|\Theta_*(e^{i\alpha})|^2} = \frac{\sigma^2\prod_{j=1}^{r}|z_j|^2}{2\pi}.$$

Donc η est un bruit blanc de variance $\sigma^2\prod_{j=1}^{r}|z_j|^2$ et $X_t = \Theta_*(B)\eta_t$ est une moyenne mobile d'ordre q inversible. Cependant, si $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blance fort, ce n'est plus le cas de $(\eta_t)_{t\in\mathbb{Z}}$ en général.

3. On rappelle que si $(X_t)_{t\in\mathbb{Z}}$ est une moyenne mobile d'ordre q, alors

$$\text{Var}\,(X_t) = \sigma^2\left(1+\theta_1^2+\cdots+\theta_q^2\right)$$

et sa fonction d'autocovariance est donnée par

$$\text{Cov}\,(X_t, X_{t+h}) = \sigma^2\left(-\theta_{|h|}+\theta_{|h|+1}\theta_1+\cdots+\theta_q\theta_{q-|h|}\right)\mathbb{1}_{|h|\le q}.$$

En particulier les autocorrélations vérifient $\rho(h) = 0$ pour $|h| \ge q+1$. On peut montrer que cette dernière propriété est caractéristique des processus $MA(q)$: tout processus stationnaire au second ordre tel que $\rho(h) = 0$ pour $|h| \ge q+1$ peut s'écrire comme un processus moyenne mobile d'ordre q.

4. Les autocorrélations partielles d'un processus $MA(q)$ sont plus difficiles à exprimer. En général, elles ne sont pas nulles à partir d'un certain rang mais on sait qu'elles décroissent exponentiellement vite vers 0 avec h.

2.9.3 Les processus AR

Définition 14 *On appelle processus autorégressif d'ordre p, noté AR(p), tout processus stationnaire $(X_t)_{t\in\mathbb{Z}}$ qui satisifait aux équations*

$$X_t = c + \sum_{i=1}^{p}\phi_i X_{t-i} + \varepsilon_t, \quad \forall t\in\mathbb{Z},$$

où $c, \phi_i \in \mathbb{R}$ et $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc de variance σ^2.

Proposition 7 *Lorsque toutes les racines du polynôme Φ tel que $\Phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$ sont à l'extérieur du disque unité, il existe une unique solution stationnaire $(X_t)_{t\in\mathbb{Z}}$ au système d'équations donné dans la définition précédente. Cette solution est une moyenne mobile infinie (MA(∞)) de la forme $X_t = m + \sum_{i=0}^{+\infty} \psi_i \varepsilon_{t-i}$ avec $\sum_{i=0}^{+\infty} |\psi_i| < +\infty$. On a en particulier $\sum_{i=0}^{+\infty} \psi_i B^i = \Phi(B)^{-1}$.*

On déduit de cette proposition que lorsque toutes les racines de Φ sont situées à l'extérieur du disque unité, la variable aléatoire $c + \sum_{i=1}^{p} \phi_i X_{t-i}$ est la projection de X_t sur le sous-espace vectoriel fermé engendré par $1, X_{t-1}, X_{t-2}, \ldots$. Lorsque $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc fort, on a

$$\mathbb{E}\left(X_t | \sigma\left(X_{t-1}, X_{t-2}, \ldots\right)\right) = c + \sum_{i=1}^{q} \phi_i X_{t-i}.$$

Remarquons aussi que $m = \mathbb{E}(X_t) = \frac{c}{1-\sum_{j=1}^{p} \phi_j}$.

Exemple : le processus AR(1)

$$X_t = c + \phi X_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z}.$$

– Plaçons nous d'abord dans le cas $|\phi| < 1$. Il existe alors une unique solution stationnaire. Posons $m = \mathbb{E}(X_t)$. Comme

$$(1 - \phi B)^{-1} = \sum_{j=0}^{+\infty} \phi^j B^j,$$

on a

$$X_t = \sum_{j=0}^{+\infty} \phi^j \left(\varepsilon_{t-j} + c\right) = \frac{c}{1-\phi} + \sum_{j=0}^{+\infty} \phi^j \varepsilon_{t-j}.$$

D'après l'équation, on a $m = \frac{c}{1-\phi}$. On peut calculer la fonction d'autocovariance γ à partir de la solution MA(∞). Mais on peut aussi remarquer que si $h > 0$,

$$\gamma(h) = \mathrm{Cov}\left(X_{t+h}, X_t\right) = \phi \mathrm{Cov}\left(X_{t+h-1}, X_t\right) = \phi\gamma(h-1).$$

On en déduit que $\gamma(h) = \phi^h \gamma(0)$ et $\gamma(0) = \phi^2\gamma(0) + \sigma^2$ (toujours d'après l'équation). On en déduit que

$$\gamma(h) = \phi^h \frac{\sigma^2}{1-\phi^2}, \quad h \geq 0.$$

– Si maintenant $|\phi| = 1$, par exemple $\phi = 1$, il n'existe pas de solution stationnaire. En effet, on a dans ce cas pour $t > 0$,

$$X_t - X_0 = \sum_{i=1}^{t} (X_i - X_{i-1}) = tc + \sum_{i=1}^{t} \varepsilon_i.$$

On en déduit que $\mathrm{Var}\,(X_t - X_0) = \sigma^2 t$ est non borné en t. On peut aisément montrer que $\mathrm{Var}\,(X_t - X_0)$ est toujours borné en t pour un processus stationnaire.

– Enfin, supposons que $|\phi| > 1$. A partir de la relation

$$(I - \phi B)^{-1} = -\sum_{i=1}^{+\infty} \phi^{-i} B^{-i},$$

on obtient

$$X_t = -\sum_{i=1}^{+\infty} \phi^{-i} (c + \varepsilon_{t+i}), \quad t \in \mathbb{Z}.$$

Cette dernière repésentation n'est pas naturelle car (ε_t) n'est plus le processus d'innovation de $(X_t)_{t\in\mathbb{Z}}$. On peut alors changer le bruit blanc (comme pour les moyennes mobiles non inversibles). La densité spectrale f_X vérifie

$$|1 - \phi e^{i\alpha}|^2 f_X(\alpha) = \frac{\sigma^2}{2\pi}.$$

Comme

$$|1 - \phi e^{i\alpha}| = |\phi| \cdot |1 - \phi^{-1} e^{i\alpha}|,$$

on pose $\eta_t = X_t - \frac{1}{\phi} X_{t-1}$ et on a

$$f_\eta(\alpha) = \frac{\sigma^2}{2\pi|\phi|}.$$

On se retrouve dans le cas standard où $(\eta_t)_t$ est le processus d'innovation de $(X_t)_t$ (mais on peut perdre l'hypothèse initiale de bruit blanc fort).

Fonction d'autocorrélation d'un AR(p) et équations de Yule-Walker. Soit un processus AR(p) stationnaire défini par

$$X_t = c + \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t, \quad t \in \mathbb{Z},$$

et pour lequel les racines de Φ sont à l'extérieur du disque unité. Si ρ désigne la fonction d'autocorrélation, on a

$$\rho(h) = \sum_{j=1}^{p} \phi_j \rho(h - j), \quad h = 1, 2, \ldots.$$

On sait d'après le résultat sur les suites linéaires récurrentes que $\rho(h)$ s'obtient en cherchant les racines λ non nulles de l'équation

$$\lambda^p - \sum_{j=1}^{p} \phi_j \lambda^{p-j} = 0 \Longleftrightarrow 1 - \sum_{j=1}^{p} \phi_j \lambda^{-j} = 0.$$

Donx $1/\lambda$ est racine de Φ, ce qui entraine $|\lambda| < 1$. On en déduit que les autocorrélations du processus AR convergeront toujours vers 0 à vitesse exponentielle lorsque $h \to +\infty$. Remarquons que matriciellement, on obtient

$$\begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} = [\rho(i-j)]^{-1}_{1\le i,j\le p} \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(p) \end{pmatrix}.$$

On obtient donc une première façon d'estimer en pratique les coefficients ϕ_j du modèle (il suffit de remplacer les autocorrélations par les autocorrélations empiriques). Remarquons que la variance $\gamma(0)$ de la série est donnée par

$$\begin{aligned}\gamma(0) &= \sum_{j=1}^{p} \phi_j \mathrm{Cov}\left(X_t, X_{t-j}\right) + \mathrm{Cov}\left(X_t, \varepsilon_t\right) \\ &= \sum_{j=1}^{p} \phi_j \gamma(j) + \sigma^2.\end{aligned}$$

Fonction d'autocorrélation partielle d'un processus AR(p). Soit un processus AR(p) défini par

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t, \quad t \in \mathbb{Z},$$

pour lequel les racines de Φ sont à l'extérieur du disque unité. Les autocorrélations partielles peuvent se calculer à l'aide de l'algorithme de Durbin. Quitte à centrer le processus, on supposera que $c = 0$. On voit alors que $r(h) = 0$ si $h \geq p + 1$ (la projection de X_t sur $Vect\,(X_{t-1}, \ldots, X_{t-h})$ est $\sum_{i=1}^{p} \phi_i X_{t-i}$ et le coefficient de X_{t-h} est nul) et que $r(p) = \phi(p)$. Les autocorrélations partielles empiriques permettent donc d'avoir une idée de l'ordre p d'un processus AR (il suffit de voir si les autocorrélations partielles empiriques semblent négligeables lorsque $h \geq p + 1$).

2.9.4 Les processus ARMA

Définition 15 *On dira qu'un processus $(X_t)_{t\in\mathbb{Z}}$ est un processus ARMA(p, q) si*

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = c + \varepsilon_t - \sum_{h=1}^{q} \theta_h \varepsilon_{t-h}, \quad t \in \mathbb{Z},$$

où

- *le processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc de variance σ^2,*
- *ϕ_p et θ_q sont non nuls,*
- *les polynômes $\Phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$ et $\Theta(B) = 1 - \sum_{h=1}^{q} \theta_h B^h$ ont toutes leurs racines de module différent de* 1.

Lorsque les racines des polynômes $\Phi(B)$ et $\Theta(B)$ sont toutes à l'extérieur du disque unité et que les deux polynômes n'ont pas de racine commune, on dira que la représentation ARMA est minimale.

Lorsque $\Phi(B)$ et $\Theta(B)$ ont des racines en commun, on peut diminuer les p et q de l'ARMA.

Exercice 7 On suppose que $X_t - \phi X_{t-1} = c + \varepsilon_t - \phi\varepsilon_{t-1}$ avec $|\phi| < 1$. Montrer que $X_t = \frac{c}{1-\phi} + \varepsilon_t$.

Exemple. Considérons un processus ARMA(1, 1) de représentation minimale

$$X_t - \phi X_{t-1} = \varepsilon_t - \theta\varepsilon_{t-1}.$$

On a donc $|\phi| < 1$ et $|\theta| < 1$. Déterminons le développement en moyenne mobile infinie. On a

$$\begin{aligned} X_t &= \sum_{j=0}^{+\infty} \phi^j \left(\varepsilon_{t-j} - \theta\varepsilon_{t-j-1}\right) \\ &= \varepsilon_t + \sum_{j=1}^{+\infty} \left(\phi^j - \theta\phi^{j-1}\right)\varepsilon_{t-j} \\ &= \varepsilon_t + (\phi-\theta)\sum_{j=1}^{+\infty} \phi^{j-1}\varepsilon_{t-j}. \end{aligned}$$

On peut directement calculer la fonction d'autocovariance ou d'autocorrélation à partir de cette expression. La variance $\gamma(0)$ vaut

$$\begin{aligned} \gamma(0) &= \sigma^2\left(1 + (\phi-\theta)^2\sum_{j=1}^{+\infty}\phi^{2(j-1)}\right) \\ &= \sigma^2\left(1 + \frac{(\phi-\theta)^2}{1-\phi^2}\right). \end{aligned}$$

De plus $h > 0$,

$$\mathrm{Cov}\,(X_t, X_{t+h}) = \sigma^2\left((\phi-\theta)\phi^{h-1} + (\phi-\theta)^2\phi^h\sum_{j=1}^{+\infty}\phi^{2(j-1)}\right).$$

On voit alors que

$$\gamma(h) = \sigma^2\left((\phi-\theta)\phi^{h-1} + \frac{(\phi-\theta)^2\phi^h}{1-\phi^2}\right).$$

Quelques propriétés

1. Tout processus ARMA est faiblement stationaire. De plus pour une représentation ARMA minimale, le processus s'exprime soit sous la forme d'une moyenne mobile infinie soit sous la forme d'un processus AR avec une infinité de retards. Pour le voir, on part de la représentation

$$\phi(B)X_t = c + \Theta(B)\varepsilon_t. \tag{2.5}$$

Les hypothèses faites sur les polynômes entrainent que $\Phi(B)$ et $\Theta(B)$ sont inversibles. On pose $\Psi(B) = \Phi(B)^{-1}\Theta(B)$. Il existe alors des coefficients $\left(\psi_j\right)_{j\geq 0}$ et $\left(\eta_j\right)_{j\geq 0}$ tels que $\eta_0 = \psi_0 = 1$, $\sum_{j=0}^{+\infty}|\psi_j| < +\infty$, $\sum_{j=0}^{+\infty}|\eta_j| < +\infty$ et

$$\Psi(B) = \sum_{j=0}^{+\infty}\psi_j B^j, \quad \Psi(B)^{-1} = \Theta(B)^{-1}\Phi(B) = \sum_{j=0}^{+\infty}\eta_j B^j.$$

On a alors, en appliquant l'opérateur $\phi(B)^{-1}$ de part et d'autre de l'égalité (2.5),

$$X_t = \Phi(B)^{-1}c + \Psi(B)\varepsilon_t = m + \varepsilon_t + \sum_{j=1}^{+\infty} \psi_j \varepsilon_{t-j}.$$

Ici m désigne la moyenne $\mathbb{E}(X_t)$. De plus, en appliquant l'opérateur $\Theta(B)^{-1}$ de part et d'autre de l'égalité (2.5), on a

$$\varepsilon_t = -\Theta(B)^{-1}c + \Psi(B)^{-1}X_t = d + X_t + \sum_{j=1}^{+\infty} \eta_j X_{t-j},$$

où d est le produit de c par la somme des coefficients de $\Theta(B)^{-1}$. Ceci correspond à un processus autorégressif avec un nombre infini de retards ($p = +\infty$).
L'intérêt des modèles ARMA est d'être plus parcimonieux : pour une qualité d'ajustement aux données comparable, il faut en général considérer des processus AR ou MA avec un ordre p ou q plus élevé (et donc plus de coefficients sont à estimer).

2. Remarquons que pour un processus ARMA, on a $m = \mathbb{E}(X_t) = \frac{c}{1-\sum_{j=1}^{p}\phi_j}$. De plus, avec les notations de la définition, on a $\Phi(B)(X_t - m) = \Theta(B)\varepsilon_t$. Le processus recentré devient un processus ARMA sans intercept.

3. La densité spectrale f d'un processus ARMA vérifie $f(\alpha) = \frac{\sigma^2}{2\pi}\frac{\Theta\left(e^{i\alpha}\right)}{\Phi\left(e^{i\alpha}\right)}$. Comme nous l'avons déjà vu pour les processus AR et MA, il est toujours possible de se ramener à une représentation ARMA minimale en changeant le bruit blanc. Pour cela, il suffit d'échanger les racines des deux polynômes qui sont à l'intérieur du disque unité avec leur inverse, à partir de l'expression de la densité spectrale.

4. Lorsque le processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc gaussien, le processus $(X_t)_{t\in\mathbb{Z}}$ est un processus gaussien. En effet, en utilisant la représentation en moyenne mobile infinie, tout vecteur $Y_t = (X_t, X_{t+1}, \ldots, X_{t+\ell})$ s'écrit $\lim_{N\to+\infty} Y_t^{(N)}$ (la limite est dans $\mathbb{L}^2$) avec

$$Y_t^{(N)} = \left(m + \sum_{j=0}^{N} \eta_j \varepsilon_{t-j}, \ldots, m + \sum_{j=0}^{N} \eta_j \varepsilon_{t+\ell-j}\right).$$

Le vecteur $Y_t^{(N)}$ étant gaussien pour tout N, on en déduit que le vecteur Y_t est aussi gaussien.

5. Les propriétés des fonctions d'autocorrélation et d'autocorrélation partielle d'un ARMA sont plus difficiles à obtenir. On peut noter que d'après l'équation du processus, on a

$$\gamma(h) = \sum_{j=1}^{p} \phi_j \gamma(h-j), \quad h \geq q+1.$$

Ceci entraine que $\gamma(h)$ tend vers 0 très rapidement, comme dans le cas des processus AR. Les autocorrélations partielles convergent également rapidement vers 0 mais ne sont pas nulles à partir d'un certain rang (comme les processus MA). On peut déterminer les premières valeurs $\gamma(0), \ldots, \gamma(q)$ en utilisant la représentation de X_t sous la forme d'une moyenne mobile infinie.

2.9.5 Prévision des processus ARMA

Dans cette section, nous expliquons comment obtenir des formules de prévision à partir des équations du modèle. Nous noterons $\hat{X}_T(k)$ la prévision linéaire optimale construite à partir des coordonnées $X_T, X_{T-1}, \ldots$ (dans un premier temps nous considérons tout le passé du processus). Commençons par trois exemples. Pour simplifier, nous supposerons que la moyenne m du processus vaut 0.

1. Considérons d'abord le processus AR(1), $X_t = \phi X_{t-1} + \varepsilon_t$ où $|\phi| < 1$. En itérant l'équation, on a
$$X_{T+k} = \phi^k X_T + \varepsilon_{T+k} + \phi\varepsilon_{T+k-1} + \cdots + \phi^{k-1}\varepsilon_{T+1}.$$
Par orthogonalité, il est facile de vérifier que $\hat{X}_T(k) = \phi^k X_T$.

2. Considérons ensuite un processus MA(1), $X_t = \varepsilon_t - \theta\varepsilon_{t-1}$. Commençons par le cas $k = 1$. On a $\hat{X}_T(1) = -\theta\varepsilon_T$, mais pour exprimer cette prévision à partir des X_{T-j}, on utilise l'expression
$$\varepsilon_T = \sum_{j=0}^{+\infty} \theta^j X_{T-j}.$$
On a donc
$$\hat{X}_T(1) = -\sum_{j=0}^{+\infty} \theta^{j+1} X_{T-j}.$$
Pour $k \geq 2$, on sait que X_{T+k} est orthogonal à chacune des variables X_{T-j} pour $j \geq 0$. On a donc $\hat{X}_T(k) = 0$.

3. Prenons ensuite un processus ARMA(1, 1) : $X_t - \phi X_{t-1} = \varepsilon_t - \theta\varepsilon_{t-1}$ où $|\phi| < 1$ et $|\theta| < 1$. En posant $Z_t = \varepsilon_t - \theta\varepsilon_{t-1}$, on a
$$X_{T+k} = \phi^k X_T + Z_{T+k} + \phi Z_{T+k-1} + \cdots + \phi^{k-1} Z_{T+1}.$$
Il est alors clair que $\hat{X}_T(k) = \phi^k X_T - \theta\phi^{k-1}\varepsilon_T$. On utilise alors la formule
$$\varepsilon_T = \sum_{j=0}^{+\infty} \theta^j \left(X_{T-j} - \phi X_{T-j-1}\right) = X_T + \sum_{j=1}^{+\infty} \left(\theta^j - \phi\theta^{j-1}\right) X_{T-j}.$$
On obtient au final :
$$\hat{X}_T(k) = \left(\phi^k - \theta\phi^{k-1}\right) \sum_{j=0}^{+\infty} \theta^j X_{T-j}.$$

Passons maintenant au cas général. On considère le processus ARMA(p, q), $\Phi(B)X_t = \Theta(B)\varepsilon_t$. Déterminons des formules (générales mais non explicites en fonction des coefficients) pour la prévision lorsque la représentation est minimale. On note $\Psi(B) = \Phi(B)^{-1}\Theta(B)$. On rappelle que $\psi_0 = 1$.

Les autres coefficients peuvent être déterminés récursivement en utilisant l'identité $\phi(B)\Psi(B) = \Theta(B)$. On a alors

$$\begin{aligned} X_{T+k} &= \sum_{j=0}^{+\infty} \psi_j \varepsilon_{T+k-j} \\ &= \sum_{j=0}^{k-1} \psi_j \varepsilon_{T+k-j} + \sum_{j=k}^{+\infty} \psi_j \varepsilon_{T+k-j} \\ &= \sum_{j=0}^{k-1} \psi_j \varepsilon_{T+k-j} + \sum_{j=0}^{+\infty} \psi_{j+k} \varepsilon_{T-j}. \end{aligned}$$

En utilisant les relations d'orthogonalité, on en déduit que

$$\hat{X}_T(k) = \sum_{j=0}^{+\infty} \psi_{j+k} \varepsilon_{T-j}.$$

On peut vérifier que l'opérateur associé s'écrit $\hat{X}_T(k) = \left[B^{-k}\Psi(B)\right]_+ \varepsilon_T$ où $[\cdot]_+$ signifie que l'on ne garde que les puissances positives de B dans le développement en série. Ensuite, on a

$$\varepsilon_T = \Theta(B)^{-1}\phi(B)X_T = \Psi(B)^{-1}X_T.$$

Les coefficients η_j de $\Psi(B)^{-1}$ peuvent être également trouvés récursivement à partir de l'identité $\Theta(B)\Psi(B)^{-1} = \phi(B)$. On obtient alors la formule

$$\hat{X}_T(k) = \left[B^{-k}\Psi(B)\right]_+ \Psi(B)^{-1}X_T. \tag{2.6}$$

De plus (en revenant au développement MA(∞) initial du processus), l'erreur de prévision est

$$e_T(k) = X_{T+k} - \hat{X}_T(k) = \sum_{j=0}^{k-1} \psi_j \varepsilon_{T+k-j}$$

et la variance de cette erreur est

$$\mathbb{E}\left(|e_T(k)|^2\right) = \sigma^2 \sum_{j=0}^{k-1} \psi_j^2.$$

Remarquer que la variance de l'erreur de prévision est plus grande que la variance du bruit ε et qu'elle augmente avec k (la prévision est de moins bonne qualité lorsque k augmente). Lorsque $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc gaussien (en particulier c'est une suite de variables aléatoires i.i.d), $e_T(k)$ est une variable aléatoire gaussienne centrée et de variance $\sigma^2 \sum_{j=0}^{k-1} \psi_j^2$. En notant $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi $\mathcal{N}(0,1)$, on en déduit que

$$X_{T+k} \in \left[\hat{X}_T(k) - \sigma q_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=0}^{k-1} \psi_j^2}, \hat{X}_T(k) + \sigma q_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=0}^{k-1} \psi_j^2}\right]$$

avec probabilité $1 - \alpha$.

Remarques

1. En pratique, l'équation (2.6) n'est pas utilisable car on ne dispose que des réalisations des variables $X_1, \ldots, X_T$ et les coefficients de l'ARMA sont inconnus. Pour les coefficients, il suffit de les estimer (voir le chapitre suivant). Ensuite, on peut toujours tronquer le développement de la série qui donne la prévision en fonction de $X_T, X_{T-1}, \ldots$ en mettant 0 pour $X_0, X_{-1}, \ldots$. Par exemple, pour la moyenne mobile d'ordre 1, on ne garde que $-\sum_{j=0}^{T-1} \theta^{j+1} X_{T-j}$. Vu que les coefficients de la série sont sommables, on obtient une prévision approchée qui est sensiblement la même que la prévision exacte si T est grand. Mais la prévision approchée peut être mauvaise si T est petit et si les racines des polynômes sont proches de 1. Il existe en fait des algorithmes de prévision qui donne les prévisions exactes uniquement à partir de $X_1, X_2, \ldots, X_T$. Deux de ces algorithmes sont détaillés dans [2], Chapitre 5. Dans le cas de bruits gaussien, l'intervalle de confiance pour la prévision peut aussi être estimé à partir d'estimateurs de σ^2 et de ψ_j pour $1 \leq j \leq k-1$ (les η_j s'expriment en fonction des θ_n et des ϕ_ℓ).
2. Nous avons ignoré la présence d'un intercept dans la formulation du processus ARMA (et donc supposé que le processus était de moyenne nulle). Mais comme le processus recentré $(X_t - m)_{t\in\mathbb{Z}}$ est un ARMA sans intercept, on déduit aisément l'analogue de (2.6) :

$$\hat{X}_T(k) = m + \left[B^{-k}\Psi(B)\right]_+ \Psi(B)^{-1}\left(X_T - m\right).$$

Exercice 8 Donner une expression des variances des erreurs de prévisions théoriques (en conservant un passé infini) pour les processus AR(1), MA(1).

2.9.6 Les ordres p et q des processus ARMA

En pratique, on détermine d'abord les ordres p et q du modèle ARMA (le choix dépend des données) pour ensuite estimer les paramètres. Les propriétés vues dans ce chapitre permettent déjà d'avoir une information à ce sujet pour les processus AR ou MA.

- Pour les processus AR(p), les autocorrélations convergent vers 0 à vitesse géométrique alors que les autocorrélations partielles sont nulles à partir du rang $p+1$. Pour un jeu de données, les autocorrélations partielles empiriques permettent donc d'avoir une idée des valeurs possibles pour p.
- Pour les processus MA(q), les autocorrélations partielles convergent vers 0 à vitesse géométrique alors que les autocorrélations sont nulles à partir du rang $q+1$. Les autocorrélations empiriques permettent d'avoir une idée des valeurs possibles pour q.

Pour les processus ARMA(p, q), la situation est moins simple puisque les autocorrélations ou les autocorrélations partielles convergent vers 0 à vitesse géométrique mais ne sont pas nulles à partir d'un certain rang. Pour une représentation ARMA(p, q) minimale, il existe une caractérisation des ordres p et q à l'aide de certains déterminants de matrices de corrélations qui peuvent être estimés (c'est la méthode du coin présentée dans [3] Chapitre 5). Une autre approche proposée par les logiciels est de sélectionner un modèle ARMA à partir des critères d'information du type AIC ou BIC et dont nous reparlerons au prochain chapitre.

Chapitre 3

Statistique inférentielle dans les modèles ARMA

3.1 Moindres carrés et vraisemblance gaussienne pour les modèles ARMA

Il y a deux possibilités pour présenter les résultats d'estimation d'un processus ARMA. La première est de supposer le modèle avec intercept (en général les séries temporelles sont non centrées). La deuxième est de recentrer les variables par la moyenne empirique et de modéliser les nouvelles variables par un processus ARMA centré (donc sans intercept). Les logiciels permettent d'inclure ou non un intercept pour l'ajustement. Pour cette exposition, nous choisirons la deuxième solution. A partir de maintenant, nous supposerons que polynômes Θ et Φ des processus ARMA considérés n'ont pas de racine en commun et ont toutes leurs racines à l'extérieur du disque unité.

3.1.1 Estimation des coefficients d'un AR

Nous avons évoqué au chapitre précédent la possibilité d'estimer les paramètres d'un AR à l'aide des équations de Yule-Walker. Supposons que les données $X_1, \dots, X_T$ aient été générées par

$$X_t = \sum_{j=1}^{p} \phi_{0,j} X_{t-j} + \varepsilon_t, \quad t \in \mathbb{Z}, \quad \text{Var}\,(\varepsilon_1) = \sigma_0^2. \tag{3.1}$$

Notons par γ_0 la fonction d'autocovariance du processus. Nous avons, en posant $\Gamma_{0,p} = [\gamma_0(i-j)]_{1\le i,j\le p}^{-1}$ et $\gamma_{0,p} = \begin{pmatrix} \gamma_0(1) \\ \vdots \\ \gamma_0(p) \end{pmatrix}$,

$$\phi_0 = \begin{pmatrix} \phi_{0,1} \\ \vdots \\ \phi_{0,p} \end{pmatrix} = \Gamma_{0,p}^{-1} \gamma_{0,p}.$$

En remplaçant les autocovariances par les autocovariances empiriques $\hat{\gamma}$, on obtient un estimateur $\hat{\phi}$ qui est consistant lorsque la suite $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc fort. La consistance peut se prouver à l'aide du Théorème 2 du Chapitre 2 : les covariances empiriques convergent p.s et la continuité de l'inverse d'une matrice (si $A_T \to A$ avec A inversible, alors A_T est inversible à partir d'un certain rang et $A_t^{-1} \to A^{-1}$) assure la converge p.s de l'estimateur. De plus la variance du bruit, est donnée par la formule (voir la formule de la variance de l'erreur de prévision, Chapitre 2, sous-section 2.4.2)

$$\sigma_0^2 = \gamma_0(0) - \gamma_{0,p}'\Gamma_{0,p}^{-1}\gamma_{0,p} = \gamma_0(0) - \gamma_{0,p}'\phi_0.$$

Une estimation $\hat{\sigma}^2$ de σ_0^2 peut être déduite en remplaçant les covariances par les covariances empiriques.
Les équations de Yule-Walker sont basées sur des relations d'orthogonalité ou encore sur la minimisation de l'application

$$(\phi_1, \ldots, \phi_p) \mapsto \mathbb{E}\left| X_{p+1} - \sum_{j=1}^{p} \phi_j X_{p+1-j} \right|^2 .$$

On peut se demander ce qu'il advient si on minimise directement le critère empirique

$$\left(\phi_1, \ldots, \phi_p\right) \mapsto \frac{1}{n}\sum_{t=p+1}^{T}\left(X_t - \sum_{j=1}^{p} \phi_j X_{t-j} \right)^2 . \tag{3.2}$$

Minimiser (3.2) revient à utiliser les moindres carrés ordinaires. Cette méthode est naturelle puisque lorsque le bruit blanc est fort, $\mathbb{E}\left(X_t | X_{t-1}, \ldots, X_{t-p}\right) = \sum_{j=1}^{p} \phi_{0,j} X_{t-j}$. On notera la différence avec les équations de Yule-Walker : au lieu de minimiser en espérance et de remplacer ensuite les covariances par des estimateurs, on minimise ici directement le critère empirique correspondant. On peut montrer que le comportement asymptotique de ce nouvel estimateur est le même que celui de $\hat{\phi}$. L'estimateur $\hat{\phi}$ vérifie en particulier la convergence en loi suivante.

Théorème 5 *Lorsque $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc fort, alors*

$$\lim_{T\to+\infty} \sqrt{T}\left(\hat{\phi} - \phi_0\right) = \mathcal{N}_p\left(0, \sigma_0^2\Gamma_{0,p}^{-1}\right).$$

De plus $\hat{\sigma}^2 \to \sigma_0^2$ p.s.

Exercice 9 Pour simplifier les notations, (ϕ, σ^2) désigne la valeur des paramètres. Vérifier que $\sigma^2\Gamma_1^{-1} = \left(1 - \phi_1^2\right)$ et que $\sigma^2\Gamma_2^{-1} = \begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix}$. Pour les processus AR(2), montrer que la condition "$1 - \phi_1 B - \phi_2 B^2$ a ses racines à l'extérieur du disque unité" est équivalente à

$$\begin{cases} |\phi_2| < 1 \\ \phi_2 < 1 - \phi_1 \\ \phi_2 < 1 + \phi_1 \end{cases}$$

Remarque. Lorsque le bruit blanc $(\varepsilon_t)_{t\in\mathbb{Z}}$ est gaussien, le vecteur des observations $(X_1,\dots,X_T)$ est gaussien. Il est alors possible de calculer la vraisemblance de l'échantillon (voir aussi le paragraphe suivant pour les ARMA) mais une méthode plus simple est de calculer la densité conditionnelle de $\left(X_{p+1},\dots,X_T\right)|X_1,\dots,X_p$ qui est gaussienne (on parle alors de vraisemblance conditionnelle). On utilise la formule des conditionnements successifs. Sachant que pour tout $t\in\mathbb{Z}$, la loi conditionnelle de $X_t|X_{t-1},\dots,X_{t-p}$ est une loi gaussienne de moyenne $\sum_{j=1}^p \phi_{0,j}X_{t-j}$ et de variance σ_0^2, on a

$$\begin{aligned} f_{(X_{p+1},\dots,X_T)|X_1,\dots,X_p}\left(X_{p+1},\dots,X_T|X_1,\dots,X_p\right) &= \prod_{t=p+1}^{T} f_{X_t|X_{t-1},\dots,X_{t-p}}\left(X_t|X_{t-1},\dots,X_{t-p}\right) \\ &= \left(2\pi\sigma_0^2\right)^{-\frac{T-p}{2}} \exp\left(-\sum_{t=p+1}^{T}\frac{\left(X_t-\sum_{j=1}^p\phi_{0,j}X_{t-j}\right)^2}{2\sigma_0^2}\right). \end{aligned}$$

En passant au logarithme, on est amené à minimiser

$$\left(\phi_1,\dots,\phi_p,\sigma^2\right)\mapsto\sum_{t=p+1}^{T}\left\{\frac{\left(X_t-\sum_{j=1}^p\phi_jX_{t-j}\right)^2}{\sigma^2}+\ln\left(\sigma^2\right)\right\}.$$

On voit donc qu'à σ^2 fixé, il faut minimiser le critère des moindres carrés. Comme pour le modèle de régression linéaire, on voit que l'estimateur du maximum de vraisemblance des coefficients de régression de l'AR coïncide avec celui des MCO. Les estimateurs du maximum de vraisemblance conditionnel $\left(\widetilde{\phi}_{EMV},\widetilde{\sigma}^2_{EMV}\right)$ vérifient

$$\widetilde{\phi}_{EMV}=\arg\min_{\phi}\sum_{t=p+1}^{T}\left(X_t-\sum_{j=1}^p\phi_jX_{t-j}\right)^2,\quad \widetilde{\sigma}^2_{EMV}=\frac{1}{T-p}\sum_{t=p+1}^{T}\left(X_t-\sum_{j=1}^p\widetilde{\phi}_{EMV,j}X_{t-j}\right)^2.$$

Exemple. La série analysée est le niveau du lac Huron relevé chaque année entre 1875 et 1972 (Figure 3.1). Les ACF et PACF semblent a priori compatibles avec un AR(2). Les paramètres estimés sont $\left(\hat{\phi}_1,\hat{\phi}_2,\hat{\sigma}^2\right)=(1.0421,-0.2483,0.4808)$ et les racines correpondantes du polynôme Φ sont estimées par environ 1.49 et 2.71. Le premier diagnostic est de regarder l'ACF des résidus $\hat{\varepsilon}_t=X_t-\hat{\phi}_1X_{t-1}-\hat{\phi}_2X_{t-2}$ qui doivent normalement présenter des corrélations très faibles.

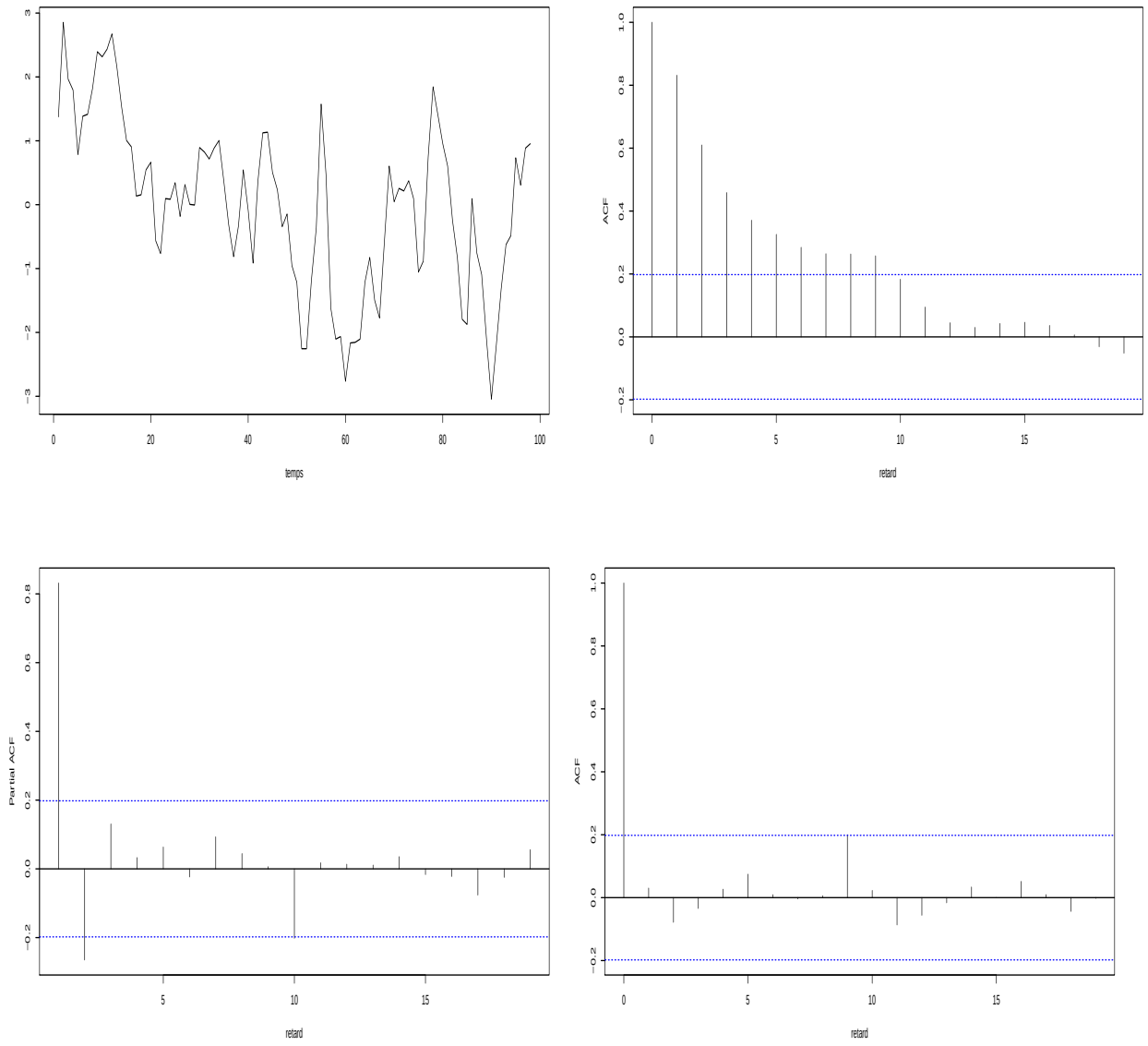


FIGURE 3.1 – Etude de la série du niveau du lac Huron (les variables sont recentrées)

3.1.2 Estimation des coefficients d'un ARMA

Nous noterons

$$\beta_0 = \left(\phi_0, \theta_0, \sigma_0^2\right) = \left(\phi_{0,1}, \ldots, \phi_{0,p}, \theta_{0,1}, \ldots, \theta_{0,q}, \sigma_0^2\right)$$

les paramètres qui correspondent à la loi des données $X_1, \ldots, X_T$.

Méthode du maximum de vraisemblance. L'estimation des coefficients d'un ARMA par maximum de vraisemblance est plus délicate. Nous ne détaillerons que sommairement le calcul de la

vraisemblance. Une méthode de vraisemblance conditionnelle (qui généralisera celle présentée pour les processus AR) sera détaillée dans le paragraphe suivant. Dans le cas d'un bruit blanc fort gaussien, le vecteur des observations $(X_1, \ldots, X_T)$ est gaussien. On peut a priori calculer la vraisemblance exacte. La densité des observations s'écrit

$$f_{X_1,\ldots,X_T}(x_1,\ldots,x_T) = f_{X_T|X_{T-1},\ldots,X_1}(x_T|x_{T-1},\ldots,x_1)\, f_{X_{T-1}|X_{T-2},\ldots,X_1}(x_{T-1}|x_{T-2},\ldots,x_1)\cdots f_{X_2|X_1}(x_2|x_1)\, f_{X_1}(x_1).$$

Chacune des densités de l'équation ci-dessus est gaussienne mais les paramètres de ces lois gaussiennes ne sont pas directement exprimables à partir des coefficients de l'ARMA. En notant $\hat{X}_t(\phi_0, \theta_0)$ la projection de X_t sur $Vect(X_{t-1}, \ldots, X_1)$ (qui coïncide avec l'espérance conditionnelle dans le cas gaussien) et $r_t(\phi_0, \theta_0) = \frac{\mathbb{E}\left|X_t - \hat{X}_t(\phi_0,\theta_0)\right|^2}{\sigma_0^2}$, la loi conditionnelle de $X_t|X_{t-1}, \ldots, X_1$ est une loi gaussienne de moyenne $\hat{X}_t(\beta_0, \theta_0)$ et de variance $\sigma_0^2 r_t(\phi_0, \theta_0)$ (dans un vecteur gaussien $(X, Y) \in \mathbb{R} \times \mathbb{R}^d$, la loi conditionnelle de $X|Y$ est gaussienne, la moyenne coïncide alors avec la projection linéaire de X sur $\mathrm{Vect}_{\mathbb{L}^2}(Y)$ et la variance coïncide avec la variance de l'erreur de prévision). On a donc (en posant $r_1(\phi_0, \theta_0) = \frac{\mathbb{E}(X_1^2)}{\sigma_0^2}$ et $\hat{X}_1(\phi_0, \theta_0) = 0$)

$$f_{X_1,\ldots,X_T}(X_1,\ldots,X_T) = \left(2\pi\sigma_0^2\right)^{-\frac{T}{2}} \left(r_1(\phi_0,\theta_0)\cdots r_T(\phi_0,\theta_0)\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{t=1}^{T}\frac{\left(X_t - \hat{X}_t(\phi_0,\theta_0)\right)^2}{r_t(\phi_0,\theta_0)}\right).$$

Posons

$$S(\phi,\theta) = \sum_{t=1}^{T}\frac{\left(X_t - \hat{X}_t(\phi,\theta)\right)^2}{r_t(\phi,\theta)}.$$

On peut alors montrer que l'estimateur du maximum de vraisemblance $\left(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2\right)$ vérifie $\hat{\sigma}^2 = \frac{1}{T}S\left(\hat{\phi}, \hat{\theta}\right)$ et $\left(\hat{\phi}, \hat{\theta}\right)$ minimise

$$L(\phi,\theta) = \ln\left(\frac{1}{T}S(\phi,\theta)\right) + \frac{1}{T}\sum_{t=1}^{T}\ln r_t(\phi,\theta).$$

Le fait que L ne dépende que de ϕ et θ (et pas de σ^2) est justifié : $r_t(\phi, \theta)$ ne dépend pas de σ^2 car cette quantité correspond à l'erreur de prévision lorsque le bruit ε a pour variance 1, de plus $\hat{X}_t(\phi, \theta)$ se définit à partir des autocorrélations entre les variables X_t et ces autocorrélations (contrairement aux autocovariances) ne dépendent plus de σ^2.

Il existe des algorithmes pour un calcul récursif des projections $\hat{X}_t(\phi, \theta)$ et des variances $r_t(\phi, \theta)$ en fonction des paramètres (voir en particulier [2] Chapitres 5 et 8). Le problème d'optimisation de la vraisemblance est alors complexe car fortement non linéaire en (ϕ, θ).
Il existe aussi une solution alternative du type "moindres carrés". Pour un ARMA, on peut vérifier que $r_t(\theta_0, \phi_0) \to 1$ lorsque $t \to +\infty$. Le terme $\frac{1}{T}\sum_{t=1}^{T}\ln r_t(\phi_0, \theta_0)$ converge donc vers 0 lorsque $T \to +\infty$. Les estimateurs $\left(\widetilde{\phi}, \widetilde{\theta}\right)$ des moindres carrés sont alors définis en minimisant $(\phi, \theta) \mapsto S(\phi, \theta)$. Par analogie avec la régression linéaire standard, on définit $\widetilde{\sigma}^2 = \frac{S\left(\widetilde{\phi},\widetilde{\theta}\right)}{T-p-q}$ (on peut montrer que la loi de $\frac{\widetilde{\sigma}^2}{\sigma_0^2}$ est approximativement une $\chi^2(T - p - q)$).

Méthode de la vraisemblance conditionnelle. Cette méthode est plus simple que celle de la vraisemblance exacte. Le principe est de calculer la densité conditionnelle de

$$\left(X_T, \ldots, X_{p+1}\right) | X_p, \ldots, X_1, \varepsilon_p, \ldots, \varepsilon_{p-q+1}$$

qui est celle d'une loi gaussienne. On fixe ensuite des valeurs intiales $\varepsilon_p = z_p, \ldots, \varepsilon_{p-q+1} = z_{p-q+1}$ pour les bruits. Toujours par la formule des conditionnements successifs, on est amené à minimiser

$$\left(\phi, \theta, \sigma^2\right) \mapsto \sum_{t=p+1}^{T} \left\{ \frac{e_t^2(\phi, \theta)}{\sigma^2} + \ln\left(\sigma^2\right) \right\}$$

où

$$e_t(\phi, \theta) = X_t - \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j e_{t-j}(\phi, \theta), \quad p+1 \leq t \leq T.$$

Pour tout θ, ϕ, on pose $e_i(\phi, \theta) = z_i$ si $p - q + 1 \leq i \leq p$.
Examinons l'exemple d'un processus MA(1), $X_t = \varepsilon_t - \theta_0 \varepsilon_{t-1}$. Alors on sait que

$$\varepsilon_{t-1} = X_{t-1} + \theta_0 X_{t-2} + \cdots + \theta_0^{t-2} X_1 + \theta_0^{t-1} \varepsilon_0.$$

La loi de $X_t | X_{t-1}, \ldots, X_1, \varepsilon_0$ est une loi gaussienne de moyenne

$$-\sum_{j=1}^{t-1} \theta_0^j X_{t-j} - \theta_0^t \varepsilon_0$$

et variance σ_0^2. On a alors

$$e_t(\theta) = X_t + \sum_{j=1}^{t-1} \theta^j X_{t-j} + \theta^t \varepsilon_0.$$

Les trois méthodes proposées donnent des estimateurs qui ont les mêmes propriétés asymptotiques. Ce qui est intéressant ici, c'est que les propriétés asymptotiques sont valables même si le bruit est un bruit blanc fort non gaussien.

Théorème 6 *Supposons que $(\varepsilon_t)_{t \in \mathbb{Z}}$ soit un bruit blanc fort. Alors les estimateurs $\hat{\phi}, \hat{\theta}$ et $\hat{\sigma}^2$ sont fortement consistants. De plus, on a*

$$\sqrt{T}\left(\hat{\phi} - \phi_0, \hat{\theta} - \theta_0\right) \to \mathcal{N}_2\left(0, V(\phi_0, \theta_0)\right),$$

avec

$$V\left(\phi_0, \theta_0\right) = \sigma_0^2 \begin{pmatrix} \mathbb{E}\left(U_t U_t'\right) & -\mathbb{E}\left(U_t V_t'\right) \\ -\mathbb{E}\left(V_t U_t'\right) & \mathbb{E}\left(V_t V_t'\right) \end{pmatrix}^{-1}.$$

Ici, on a $U_t = \left(Y_{t-1}, \ldots, Y_{t-p}\right)'$ et $V_t = \left(Z_{t-1}, \ldots, Z_{t-q}\right)'$ avec

$$\Phi_0(B) Y_t = \varepsilon_t, \quad \Theta_0(B) Z_t = \varepsilon_t, \quad t \in \mathbb{Z}.$$

Lorsque le processus est supposé être un AR ou un MA (cas $q = 0$ ou $p = 0$), on convient que $V(\phi_0) = \sigma_0^2 \mathbb{E}^{-1}(U_t U_t')$ ou $V(\theta_0) = \sigma_0^2 \mathbb{E}^{-1}(V_t V_t')$.
Les méthodes d'optimisation de la vraisemblance peuvent être assez sensibles à la valeur initiale utilisée par les algorithmes. En général, une estimation préliminaire des coefficients est utilisée (on pourra consulter [3], Chapitre 5 et [2], Chapitre 8 pour plus de détails).

Exercice 10 Dans le cas du processus ARMA(1, 1), vérifier que la variance asymptotique est donnée par (en notant simplement ϕ et θ les paramètres)

$$\frac{1-\phi\theta}{(\phi-\theta)^2}\begin{pmatrix}\left(1-\phi^2\right)(1-\phi\theta) & \left(1-\phi^2\right)\left(1-\theta^2\right)\\ \left(1-\theta^2\right)\left(1-\phi^2\right) & \left(1-\theta^2\right)(1-\phi\theta)\end{pmatrix}.$$

Construction d'intervalles de confiance pour les coefficients. En notant $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$, on déduit du résultat précédent que pour tout $j = 1, \ldots, p$, l'intervalle

$$\left[\hat{\phi}_j - \frac{1}{\sqrt{n}}\sqrt{V\left(\hat{\phi},\hat{\theta}\right)_{j,j}}q_{1-\frac{\alpha}{2}}, \hat{\phi}_j + \frac{1}{\sqrt{n}}\sqrt{V\left(\hat{\phi},\hat{\theta}\right)_{j,j}}q_{1-\frac{\alpha}{2}}\right]$$

est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour ϕ_j. Le même type de résultat peut être utilisé pour θ_j. Un test de significatité des coefficients s'en déduit (on rejette la nullité d'un coefficient si 0 n'est pas dans l'intervalle de confiance). Des régions de confiance pour plusieurs paramètres peuvent également se déduire de la normalité asymptotique des estimateurs.

3.2 Les processus ARIMA

En pratique la stationnarité n'est pas toujours réaliste. La présence d'une tendance, ou d'une saisonnalité ou bien encore d'un effet marche aléatoire (la marche aléatoire correspond au cas où $X_t = X_{t-1} + U_t$ avec $(U_t)_t$ stationnaire) conduisent à la non stationnarité des séries. Nous avons vu au Chapitre 1 comment estimer la tendance et la saisonnalité. Dans une optique de prévision des séries temporelles, il est souvent possible de modéliser la série par un ARMA après avoir différencié la série. Par exemple une série du type $X_t = a + bt + U_t$ aura ses différences premières $(1-B)X_t$ stationnaires et on peut essayer d'ajuster un ARMA sur les différences. On peut également utiliser l'opérateur $1 - B^k$ pour enlever une composante saisonnière de période k. L'incorporation de la saisonnalité sera traitée au paragraphe suivant. Dans ce paragraphe, nous introduisons les modèles ARIMA (autoregressive integrated moving average).

Définition 16 *On dit que $(X_t)_{t\geq 1}$ est processus ARIMA(p,d,q) (avec d un entier ≥ 1) si $(1-B)^d X_t = Y_t$ pour $t \geq d+1$ avec $(Y_t)_{t\in\mathbb{Z}}$ un processus ARMA(p,q) et*

$$\phi(B)Y_t = c + \Theta(B)\varepsilon_t, \quad Cov(\varepsilon_s, X_i) = 0, \quad 1 \leq i \leq d.$$

Lorsque $d = 0$, on retrouve les processus ARMA. Lorsque $d = 1$, on a des équations du type $X_t = X_{t-1} + Y_t$ si $t \geq 2$ où $\phi(B)Y_t = c + \Theta(B)\varepsilon_t$. Dans le cas d quelconque, on a

$$X_t = Y_t - \sum_{j=1}^{d}\binom{d}{j}(-1)^j X_{t-j}, \quad t \geq d+1. \tag{3.3}$$

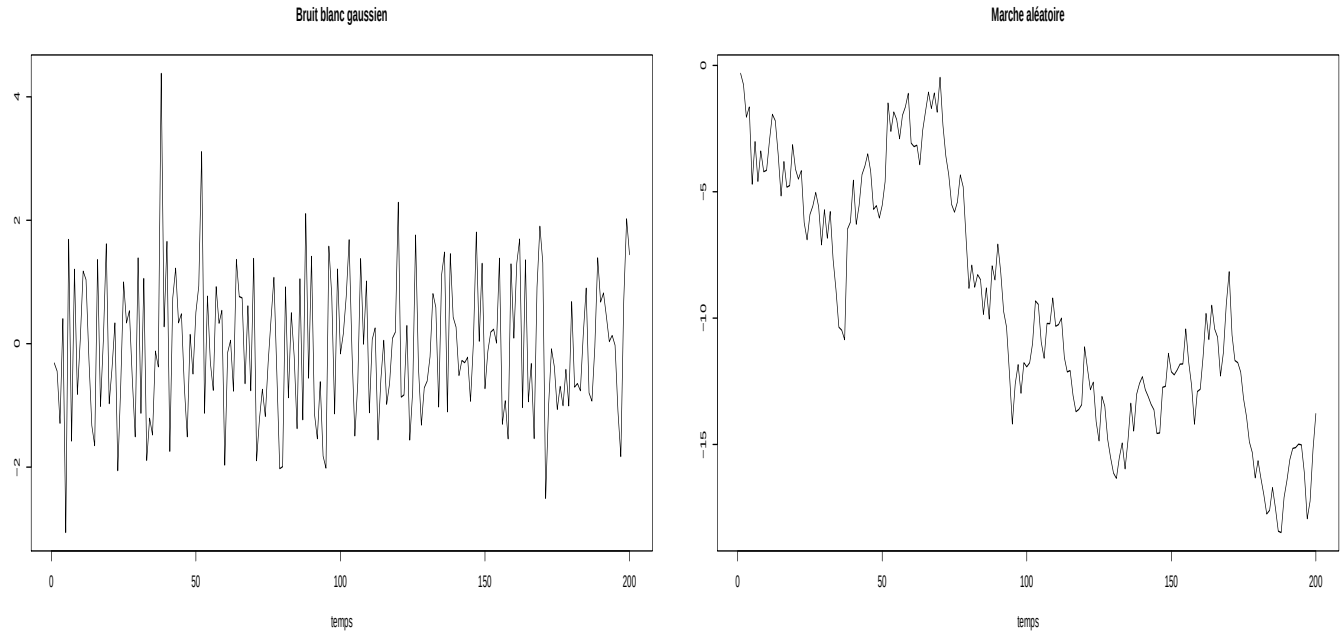


FIGURE 3.2 – Trajectoire d'un bruit blanc gaussien et marche aléatoire associée (bruit blanc intégré à l'ordre 1)

Remarques

1. Pour estimer les paramètres d'un ARIMA(p, d, q), on applique les procédures d'estimation pour les ARMA au processus $(1 - B)^d X_t$.
2. Pour le problème de la prévision, on suppose pour simplifier que $c = 0$. Posons

$$F_T = \text{ Vect } (X_1, \ldots, X_T) = \text{ Vect } (X_1, \ldots, X_d, Y_{d+1}, \ldots, Y_T).$$

Comme les X_i et les Y_s sont décorrélées lorsque $1 \leq i \leq d$, on a $p_{F_T}(Y_{T+h}) = \widetilde{Y}_T(h)$ où $\widetilde{Y}_T(h)$ désigne la projection de Y_{T+h} sur le sous-espace vectoriel de $\mathbb{L}^2$ engendré par $Y_{d+1}, \ldots, Y_T$. On a alors d'après (3.3),

$$\hat{X}_T(h) = \widetilde{Y}_T(h) - \sum_{j=1}^{d} \binom{d}{j} (-1)^j \hat{X}_T(h - j)$$

avec la convention que $\hat{X}_T(h - j) = X_{T+h-j}$ si $h \leq j$. On peut alors calculer les prévisions successives à partir des calculs des prévisions pour les ARMA. Pour calculer la variance de l'erreur de prévision, on peut partir de l'équation

$$(1 - B)^d \Phi(B) X_{T+h} = \Theta(B) \varepsilon_{T+h}.$$

On peut alors montrer (nous l'admettrons) que l'erreur de prévision peut être approchée pour T grand par $\sum_{j=0}^{h-1} \eta_j^* \varepsilon_{T+h-j}$ où

$$\sum_{j=0}^{+\infty} \eta_j^* z^j = \left[(1-z)^d \Phi(z)\right]^{-1} \Theta(z), \quad |z| < 1.$$

Les intervalles de confiance pour les prévisions se construisent alors comme pour les ARMA.

Attention à ne pas sur-différencier une série temporelle. Une différenciation excessive peut avoir une d'incidence sur la prévision des valeurs futures. La Figure 3.3 montre ce problème pour un processus AR(1) définie par $X_t = 0.5X_{t-1} + \varepsilon_t$. Lorsque cette série pourtant stationnaire est différenciée deux fois, le modèle ajusté par le logiciel R (en utilisant la fonction ARIMA) est un AR(4) et les intervalles de confiance pour la prévision deviennent bien plus larges.

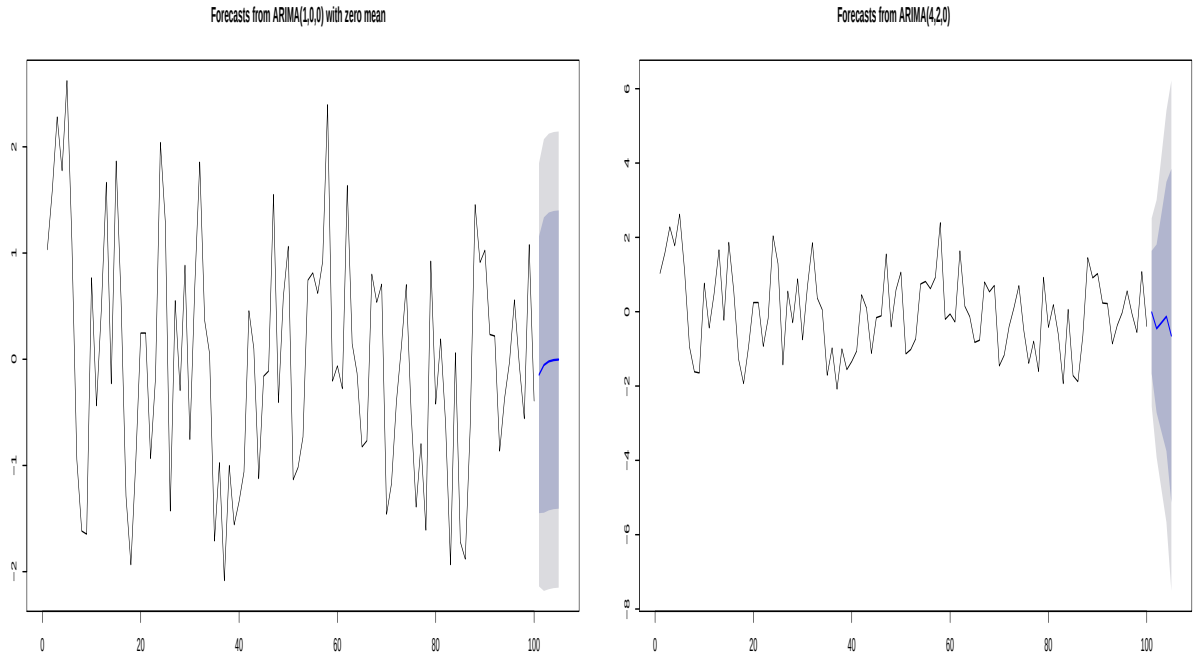


FIGURE 3.3 – Graphe de la série puis des prévisions et des intervalles de confiance pour les prévisions (série initiale à gauche et différenciée deux fois à droite)

La sur-différenciation peut aussi causer des problèmes d'inversibitité.

Exercice 11 Soit $X_t = \varepsilon_t - \theta\varepsilon_{t-1}$ une moyenne mobile inversible. Déterminer les racines du polynôme de la moyenne mobile $(1-B)X$.

Il faut donc toujours procéder pas à pas pour trouver l'ordre de différenciation approprié (en complément de tests que nous verrons plus loin dans ce chapitre).

3.3 Les processus ARIMA saisonniers

3.3.1 Définition des processus SARIMA

Au Chapitre 1, nous avons introduit la décomposition classique $X_t = m_t + s_t + U_t$ en tendance, saisonnalité et composante aléatoire. L'application de la régression sur les données SNCF a montré qu'il était parfois difficile de supposer que la composante saisonnière se répète à l'identique pour chaque période. Les modèle ARIMA saisonniers (SARIMA) permettent de prendre en compte l'aléatoire qui peut apparaitre à ce niveau.
Prenons l'exemple de données mensuelles et considérons le processus ARMA défini par $\Gamma\left(B^{12}\right) X_t = \Xi\left(B^{12}\right) U_t$ où $(U_t)_{t\in\mathbb{Z}}$ est un bruit blanc et

$$\Gamma(B) = 1 - \sum_{j=1}^{P} \gamma_j B^j, \quad \Xi(B) = 1 - \sum_{j=1}^{Q} \chi_j B^j.$$

On peut alors modéliser le comportement de la suite pour chaque mois de l'année : la suite des variables qui correspondent aux mois de janvier suit un ARMA, ainsi que celle qui correspondent aux mois de février... Par exemple pour les mois de janvier, en posant $Y_n = X_{12n+1}$ et $V_n = U_{12n+1}$, on a $\Gamma(B)Y_n = \Xi(B)V_n$. Mais la modélisation n'est pas réaliste puisque les différentes séries (mois de janvier, de février...) sont décorrélées deux à deux. L'idée est alors introduire de la corrélation dans le bruit. On supposera en plus que $\Phi(B)U_t = \Theta(B)\varepsilon_t$ avec $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc de variance σ^2. Au final, on a

$$\Gamma\left(B^{12}\right)\Phi(B)X_t = \Xi\left(B^{12}\right)\Theta(B)\varepsilon_t.$$

Le modèle reste stationnaire. On peut alors inclure des différences du type $1 - B$ ou $1 - B^{12}$.

Définition 17 *On dit que $(X_t)_{t\geq 1}$ est un SARIMA$(p, d, q) \times (P, D, Q)_k$ si*

$$\left(1 - B^{12}\right)^D \Gamma\left(B^k\right)(1 - B)^d \Phi(B)X_t = c + \Xi\left(B^k\right)\Theta(B)\varepsilon_t.$$

Les différents polynômes Φ, Θ, Γ et Ξ satisfont les hypothèses habituelles et sont donnés par

$$\Gamma(B) = 1 - \sum_{j=1}^{P} \gamma_j B^j, \quad \Xi(B) = 1 - \sum_{j=1}^{Q} \chi_j B^j, \quad \Phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j, \quad \Theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j.$$

En pratique D excède rarement 1. Les entiers d et D sont choisis de sorte à avoir une série $(1 - B)^d \left(1 - B^k\right)^D X_t$ stationnaire en apparence (graphe, autocorrélations empiriques...). Les choix des ordres p, q, P et Q peuvent se deviner (surtout dans le cas de processus AR ou MA) en regardant les autocorrélations (simples et partielles) empiriques pour les retards $1, 2, \ldots, k-1$ (qui se comporte un peu comme celles d'un ARMA(p, q)) et pour les retards $k, 2k, 3k \ldots$ (celles-ci doivent être compatibles avec un ARMA(P, Q)). Par exemple, des autocorrélations partielles empiriques quasi-nulles à partir du rang ks suggère de prendre $P = s - 1$ et $Q = 0$. Les problèmes d'estimation et de prévision se traitent alors comme pour les processus ARIMA.

Remarque. En pratique, il vaut mieux commencer par appliquer l'opérateur $1 - B^k$ et regarder si la série est stationnaire avant d'éventuellement différencier en plus à l'ordre 1 (application de l'opérateur $1 - B$). Comme

$$1 - B^k = (1 - B)(1 + B + \cdots + B^{k-1}),$$

on voit que $1 - B^k$ contient déjà une différenciation à l'ordre 1.

3.3.2 Exemple sur des données de températures

Nous illustrons la modélisation SARIMA sur des données de températures à partir d'une série d'observations mensuelles disponible sous le logiciel R. On pourra aussi consulter [1] pour des comparaisons avec la méthode de la régression linéaire (estimation de la saisonnalité telle qu'elle a été faite au Chapitre 1). La Figure 3.5 montre que les données sont non stationnaires (présence d'une saisonnalité) mais que l'application de $I - B^{12}$ permet de stationnariser la série. Au vu de l'ACF et de la PACF, il semble raisonnable de modéliser la série par un SARIMA$(1, 0, 0)(2, 1, 0)_{12}$. Les résidus obtenus après estimation ne semblent pas montrer de corrélation significative. Enfin, nous avons estimé la série entre janvier 1937 et décembre 1938 pour comparer les valeurs prédites sur l'année 1939 avec les valeurs réelles.

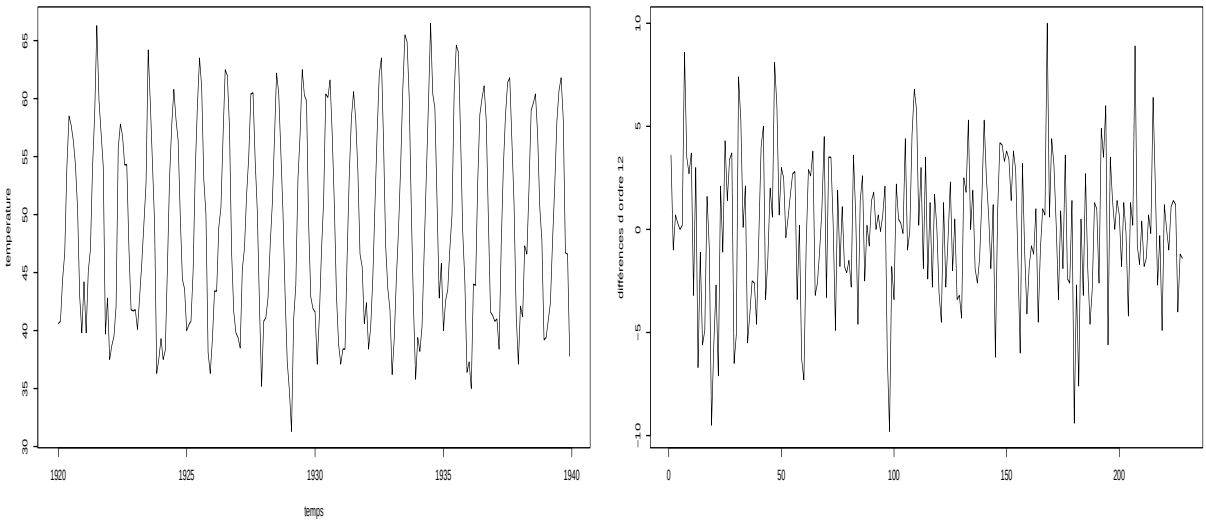


FIGURE 3.4 – Série initiale des températures mensuelles moyennes à Nottingham Castle entre janvier 1920 et décembre 1939 puis série différenciée à l'ordre 12

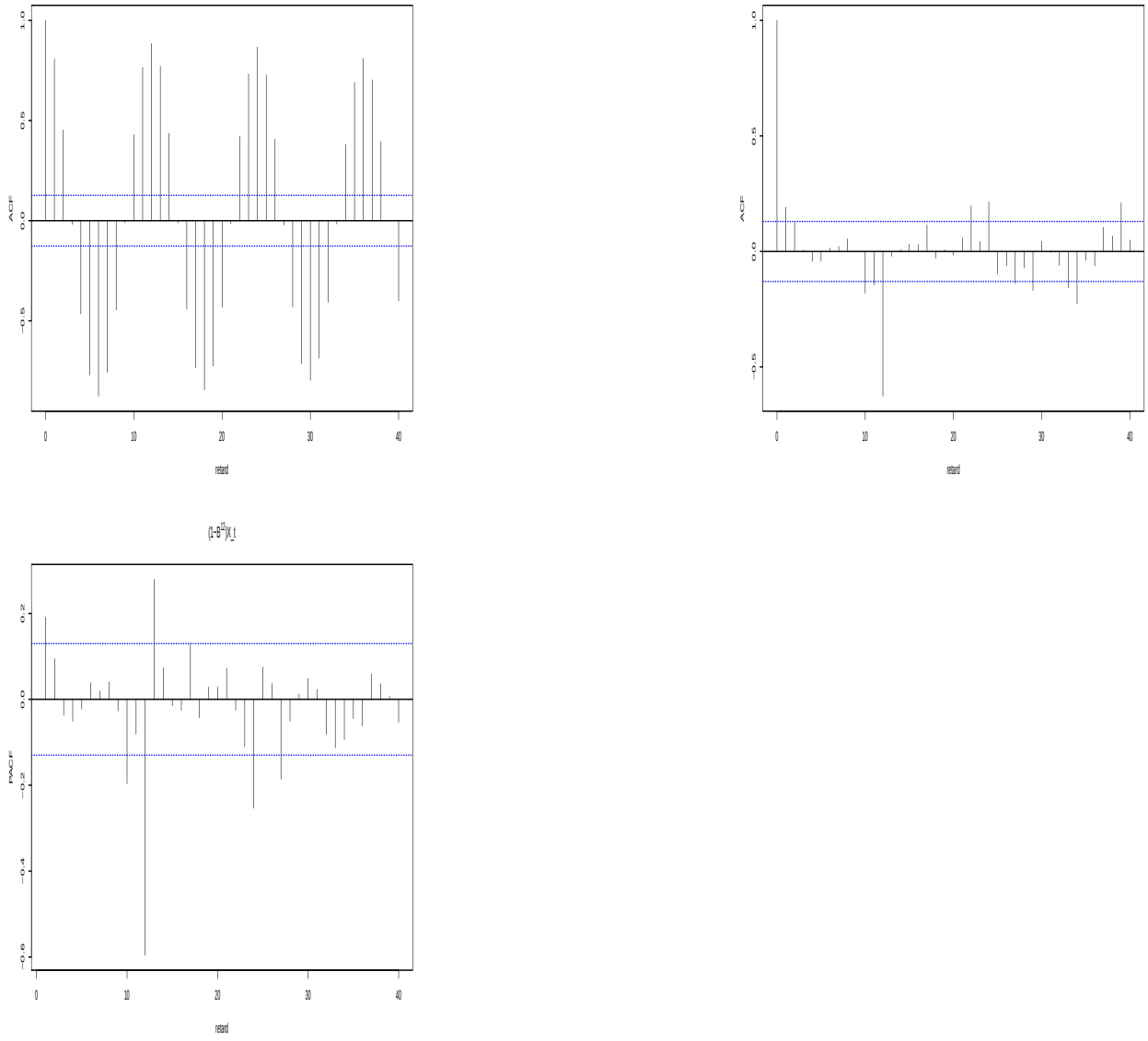
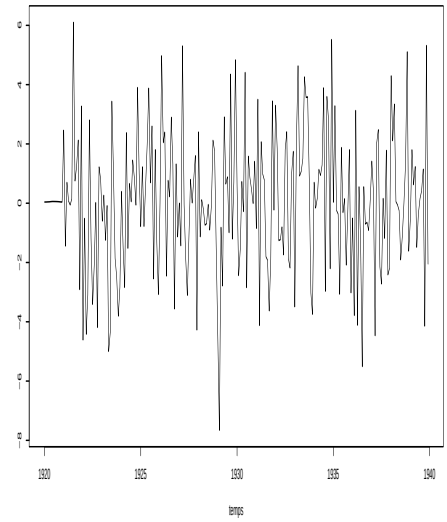
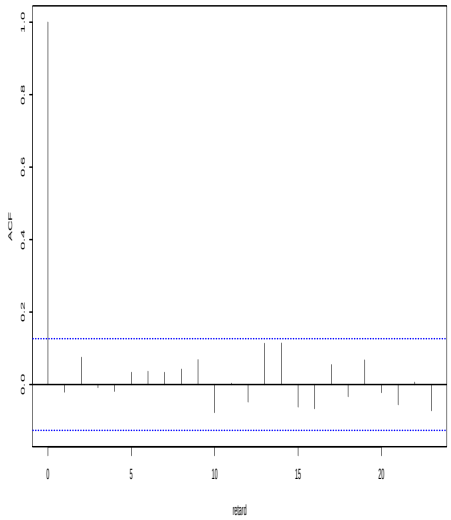


FIGURE 3.5 – ACF de la série brute, ACF et PACF de la série différenciée à l'ordre 12





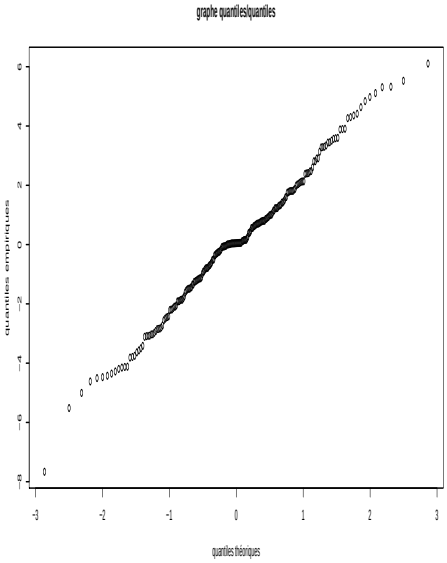


FIGURE 3.6 – graphe des résidus, ACF des résidus et comparaison des quantiles des résidus avec ceux d'une gaussienne

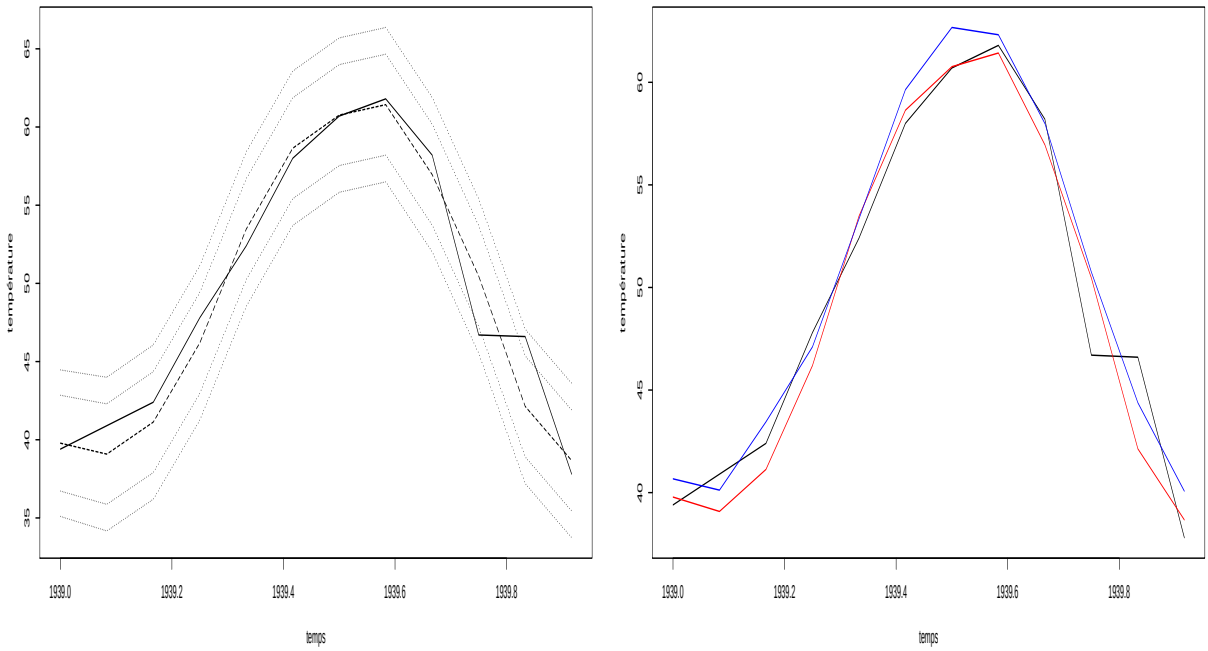


FIGURE 3.7 – Prévision des températures pour l'année 1939 (à gauche, les vraies valeurs sont en trait plein, les prévisions en pointillés ainsi que les intervalles de confiance à 80 et 95 %) et comparaison avec la méthode de Holt-Winters (à droite, courbe bleue pour H-W et rouge pour le SARIMA)

3.4 L'approche de Box et Jenkins

L'approche de Box et Jenkins est une méthode générale pour choisir un modèle et construire des prévisions à l'aide des processus ARMA. Elle est basée sur différentes étapes dont quelques-unes ont déjà été discutées dans ce chapitre. Nous la décrivons pour les processus ARIMA mais elle se généralise aussi aux processus SARIMA.

3.4.1 Choix du triplet (p, d, q) et estimation des paramètres

La première étape est de choisir des valeurs raisonnables pour les ordres p, d et q à partir de statistiques descriptives (graphiques de la série, autocorrélations empiriques).

– Le premier ordre à choisir est l'ordre de différenciation d. On peut suspecter la non-stationnarité de la série à l'aide des autocorrélations empiriques (typiquement la décroissance de l'ACF n'est pas rapide). On peut alors différencier la série (application de l'opérateur $1 - B$). On peut aussi regarder si la série semble linéaire ou quadratique en moyenne (on rappelle que l'opérateur $(1 - B)^k$ permet de transformer un polynôme de degré k en une constante). Des test dits de stationnarité sont diponibles (en complément, nous y reviendrons plus loin). Il

faut cependant faire attention à la sur-différenciation qui peut conduire à des intervalles de prévision trop larges ou à des ARMA non inversibles (les calculs de vraisemblance et les résultats théoriques des estimateurs sont basés sur l'inversibilité). Le cas $d > 2$ est rare en pratique.

Remarque. Il n'est pas toujours possible d'atteindre la stationnarité ainsi. Pour certaines séries qui présentent une tendance de type exponentielle ou un variance qui changent avec le temps, il est parfois nécessaire d'appliquer la transformation logarithmique ou plus généralement une transformation de Box-Cox : $T_\lambda(X_t) = \frac{X_t^\lambda - 1}{\lambda}$, où $\lambda > 0$. Le cas logarithmique correspond au cas $\lambda = 0$. On remarquera que $T_{1/2}$ correspond à la racine carrée et T_1 à l'identité (modulo un décalage et une homothétie). Certains logiciels proposent d'estimer λ par maximum de vraisemblance en même temps que les autres paramètres du modèle. Attention cependant, dans le cas d'une transformation logarithmique, la prévision linéaire optimale n'est pas $\exp\left(\hat{Y}_T(k)\right)$ lorsque $Y_t = \ln(X_t)$ est modélisé par un ARMA (voir [3] page 203 pour une explication de ce problème).

– Ensuite des ordres p et q peuvent être choisis à l'aide des autocorrélations et des autocorrélations partielles. Si des modèles AR ou MA n'apparaissent pas clairement, on peut choisir des bornes supérieures pour p et q en regardant les retards pour lesquels les autocorrélations et autocorrélations partielles ne semblent plus significatives (cette règle est loin d'être absolue).
– On passe ensuite à la phase d'estimation avec les méthodes discutées précédemment.

3.4.2 Diagnostic

Cette phase consiste à examiner les résidus de l'estimation qui sont définies par

$$\hat{\varepsilon}_t = \hat{\Theta}(B)^{-1}\hat{\Phi}(B)\,(1-B)^d\,X_t.$$

Une première étape est de tracer l'ACF des résidus, c'est à dire

$$h \mapsto \hat{\rho}_\varepsilon(h) = \frac{\sum_{t=1}^{T-h}\hat{\varepsilon}_t\hat{\varepsilon}_{t+h}}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2}$$

et de regarder si on a le comportement d'un bruit blanc. On peut alors choisir de diminuer p ou q ou de forcer des coefficients à être nuls en utilisant les test de significativité des coefficients ($t-$tests). Si on est amené à augmenter p et q, il faut faire attention : un ARMA(p,q) admet une infinité de représentation ARMA$(p+1, q+1)$ (il suffit de multiplier par le même polynôme de degré 1 de chaque côté de l'équation) et les paramètres ne sont plus identifiables. Il vaut mieux d'abord augmenter p (ou q) et tester si $\phi_{p+1} = 0$.
Enfin, le test de Portemanteau ou test de Ljung-Box permet de tester l'hypothèse bruit blanc. La statistique correspondante est

$$Q_m = T(T+2)\sum_{h=1}^{m}\frac{\hat{\rho}_\varepsilon^2(h)}{T-h}$$

dont la loi converge (sous l'hypothèse d'indépendance des bruits) vers une loi $\chi^2(m-p-q)$. Cependant, ce test a le désavatange d'accepter des modèles peu intéressants du point de vue de l'ajustement. On s'en sert plutôt pour disqualifier certains modèles. Notons aussi que le nombre m de retards est à fixer (on prend m habituellement beaucoup plus grand que $p+q$ mais ça dépend aussi de T).

3.4.3 Sélection de modèles et prévision

Lorsque plusieurs modèles passent l'étape précédente, on peut suivant l'objectif les départager avec d'autres critères. On peut par exemple comparer la qualité prédictive. Il suffit pour cela de partager l'échantillon en deux. Une première partie $X_1, \ldots, X_{T-m}$ sert à calibrer le modèle, la deuxième partie sert à comparer un critère d'erreur, par exemple $\sum_{k=1}^{m} \left(X_{T-m+k} - \hat{X}_{T-m}(k)\right)^2$ (on peut aussi remplacer le carré par la valeur absolue mais le carré est plus naturel ici).
Comme en régression linéaire standard, il existe des critères d'information du type AIC ou BIC qui choisissent les ordres p et q qui minimisent l'opposé de la log-vraisemblance maximisée + une pénalité qui dépend du nombre de paramètres $p+q$.
Il n'est pas exclu que plusieurs modèles passent ces étapes : par exemple pour les données du Lac Huron, les deux modèles AR(2) et ARMA(1, 1) valident l'étape diagnostic et les AIC/BIC sont comparables.
L'étape de prévision des valeurs futures se fait alors comme indiqué dans les sections précédentes. On remarquera que dans les formules pour la prévision (expression de la prévision et intervalles de confiance), l'incertitude liée au remplacement des paramètres par les paramètres estimés n'est pas contrôlée.

3.4.4 Exemple d'utilisation d'un ARMA

Nous considérons ici un jeu de données disponible sous SAS : les totaux mensuels de produits en acier expédiés des aciéries américaines entre 1984 et 1992. On a donc $T = 96$. Les autocorrélations empiriques ne contredisent pas la stationnarité de la série. Une autocorrélation importante apparait pour $h = 12$ mais diminue considérablement pour $h = 24$. Il n'y pas besoin de différencier la série pour tenir compte de cette dépendance saisonnière. Par contre, les premières autocorrélations ou autocorrélations partielles empiriques ont l'air significatives (faire attention, on sait juste que les intervalles de confiance en pointillés bleues contiennent les autocorrélations empiriques d'un bruit blanc fort avec probabilité $0,95$).

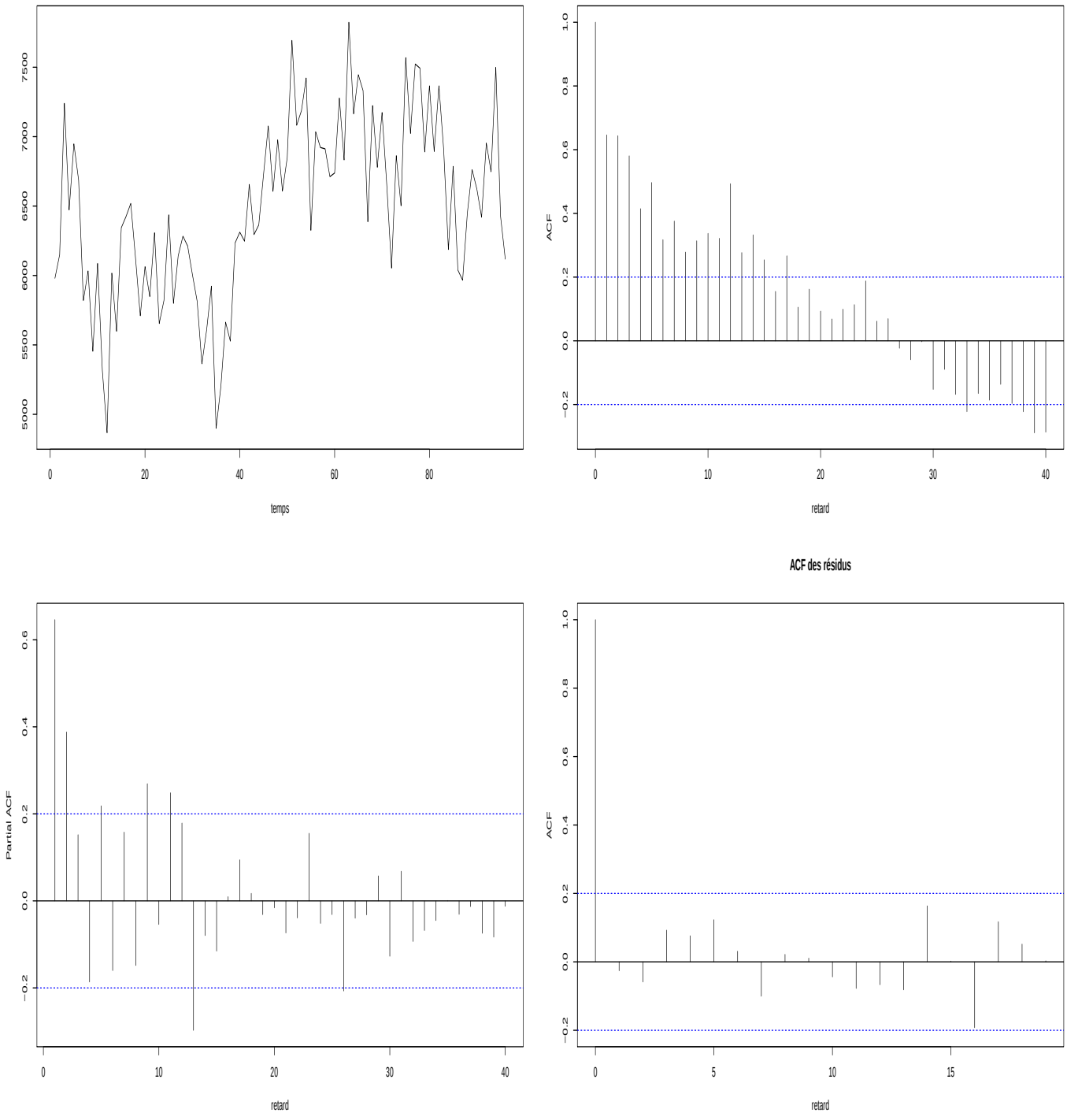


FIGURE 3.8 – Graphe des données, ACF/PACF de la série et ACF des résidus pour le modèle retenu

Il est donc intéressant d'inclure une partie autorégressive et une partie moyenne mobile pour les premiers retards afin de limiter le nombre de coefficients. On regarde des modèles du type

$$\left(1 - \gamma B^{12}\right) \Phi(B) X_t = \Theta(B)\varepsilon_t,$$

en limitant les ordres p et q à 3. Les étapes de diagnostic conduisent à considérer un modèle SARIMA$(2,0,3) \times (1,0,0)_{12}$ mais pour lequel le premier coefficient AR n'est pas déclaré significatif, le troisième non plus et le deuxième coefficient MA non plus. On fixe ces coefficients à 0. Les coefficients du modèle retenus sont donnés ci-dessous (attention, la fonction Arima de R suppose que le deuxième polynôme est donné par $\Theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$, il faut donc prendre l'opposé des coefficients pour se ramener à la formulation adoptée dans ce cours).

```
Series: data
ARIMA(2,0,3)(1,0,0)[12] with non-zero mean

Coefficients:
      ar1     ar2     ma1  ma2     ma3    sar1  intercept
        0  0.5287  0.5810    0  0.4165  0.6542  6410.9715
s.e.    0  0.0943  0.0907    0  0.0871  0.0816   324.8637

sigma^2 estimated as 103551:  log likelihood=-694.81
AIC=1401.63   AICc=1403.28   BIC=1422.14
```

FIGURE 3.9 – Résultats de la sortie R

Le modèle retenu est aussi celui qui minimise le BIC/AIC pour les contraintes choisies sur p et q. Les autocorrélations empiriques des résidus semblent négligeables et la $p-$valeur du test de Ljung-Box est supérieure à 0.5 pour les premiers retards.

Remarque. En série temporelle, les résultats sont souvent à interpréter avec prudence lorsque la taille de l'échantillon est modéré : les tests et les méthodes d'estimation et de sélection des ordres sont justifiés asymptotiquement et servent d'indicateurs qu'il faut compléter par des analyses graphiques (utiliser l'approche de Box et Jenkins). Ici les différents critères vont largement dans le même sens mais ce ne sera pas toujours le cas. Parfois plusieurs modèles peuvent être retenus.

3.5 Tests de non-stationnarité

Une série temporelle peut exhiber diverses formes de non-stationnarité. La non-stationnarité peut provenir d'une tendance ou d'une saisonnalité, de coefficients ARMA non stables dans le temps (les coefficients du modèle peuvent changer au cours du temps) ou encore la présence de la racine 1 dans le polynôme $\Phi(B)$ d'autorégression. Concernant le troisième cas, l'exemple le plus est celui de la marche aléatoire $X_t = X_{t-1} + \varepsilon_t$ où $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc, puisque $\Phi(B) = 1 - B$. Tester la présence de la racine 1 dans le polynôme d'autorégression (on parle de test de racine

unité) peut par exemple aider à décider si une série temporelle est un ARIMA ou plus simplement un ARMA, autrement dit si une différenciation supplémentaire est nécessaire pour atteindre la stationnarité (en complément du graphe des autocorrélations empiriques). De plus, lorsqu'une série temporelle semble avoir une moyenne $\mathbb{E}(X_t) = a + bt$, on peut aussi se demander si la série est stationnaire à une tendance linéaire près ou si la non-stationarité apparente est due à la présence d'une marche aléatoire (voir ci-dessous). Il faut aussi faire attention à ces racines unité à cause des régressions dites fallacieuses. Par exemple, si $Y_t = Y_{t-1} + \varepsilon_t$ et $X_t = X_{t-1} + \eta_t$ sont deux marches aléatoires indépendantes, considérer le modèle de régression $Y_t = \beta_0 + \beta_1 X_t + \kappa_t$ conduit à une estimation de β_1 par MCO significativement non nulle (alors qu'aucun lien n'existe entre les séries).

Le comportement des estimateurs en présence de racine unitaire est très différent du cas où les processus en jeu sont stationnaires. Par exemple, dans le modèle $X_t = \phi X_{t-1} + \varepsilon_t$, supposons les données correspondent au cas de la marche aléatoire (cas $\phi_0 = 1$). L'estimateur des MCO vérifie

$$\hat{\phi} = \frac{\sum_{t=2}^{T} X_{t-1} X_t}{\sum_{t=2}^{T} X_{t-1}^2} \Leftrightarrow \hat{\phi} - 1 = \frac{\sum_{t=2}^{T} X_{t-1} \varepsilon_t}{\sum_{t=2}^{T} X_{t-1}^2}.$$

Supposons que $X_0 = 0$. On peut alors montrer à l'aide de quelques calculs que

$$\text{Var}\left(\sum_{t=2}^{T} X_{t-1}\varepsilon_t\right) = \sigma^4 \frac{T(T-1)}{2}.$$

Ceci amène à regarder le comportement de $T\left(\hat{\phi} - 1\right)$ au lieu de $\sqrt{T}\left(\hat{\phi} - 1\right)$. De plus la loi limite de cette statistique n'est pas gaussienne. On peut montrer que pour un bruit blanc fort ε,

$$\lim_{T \to +\infty} T\left(\hat{\phi} - 1\right) = \frac{\frac{1}{2}\left(B_1^2 - 1\right)}{\int_0^1 B_s^2 ds} \text{ en loi,}$$

où $(B_t)_{t\in[0,1]}$ est un mouvement brownien (un mouvement brownien est un processus gaussien, nul en $t = 0$, dont les trajectoires sont continues et tel que $\text{Cov}(B_s, B_t) = \min(s, t)$).

3.5.1 Test de Dickey-Fuller

Dans ce paragraphe, $(\varepsilon_t)_{t\in\mathbb{Z}}$ désigne un bruit blanc fort de variance σ^2. Il existe en fait trois versions de ce test qui correspondent à trois situations bien distinctes.

1. **Cas** 1 : la série semble nulle en moyenne et on veut tester l'hypothèse $H_0 : \phi = 1$ pour le modèle $X_t = \phi X_{t-1} + \varepsilon_t$. Sous l'hypothèse nulle, on a une marche aléatoire. Sous l'hypothèse $H_1 : |\phi| < 1$, le processus est un AR(1). On rejette H_0 lorsque la statistique $T\left(\hat{\phi} - 1\right)$ a une valeur trop faible (on utilise la loi limite introduite précédemment et qui est tabulée dans les logiciels).
2. **Cas** 2 : la série semble avoir une moyenne constante au cours du temps (mais un moyenne non nulle). Le test de Dickey-Fuller est basé sur la décomposition

$$X_t = \mu + \phi X_{t-1} + \varepsilon_t. \tag{3.4}$$

On teste alors $H_0 : \phi = 1, \mu = 0$ contre $H_1 : |\phi| < 1$. Sous l'hypothèse nulle, on a une marche aléatoire et sous l'alternative, un AR(1) décentré. Les paramètres μ et ϕ sont estimés par par moindres carrés ordinaires. Comme dans le cas précédent, le comportement de $\hat{\phi}$ n'est pas classique sous l'hypothèse nulle de non-stationnarité. La statistique $T\left(\hat{\phi} - 1\right)$ converge en loi. La loi limite diffère du cas précédent et s'exprime aussi à partir du mouvement brownien. Il suffit alors de rejetter H_0 si la valeur de la statistique $T\left(\hat{\phi} - 1\right)$ est plus petite que le quantile d'ordre α de la loi limite. Remarquer que sous l'hypothèse H_1, la statistique précédente converge vers $-\infty$. Une autre possibilité (proposée par les logiciels) est de baser le test sur le t de Student de la régression (on divise $\hat{\phi} - 1$ par son écart type estimé). Pour le modèle (3.4), cette statistique ne suit pas une loi de Student mais sa loi asymptotique a été tabulée.

3. **Cas** 3 : la moyenne de la série semble évoluer linéairement avec le temps. Dans ce cas, le test de Dickey-Fuller est basé sur la décomposition

$$X_t = \mu + \nu t + \phi X_{t-1} + \varepsilon_t. \tag{3.5}$$

On teste alors $H_0 : \phi = 1, \nu = 0$ contre $H_1 : |\phi| < 1$. Sous H_0, on a $X_t = X_0 + \mu t + \varepsilon_t + \cdots + \varepsilon_1$. Sous l'alternative, le processus se comporte comme un AR(1) (modulo une tendance affine). Les paramètres μ, ν et ϕ sont estimés par moindres carrés ordinaires. La statistique $T\left(\hat{\phi} - 1\right)$ converge en loi et la loi limite a été tabulée. On rejette H_0 si la valeur de la statistique $T\left(\hat{\phi} - 1\right)$ est plus petite que le quantile d'ordre α de la loi limite. Le test peut être aussi basé sur la statistique de Student.

Le choix du type de test dépend de la série observée. Ce choix est important : on peut par exemple rejeter H_0 si on choisit le cas 1 alors que la série est non centrée et avec une racine unité.

3.5.2 Test de Dickey-Fuller augmenté (ADF)

Considérer un processus autorégressif d'ordre 1 n'est pas assez réaliste pour justifier la pertinence des tests du paragraphe précédent. On généralise l'approche précédente au cas de p retards. Considèrons l'analogue de (3.5) avec la décomposition

$$X_t = \mu + \nu t + \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t. \tag{3.6}$$

Lorque le polynôme $\Phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$ admet toutes ses racines à l'extérieur du disque unité, la série est stationnaire à une tendance déterministe près. Dans ce cas, on a $\sum_{j=1}^{p} \phi_j < 1$. Le cas non-stationnaire correspond alors au cas $\sum_{j=1}^{p} \phi_j = 1$ (présence d'une racine unité dans le polynôme). La décomposition (3.6) peut être réécrite

$$\Delta X_t = \mu + \nu t + \pi X_{t-1} + \sum_{j=1}^{p-1} \psi_j \Delta X_{t-j} + \varepsilon_t$$

avec $\pi = \sum_{j=1}^{p} \phi_j - 1$, $\Delta X_t = X_t - X_{t-1}$ et $\psi_j = -\sum_{i=j+1}^{p} \phi_j$ pour $1 \leq j \leq p-1$. Comme pour le cas $p = 1$, trois situations peuvent être examinées.

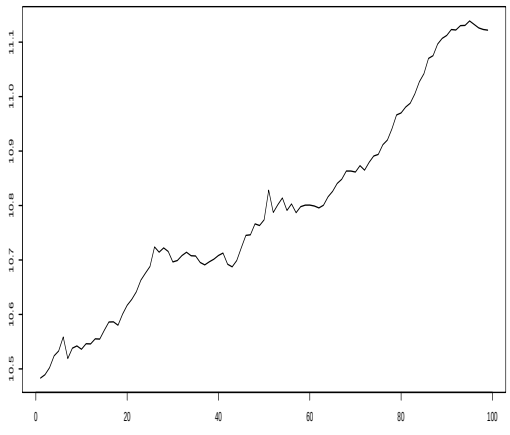
– Lorsque la série semble nulle en moyenne (on fixe $\mu = \nu = 0$ dans (3.6)), on teste $H_0 : \pi = 0$ contre $H_1 : \pi < 0$. La procédure est alors basée sur l'estimateur de π par moindres carrés ordinaires.
– Lorsque la moyenne de la série ne semble avoir de tendance (on impose $\nu = 0$ dans (3.6)), on teste $H_0 : \mu = 0, \pi = 0$ contre $H_1 : \pi < 0$. La procédure est aussi basée sur l'estimateur de π par MCO.
– Lorsque la moyenne de la série semble évoluer linéairement avec le temps, on teste $H_0 : \nu = 0, \pi = 0$ contre $H_1 : \pi < 0$. Lorsque l'alternative semble préférable, il est possible d'étudier directement la série sans la différencier (estimation par moindres carrés ou par maximum de vraisemblance).

Remarques

1. Pour choisir p (attention c'est $p - 1$ pour SAS ou R), on peut prendre un choix initial p_0 et réduire l'ordre si le dernier coefficient ψ_{p-1} n'est pas significatif dans la régression.
2. Ces tests, comme la plupart des tests, sont conservatifs, c'est-à-dire qu'ils ont tendance à garder l'hypothèse nulle (problème de puissance). C'est pour cela qu'il faut aussi se fier aux ACF.

3.5.3 Exemples

– Le premier exemple (Figure 3.10) est tiré de [1]. On voit clairement apparaitre une tendance linéaire. La fonction ur.df de R (package urca) permet d'appliquer le test de Dickey-Fuller. Il y a trois options qui correspondent aux trois cas de figures discutés précédemment, "none", "drift" et "trend". Ici, l'option "trend" est utilisée pour prendre en compte la tendance affine. On choisit $p-1=3$. La valeur du t de Student (ici -2.2389) est plus grande que le seuil -3.15 du test à 10%. La $p-$valeur est donc plus grande que 10% et on peut garder l'hypothèse de racine unité. A noter que les valeurs 3.7382 et 2.5972 correspondent aux valeurs des statistiques de Fisher pour tester $H_0 : (\mu, \nu, \pi) = (0,0,0)$ et $H_0 : (\nu, \pi) = 0$ respectivement dans la régression (on rejette H_0 si la valeur est plus grande que les valeurs données sur la ligne "phi2" et "phi3").




```
Value of test-statistic is: -2.2389 3.7382 2.5972

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.04 -3.45 -3.15
phi2  6.50  4.88  4.16
phi3  8.73  6.49  5.47
```

FIGURE 3.10 – Logarithme de la dépense de consommation trimestrielle au Royaune-Uni et sortie R pour le test de Dickey-Fuller

– Le deuxième exemple est une série assez classique : les données de trafic aérien déjà utilisées au Chapitre 1 pour illustrer la prévision par lissage exponentiel. Après avoir pris le logarithme de la série et différencié à l'ordre 12, on trace l'ACF. La décroissance des autocorrélations et le graphe de la série suggère une possible non stationnarité. Le test de Dickey-Fuller (option "drift" lorsque il n' a pas de tendance apparente) ne rejette pas la non-stationnarité à 5% mais la rejette pour $\alpha = 10\%$ ($p-1=12$ ici). Différencier la série est sans doute nécessaire au vu des différents indicateurs (ACF, graphe de la série, test). On pourra consulter [2] p. 296 pour des exemples de modèles pour la série différenciée.

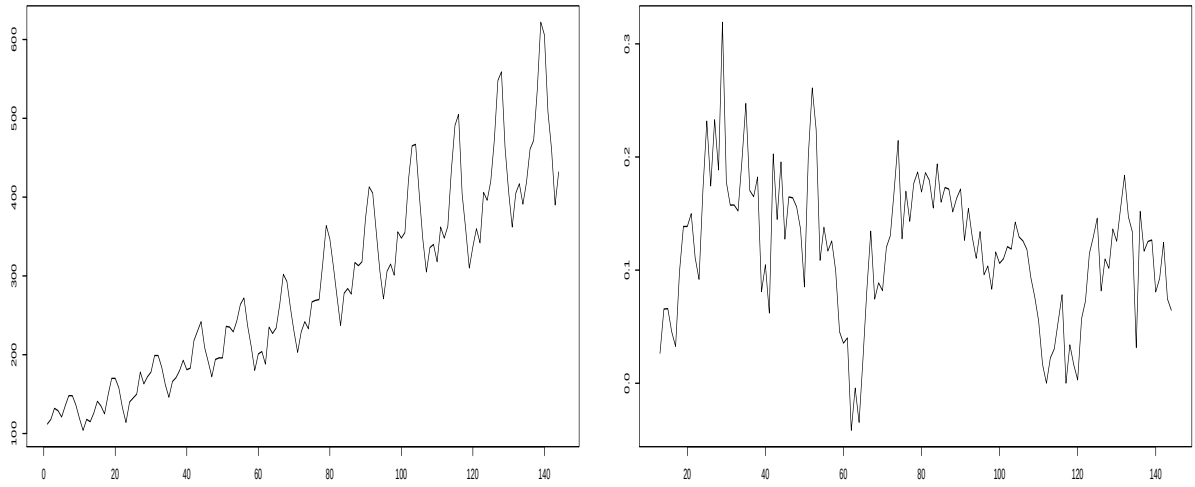
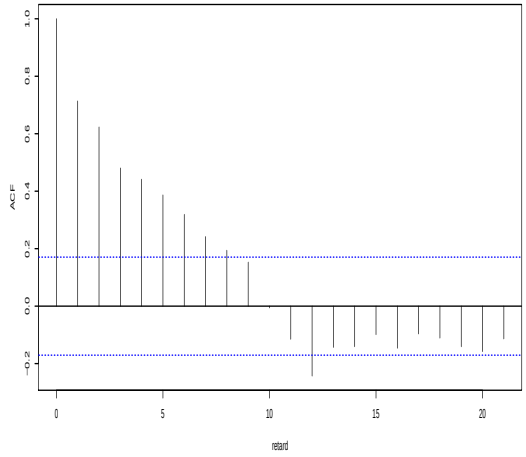


FIGURE 3.11 – Série initiale et différence d'ordre 12 du logarithme de la série



```
Value of test-statistic is: -2.7096 3.7342

Critical values for test statistics:
     1pct  5pct 10pct
tau2 -3.46 -2.88 -2.57
phi1  6.52  4.63  3.81
```

FIGURE 3.12 – ACF de la série différenciée et sortie R pour le test de Dickey-Fuller

Bibliographie

[1] Aragon, Y. (2011) *Séries temporelles avec R.* Springer.

[2] Brockwell, P. J., Davis, R. A. (2006) *Time series, theory and methods.* Second edition. Springer.

[3] Gouriéroux, C., Monfort, A. (1995) *Séries temporelles et modèles dynamiques.* Economica.

[4] Ladiray, D., Quenneville, B. (2000) *Désaisonnaliser avec la méthode X − 11.*