Master 1

Cours de Séries Chronologiques

Notes de Jean-Marc Bardet (Université Paris 1, SAMM)
modifiées par Gabriel Lang (AgroParisTech)

Plan du cours

1. Séries chronologiques ; définition et exemples
2. Modélisation déterministe, tendance et saisonalité
3. Processus aléatoires: premières définitions et propriétés.
4. Modèles ARMA. Définition et propriétés.
5. Identification d'un processus ARMA
6. Modèle ARCH
7. Prédiction pour un processus à temps discret.

References

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

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

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

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

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

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

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

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

Documents accessibles librement sur internet

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

Quelques sites internet intéressants

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

Introduction

Définition

L'objet de ce cours est l'étude de séries chronologiques, de leur modélisation et des applications de ces modèles à la prévision.

Définition 1. *Une série chronologique est une série de mesures à valeurs réelles effectuées à des temps écartés régulièrement.*

L'intervalle de temps entre deux mesures successives dépend de la série ; il peut s'agir d'un jour, d'une semaine, d'une minute... Par extension, il peut s'agir d'intervalles légèrement irréguliers (mois, trimestre, semestres, années, jours ouvrables...). Les dates sont numérotées par des entiers positifs $n = 1, 2, \ldots$. Les données sont de nature très diverse ; il peut s'agir de relevés de phénomènes d'origine naturelle (température à Paris, hauteur de la Seine sous le pont de l'Alma, débit moyen journalier de la Garonne à Toulouse, activité des taches solaires), de séries économiques (indice de la consommation, taux de chômage, prix de matières premières) ou financières (cours boursiers, taux de change).

Questions

Les principaux problémes que l'on se pose à propos de ces séries sont les suivants :

- Peut-on représenter la série par une fonction simple du temps (modélisation)?
- Peut-on prévoir une donnée future à partir des enregistrements déjà effectués (prévision)?
- Y-a-il un instant t, où la série change significativement (changement de régime, détection de rupture)?
- Peut-on déterminer des relations entre la série observée et d'autres séries plus facilement mesurables?

Dans ce cours nous nous concentrerons sur les deux premières questions.

1 Modèles déterministes pour les séries chronologiques

Cette partie est souvent omise ou vite traitée dans les livres consacrés aux séries chronologiques. Pourtant l'estimation de la tendance et de la saisonnalité est essentielle dans la plupart des travaux concrets portant sur les séries chronologiques, en particulier parce qu'elle apporte une information souvent bien plus importante que la partie bruit en vue de la prévision.
Nous nous intéressons à des ensembles de données qui ne sont pas la répétition d'une même mesure dans des conditions équivalentes, mais des mesures prises à différents moments. Contrairement à ce qui est observé dans un échantillon contrôlé, les mesures n'ont aucune raison de donner des valeurs de niveau constant. Les températures d'été à Paris seront plus hautes que les températures en hiver; la valeur d'un placement financier pourra augmenter régulièrement au cours du temps (du moins l'espère-t-on...), une série de prix sera affectée par l'inflation. Selon les séries, nous allons modéliser cet effet général du temps qui peut être selon les séries un changement monotone (croissance ou décroissance) ou une fluctuation régulière. Les méthodes élémentaires pour modéliser ces évolutions générales sont les suivantes

- Si la série montre une fluctation générale périodique et/ou une tendance mais que l'amplitude des petites oscillations est stable, on cherchera à exprimer la série comme une somme d'une fonction simple du temps et d'un bruit de mesure d'amplitude constante (cf serie El niño, figure **??**) ;
- Si les oscillations autour de l'évolution générale sont d'amplitude croissante avec le temps, on proposera un modèle multiplicatif (cf serie indice de prix du blé, figure **??**) ;
- Si le niveau moyen de la série semble fixe et les oscillations autour de ce niveau moyen ne change pas d'amplitude, on considèrera qu'on a affaire à une série aléatoire stationnaire. Dans ce cas, le seul traitement consiste à calculer le niveau moyen.

Figure 1: Température moyenne de surface de la mer, phénomène El Niño en région 3 (90°W à 150°W) et en région 3.4 (120°W à 170°W)

1.1 Modèles additifs

Dans toute la suite, l'observation de la série X à l'instant t sera notée $X(t)$. La décomposition la plus simple revient à considérer qu'il existe un phénomène sous-jacent, déterministe, évoluant lentement qui peut se décrire par une fonction $f(t)$ et que la mesure X_t est le résultat de l'addition) de cette fonction et d'un bruit de mesure aléatoire ε_t :

$$X_t = f(t) + \varepsilon_t. \tag{1}$$

La fonction f est choisie dans une classe de fonctions simples et le bruit est une série de fluctuations aléatoires indépendantes les unes des autres d'espérance nulle et de petite amplitude.

1.1.1 Tendance et composante saisonnière

Définition 2. *La fonction $f(t)$ est habituellement découpée en deux termes :*

$$f(t) = a(t) + S(t),$$

où $t \mapsto a(t)$ et $t \mapsto S(t)$ sont deux fonctions avec

1. *si $t \mapsto S(t)$ est une fonction périodique non nulle de période $r > 0$ telle que $\sum_{i=1}^{r} S(i) = 0$, alors S est la composante saisonnière, saisonnalité de X,*
2. *si $t \mapsto a(t)$ est non nulle, a est la tendance de X.*

Soit X une série chronologique ayant pour tendance a et pour saisonnalité S. On appelle :

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

Figure 2: Indice agrégé du prix du blé en Europe : série brute et différence des logarithmes

La tendance est une fonction d'une famille simple décrite par peu de paramètres : fonction affine (droite), polynôme de degré faible, ou fonction exponentielle. Le choix des familles de fonctions parmi lesquelles on cherche la fonction f se fait souvent à vue, en regardant les graphiques des données sans autre justification théorique : certaines saisonalités sont visibles, d'autres apparaissent plus nettement après certaines transformations des données brutes (transformée de Fourier appelée périodogramme). Une tendance croissante régulière sera modélisée par une tendance affine, par un polynôme d'ordre 2 si on observe un certain creusement... Il n'existe pas de critère objectif pour faire le choix de cette famille.
Le choix de la période de la saisonalité est orienté par des connaissances a priori sur la série. Par exemple, les phénomènes liés aux climats, ainsi qu'une grande partie des séries économiques ont une périodicité annuelle. Mais une série de consommation électrique présente une saisonnalité hebdomadaire en plus.

1.2 Modèles multiplicatifs

Il est aussi possible que la série soit soumise a des fluctuations dont l'amplitude croît en fonction du temps. En ce cas on envisagera une tendance multiplicative:

Définition 3. *On dira que X possède une tendance multiplicative si $X_t = d(t)\,(m + u(t))$ pour tout $t \in T$, où $u = (u_t)_{t\in T}$ est une suite de variables aléatoires centrée et de variance égale à 1 et $\sigma(\cdot)$ est une fonction positive.*

On peut transformer cette série en appliquant la fonction logarithme :

$$Y_t = \log(X_t) = \log(d(t)) + \log(m) + \log(1 + u_t/m),$$

et le modèle mmultiplicatif est approché par un modèle additif. Un exemple de série à comportement multiplicatif est une série de valeur d'actifs financiers. L'investisseur n'est pas intéressé par le prix mais par le rendement entre deux dates défini par le ratio :

$$Y_t = \frac{X_{t+1} - X_t}{X_t}.$$

Lorsque les dates sont rapprochées et que le rendement est petit par rapport à 1, cette quantité est proche du log-ratio :

$$r_n = \log\left(\frac{X_{n+1}}{X_n}\right) = \log(X_{n+1}) - \log(X_n).$$

C'est cette quantité transformée qui sera modélisée par un modèle additif, car elle présente généralement des fluctuations d'amplitude plus constante. La figure **??** présente un indice du prix du blé en Europe depuis les années 1500. La première figure représente la séie brute, la deuxiéme la série des log-ratio. on observe que le niveau moyen du log-ratio est relativement constant et que les fluctuations sont d'amplitude relativement constante.

1.3 Exemple d'estimation de la tendance

Les séries à comportement multiplicatif étant généralement transformées en séries à comportement additif, nous ne traiterons que ces dernières. Supposons pour simplifier que la tendance est une droite et que les fluctuations sont centrées et petites autour de cette tendance :

$$X_t = at + b + \varepsilon_t$$

Il y a deux méthodes élémentaires pour déterminer a et b: la régression linéaire simple et la méthode de filtrage. Nous présentons ces méthodes sur le cas simple de la tendance linéaire puis nous les généraliserons.

1.3.1 Régression linéaire simple

La première consiste à choisir une mesure de l'écart entre les points observés et les points de la droite proposée $D(a,b)$ d'équation $ct+d$(on calculera en pratique la somme du carré des écarts) et de minimiser cette distance :

$$(\hat{a},\hat{b}) = \text{Argmin}_{a,b}\mathcal{D}(D(a,b),X)$$

où $\mathcal{D}(D(a,b),X) = \sum_{i=1}^{n}(X_i - D(a,b)_i)^2 = \sum_{i=1}^{n}(X_i - ai - b)^2$.
Deux méthodes de résolution de ce problème sont possibles; la première est analytique et consiste à rechercher le minimum de la fonction en dérivant par rapport à chacune des variables et à rechercher les valeurs qui annulent les dérivées; la fonction étant quadratique, on déterminera ainsi un point minimum.

$$\frac{\partial\mathcal{D}(D(a,b),X)}{\partial a} = = \sum_{i=1}^{n}(X_i - ai - b)i$$
$$\frac{\partial\mathcal{D}(D(a,b),X)}{\partial a} = = \sum_{i=1}^{n}(X_i - ai - b)$$

Nous obtenons un système linéaire que nous résolvons par substitution

$$\sum_{i=1}^{n}(X_i - \hat{a}i - \hat{b})i = 0,$$
$$\sum_{i=1}^{n}(X_i - \hat{a}i - \hat{b}) = 0.$$

Si on note $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ et $\bar{i} = \frac{1}{n}\sum_{i=1}^{n}i = \frac{n-1}{2}$, la deuxième équation donne

$$\hat{b} = \bar{X} - \hat{a}\bar{i}.$$

En substituant $\hat{b}$ dans la première équation et en soustrayant $\bar{i}$ fois la deuxième équation, on obtient

$$\hat{a} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(i - \bar{i})}{\sum_{i=1}^{n}(i - \bar{i})^2},$$

ce qui permet de déterminer $\hat{b}$.
L'autre méthode de résolution est géométrique. On considère les vecteurs

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},\ Y = \begin{pmatrix} 1 \\ \vdots \\ n \end{pmatrix},\ Z = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Le problème de minimiser la distance euclidienne correspondant à $\mathcal{D}$ entre X et $aY+bZ$ revient à rechercher le projeté orthogonal de X sur le plan vectoriel engendré par Y et Z. Ce projeté noté $PX = \hat{a}Y + \hat{b}Z$ est défini par les deux propriétés suivantes :

$$(X - PX) \perp Y$$
$$(X - PX) \perp Z,$$

soit, traduit en terme de produit scalaire

$$\sum_{i=1}^{n} (X_i - \hat{a}i - \hat{b})i = 0$$
$$\sum_{i=1}^{n} (X_i - \hat{a}i - \hat{b}) = 0,$$

ce qui redonne le même système d'équations.
La procédure permet donc bien d'identifier $\hat{a}$ et $\hat{b}$. On appelle cette méthode régression linéaire de X.

1.3.2 Filtrage

Une deuxième méthode consiste à appliquer un opérateur de différence sur la série des X_i:

$$Y_i = X_{i+1} - X_i, \text{ pour } i = 1, \ldots, n-1$$

La série Y est à peu près constante et égale à a. Sa moyenne empirique est un bon estimateur de a. Ayant identifié a, on soustrait ai à tous les X_i et on obtient une série de moyenne empirique proche de b. On appelle cette méthode filtrage de la série X par l'opérateur linéaire de différence.
Les deux méthodes ne sont pas de qualitééquivalente. La régression linéaire donne des estimateurs dont les propriétés sont connues dans le cadre du modèle choisi. Nous rappelerons ces propriétés dans le cadre général de la régression linéaire multiple ci-dessous. La qualité des estimateurs obtenus par filtrage est moins sûre. Ils seront utilisés comme pis-aller, lorsque que le cadre de modèle est moins précis. Nous allons généraliser maintenant ces deux méthodes à des tendances et saisonnalités plus complexes.

1.4 Estimation semi-paramétrique par régression

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

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

où $g_i(t) = \mathbb{I}_{\{t=i,\ [r]\}}$ sont r-périodiques (on a ainsi $\sum_{i=1}^{r} S(i) = 0$).
Nous utilisons les notations vectorielles suivantes :

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},\ F_i = \begin{pmatrix} f_i(1) \\ \vdots \\ f_i(n) \end{pmatrix},\ G_i = \begin{pmatrix} g_i(1) - g_r(1) \\ \vdots \\ g_i(n) - g_r(n) \end{pmatrix},\ U = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Le modèle de régression s'écrit alors vectoriellement:

$$X = \sum_{i=1}^{k} a_i F_i + \sum_{i=1}^{r} s_i G_i + U,$$

et même de manière plus compacte

$$X = Za + U,$$

avec $Z = (F_1 \cdots F_k G_1 \cdots G_{r-1})$ et $a = \begin{pmatrix} a_1 \\ \vdots \\ a_k \\ s_1 \\ \vdots \\ s_r \end{pmatrix}$

Proposition 1. *On peut estimer les coefficients (a_i) et (s_i) par une régression par moindres carrés en minimisant une distance dans $\mathbb{R}^n$:*

$$\| X - Za \|^2 .$$

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

Pour montrer le premier point, on constate que la minimisation de la distance euclidienne correspond comme précédement à la détermination de la projection orhogonale du vecteur X sur le plan engendrés par les F_i et les G_i. Les conditions d'orthogonalité définissant cette projection sont $^t F_i(X - Z\hat{a}) = 0$ pour tout $i = 1, \ldots, k$ et $^t G_j(X - Z\hat{a}) = 0$ pour tout $j = 1, \ldots, d$, ce qui peut se résumer par

$$^t Z(X - Z\hat{a}) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

On utilise alors le résultat suivant: la matrice $^t ZZ$ est inversible si et seulement si les vecteurs colonnes de Z forment un système libre ; dans ce cas $\hat{a} = (^t Z\, Z)^{-1\,t} ZX$ est l'unique solution de l'équation précédente. La démonstration du deuxième point est analogue pour la distance correspondant à la matrice définie positive Σ. On déduit aisément de ces expressions un premier résultat de convergence pour les estimateurs de la tendance et de la saisonnalité:

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

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

Remarque.

- On doit bien vérifier a priori que les fonctions de régression génèrent un système libre de vecteurs colonnes.
- Ces estimateurs sont sans biais.
- Ces estimateurs sont linéaires au sens où ils s'écrivent comme des combinaisons linéaires des X_i. L'estimateur des moindres carrés ordinaires réalise la plus petite variance parmi tous les estimateurs linéaires sans biais.

- Une telle régression nécessite un choix a priori des fonctions f_i. Par exemple, on peut considérer que $a(\cdot)$ est un polynôme (typiquement $f_i = t^i$), ou bien que $a(\cdot)$ est un polynôme trigonométrique (typiquement $f_i(t) = \sin(i\,t)$ ou bien $f_i(t) = \cos(i\,t)$). En pratique, on utilisera plutôt une régression polynomiale lorsque les données semblent prendre une certaine direction, alors que la modélisation par un polynôme trigonométrique permet d'avoir des prédictions qui restent dans le même ordre de grandeur que les données connues. Pour aller un peu plus loin, on peut même essayer de décomposer la tendance $a(t)$ dans une certaine base de fonction(ce qui peut être fait par la transformée de Fourier ou avec une base d'ondelettes par exemple).

Comment choisir le nombre k de fonctions f_i considérées ? Cela revient dans le cas polynomial à choisir le degré maximal du polynôme que l'on utilise dans la régression. Il est clair que ce nombre k doit être nettement inférieur à $N - r$, sinon les différents paramètres a_i ne pourront pas être correctement estimés. Une technique efficace pour obtenir "mathématiquement" un choix "optimal" de k est d'utiliser un critère de sélection de modèle. Le **critère BIC** (Bayesian Information Criterium) que nous présenterons plus tard dans ce cours est un choix intéressant. Il consiste à mettre en balance la qualité de la représentation de la série par la tendance avec le nombre de paramètres de la tendance. En pratique, on cherche la valeur de k qui maximise le critère approché :

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

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

1.4.1 Cas particulier de l'estimation de la saisonnalité

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

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

et

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

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

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

1.5 Application d'un filtre linéaire

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

Définition 4. *Soit $X = (X_k, k \in \mathbb{Z})$ un processus à temps discret. On appelle filtre linéaire une famille de réels $a = (a_i)_{n \in I}$, où $I \subset \mathbb{Z}$, et la série filtrée est $Y = (Y_k, k \in \mathbb{Z})$ telle que $Y_k = \sum_{i \in I} a_i X_{k-i}$.*

Dans cette partie, nous appliquerons les filtres aux observations de la série et nous n'utiliserons que des filtres de longueur finie. Nous verrons plus tard que l'on peut également considérer des filtres de longueur infinie pour construire les modèles dont nous avons besoin.

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

Propriété 3. *On considère les filtres linéaires particuliers suivants:*

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

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

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

Propriété 4. *Quelque soit le filtre linéaire utilisé, si Y est le processus à temps discret filtré, on peut reconstruire le processus original X à partir de Y (+ les premières valeurs de X).*

Remarque.

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

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

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

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

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

1.6 Estimation non-paramétrique de la tendance

On supposera donc ici que la série n'admet pas de composante saisonnière, et que l'on a $X_t = a(t) + u_t$ pour $t \in \{t_1, \cdots, t_N\}$. Plutôt que de poser a priori un modèle pour la fonction a comme cela a été fait avec la régression, on peut estimer directement cette fonction avec une méthode non-paramétrique. Dans ce type de méthode, la classe des fonctions envisagées est plus large : au lieu de polynômes on considère par exemple l'ensemble des fonctions dérivables à dérivée bornée. On présente ici deux types de méthodes:

• Méthode d'estimation par noyau.

L'idée est d'approcher la valeur de a en un temps t_0 par une moyenne pondérée des valeurs pour des temps proches de t_0, comme si a était localement constante. L'exemple le plus simple est la moyenne arithmétique mobile ($\widehat{a}_\ell(t) = \frac{1}{2k+1}\sum_{i=-k}^{k} X_{t+i}$), où on approche $a(t)$ en moyennant sur les valeurs de temps proche de t. L'intérêt réside dans le fait que la variance du bruit u_t baisse lorsqu'on le moyenne, ce qui se montre facilement dans le cas où le bruit u est un bruit blanc. Mais il faut choisir k : trop grand, on lissera trop jusqu'à n'obtenir qu'une constante pour tout t, trop petit, on ne fera guère mieux que d'estimer $a(t)$ par X_t, et la fonction a sera donc très irrégulière. On pourra également considérer des pondérations à décroissance exponentielle qui prennent en compte toutes les observations disponibles : pour $\beta \in [0,1]$, on définit:

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

L'influence est plus grande pour les données correspondant aux dates les plus proches de t. Il n'y a plus de séparation brutale entre observations prises en compte ou ignorées dans la moyenne mais une décroissance progressive de l'influence des observations. Plus β est proche de 1, plus le nombre de données ayant une influence sur la moyenne calculée est grand.
Le choix de β peut se faire par validation croisée. La méthode consiste à supprimer le terme correspondant à l'observation au temps t :

$$\widehat{a}_\beta^{(t)}(t) = \sum_{i=-t+1, i\neq 0}^{n-t} \frac{\beta^{|t-i|} X_{t+i}}{\sum_{i=-t+1, i\neq 0}^{n-t} \beta^{|t-i|}}.$$

Cela revient à regarder si la prise en compte des points voisins permet d'approcher l'observation X_i de façon satisfaisante ; on choisit la valeur de β qui minimise l'erreur de cette approximation :

$$\widehat{\beta}_n = \text{Argmin}_\beta \sum_{i=1}^{n} (\widehat{a}_\beta^{(i)}(i) - X_i)^2.$$

Les estimateurs à noyau sont une extension de ces méthodes de moyennes mobiles. On construit une moyenne pondérée dont les poids sont calculés à partir d'une fonction K appelée noyau. L'étalement de la moyenne de part et d'autre du point t est réglé par un paramètre h appelé paramètre de fenêtre. Les estimateurs à noyau K généralisent les méthodes de moyennes mobiles sur deux points :

- La fonction K permet de définir des poids qui ne sont pas tous égaux dans la moyenne mobile ; comme dans le cas de la moyenne exponentielle, ces poids sont en général décroissants quand on s'éloigne du temps t.
- La taille de fenêtre h qui décide du nombre de données prises en compte dans la moyenne mobile, peut être optimisée en fonction des données examinées par des procédures de validation croisée.

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

Pour un noyau K et un paramètre de fenêtre h, on définit l'estimateur de la tendance a par:

$$\widehat{a}_{n,h}(t) = \frac{\frac{1}{nh}\sum_{j=1}^{n} X_j K\big(\frac{t-j}{h}\big)}{\frac{1}{nh}\sum_{j=1}^{n} K\big(\frac{t-j}{h}\big)}.$$

Sous certaines hypothèses sur la vitesse avec laquelle h converge vers 0 en fonction de n, on peut montrer que $\widehat{a}_{n,h}(t) \to a(t)$ pour tout t lorsque a est suffisamment régulière. L'idée de la démonstration est que pour $x_0 \in R$ et $h > 0$ dit taille de fenêtre, $\frac{1}{h} K\big(\frac{x - x_0}{h}\big)$ converge vers une masse de Dirac lorsque $h \to 0$, dans le sens où: $\int_R \frac{1}{h} K\big(\frac{x - x_0}{h}\big) f(x)dx \to f(x_0)$ quand $h \to 0$, pour toute fonction f.
Comme dans le cas des poids exponentiels, on peut également estimer de façon automatique une taille de fenêtre $\widehat{h}$ adaptée aux données et optimale en un certain sens: pour ce faire, on utilise par exemple le principe de la validation croisée, c'est-à-dire que pour $h > 0$ fixé et pour chaque $i = 1, \cdots, n$, on calcule $\widehat{a}_{N,h}^{(i)}(i)$ à partir de l'échantillon $(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$. Puis on calcule l'écart entre $\widehat{a}_{n,h}^{(i)}(i)$ et X_i et on choisit la valeur de h qui minimise les écarts :

$$\widehat{h} = \text{Argmin}_{h>0} \sum_{i=1}^{n} (\widehat{a}_{n,h}^{(i)}(i) - X_i)^2.$$

• Régressions localisées

Pour estimer la tendance, il est aussi possible d'utiliser des **régressions localisées** de type Spline, Loess ou Lowess. Celles-ci s'obtiennent en fixant une taille de fenêtre et on fait une régression polynomiale (pour les Splines, de degré 3) ou linéaire mais pondérées (pour Loess ou Lowess) de X dans la fenêtre que l'on fait glisser. En fait ce sont également des estimations non-paramétriques, au sens où elles ne supposent pas connues les fonctions pouvant composer la tendance (comme dans le cas de la régression), mais juste une certaine régularité de la fonction tendance.

Définition 5 (Procédure de calcul d'un estimateur LOWESS (LOcally WEighted Scatterplot Smoothing)). *On se donne une observation $(X_1, \cdots, X_n)$, un nombre d'itération j, une proportion p entre 0 et 1, une fonction de poids initiale et une méthode de modification des poids $u(x) = (15/16)(1-x)^2 \mathbb{I}_{|x|<1}$*

- *pour chaque point i de l'échantillon, on effectue une régression linéaire pondérée sur les pn points les plus proches de i dans l'échantillon.*

$$(\hat{a}_i, \hat{b}_i) = Argmin_{a,b} \sum_{k=1}^{pn} W_{i,k}(X_i - ai - b)^2,$$

  où k est un des pn plus proches voisins de i et le poids correspondant $W_{i,k} = K(|i-k|/p)/\sum_k K(|i-k|/p)$.

- *On calcule les erreurs de prévision $\hat{\varepsilon}_i = \hat{a}_i i + \hat{b}_i - X_i$.*
- *On définit le modificateur de poids de chaque point par $\delta_i = u(\hat{\varepsilon}_i/6m)$ où m est la médiane des $\hat{\varepsilon}_i$.*
- *On modifie les poids $W'_{i,k} = \delta_i W_{i,k}/\sum_k \delta_i W_{i,k}$ et on recalcule la régression linéaire pondérée de ces nouveaux poids sur les pn points les plus proches de i.*

On réitère la procédure de calcul des poids et les régressions correspondante j fois. L'estimateur LOWESS de la tendance est donné par la dernière régression

$$\hat{a}^{LOWESS}(i) = \hat{a}_i i + \hat{b}_i$$

Cette méthode permet de d'adapter les poids aux données de manière robuste, au sens où le déplacement de quelques points dans l'observation ne modifie pas l'estimateur obtenu de façon importante. Mais comme dans les méthodes non paramétriques précédentes, la proportion p de données prises en compte a une influence très forte ; si elle est proche de 1, l'estimateur de la tendance est une droite de régression, si elle baisse on obtient une courbe lisse qui suit les données.
Pour conclure, il faut noter que la généralité des estimateurs non-paramétriques se paye souvent par une moins bonne vitesse de convergence de $\widehat{a}_n$ vers a que pour les estimateurs paramétriques. Par exemple, si la vraie tendance a est un polynôme, la régression multiple polynomiale est plus efficace.

1.7 Prévision

L'utilisation de modèles déterministes pour la prévision est élémentaire. Une fois déterminée la fonction f à partir des données observées $X_1, \cdots, X_n$, on propose une prévision à la prochaine date par $\hat{X}_{n+1} = f(n+1)$. La précision de cette prévision peut elle aussi être évaluée par une estimation de la varience de $\hat{\varepsilon}_{n+1} = \hat{X}_{n+1} - f(n+1)$. Cette estimation est réalisée par la variance empirique de la série des écarts $\hat{\varepsilon}_i = X_i - f(i)$, considèrée comme une série de variables aléatoires indépendantes de même loi. La donnée de cette variance empirique permettra de proposer un intervalle de confiance autour de la prévision.

2 Modèles aléatoires

Les méthodes précédentes font l'hypothèse que les écarts à la fonction f sont des séries de variables aléatoires indépendantes ; si cette hypothèse est vraie, la prévision que nous avons est la meilleure réalisable, car on ne peut rien prévoir de la série des ε_i, de même qu'on peut prévoir le tirage suivant lorsqu'on lance successivement un dé. Si au contraire les résidus ne sont pas indépendants, il est possible de prévoir en

partie les résidus futurs et donc de diminuer encore l'imprécision de la prévision. Nous allons maintenant construire rigoureusement des modèles aléatoires dépendants pour décrire ces séries de résidus dépendants, ce qui nécessite un rappel sur les probabilités élémentaires et l'introduction de définitions nouvelles, en particulier celle de processus aléatoire.

2.1 Variable aléatoire

On rappelle qu'un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ se compose d'un ensemble Ω, d'un ensemble $\mathcal{A}$ (appelé tribu) de parties de Ω appelées événements, stable par intersection, union infinie et complémentarité et d'une mesure positive $\mathbb{P}$ de masse 1.

Définition 2. *Une variable aléatoire réelle X est une fonction mesurable de $(\Omega, \mathcal{A})$ dans $(\mathbb{R}, \mathcal{B})$ où $\mathcal{B}$ est la tribu de $\mathbb{R}$ engendrée par les intervalles ouverts. Une fonction f est dite mesurable si toute image réciproque par f d'une partie de $\mathbb{R}$ appartenant à $\mathcal{B}$ appartient à $\mathcal{A}$.*

Une variable aléatoire X est définie comme une fonction. Mais elle est surtout utilisée pour définir une loi de probabilité sur $\mathbb{R}$ par transport de Ω sur $\mathbb{R}$. On définit une probabilité $\mathbb{P}'$ sur les intervalles I de $\mathbb{R}$ par $\mathbb{P}'(I) = \mathbb{P}(X^{-1}(I)) = \mathbb{P}\{\omega; X(\omega) \in I\}$.

2.2 Vecteur aléatoire

Nous pouvons généraliser la définition précédente pour définir conjointement plusieurs variables aléatoires :

Définition 3. *Un vecteur aléatoire réel de dimension k X est une fonction mesurable de $(\Omega, \mathcal{A})$ dans $(\mathbb{R}^k, \mathcal{B})$ où $\mathcal{B}$ est la tribu de $\mathbb{R}^k$ engendrée par les pavés ouverts.*

Les coordonnées du vecteur aléatoire sont des variables aléatoires.

2.3 Lois de probabilité

Puisqu'une loi de probabilité sur un espace quelconque peut être transportée sur $\mathbb{R}$ par une variable aléatoire X, la définition d'une variable aléatoire est souvent réduite à la donnée de la loi de probabilité sur $\mathbb{R}$ qu'elle induit. Cette probabilité induite est entièrement déterminée par la donnée des probabilités de chaque intervalle ouvert de $\mathbb{R}$; il suffit en pratique de spécifier la fonction $F(x) = \mathbb{P}(]-\infty, x])$, fonction croissante de x. Parmi les lois de probabilité sur $\mathbb{R}$, on distingue

- Les lois discrètes dont la fonction de répartition est constante par morceaux et saute en un nombre dénombrable de points (une fonction croissante ne peut pas sauter plus souvent que celà). L'ensemble de ces points est appelé support de la probabilité. Ces lois sont spécifiées en donnant l'amplitude du saut en chaque point du support. L'exemple de loi de ce type est le résultat du tirage d'un dé ;
- Les lois de probabilité continues qui donnent une probabilité nulle à tout ensemble ne contenant pas un intervalle ouvert de $\mathbb{R}$. Les lois de probabilités continues (sous-entendu par rapport à la mesure de Lebesgue sur $\mathbb{R}$) se définissent par leur densité par rapport à cette mesure. Il y a analogie avec la répartition de la masse sur une tige matérielle inhomogène ; certaines parties peuvent être plus pesantes que d'autres, mais la masse de chaque point est nulle. On définit en chaque point une densité ; la masse d'un segment de la tige est donné par l'intégrale de la densité sur ce segment.

Il existe de plus des cas mixtes de lois continues et discrètes ainsi que des lois de probabilités dont les supports ne sont ni des intervalles ni des points de $\mathbb{R}$.

2.4 Dépendance des variables aléatoires

Deux variables aléatoires réelles X et Y sont indépendantes si et seulement si pour tout intervalle A et B de $\mathbb{R}$:

$$\mathbb{P}(X \in A \text{ et } Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

La dépendance de deux variables se mesure par la covariance:

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

- Pour des variables d'espérance nulle, la covariance est un produit scalaire. Deux variables seront dites orthogonales si leur covariance est nulle.
- Deux variables indépendantes sont orthogonales.
- Deux variables orthogonales ne sont pas toujours indépendantes (même si ce sont des variables de loi gaussienne).

Soit $X = (X_1, \ldots, X_p)$ un vecteur aléatoire ; on associe à X la matrice de covariance Σ formée des $\mathrm{cov}(X_i, X_j)$ pour i et j variant de 1 à p. Cette matrice est symétrique et positive. Si $(a_1, \ldots, a_p)$ et $(b_1, \ldots, b_p)$ sont des vecteurs réels :

$$\mathrm{cov}(a_1X_1 + \cdots + a_pX_p, b_1X_1 + \cdots + b_pX_p) = (a_1, \ldots, a_p)\Sigma\, {}^t(b_1, \ldots, b_p).$$

Proposition 2. *Soit X un vecteur aléatoire de dimension d et P une matrice de taille $d \times d$. Si Y est un vecteur aléatoire tel que $Y = PX$, alors la matrice de covariance de Y est $P\Sigma\, {}^tP$.*

Remarque.

La confusion de la variable aléatoire avec la loi de probabilité qu'elle induit sur $\mathbb{R}$ ne pose pas de problème en général. Il faut cependant remarquer que si on définit plusieurs variables aléatoires conjointement comme coordonnées d'un vecteur aléatoire, on ne peut plus considérer comme équivalent de spécifier les lois induites par chacune des variables sur $\mathbb{R}$. La donnée de cet ensemble de lois appelées lois marginales d'ordre 1 du vecteur ne donne aucune information sur les relations de dépendance ou même sur la structure de covariance entre les variables coordonnées.

2.5 Loi gaussienne $\mathcal{N}(\mu, \sigma^2)$

Nous rappelons la définition de la loi de probabilité gaussienne ou normale que nous utiliserons le plus par la suite. Cette loi est continue et sa densité est :

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

où $\mu \in \mathbb{R}$ et $\sigma > 0$. Tous les moments de la loi sont finis

$$\int_{-\infty}^{+\infty} x^p f(x)dx < +\infty,$$

pour tout $p \in \mathbb{N}$. En particulier, l'espérance vaut μ et la variance σ^2.

Propriété 5. *Stabilité de la loi gaussienne.*
On considère une suite infinie de variables indépendantes $(X_i)_{i\in\mathbb{N}}$ de lois gaussiennes.

- *La loi de $Y = a_1X_1 + a_2X_2$ est gaussienne.*
- *La série $Y = \sum_{i\in\mathbb{N}} a_iX_i$ est convergente dans $\mathbb{L}^2$ dès que $\sum_{i\in\mathbb{N}} |a_i| < \infty$. Y est une variable aléatoire gaussienne.*

La stabilité de la loi normale est à l'origine de la propriété suivante :

Théorème 1. *Théorème de la limite centrale :*
Soit une collection de n variables $(X_1, ...X_n)$ indépendantes de même loi, d'espérance nulle et de variance finie σ^2. Soit S_n la somme de ces variables. Alors la loi de $S_n/\sqrt{n}$ tend vers la loi $\mathcal{N}(0, \sigma^2)$.

Ce résultat s'interprète de la façon suivante : si un phénomène est le résultat de l'addition d'une multitude de petites variations aléatoires indépendantes de même variabilité, ce phénomène suit une loi normale. Il peut s'agir de la taille d'un individu dans une population, d'un cours boursier résultat d'un grand nombre d'ordre d'achat et de ventes...

2.6 Vecteur aléatoire gaussien

Soit p un entier et $(X_1, \ldots, X_p)$, p variables aléatoires.

Définition 4. *$(X_1, \ldots, X_p)$ forment un vecteur aléatoire gaussien si et seulement si pour tout vecteur réel $(a_1, \ldots, a_p)$, la variable $a_1X_1 + \cdots + a_pX_p$ suit une loi gaussienne.*

La notion de vecteur gaussien définit non seulement la loi des variables coordonnées mais également la relation de dépendance entre ces variables :

- Les variables coordonnées ont toutes une loi gaussienne.
- Un vecteur formé de variables coordonnées gaussiennes n'est pas nécessairement gaussien.
- Un vecteur formé de variables coordonnées gaussiennes orthogonales n'est pas nécessairement gaussien.
- Un vecteur formé de variables coordonnées gaussiennes indépendantes est gaussien.
- Si dans un vecteur gaussien, les coordonnées sont orthogonales, alors elles sont indépendantes.
- Le vecteur image d'un vecteur gaussien par une transformation linéaire est gaussien.
- Tout vecteur gaussien X peut être écrit comme la transformation linéaire d'un vecteur de variables gaussiennes indépendantes.
- Tout vecteur gaussien X peut être écrit comme la transformation linéaire d'un vecteur de variables gaussiennes indépendantes de variance 1.

Les deux derniers résultats sont une conséquence de la diagonalisation des matrices symétriques définies positives. Supposons pour simplifier que les espérances des variables coordonnées de X sont nulles ; la matrice Σ est réelle et symétrique, donc on peut trouver une matrice P orthonormale ($^tP = P^{-1}$) et une matrice diagonale D telles que $\Sigma = {}^tPDP$. Posons $Y = PX$. Le vecteur Y est un vecteur gaussien et sa matrice de covariance est D qui est diagonale. Les coordonnées de Y sont orthogonales et donc indépendantes. Définissons $D^{-1/2} = diag(d_i^{-1/2})$ où les d_i sont les coefficients diagonaux de D, c'est à dire les valeurs propres (toutes strictement positives) de Σ. Alors $Z = D^{1/2}Y$ est bien un vecteur gaussien de matrice de covariance la matrice identité et $X = P^{-1}(D^{-1/2})^{-1}Z$.
La loi du vecteur gaussien a une densité généralisant la loi gaussienne. Soit $x = {}^t(x_1, \ldots x_d)$, Σ une matrice symétrique définie positive et M un vecteur réel de taille d :

$$f(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{{}^t(x-M)\Sigma^{-1}(x-M)}{2}\right)$$

où $\det(\Sigma)$ désigne le déterminant (nécessairement positif) de Σ.

2.7 Processus aléatoires: définition et propriétés

Après élimination de la partie déterministe deux cas sont possibles : ou bien la série résultante est formée de variables indépendantes, ou bien les variables semblent corrélées. Il peut être intéressant de prendre en compte dans le modèle cette dépendance. Il faut savoir modéliser une série de variables aléatoires dépendantes en nombre croissant. Pour cela, on fait appel à la notion de processus stochastique qui généralise les notions de variables et vecteurs aléatoires.

Définition 5. *Soit T un ensemble d'indices infini ; $T = \mathbb{N}$, $\mathbb{Z}$ ou $\mathbb{R}$. Soit $(\Omega, \mathcal{A})$ un espace probabilisable. On appelle processus stochastique sur T à valeur dans $\mathbb{R}$ une collection $(X(t))_{t\in T}$ de variables aléatoires indicées par T.*

Proposition 3. *Un processus stochastique sur T définit une loi de probabilité sur $(\mathbb{R}^T, \mathcal{C}_T)$, c'est-à-dire sur l'ensemble des fonctions $f : T \to \mathbb{R}$. $\mathcal{C}_T$ est la tribu engendrée par les ensembles de fonctions réelles $\{f \in \mathbb{R}^T \ ; \ f(t_0) \in B_0\}$ indicés par t_0 de T et B_0 ensemble borélien de $\mathbb{R}$.*

La tribu $\mathcal{C}_T$ appelée tribu cylindrique est la plus petite tribu qui rend toutes les applications coordonnées π_t mesurables. L'application coordonnée π_t est l'application qui à toute fonction f de T sur $\mathbb{R}$ fait correspondre le réel $f(t)$.
Remarque : On ne peut pas définir une loi de probabilité sur $\mathbb{R}^T$ comme on fait pour définir une probabilité discrète (le nombre de fonctions est infini indénombrable) ni par une densité par rapport à une mesure de référence comme la mesure de Lebesgue. Une telle mesure sur l'ensemble $\mathbb{R}^T$ n'existe pas comme dans le cas de $\mathbb{R}^d$. La méthode pour définir une probabilité sur $\mathbb{R}^T$ consiste à définir des probabilités sur les ensembles générant la tribu cylindrique. Il suffit pour cela de définir les lois des vecteurs $(f(t_1), \dots f(t_p))$ de toute taille p et pour tout $(t_1, ...t_p)$ éléments de T (appellées lois marginales de dimension finie du processus) en s'assurant que ces lois sont cohérentes entre elles (c'est-à-dire par exemple que la loi de la variable aléatoire $f(t_1)$ se déduit de la loi du couple $(f(t_1), f(t_2))$). Toute méthode qui permet de définir une telle famille de probabilités cohérentes définit bien un processus stochastique (théorème de Kolmogorov).

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

Définition 6. *Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité et $X = (X_t,\, t \in T)$ un processus stochastique sur T à valeurs dans $\mathbb{R}$*

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

Pour modéliser une série chronologique, on utilisera un processus aléatoire à valeur réellle indicé par $\mathbb{Z}$ ou $\mathbb{N}$, appelé également processus à temps discret; quand ce n'est pas le cas, notamment lorsque $T = \mathbb{R}$, on parle de processus à temps continu.

Définition 7. *On appelle $X = (X_t,\, t \in T)$ processus aléatoire gaussien , un processus aléatoire tel que $\forall k \in \mathbb{N}^*$, $\forall (t_1, \cdots, t_n) \in T^n$, $(X_{t_1}, \cdots, X_{t_n})$ est un vecteur gaussien.*

Un processus gaussien est donc entièrement défini par ses fonctions espérance $m(t) = \mathbb{E}X(t)$ et covariance $\gamma(s,t)$.

Définition 8. *On dit qu'un processus aléatoire $X = (X_t,\, t \in T)$ est strictement stationnaire lorsque X est invariant en distribution par toute translation du temps, c'est-à-dire que $\forall n \in \mathbb{N}^*$, $\forall (t_1, \cdots, t_n) \in T^n$, $\forall c \in T$, $(X_{t_1}, \cdots, X_{t_n})$ à la même distribution que $(X_{t_1+c}, \cdots, X_{t_n+c})$.*

Remarque.

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

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

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

Définition 9. *On dit qu'un processus aléatoire $X = (X_t,\, t \in T)$ est stationnaire à l'ordre 2 lorsque:*

- *son espérance $m(t)$ est constante*
- *sa covariance $cov(X_s, X_t)$ est une fonction de $|t - s|$.*

Remarque.

On pourrait penser qu'il est plus difficile d'être stationnaire strict que stationnaire d'ordre 2. Cependant la stationnarité d'ordre 2 requiert d'avoir des moments d'ordre 2 ce qui n'est pas demandé par la stationnarité stricte. On peut ainsi montrer que pour un ARCH(p) (on verra plus loin la définition d'un tel processus), les conditions de stationnarité d'ordre 2 sont plus fortes que celles de stationnarité stricte.
Le cas des processus gaussiens est particulier car la loi d'un vecteur gaussien ne dépend que de son espérance et de sa fonction de covariance. On peut donc en déduire que

Propriété 6. *Si X est un processus gaussien, alors (Stationnarité stricte $\Longleftrightarrow$ Stationnarité d'ordre 2).*

Définition 10. *Bruits blancs*

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

2.7.1 Processus gaussiens généraux

Pour définir un processus gaussien sur T, il suffit de se donner une fonction espérance $m(t)$ et une fonction d'autocovariance $\gamma(t_1, t_2)$ définie sur T^2 qui soit de type positif.

Définition 6. *Une fonction $\gamma(t_1, t_2)$ est de type positif si elle est symétrique et si pour tout entier p et p-uplet $(t_1, \ldots, t_p)$ de T et $(u_1, \ldots, u_p)$ de $\mathbb{R}$:*

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\gamma(t_i, t_j)u_i u_j \geq 0.$$

Cette définition signifie que la matrice des $\gamma(t_i, t_j)_{i=1,\ldots,p, j=1,\ldots,p}$ est une matrice symétrique définie positive. On peut donc définir un vecteur gaussien $(X(t_1), ..., X(t_p))$, ayant pour espérance $(m(t_1), ..., m(t_p))$ et pour matrice de covariance cette matrice, ce qui suffit pour définir le processus gaussien.

2.7.2 Transformé instantané d'un processus

Si X est un processus sur T et si f est une fonction mesurable, alors Y definie par $Y_t = f(X_t)$ est un processus sur T. Il suffit de voir que les marginales finies sont bien définies par la relation qui lie X_t et Y_t et que les marginales de dimension finies restent cohérentes par la transformation par f ; il existe donc un processus Y qui correspond à ces marginales.

2.7.3 Processus linéaires

Soit $(\varepsilon(t))_{t\in\mathbb{Z}}$ un bruit blanc fort de variance finie. Soit $(a_t)_{t\in\mathbb{N}}$ une suite de coefficients réels tels que $\sum_{s\in\mathbb{N}} |a_s| < \infty$. On définit un processus X par

$$X_t = \sum_{s\in\mathbb{N}} a_s \varepsilon_{t-s}.$$

Si les coefficients a_t sont nuls à partir d'un certain rang, cette définition est une généralisation simple du cas de la transformation instantanée. Si les coefficients a_t ne sont pas nuls à partir d'un certain rang, il faut montrer que la série a un sens pour une forme de convergence de suite de variables aléatoires. Nous reviendrons sur ce problème dans la partie consacrée aux filtres linéaires.

2.7.4 Processus linéaires gaussiens

Nous allons définir une famille de processus sur $\mathbb{N}$ permettant de décrire le processus des résidus. Ce modèle est adapté au problème de la prévision à court terme. Pour alléger les notations nous allons noter X_i la variable aléatoire correspondant à l'instant i de $\mathbb{N}$. Soit $(\varepsilon_i)_{i\in\mathbb{N}}$ un bruit blanc gaussien centré. Soit X_0 une variable aléatoire gaussienne centrée. Soit ρ un réel. Pour $i > 0$, on définit récursivement un processus gaussien centré autorégressif d'ordre 1 (AR1) par

$$X_{i+1} = \rho X_i + \varepsilon_{i+1}. \tag{2}$$

Nous reviendrons dans le détail sur cet exemple où tous les calculs peuvent être menés explicitement.

2.7.5 Processus autorégressif d'ordre p

On peut généraliser le processus autorégressif d'ordre 1 à un processus d'ordre p ; on se donne ε le bruit blanc précédent, $(X_0, ..., X_{p-1})$ un vecteur gaussien, $(a_0, \ldots, a_p)$ une suite de coefficients réels et on définit le modèle AR(p) par

$$a_0 X_i + a_1 X_{i-1} + \cdots + a_p X_{i-p} = \varepsilon_i.$$

Le polynôme $P(z) = a_p z^p + a_{p-1} z^{p-1} + \cdots + a_0$ est le polynôme associé au processus précédent. L'existence d'un processus stationnaire, obtenu pour un bon choix du vecteur initial $(X_0, ..., X_{p-1})$ est assurée lorsque toutes les racines complexes ou réelles de ce polynôme sont de module différent de 1.

2.7.6 Processus de moyenne mobile d'ordre q

On peut définir le processus de moyenne mobile d'ordre q (MAq) à partir du bruit blanc ε et de $(b_0, \ldots, b_q)$ une suite de coefficients réels par

$$X_i = b_0 \varepsilon_i + b_1 \varepsilon_{i-1} + \cdots + b_q \varepsilon_{i-q}.$$

Le polynôme $Q(z) = b_0 z^q + b_1 z^{q-1} + \cdots + b_q$ est le polynôme associé au processus précédent. Le processus MAq est toujours stationnaire par définition.

2.7.7 Processus ARMA(p, q)

On généralise les processus suivants par le processus ARMA(p, q) défini à partir du bruit blanc ε, d'un vecteur gaussien $(X_0, ..., X_{p-1})$, de deux suites $(a_0, \ldots, a_p)$ et $(b_0, \ldots, b_q)$ de coefficients réels par

$$a_0 X_i + a_1 X_{i-1} + \cdots + a_p X_{i-p} = b_o \varepsilon_i + b_1 \varepsilon_{i-1} + \cdots + b_q \varepsilon_{i-q}.$$

Par convention on fixe $a_0 = b_0 = 1$. On considère les polynômes P et Q précédents et on les simplifie quand ils ont des racines communes. L'existence d'un processus stationnaire, obtenu pour un bon choix du vecteur initial $(X_0, ..., X_{p-1})$ est assurée lorsque toutes les racines de P sont de module strictement supérieur à 1 et celles de Q de module supérieur à 1.

2.7.8 Processus integré ARIMA

La série chronologique considérée peut, même après soustraction d'une tendance, apparaître comme variant au cours du temps de la façon suivante : l'espérance reste apparemment nulle, mais la variabilité s'accroît au cours du temps. La variance empirique des données calculée sur un intervalle augmente avec le temps. Une croissance de la variance peut être le signe que la série observée correspond à un cumul d'une série stationaire; l'accumulution des petites fluctuations de la série d'origine a une variabilité qui s'accroît au cours du temps. Le modèle le plus simple ayant cette propriété est la marche aléatoire. Partant d'un bruit blanc ε centré de variance 1, on définit le processus de marche aléatoire

$$X_t = \sum_{i=1}^{t} \varepsilon_i.$$

Les variables X_t ne sont pas indépendantes et leur variance vaut t. Par contre, par définition, le processus des différences $X_t - X_{t-1}$ est un processus indépendant. Selon cet exemple, lorsque l'on observe une série dont la variance semble s'accroître, on peut calculer les différences discrètes de la série pour rechercher une série stationnire qu'on pourra à son tour modéliser comme un ARMA. Cette opération de différentiation peut être opérée plusieurs fois de suite. Les processus qui donnent des processus ARMA après différentiations forment la classe des processus ARIMA(p, d, q), où d est le nombre de fois qu'il faut différencier le processus pour obtenir un processus ARMA (p, q).

3 Prévision linéaire d'un processus stationnaire à l'ordre 2

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

Nous nous intéressons à la prévision de la valeur X_t du processus à l'instant par une fonction des valeurs du processus aux instants précédents $(X_i)_{i<t}$. Cela consiste à projeter la variable X_t sur l'espace vectoriel engendré par toutes les fonctions mesurables des variables précédentes. Ce calcul est en général impraticable en raison du grand nombre de fonctions à considérer. C'est pourquoi nous considèrerons uniquement les fonctions linéaires des variables précédentes et non toutes les fonctions. Cette idée correspond à un modèle de régression linéaire X_t sur les $(X_i)_{i<t}$. Il faut cependant remarquer que les variables régresseurs sont en nombre infini et que la notion de projection classique doit être étendue à un espace vectoriel infini.

3.1 Espace de Hilbert

On se place dans le contexte suivant :

Définition 11. *Soit H un espace vectoriel de dimension éventuellement infinie muni d'un produit scalaire ; si l'espace vectoriel est complet par rapport à la norme $\|\cdot\|_2$ induite par le produit scalaire, on dit que H est un espace de Hilbert.*

Le fait que l'espace soit complet permet d'affirmer que les séries $\sum_{i=0}^{\infty} e_i$ sont convergentes dès que $\sum_{i=0}^{\infty} \|e_i\|_2$ est convergente. On montre grâce à cette propriété que la projection orthogonale sur un sous-espace vectoriel de dimension finie se généralise à la projection sur les sous-espaces vectoriels fermés de dimension infinie. Par définition, un sous-espace vectoriel fini contient toute combinaison linéaire de ses éléments; un sous-espace vectoriel fermé de dimension infinie doit en plus contenir toutes les limites des suites constituées de ces combinaisons linéaires.

Propriété 7. *Soit V un sous-espace vectoriel fermé de H et $x \in H$.*

1. *Il existe un unique v dans V tel que $\|x - v\| = \inf_{y \in V} \|x - v\|$.*
2. *Ce vecteur v est également caractérisé de manière unique par $v \in V$ et $x - v \in V^{\perp}$.*

Le vecteur v est appelé projeté de x sur V et noté $P_V(x)$.
On a immédiatement la propriété $\|x\|^2 = \|P_V(x)\|_2^2 + \|x - P_V(x)\|^2$

Preuve Elle est fondée sur l'égalité du parallèlogramme caractéristique des normes hilbertiennes:

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

1. Prouvons l'existence. Notons $d = \inf_{y \in V} \|x - v\|^2$. Soit y_n une suite dans V approchant cet infimum. Pour tout indice m et n de cette suite $(y_m + y_n)/2 \in V$. Par l'égalité du parallèlogramme

$$\begin{aligned} 0 \leq \|y_m - y_n\|^2 &= -\|y_m + y_n) - 2x\|^2 + 2\|y_m - x\|^2 + 2\|y_n - x\|^2 \\ &\leq -4\|y_m + y_n)/2 - x\|^2 + 2\|y_m - x\|^2 + 2\|y_n - x\|^2 \\ &\leq -4d + 2\|y_m - x\|^2 + 2\|y_n - x\|^2 \end{aligned}$$

   Les deux derniers termes tendent vers $4d$ quand m et n tendent vers l'infini donc y_n est une suite de Cauchy qui converge vers une limite v dans V car V est fermé.
   Prouvons l'unicité. Soit v_1 et v_2 deux solutions. Par la même relation on a

$$\begin{aligned} 0 \leq \|v_1 - v_2\|^2 &\leq -4\|v_1 + v_2)/2 - x\|^2 + 2\|v_1 - x\|^2 + 2\|v_2 - x\|^2 \\ &\leq -4d + 4d \end{aligned}$$

   ce qui prouve que $v_1 = v_2$.

2. Si $v \in V$ est tel que $x - v \in V^{\perp}$, alors pour tout $y \in V$, $\|x - y\|^2 = \|x - v\|^2 + \|v - y\|^2 \geq \|x - v\|^2$. Donc v est bien l'unique vecteur réalisant l'infimum.
   Si $v \in V$ et $x - v$ n'appartient pas à $V^{\perp}$, on choisit $y \in V$ tel que $\|y\| = 1$ et $\langle x - v, y\rangle = a \neq 0$ et on définit $w = v + ay$. On vérifie que

$$\|x - w\|^2 = \|x - v - ay\|^2 = \|x - v\|^2 - 2a\langle x - v, y\rangle + a^2 = \|x - v\|^2 - a^2 \leq \|x - v\|^2.$$

   Donc v ne réalise pas l'infimum de distance.

3.2 Espérance conditionnelle et prédicteur linéaire

Nous allons utiliser le fait que l'ensemble des variables aléatoires de carré intégrable $\mathbb{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, noté par la suite simplement $\mathbb{L}^2$ est un espace de Hilbert pour le produit scalaire $\langle X, Y \rangle = \mathbb{E}(XY)$.

Pour un processus stationnaire à l'ordre 2, on définit deux espaces vectoriels fermés de variables engendrés par les variables passées à l'instant t :

- le sous espace vectoriel fermé V_t engendré par les fonctions mesurables d'un nombre fini de variables prises parmi les variables $(X_{t-i})_{i=1\ldots,\infty}$,
- le sous espace vectoriel fermé L_t engendré par les combinaisons linéaires d'un nombre fini de variables prises parmi les variables $(X_{t-i})_{i=1\ldots,\infty}$.

Définition 7. *Espérance conditionnelle et prédiction linéaire*

- *La projection de X_t sur le sous espace vectoriel V_t s'appelle l'espérance conditionnelle de X_t connaissant les $(X_{t-i})_{i=1\ldots,\infty}$. Nous la noterons $\mathbb{E}(X_t|(X_{t-i})_{i=1\ldots,\infty})$,*
- *La projection de X_t sur le sous espace vectoriel L_t s'appelle le prédicteur linéaire de X_t par les $(X_{t-i})_{i=1\ldots,\infty}$. Nous la noterons $\hat{X}_t$.*

L'espérance conditionnelle représente la meilleure prédiction dans la mesure où elle tire profit de façon parfaite de l'information contenue dans les observations passées. Le prédicteur linéaire ne représente qu'une approximation de cette meilleure prédiction.

Proposition 4. *Les projections précédentes sont caractérisées par*

- *Pour tout entier d et toute variable $W = f(X_{i_1}, \ldots, X_{i_d})$,*

$$\mathbb{E}\left(W\mathbb{E}(X_t|(X_{t-i})_{i=1\ldots,\infty})\right) = \mathbb{E}\left(WX_t\right).$$

  Cela signifie que $X_t - \mathbb{E}\left(X_t|(X_{t-i})_{i=1\ldots,\infty}\right)$ est indépendante des variables $(X_{t-i})_{i=1\ldots,\infty}$.

- *Pour tout indice $i > 0$*

$$\mathbb{E}\left(X_{t-i}\hat{X}_t\right) = \mathbb{E}\left(X_{t-i}X_t\right).$$

  Cela signifie que $X_t - \hat{X}_t$ est orthogonale aux variables $(X_{t-i})_{i=1\ldots,\infty}$.

Cette caractérisation montre la difficulté de calculer explicitement l'espérance conditionnelle qui correspond à une infinité non dénombrable de conditions (même quand le nombre d est fixé) et la relative simplicité du calcul du prédicteur linéaire qui correspond au calcul de la solution d'un système linéaire (mais de taille infinie dénombrable quand même). Remarquons toutefois que le prédicteur $\hat{X}_t$ est également la limite dans H des projetés de X_t sur les sous-espaces de dimension finie croissante engendrés par les variables $(X_{t-i})_{i=1\ldots,n}$ quand n tend vers l'infini. Cette remarque ainsi qu'un algorithme récursif de calcul des projections sur ces espaces permet de calculer le prédicteur de façon performante avec une précision contrôlée. Nous chercherons à nous placer dans des situations où le calcul de l'espérance conditionnelle n'est pas nécessaire, car celle-ci est strictement égale au prédicteur linéaire :

- si la meilleure fonction est effectivement une combinaison linéaire des observations passées, les deux projections sont égales. Ce sera le cas pour une série générée par un modèle AR(p), même non gaussien.
- si le processus est gaussien, les deux projections sont toujours égales. En effet, au sein d'un processus gaussien la condition d'indépendance et la condition d'orthogonalité sont équivalentes.

3.3 Processus singulier et régulier

Il existe deux cas particuliers où le calcul du projeté est immédiat :

- Si L_t est un espace vectoriel constant qui ne dépend pas de t, alors toute variable X_t appartient à L_t. Le prédicteur $\hat{X}_t$ est alors égal à X_t. La prévision est donc exacte à chaque étape.
- A l'opposé, si le processus est un bruit blanc, X_t est par définition orthogonal à L_t et le prédicteur est égal à 0.

Définition 12. *Si l'espace L_t est un espace vectoriel indépendant de t, on dit que le processus est singulier.*

Un exemple de processus singulier est un processus aléatoire constant $X_t = Z$ ou Z est une variable aléatoire. Un autre exemple est $X_t = cos(t)Z$. L'aléa de ces processus singuliers est tiré une fois pour toute avant que leur évolution commence. En revanche, on appellera processus régulier un processus où aucun tirage aléatoire n'a lieu avant que l'évolution soit lancée :

Définition 13. *Si l'intersection $\cap_{t\in\mathbb{Z}} L_t$ est vide, on dit que le processus est régulier. Dans ce cas, le processus peut être représenté comme une moyenne mobile infinie. Il existe un unique bruit blanc faible $(\varepsilon_t)_{t\in\mathbb{Z}}$ et une famille de réels $(a_i)_{i\in\mathbb{N}}$ avec $a_0 = 1$ et $\sum_{k\in\mathbb{N}} |a_k|^2 < \infty$ tels que*

$$X_t = \sum_{i=0}^{\infty} a_i \, \varepsilon_{t-i} \quad \textit{pour tout } t \in \mathbb{Z}$$

Le processus ε peut être défini par $\varepsilon_t = X_t - \hat{X}_t$. C'est l'erreur de la prédiction linéaire. On appelle le processus ε le processus d'innovation du processus X. La série est appelée représentation causale du processus.

Nous pouvons de plus énoncer le résultat général suivant :

Théorème 2 (Décomposition de Wold)**.** *Soit $X = (X_t)_{t\in\mathbb{Z}}$ un processus à temps discret stationnaire d'ordre 2. Alors il existe une décomposition*

$$X_t = X_t^r + X_t^s. \tag{3}$$

où X^r est un processus régulier et X^s est un processus singulier. Ces deux processus sont orthogonaux : pour tout t et t', $X_t^r \perp X_{t'}^s$.

Ceci permet d'avoir une représentation générale d'un processus stationnaire du second ordre en utilisant la représentation du processus régulier. Cependant le bruit blanc est faible et la représentation n'est fidèle que pour la structure de covariance. Les lois jointes du processus ne sont pas représentées. Cette faiblesse interdit la plupart des calculs nécessaires à l'établissement de résultats statistiques.

3.4 Fonction d'autocorrélation et densité spectrale d'un processus stationnaire du second ordre

Les fonctions suivantes caractérisent les projections orthogonales entre variables coordonnées du processus. Elles ne sont pas suffisantes pour décrire la loi du processus sauf quand celui-ci est gaussien. Soit $X = (X_k)_{k\in\mathbb{Z}}$ un processus stationnaire au second ordre.

Définition 14. *Autocovariance et autocorrélation.*

- *On appelle fonction d'autocovariance de X la fonction $\gamma(k) = \mathbb{E}(X_0 X_k)$ pour $k \in \mathbb{Z}$. Ainsi, $r(0)$ est la variance de la série.*
- *On appelle fonction d'autocorrélation $\rho(k) = \gamma(k)/\gamma(0)$. On a $-1 \leq \rho(k) \leq 1$ pour $k \in \mathbb{Z}$.*

Ces fonctions correspondent respectivement au produit scalaire et au cosinus de l'angle entre deux variables. La fonction suivante correspond aux calculs successifs de projection d'une variable sur les k variables passées les plus proches:

Définition 8. *Autocorrélation partielle.*
La fonction d'autocorrélation partielle d'un processus stationnaire X est la fonction $\alpha(k)$ telle que

- $\alpha(0) = 1$,
- $\alpha(1)$ *est le coefficient de correlation de X_i avec X_{i-1},*
- *pour calculer le coefficient d'autocorrelation partielle d'ordre k, noté $\alpha(k)$, on projette X_i sur l'espace vectoriel engendré par $X_{i-1}, \ldots X_{i-k+1}$ et on considère la variable résiduelle de cette projection e_0; on projette X_{i-k} sur le même espace et on considère la variable résiduelle de cette projection e_k. Le coefficient d'autocorrélation d'ordre k est le coefficient de corrélation de ces deux variables résiduelles.*

Il existe une définition plus simple mais qui n'est pas symétrique des deux variables et n'explique pas le nom de corrélation partielle. Le coefficient $\alpha(k)$ est égal à la coordonnée sur X_{i-k} du projeté de X_i sur l'espace vectoriel engendré par $X_{i-1}, \dots X_{i-k}$.
Pour finir nous introduisons la transformée de Fourrier de la fonction d'autocovariance. Cette fonction à une expression directe en fonction des paramètres dans les modèles ARMA(p, q).

Définition 15. *Densité spectrale.*
S'il existe une fonction $f : [-\pi, \pi[\to \mathbb{R}^+$ *telle que* $\forall k \in \mathbb{Z}$, $\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda$, *alors on dit que* X *admet une densité spectrale* f.

Exemple.

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

Propriété 8. *Existence et propriété de la densité spectrale.*

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

Propriété 9. *Soit* $X = (X_k)_{k \in \mathbb{Z}}$ *un processus du second ordre à temps discret centré stationnaire. Si la densité spectrale* f *existe sur* $[-\pi, \pi[$ *alors* f *est paire et* $f(\lambda) = \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} \gamma(k) e^{-ik\lambda}$ *pour tout* $\lambda \in [-\pi, \pi[$.

Grâce à la densité spectrale, il est possible de donner une condition de non-singularité du processus :

Théorème 3 (Formule de Kolmogorov)**.** *Soit* $X = (X_t)_{t \in \mathbb{Z}}$ *un processus à temps discret stationnaire d'ordre 2 tel que* $F = \int_{-\pi}^{\pi} \log(f(\lambda)) d\lambda > -\infty$. *Alors la partie régulière de* X *est non nulle.*

On peut montrer la relation $\|X_t - \hat{X}_t\|_2^2 = 2\pi \exp(F/2\pi)$. La relation $F > -\infty$ est donc équivalente à dire que la prévision n'est pas exacte, donc que le processus n'est pas singulier.

3.5 Algorithmes de calcul du prédicteur

On considère un processus stationnaire du second ordre régulier X de fonction d'autocovariance γ connue. Le calcul du prédicteur linéaire peut être approché par la projection sur des espaces de plus en plus grand engendrés par les n plus proches observations dans le passé. Il correspond à la résolution d'un système linéaire. Nous présentons une résolution itérative de ce calcul appelé algorithme de Durbin-Levinson.

3.5.1 Algorithme de Durbin-Levinson

On note $\hat{X}_t^n$ la projection de X_t sur l'espace L_t^n engendré par les variables $(X_{t-1}, \dots, X_{t-n})$. les coefficients $\phi_{n,i}$ sont les coordonnées de cette projection :

$$\hat{X}_t^n = \phi_{n,1} X_{t-1} + \dots + \phi_{n,n} X_{t-n}$$

De plus, on note $v_n = \|X_t^n - X_t\|^2$, la variance de l'erreur de cette projection. Par définition, $v_0 = \gamma(0)$.

Proposition 5. *Supposons qu'on ait calculé les coefficients* $\phi_{n-1,i}$ *et* v_{n-1}. *Les coefficients* $\phi_{n,i}$ *et* v_n *se calculent explicitement par les formules suivantes :*

$$\begin{aligned}
\phi_{n,n} &= \frac{1}{v_{n-1}} \left(\gamma(n) - \sum_{i=1}^{n-1} \phi_{n-1,i} \gamma(i) \right) \\
\phi_{n,i} &= \phi_{n-1,i} - \phi_{n,n} \phi_{n-1,n-i} \text{ pour } i = 1, \dots, n-1 \\
v_n &= (1 - \phi_{n,n}^2) v_{n-1}
\end{aligned}$$

Preuve : on décompose l'espace L_t^n sous forme de la somme de $K_1 = L_t^{n-1}$ espace engendré par les variables $(X_{t-1}, \dots, X_{t-n+1})$ et de l'espace K_2, orthogonal à K_1 engendré par $X_{t-n} - P_{K_1}(X_{t-n})$. On écrit alors :

$$X_t^n = P_{K_1}(X_t) + P_{K_2}(X_t) = X_t^{n-1} + \lambda (X_{t-n} - P_{K_1}(X_{t-n})).$$

Calcul de la projection sur K_2 : On rappelle que si D est une droite vectorielle et y un vecteur non nul de D, la projection sur D d'un vecteur x quelconque de l'espace est égale à λy avec $\lambda = \langle x, y\rangle / \|y\|^2$. K_2 étant une droite vectorielle engendrée par $X_{t-n} - P_{K_1}(X_{t-n})$, le coefficient λ vaut donc

$$\lambda = \frac{< X_{t-n} - P_{K_1}(X_{t-n}), X_t >}{\|X_{t-n} - P_{K_1}(X_{t-n})\|^2}.$$

Pour calculer $P_{K_1}(X_{t-n})$, on remarque que la matrice de covariance de $(X_t, X_{t-1}, \ldots, X_{t-n+1})$ est identique à celle de $(X_{t-n}, X_{t-n+1}, \ldots, X_{t-1})$. Les coefficients de projection ne dépendent que de la structure de covariance donc

$$P_{K_1}(X_{t-n}) = \phi_{n-1,1}X_{t-n+1} + \cdots + \phi_{n-1,n-1}X_{t-1} = \sum_{i=1}^{n-1} \phi_{n-1,n-i}X_{t-i},$$

et l'erreur de cette projection $\|X_{t-n} - P_{K_1}(X_{t-n})\|^2$ est égale à v_{n-1}. On en déduit que

$$\lambda = \frac{1}{v_{n-1}} \left(\gamma(n) - \sum_{i=1}^{n-1} \phi_{n-1,i}\gamma(i) \right).$$

Identification des coefficients $\phi_{n,i}$:

$$X_t^n = \sum_{i=1}^{n-1} \phi_{n-1,i}X_{t-i} + \lambda(X_{t-n} - \sum_{i=1}^{n-1} \phi_{n-1,n-i}X_{t-i}) = \sum_{i=1}^{n-1} (\phi_{n-1,i} - \lambda\phi_{n-1,n-i})X_{t-i} + \lambda(X_{t-n}).$$

Cette formule permet d'identifier $\phi_{n,n}$ à λ et donne l'expression des autres $\phi_{n,i}$. Il reste à calculer v_n.
Calcul de v_n :

$$\begin{aligned} v_n &= \|X_t - P_{K_1}(X_t) - P_{K_2}(X_t)\|^2 \\ &= \|X_t - P_{K_1}(X_t)\|^2 + \|P_{K_2}(X_t)\|^2 - 2 < X_t - P_{K_1}(X_t), P_{K_2}(X_t) > \\ &= v_{n-1} + \lambda^2 v_{n-1} - 2\lambda^2 v_{n-1} \\ &= v_{n-1}(1 - \lambda^2). \end{aligned}$$

Remarques:

- Cet algorithme calcule les coefficients d'autorégression de la série sur elle-même, et propose une représentation de la série par un modèle AR(n) avec une valeur n aussi grande que l'on veut et limité en pratique par la taille de l'échantillon disponible. Tout processus stationnaire du second ordre causal régulier peut être représenté par une moyenne mobile infinie; le résultat précédent implique qu'il peut également être représenté par un modèle autorégressif de taille infinie. Les deux représentations ne sont pas égales, mais possédent seulement la même structure de covariance.
- L'algorithme permet de calculer le prédicteur pour tout modèle stationnaire du second ordre régulier lorsque la fonction de covaraince du modèle est connue. On peut par ailleurs estimer cette covariance à partir des données par l'estimateur empirique de la covariance. Nous disposons donc déjà d'une méthode de prévision adaptable à toutes les séries d'observations stationnaires et facile à programmer en pratique. Mais cette méthode prend bien en compte tous les modèles possibles (méthode non paramétrique) au prix d'une moindre efficacité statistique par rapport à des méthodes ou un choix de modèle paramétrique a été effectué. C'est pour cette raison que nous allons utiliser la classe des modèles ARMA(p, q) et développer des méthodes spécifiques de maximum de vraisemblance adaptées à ces méthodes.

Nous allons maintenant donner un autre algorithme de calcul du prédicteur fondé sur la représentation en moyenne mobile infinie.

3.5.2 Algorithme des innovations

Dans l'algorithme précédent, nous avions découpé l'espace L_t^n des n variables passées en deux espaces vectoriels orthogonaux. Dans cette méthode, nous le découpons en n espaces orthogonaux (orthogonalisation

de Gram-Schmidt de la base formée par les X_i) : on pose $e_{t-n} = X_{t-n}$, $e_{t-n+1} = X_{t-n+1} - P_n(X_{t-n+1})$, $e_{t-i} = X_{t-i} - P_{i+1}(X_{t-i})$ où P_i est la projection sur l'espace engendré par les variables X_{t-i} à X_{t-n}. Les $(e_{t-i})_{i=1,\dots,n}$ forment une base orthogonale de L_t^n. On note $\theta_{n,i}$ la i-ième coordonnée de $\hat{X}_t^n$ sur cette base :

$$\begin{aligned}\hat{X}_t^n =& \theta_{n,1}e_{t-1} + \cdots + \theta_{n,n}e_{t-n}\\ v_n =& \|\hat{X}_t^n - X_t\|^2.\end{aligned}$$

Proposition 6. *Les projections intermédiaires s'expriment par rapport à la famille de coefficients $\theta_{i,j}$ suivant la relation :*

$$P_{n-i+1}(X_{t-n+i}) = \hat{X}_{t-n+i}^i = \theta_{i,1}e_{t-n+i-1} + \cdots + \theta_{i,i}e_{t-n} = \sum_{j=0}^{i-1}\theta_{i,i-j}e_{t-n+j} \tag{4}$$

de plus $\|e_{t-n+i}\|^2 = v_i$.

Preuve : $P_{n-i+1}(X_{t-n+i})$ est le projeté de X_{t-n+i} sur les i variables les plus proches de son passé. Le processus X étant stationnaire au second ordre, les covariances sont les mêmes qu'entre X_t et les variables X_{t-1} à X_{t-i} et la base orthonormée est construite de la même façon, donc les coefficients de projection sont égaux. La stationarité implique de même que $\|e_{t-n+i}\|^2 = \|X_{t-n+i} - \hat{X}_{t-n+i}^i\| = \|X_t - \hat{X}_t^i\| = v_i$.

Proposition 7. *Supposons que pour $j < n$, on ait calculé v_j et les coefficients $\theta_{j,i}$ pour $i \leq j$. Les coefficients $\theta_{n,i}$ et v_n se calculent explicitement par les formules suivantes :*

$$\begin{aligned}v_0 &= \gamma(0)\\ \theta_{n,n} &= \frac{\gamma(n)}{v_0}\\ \theta_{n,n-i} &= \frac{1}{v_i}\left(\gamma(n-i) - \sum_{j=0}^{i-1}\theta_{i,i-j}\theta_{n,n-j}v_j\right)\\ v_n &= \gamma(0) - \sum_{j=0}^{n-1}\theta_{n,n-j}^2\ v_j\end{aligned}$$

Preuve : nous calculons d'abord $\theta_{n,n}$:

$$\theta_{n,n} = \frac{\langle \hat{X}_t | X_{t-n}\rangle}{\|X_{t-n}\|^2} = \frac{\gamma(n)}{v_0}.$$

Puis nous continuons dans l'ordre à calculer $\theta_{n,n-1}$, $\theta_{n,n-2}$... On identifie donc $\theta_{n,n-i}$ par:

$$\theta_{n,n-i} = \frac{\langle X_t | X_{t-n+i} - P_{n-i+1}(X_{t-n+i})\rangle}{\|X_{t-n+i} - P_{n-i+1}(X_{t-n+i}\|^2} = \frac{\langle X_t | X_{t-n+i} - P_{n-i+1}(X_{t-n+i})\rangle}{v_i}.$$

donc

$$\theta_{n,n-i}v_i = \langle X_t | X_{t-n+i}\rangle - \langle X_t | \sum_{j=1}^{i}\theta_{i,i-j}e_{t-n+j}\rangle = \gamma(n-i) - \sum_{j=1}^{i}\theta_{i,i-j}\theta_{n,n-j}v_j.$$

Pour calculer v_n, on utilise le théorème de Pythagore :

$$v_n = \|X_t\|^2 - \|X_t^n\|^2 = \gamma(0) - \sum_{j=1}^{n}\theta_{n,n-j}^2 v_j.$$

3.5.3 Prévision récursive pour une série chronologique

Soit une série chronologique d'observations $X_1, \dots, X_n$. On cherche à prévoir la valeur prochaine de la série. On suppose que la série est générée par un modèle stationnaire de fonction d'autocovariance connue. Nous allons modifier l'algorithme des innovations pour le prédicteur de X_n en fonction des valeurs précédentes (ici la date de la valeur à prévoir augmente comme la longueur des données prises en compte n) et pour un processus non nécessairement stationnaire de fonction d'autocovariance $\gamma(i,j)$. On rappelle que $v_j = \|X_j - \hat{X}_j\|^2$:

Proposition 8. *Le prédicteur linéaire de X_n est défini récursivement par*

$$\hat{X}_1 = 0$$
$$\hat{X}_n = \sum_{j=1}^{n-1} \theta_{n,j}(X_{n-j} - \hat{X}_{n-j}).$$

où les coefficients sont calculés par

$$v_1 = \gamma(1,1)$$
$$\theta_{n,n-k} = \frac{1}{v_k}\left(\gamma(n,k) - \sum_{j=1}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right)$$
$$v_n = \gamma(n,n) - \sum_{j=1}^{n-1} \theta_{n,n-j}^2 \; v_j$$

Preuve : La preuve utilise la même méthode d'orthogonalisation que l'algorithme des innovations prouvé précédemment. Mais le processus n'étant pas stationnaire, il est maintenant impossible de faire glisser les dates en conservant la même matrice de covariance et les même coefficients de projection. Le calcul n'est possible que pour une date de prévision qui avance avec n pour que les vecteurs de la base d'orthogonalisation restent fixes quand n varie. Par définition de la projection, $(X_1 - \hat{X}_1, \ldots, X_j - \hat{X}_j, \ldots, X_{n-1} - \hat{X}_{n-1})$ forment une base orthogonale non normée. Les calculs des coordonnées sur cette base sont obtenues par projection orthogonale sur chacun des axes de la base.

$$\theta_{n,n-k} = \frac{\left\langle X_n, X_k - \hat{X}_k \right\rangle}{\|X_k - \hat{X}_k\|^2} = \frac{1}{v_k}\left\langle X_n, X_k - \hat{X}_k \right\rangle = \frac{1}{v_k}\left(\gamma(n,k) - \left\langle X_n, \hat{X}_k \right\rangle\right). \tag{5}$$

Remplaçons $\hat{X}_k$ par son expression $\sum_{j=1}^{k-1} \theta_{k,j}(X_{k-j} - \hat{X}_{k-j}) = \sum_{j=1}^{k-1} \theta_{k,k-j}(X_j - \hat{X}_j)$.

$$\theta_{n,n-k} = \frac{1}{v_k}\left(\gamma(n,k) - \sum_{j=1}^{k-1} \theta_{k,k-j}\left\langle X_n, X_j - \hat{X}_j \right\rangle\right).$$

En remplaçant k par j dans (**??**), on obtient

$$\left\langle X_n, X_j - \hat{X}_j \right\rangle = \theta_{n,n-j}v_j$$

Soit

$$\theta_{n,n-k} = \frac{1}{v_k}\left(\gamma(n,k) - \sum_{j=1}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right).$$

Par le théorème de projection, on obtient

$$v_n = \|X_n\|^2 - \|\hat{X}_n\|^2 = \gamma(n,n) - \|\hat{X}_n\|^2.$$

On utilise l'orthogonalité de la base $(X_1 - \hat{X}_1, \ldots, X_j - \hat{X}_j, \ldots, X_{n-1} - \hat{X}_{n-1})$ pour calculer

$$\|\hat{X}_n\|^2 = \sum_{j=1}^{n-1} \theta_{n,n-j}^2 v_j \qquad \square$$

Nous appliquons cet algorithme à une renormalisation du processus X. Soit m le maximum de p et q. On définit Y par :

$$Y_t = X_t/\sigma \qquad t = 1, \ldots, m$$
$$Y_t = P(B)X_t/\sigma \qquad t > m$$

le processus Y n'est pas stationnaire mais sa fonction d'autocovariance se déduit de celle de X. Soit $i \leq j$,

$$\begin{aligned}
\gamma_Y(i,j) &= \sigma^{-2}\gamma_X(j-i) && \text{si } j \leq m \\
\gamma_Y(i,j) &= \sigma^{-2}(\gamma_X(j-i) - \sum_{k=1}^{p} a_k \gamma_X(k+i-j)) && \text{si } i \leq m \leq j \leq 2m \\
\gamma_Y(i,j) &= \sum_{k=1}^{q-j+i} b_k b_{k+j-i} && \text{si } i > m
\end{aligned}$$

On applique alors l'algorithme des innovations de la proposition **??** à Y ce qui nous donne les coefficients v_i et $\theta_{i,j}$ définissant le prédicteur de Y et l'erreur de prédiction :

$$\begin{aligned}
\hat{Y}_{n+1} &= \sum_{j=1}^{n} \theta_{n,j}(Y_{n+1-j} - \hat{Y}_{n+1-j}) && \text{si } n < m \\
\hat{Y}_{n+1} &= \sum_{j=1}^{q} \theta_{n,j}(Y_{n+1-j} - \hat{Y}_{n+1-j}) && \text{si } n \geq m \\
v_n &= \mathbb{E}(Y_{n+1} - \hat{Y}_{n+1})^2
\end{aligned}$$

Il ne reste plus pour conclure qu'à revenir au processus X de départ :

Proposition 9.

$$\begin{aligned}
\hat{X}_{n+1} &= \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) && \textit{si } n < m \\
\hat{X}_{n+1} &= a_1 X_n + \cdots + a_p X_{n+1-p} + \sum_{j=1}^{q} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) && \textit{si } n \geq m
\end{aligned}$$

et $\mathbb{E}(X_{n+1} - \hat{X}_{n+1})^2 = v_n \sigma^2$.

Remarque : les coefficients v_i et $\theta_{i,j}$ ne dépendent pas de σ^2. Cette propriété sera utile pour estimer les paramètres des polynômes indépendamment du paramètre de variance.

4 Un exemple : le processus AR(1) linéaire gaussien

Nous allons reprendre le plus simple des modèles afin d'illustrer toute la démarche de prévision. Soit $(\varepsilon_i)_{i\in\mathbb{N}}$ un bruit blanc gaussien centré et de variance σ^2. Soit X_0 une variable aléatoire gaussienne centrée de variance σ_0^2. Soit ρ un réel. Pour $i > 0$, on définit récursivement un processus gaussien centré (AR1) par

$$X_{i+1} = \rho X_i + \varepsilon_{i+1}. \tag{6}$$

4.1 Existence du processus

En substituant dans l'équation récursive (**??**), on montre que les composantes X_i du processus s'écrivent comme une combinaison linéaire des $(\varepsilon_j)_{j\leq i}$ et de X_0 :

$$X_i = \varepsilon_i + \rho\varepsilon_{i-1} + \rho^2\varepsilon_{i-2} + \ldots + \rho^{i-1}\varepsilon_1 + \rho^i X_0$$

Le vecteur $(X_0, \ldots, X_i)$ est donc un vecteur gaussien et les marginales finies sont cohérentes. Il suffit pour identifier la loi du processus de calculer la fonction d'autocovariance $\Gamma_{i,j} = \mathrm{cov}(X_i, X_j)$ pour $i \leq j$.

$$\begin{aligned}
\Gamma_{i,j} &= \rho^{j-i}\sigma^2 + \rho^{j-i+2}\sigma^2 + \rho^4\sigma^2 + \ldots + \rho^{i+j-2}\sigma^2 + \rho^{i+j}\sigma_0^2 \\
&= \rho^{j-i}[(1 + \rho^2 + \rho^4 + \ldots + \rho^{2i-2})\sigma^2 + \rho^{2i}\sigma_0^2].
\end{aligned}$$

4.2 Stationnarité

Si le processus précédent défini par (**??**) est stationnaire, alors nécessairement X_0 et X_1 ont même variance. Or

$$\mathrm{var}(X_1) = \Gamma_{1,1} = \sigma^2 + \rho^2\sigma_0^2.$$

Il faut donc que les paramètres définissant le modèle vérifient $(1-\rho^2)\sigma_0^2 = \sigma^2$ ce qui impose que $|\rho| < 1$. Inversement, si cette relation entre les paramètres est vérifiée,

$$\begin{aligned}\Gamma_{i,j} &= \rho^{j-i}[(1+\rho^2+\rho^4+\ldots+\rho^{2i-2})\sigma^2 + \rho^{2i}\sigma_0^2] \\ &= \rho^{j-i}[(1+\rho^2+\rho^4+\ldots+\rho^{2i-2})(1-\rho^2) + \rho^{2i}]\sigma_0^2 \\ &= \rho^{j-i}\sigma_0^2.\end{aligned}$$

Donc $\Gamma_{i,j}$ ne dépend que de l'écart $j-i$ ce qui assure la stationnarité stricte pour un processus gaussien. De plus la fonction d'autocorrélation est très simple : il s'agit d'une suite géométrique décroissante de facteur ρ.

4.3 Choix du modèle pour une série de données

On vérifie que le modèle est cohérent avec le comportement des données de la série; on trace l'autocorrélogramme empirique des données : si on observe une décroissance géométrique du coefficient de corrélation (éventuellemnt alternée) , on peut utiliser le modèle AR(1).

4.4 Estimation des paramètres du modèle

On dispose de deux méthodes d'estimation :

- on remarque que ρ est théoriquement égal au coefficient de corrélation entre X_i et X_{i-1} et on estime le coefficient ρ par le coefficient de corrélation empirique correspondant
- on cherche le meilleur ρ qui s'adapte à la décroissance de l'autocorrélogramme empirique par une méthode de régression.

4.5 Validation des paramètres du modèle

Une fois calculé un estimateur $\hat{\rho}$, on calcule des estimateurs du bruit par $\varepsilon_i = X_i - \hat{\rho}X_{i-1}$. Puis on calcule l'autocorrélogramme de ce bruit pour vérifier l'absence de corrélation. On peut également faire passer un test d'idépendance à cette série. On peut ensuite calculer l'estimateur de la variance empirique de cette série pour estimer σ^2 Si la série de bruit est bien décorrélée on considère que le modèle est correct.

4.6 Prévision d'une valeur à partir du passé de la série

Si le modèle AR(1) est validé, on utilise le prédicteur théorique dans ce modèle ; le projeté de X_t sur le passé à l'instant t est par définition ρX_{t-1} ; c'est la meilleure prévision linéaire pour X_t connaissant le passé. On propose donc la prévision $\hat{X}_t = \hat{\rho}X_{t-1}$; On calcule un intervalle de confiance évaluant la précision de cette prévision. Au premier ordre, l'erreur de prévision vient de ε_i variable gaussienne de variance σ^2. Comme on a un estimateur de σ, on peut construire l'intervalle de confiance. En toute rigueur, il y a aussi une erreur provenant de l'erreur d'estimation de ρ mais cet effet est négligeable tant que l'on dispose de suffisamment de données.

5 Construction des processus ARMA

Cette démarche va maintenant être appliquée à la classe des processus ARMA. Nous devons tout d'abord étudier les conditions d'existence de processus vérifiant les équation ARMA puis déterminer les conditions d'existence d'une solution stationnaire. Nous introduisons les filtres linéaires qui vont nous être indispensables pour cette étude théorique.

5.1 Filtres linéaires et opérateur retard

Soit $X = (X_n)_{n\in\mathbb{Z}}$ un processus à temps discret stationnaire, soit I un ensemble d'indices de $\mathbb{Z}$ et $(a_i)_{i\in I}$ une suite de coefficients réels. On définit $Y = (Y_n)_{n\in\mathbb{Z}}$ par $Y_n = \sum_{i\in I} a_i X_{n-i}$ pour $n \in \mathbb{Z}$.

Théorème 4. *Filtres linéaires*

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

La dernière propriété définit rigoureusement les processus linéaires introduits au paragraphe **??**.

Preuve :

(1) L'existence ne pose pas de problème. Pour la stationarité, en supposant $I = -m, \cdots, m$, comme X est stationnaire, on sait que pour tout $n \in \mathbb{N}^*$, pour tout $t_1, \cdots, t_n$ dans $\mathbb{Z}$, alors pour tout $c \in \mathbb{Z}$, $(X_{t_1-m}, X_{t_1-m+1}, \cdots, X_{t_1+m}, X_{t_2-m}, \cdots, X_{t_n+m})$ a même loi que $(X_{t_1-m+c}, X_{t_1-m+1+c}, \cdots, X_{t_1+m+c}, X_{t_2-m+c}, \cdots, X_{t_n+m+c})$. Maintenant, en considérant la fonction $g : \mathbb{R}^{(2m+1)k} \to \mathbb{R}^k$ telle que $g(X_{t_1-m}, X_{t_1-m+1}, \cdots, X_{t_1+m}, X_{t_2-m}, \cdots, X_{t_n+m}) = \left(\sum_{i=-m}^m a_i X_{t_1-i}, \cdots, \sum_{i=-m}^m a_i X_{t_n-i}\right)$, g étant continue donc mesurable, on montre bien que $g(X_{t_1-m}, X_{t_1-m+1}, \cdots, X_{t_1+m}, X_{t_2-m}, \cdots, X_{t_n+m})$ à la même loi que $g(X_{t_1-m+c}, X_{t_1-m+1+c}, \cdots, X_{t_1+m+c}, X_{t_2-m+c}, \cdots, X_{t_n+m+c})$ donc $(Y_{t_1}, \cdots, Y_{t_n})$ a la même loi que $(Y_{t_1+c}, \cdots, Y_{t_n+c})$: (Y_t) est bien stationnaire.

Si X est stationnaire d'ordre 2, il est facile voir que $\mathbb{E}Y_t$ est constante. On a $\mathrm{cov}(Y_s, Y_t) = \sum_{i\in I}\sum_{i'\in I} a_i a_{i'} \mathrm{cov}(X_{t-i}, X_{s-i'}) = \sum_{i\in I}\sum_{i'\in I} a_i a_{i'} \gamma(t-s-i+i')$ en notant $\gamma(k) = \mathrm{cov}(X_0, X_k)$. Donc $\mathrm{cov}(Y_s, Y_t)$ est bien une fonction de $t-s$. De plus, on peut intervertir i et i' et du fait de la parité de γ on voit bien que $\mathrm{cov}(Y_s, Y_t)$ est une fonction de $|t-s|$.

(2) On a $\mathbb{E}|Y_t| \leq \mathbb{E}\left(\sum_{i\in I} |a_i|\,|X_{t-i}|\right) \leq \left(\sum_{i\in I} |a_i|\right) \sup_{j\in\mathbb{Z}} \mathbb{E}|X_j|$ d'après le Théorème de Lebesgue, donc $\mathbb{E}|Y_t| < \infty$: (Y_t) est bien finie avec une probabilité 1.

Pour la stationarité stricte, on procède comme précédemment en considérant des restrictions $(Y_t^{(m)})_t$ qui sont bien stationnaires. Du fait de l'existence d'une limite, on a $(Y_{t_1}^{(m)}, \cdots, Y_{t_n}^{(m)})$ qui tend dans $\mathbb{L}^1$ vers $(Y_{t_1}, \cdots, Y_{t_n})$ lorsque m tend vers l'infini, donc on a également convergence en loi. Aussi comme $(Y_{t_1}^{(m)}, \cdots, Y_{t_n}^{(m)})$ a la même loi que $(Y_{t_1+c}^{(m)}, \cdots, Y_{t_n+c}^{(m)})$, cette égalité a également lieu à la limite: (Y_t) est bien stationnaire.

(3) On a $\mathbb{E}(Y_t - Y_t^{(m)})^2 = \mathbb{E}\left(\sum_{|i|>m}\sum_{|i'|>m} a_i\, a_{i'}\, \mathbb{E}(X_{t-i}X_{t-i'}\right) \leq \left(\sum_{|i|>m} |a_i|\right)^2 \sup_{j\in\mathbb{Z}} \mathbb{E}|X_j^2|$ d'après le Théorème de Lebesgue, donc $\mathbb{E}(Y_t - Y_t^{(m)})^2 \to 0$ $(m \to \infty)$ car $\sum_{i\in I} |a_i| < \infty$: (Y_t) existe dans $\mathbb{L}^2$.

La stationarité stricte s'obtient comme dans le cas précédent et la stationarité d'ordre 2 utilise le point (1): on a $(Y_t^{(m)})_t$ qui est stationnaire d'ordre 2 et comme on a convergence dans $\mathbb{L}^2$ donc en loi, on a bien convergence de l'espérance et de la covariance des $(Y_t^{(m)})$: (Y_t) est bien aussi strictement stationnaire d'ordre 2.

(4) Dans le cas d'un processus linéaire, on peut donc obtenir l'existence et la stationarité sous la condition $\sum_{i\in I} a_i^2 < \infty$ qui est plus faible que la condition $\sum_{i\in I} |a_i| < \infty$. En effet, on a $\mathbb{E}(Y_t - Y_t^{(m)})^2 = \mathbb{E}\left(\sum_{|i|>m}\sum_{|i'|>m} a_i\, a_{i'}\, \mathbb{E}(X_{t-i}X_{t-i'}\right) = \sum_{|i|>m} a_i^2\ \mathbb{E}(X_0^2) \to 0$ $(m \to \infty)$ dès que (X_t) est un bruit blanc. La preuve de la stationarité est immédiate (voir (3)), et pour la stationarité d'ordre 2, on a clairement $\mathrm{cov}(Y_s, Y_t) = \mathbb{E}X_0^2 \sum_{i\in I} a_i a_{i+s-t}$ qui est une fonction dépendant de $|t-s|$ et qui existe (d'après Cauchy-Schwarz).

Nous introduisons maintenant le formalisme de l'opérateur retard et des opérateurs qui en dérivent. Ces opérateurs vont nous permettre de réécrire les équations ARMA(p, q) et de déterminer les conditions sous lesquelles existe une solution stationnaire à cette équation.

Définition 16. *Pour $X = (X_k)_{k\in\mathbb{Z}}$ un processus à temps discret, on définit l'opérateur retard B comme l'application linéaire qui à X associe $Y = (Y_n)_{n\in\mathbb{Z}} = B \cdot X$ tel que $Y_k = X_{k-1}$ pour $k \in \mathbb{Z}$.*

C'est un cas très particulier (seul le coefficient a_1 est non nul) du Théorème **??**. Si le processus X est stationnaire, le processus Y a la même loi que X. Nous pouvons maintenant construire des opérateurs à partir de B.

- On définit la puissance B^n comme la n-ième composée de B par lui-même : $B^n \cdot X_k = X_{k-n}$.
- On construit l'opérateur inverse de B noté B^{-1} par la relation $(B^{-1}.X)_k = X_{k+1}$. Cet opérateur vérifie bien $B^{-1}B = BB^{-1} = I_d$.
- On définit la puissance $B^{-n} = (B^{-1})^n$ et on vérifie immédiatement que $B^{-n} = (B^n)^{-1}$, d'où la définition de B^k pour $k \in \mathbb{Z}$ (en posant $B^0 = Id$).
- On peut donc définir des polynômes de B (et de B^{-1}) : $P(B) = \sum_{j=0}^{p} a_j B^j$ où $(a_i)_{0\leq i\leq p} \in \mathbb{R}^{p+1}$, $(P(B)\cdot X)_k = \sum_{j=0}^{p} a_j X_{k-j}$.

Pour résoudre l'équation ARMA, nous allons devoir inverser un polynôme de B ; cet inverse est en général représenté par une série infinie de puissance de B.

Proposition 10. *On suppose que* $\sum_{i=-\infty}^{\infty} |a_i| < \infty$, *et on définit* $f(B) = \sum_{i=-\infty}^{\infty} a_i B^i$, *alors*

- $f(B)$ *est un opérateur sur l'ensemble des processus de norme* $\mathbb{L}^1$ *bornée.*
- $f(B)$ *est un opérateur sur l'ensemble des processus de norme* $\mathbb{L}^2$ *bornée.*

La première propriété signifie que pour tout processus X de norme $\mathbb{L}^1$ bornée, $f(B).X$ est bien définie et de norme $\mathbb{L}^1$ bornée. C'est une simple traduction de la propriété 2 du théorème précédent. La deuxième propriété correspond à la propriété 3.

Tous les polynômes sont décomposables en produit de polynômes complexes du premier degré. Nous construisons les inverses de ces polynômes du premier degré.

Proposition 11. *Inversion des polynômes de degré 1. Soit* $z \in \mathbb{C}$.

- *si* $|z| < 1$, $(Id - zB)^{-1} = \sum_{i=0}^{\infty} z^i B^i$;
- *si* $|z| > 1$, $(Id - zB)^{-1} = -z^{-1}B^{-1}(Id - z^{-1}B^{-1})^{-1} = -\sum_{i=1}^{\infty} z^{-i}B^{-i}$;
- *si* $|z| = 1$, $Id - zB$ *n'est pas inversible.*

Pour prouver la première relation, on remarque que

$$(Id - zB)\sum_{i=0}^{n} z^i B^i = Id - z^{n+1}B^{n+1}.$$

Donc

$$(1 - zB)\sum_{i=0}^{\infty} z^i B^i - Id = (Id - zB)\sum_{i=n+1}^{\infty} z^i B^i - z^{n+1}B^{n+1}.$$

Si on applique l'opérateur de gauche à un processus X de norme bornée dans $\mathbb{L}^1$ ou $\mathbb{L}^2$, on constate que le résultat tend vers le processus nul quand n tend vers l'infini. Le premier membre est donc nul. La deuxième relation se prouve de la même façon. Pour montrer la troisième relation, il suffit de construire un processus borné non nul dont l'image est nulle, soit le processus X défini par $X_i = Zz^{-i}$ où Z est une variable non nulle de carré intégrable.

La densité spectrale de processus écrit à partir de filtres linéaires se déduit directement de l'expression de ces filtres:

Propriété 10. *Soit le processus à temps discret $X = (X_n)_{n\in\mathbb{Z}}$ tel que $X_n = \sum_{i=-\infty}^{\infty} a_i\varepsilon_{n-i}$ pour $n \in \mathbb{Z}$, avec $(\varepsilon_n)_{n\in\mathbb{Z}}$ un bruit blanc et $\sum_{i=-\infty}^{\infty} |a_i| < \infty$. Alors X admet une densité spectrale f telle que*

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

Propriété 11. *Soit le processus à temps discret $Y = (Y_n)_{n\in\mathbb{Z}}$ tel que $Y_n = \sum_{i=-\infty}^{\infty} a_i X_{n-i}$ pour $n \in \mathbb{Z}$, avec $X = (X_n)_{n\in\mathbb{Z}}$ un processus à temps discret stationnaire de densité spectrale f_X et $\sum_{i=-\infty}^{\infty} |a_i| < \infty$. Alors Y est un processus à temps discret stationnaire de densité:*

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

5.2 Application à la construction des processsus ARMA(p, q)

Nous considérons l'équation ARMA(p, q) introduite précédemment. Nous l'écrivons sous la forme :

$$P(B) \cdot X = Q(B) \cdot \varepsilon,$$

où $P(B) = a_0 + a_1 B + \cdots + a_p B^p$ et $Q(B) = b_0 + b_1 B + \cdots + b_q B^q$.

Proposition 12. *Propriétés de l'équation ARMA.*

- *si P n'a pas de racine de module 1 :*
  - *si P et Q ont des racines communes, on simplifie les facteurs correspondants ; on obtient une équation avec des polynômes P' et Q' ayant des racines distinctes et l'équation a une unique solution stationnaire. Cette solution peut s'écrire $X = \sum_{i=-\infty}^{\infty} c_i B^i \cdot \varepsilon$, soit $X_k = \sum_{i=-\infty}^{\infty} c_i \varepsilon_{k-i}$.*
  - *si de plus P a toutes ses racines de module supérieur à 1, l'unique solution stationnaire peut s'écrire $X = \sum_{i=0}^{\infty} c_i B^i \cdot \varepsilon$, soit $X_k = \sum_{i=0}^{\infty} c_i \varepsilon_{k-i}$. Le processus solution est écrit sous forme causale et ε est son innovation.*
  - *si P et Q ont toutes leurs racines de module supérieur à 1, le processus est de plus inversible. Les innovations peuvent s'écrire sous forme de série causale infinie des X_i : $\varepsilon_k = \sum_{i=0}^{\infty} d_i X_{k-i}$.*
- *si P a des racine de module 1 :*
  - *si toutes les racines de module 1 de P sont communes à Q, on simplifie les facteurs correspondants ; mais l'équation a une infinité de solutions stationnaires. Une seule de ces solutions est régulière et de la forme précédente.*
  - *si le polynôme P a une racine de module 1, qui n'est pas racine de Q, l'équation n'a pas de solution stationnaire.*

Preuve : Lorsque P et Q ont des racines communes de module différent de 1, on peut multiplier à gauche les deux membres par les inverses des polynômes de premier degré correpondants et donc simplifier P et Q en P' et Q'. Le polynôme $P'(B)$ obtenu est également inversible. La solution est alors $X = P'(B)^{-1}Q'(B)\varepsilon$, qui est une série de puissances positives et négatives de B. Si toutes les racines de P' sont de module supérieur à 1, les polynômes de premier degré sont du premier type décrit ($|z| < 1$) dans la proposition **??**. L'inverse de $P'(B)$ est alors une série de puissances positives de B.
Si les racines de Q sont à l'extérieur du cercle unité, l'inverse de Q s'écrit comme une série à coefficients indexés par $\mathbb{N}$ d'où le résultat.
Si chaque racine μ_i de module 1 de P est racine de Q, on ne peut plus calculer l'inverse des polynômes de premier degré correspondant. Cependant la solution de l'équation simplifiée définie par P' et Q' reste

solution de notre équation originale. On peut construire une infinité de solutions en ajoutant à cette solution un processus singulier qui est annulé par un des polynômes $Id - \mu_i^{-1}B$. Pour prouver le quatrième résultat, nous utilisons la densité spectrale des processus.

Remarque : pour utiliser un modèle ARMA pour la prévision, la représentation causale est essentielle : la valeur observée doit s'exprimer en fonction des observations passées et d'un aléa indépendant à chaque étape de temps. Nous n'utiliserons donc pas de modéle ARMA ayant des racines de module inférieur ou égal à 1 en pratique. Cela semble très restrictif d'abandonner une grande famille de modèles stationnaires. En réalité, pour chacun de ces modèles non causaux on peut construire un processus ARMA de même fonction d'autocovariance ayant une représentation causale, en remplacant dans le polynôme P les racines de module inférieure à 1 par leur inverse.

Proposition 13. *Lorsque les racines de P ne sont pas de module 1, la densité spectrale du processus ARMA stationnaire solution s'exprime directement en fonction des polynômes de l'équation ARMA correspondante :*

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

où σ^2 est la variance du bruit blanc.

La preuve de ce résultat nécessite la représentation en série de Fourier des processus aléatoires qui ne sera pas abordée dans ce cours.

- Cette proposition permet de prouver le quatrième résultat de la proposition précédente ; si un processus stationnaire était solution, il aurait nécessairement une densité spectrale de cette forme. Mais si le polynome P possède une racine de module 1, $P(e^{-i\lambda})$ s'annule sur l'intervalle $[-\pi, \pi]$ et la densité spectrale définie ici n'est pas intégrable. Or une densité spectrale est toujours intégrable, d'où l'on conclut que le processus stationnaire solution n'existe pas dans ce cas.
- Elle permet de montrer qu'on peut toujours trouver un processus causal de même fonction de covariance qu'un processus non causal. Soit X le processus stationnaire solution de $P(B)X = Q(B)\varepsilon$. Soit z une racine de P de module inférieure à 1. Supposons que z est réelle Soit P' le polynôme défini par $P'(x) = P(x)\frac{x-z^{-1}}{x-z}$. Comme
$$|e^{-i\lambda} - z| = |ze^{-i\lambda}||z^{-1} - e^{i\lambda}| = |z||e^{-i\lambda} - z^{-1}|,$$
on voit que la densité spectrale correspondant au processus stationnaire X' solution de $P'(B)X = Q(B)z^2\varepsilon$ est identique à celle de X. Si z est complexe, le polynôme étant réel, $\bar{z}$ est également racine. Soit P' le polynôme défini par $P'(x) = P(x)\frac{(x-z^{-1})(x-\bar{z}^{-1})}{(x-z)(x-\bar{z})}$. Comme $|e^{-i\lambda} - \bar{z}^{-1}| = |z^{-1}||z - e^{-i\lambda}|$, la densité spectrale correspondant au processus stationnaire X' solution de $P'(B)X = Q(B)|z|^2\varepsilon$ est identique à celle de X. Les deux processus ont alors exactement la même fonction d'autocovariance.
- On peut procéder de même pour remplacer les racines de Q de module inférieur à 1 et obtenir un processus inversible de même fonction d'autocovariance.

Nous résumons ici les conditions permettant de définir un processus ARMA régulier, causal et inversible.

Définition 17. *Soit P et Q deux polynômes de degrés respectifs p et q tels que P et Q n'ont pas de racine comlexe commune. On suppose que les racines de P et de Q sont extérieures au cercle unité. Soit également $\varepsilon = (\varepsilon_n)_{n\in\mathbb{Z}}$ un bruit blanc (fort).*

- *Un processus $X = (X_n)_{n\in\mathbb{Z}}$ tel que $P(B)\cdot X = \varepsilon$ est appelé un processus AR(p). Ceci revient à écrire qu'il existe $(a_1, \cdots, a_p) \in \mathbb{R}^p$ tel que:*
$$X_n = -a_1X_{n-1} + \cdots - a_pX_{n-p} + \varepsilon_n \quad \textit{pour tout } n \in \mathbb{Z}.$$
- *Un processus $X = (X_n)_{n\in\mathbb{Z}}$ tel que $X = Q(B)\cdot\varepsilon$ est appelé un processus MA(q). Ceci revient à écrire qu'il existe $(b_1, \cdots, b_q) \in \mathbb{R}^q$ tel que:*
$$X_n = \varepsilon_n + b_1\varepsilon_{n-1} + \cdots + b_q\varepsilon_{n-q} \quad \textit{pour tout } n \in \mathbb{Z}.$$
- *Un processus $X = (X_n)_{n\in\mathbb{Z}}$ tel que $P(B)\cdot X = Q(B)\cdot\varepsilon$ est appelé un processus ARMA(p, q). Ceci revient à écrire qu'il existe $(a_1, \cdots, a_p) \in \mathbb{R}^p$ et $(b_1, \cdots, b_q) \in \mathbb{R}^q$ tels que:*
$$X_n + a_1X_{n-1} + \cdots + a_pX_{n-p} = \varepsilon_n + b_1\varepsilon_{n-1} + \cdots + b_q\varepsilon_{n-q} \quad \textit{pour tout } n \in \mathbb{Z}.$$

5.3 Calcul de la fonction d'autocovariance pour les modèles ARMA(p, q)

Pour un modèle ARMA(p, q) causal, on peut calculer explicitement la fonction d'autocovariance à partir des coefficients des polynômes de l'équation. La méthode consiste à :

- Calculer les racines complexes du polynôme P.
- Calculer les inverses des monômes correspondants à ces racines, multiplier ces inverses entre eux, puis par le polynôme Q.

On obtient alors la représentation du processus sous forme de filtre linéaire causal de l'innovation ε:

$$X_t = \sum_{i=0}^{\infty} c_i \varepsilon_{t-i}$$

Grâce à cette représentation, on peut calculer l'autocovariance d'ordre k par

$$\gamma(k) = \sigma^2 \sum_{i=0}^{\infty} c_i c_{k+i}$$

Exemple : soit le modèle ARMA(1,1) :

$$(1 - 0,5B)X = (1 + 0,2B)\varepsilon$$

L'inverse de $(1 - 0,5B)$ est $\sum_{i=0}^{\infty}(0,5B)^i$. En multipliant par Q on obtient:

$$X_t = \varepsilon_t + \sum_{i=1}^{\infty}(0,2(0,5)^{i-1} + (0,5)^i)\varepsilon_{t-i} = \varepsilon_t + \sum_{i=1}^{\infty} 0,7(0,5)^{i-1}\varepsilon_{t-i}$$

et on peut calculer

$$\gamma_0 = \sigma^2 \left(1 + \sum_{i=1}^{\infty} 0,7^2(0,5)^{2i-2}\right).$$
$$\gamma_1 = \sigma^2 \left(0,7 + \sum_{i=1}^{\infty} 0,7^2(0,5)^{2i-1}\right) ...$$

6 Modélisation et prévision à l'aide des processus ARMA

Nous récapitulons les étapes de la prévision :

- Calcul de la tendance $f(t)$ par la méthode de régression précédente.
- Modélisation de la série résultante U_t par un processus ARMA(p, q).
- Calcul de la prévision et de son intervalle de confiance.

La première étape a été décrite précédemment. On cherche maintenant un modèle ARMA stationnaire approchant la série $U_t = X_t - f(t)$; il faut choisir les ordres p et q du modèle puis estimer les coefficients a_i et b_j ; il faut ensuite estimer les ε_t résidus correspondant et vérifier qu'ils forment bien un bruit blanc.

6.1 Choix des ordres (p, q)

Comme pour la régression multiple, il n'existe pas de meilleurs choix évidents ; plus les ordres sont grands, meilleure est l'adéquation, mais le nombre de paramètres à estimer augmente et leur estimation statistique devient imprécise. On emploie donc une règle heuristique en utilisant les corrélogrammes empiriques calculés à partir des données.

Les deux séries de coefficient sont utilisés pour la propriété suivante

Propriété 12. *Pour un processus AR(p), $\alpha_k = 0$ dès que $k > p$. Pour un processus MA(q), $\rho_k = 0$ dès que $k > q$.*

Pour estimer ces coefficients d'autocorrélation on utilise les coefficients d'autocorrélation empirique. Pour estimer le coefficient d'autocorrélation partielle, on effectue la régression des $(X_{i+k})_{i=1,\dots,n-k}$ sur les vecteurs $(X_{i+k-1})_{i=1,\dots,n-k}$, $(X_{i+k-1})_{i=1,\dots,n-k}$, jusqu'à $(X_i)_{i=1,\dots,n-k}$ et on retient le coefficient de regression sur ce dernier vecteur. L'étude théorique de ces deux estimateurs permet de construire un test de significativité qui est calculé par les logiciels statistiques ; on observe les valeurs de k à partir desquelles les corrélogrammes ne sont plus significatifs. Cela donne une borne raisonnable, généralement trop grande, pour le choix des p et q.

6.2 Identification des paramètres d'un processus ARMA(p,q) stationnaire

6.2.1 Processus AR(p)

Il est possible de calculer un estimateur des coefficients du modèle à partir de l'estimateur empirique de la covariance. L'algorithme de Durbin-Levinson permet de calculer les coefficients d'autorégression en fonction des valeurs de la fonction de covariance. En substituant dans l'algorithme les estimateurs empiriques de la covariance aux vraies valeurs de la covariance, on obtient un estimateurs des coefficients. On peut plus directement écrire la relation entre fonction de covariance et coefficients d'autoregression sous forme d'un sytème linéaire appelé équations de Yule-Walker.

Proposition 14. *Soit X le processus AR(p) définie par l'équation $X_t + a_1 X_{t-1} + \ldots + a_p X_{t-p} = \varepsilon_t$, tel que le polynôme P associé a toutes ses racines à l'extérieur du cercle unité. Alors les coefficients a_i du polynôme sont solutions du système linéaire:*

$$\begin{pmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(-p+1) \\ \gamma(1) & \gamma(q) & \cdots & \gamma(-p+2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} -\gamma(1) \\ -\gamma(2) \\ \vdots \\ -\gamma(p) \end{pmatrix}.$$

Preuve : Le polynôme P associé ayant toutes ses racines à l'extérieur du cercle unité, le bruit ε est l'innovation du processus X. Donc le produit scalaire de ε_t avec les variables X_i pour $i < t$ est nul. En effectuant successivement le produit scalaire de chaque membre de l'équation AR(p) par les variables X_{t-1} à X_{t-p}, on obtient le système des équations de Yule-Walker.
Pour estimer les coefficients a_i à partir des données, on substitue à $\gamma(k)$ l'estimateur empirique de la covariance $\hat{\gamma}(k) = \frac{1}{n-k}\sum_{i=1}^{n-k} X_i X_{i+k}$. Les solutions $\hat{a}_i$ du système correspondant sont des estimateurs consistants des coefficients :

$$\begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(-1) & \cdots & \hat{\gamma}(-p+1) \\ \hat{\gamma}(1) & \hat{\gamma}(q) & \cdots & \hat{\gamma}(-p+2) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\gamma}(p-1) & \hat{\gamma}(p-2) & \cdots & \hat{\gamma}(0) \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{pmatrix} = \begin{pmatrix} -\hat{\gamma}(1) \\ -\hat{\gamma}(2) \\ \vdots \\ -\hat{\gamma}(p) \end{pmatrix}.$$

6.2.2 Processus MA(q)

L'estimation des paramètres MA(q) s'appuie directement sur l'algorithme des innovations. Soit X le processus stationnaire correspondant à l'équation MA(q):

$$X_t = \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}.$$

L'algorithme des innovations de la proposition **??** pour la longueur q permet de calculer les valeurs des coefficients b_1 à b_q à partir de la fonction de covariance γ. Comme dans la méthode de Yule Walker, on substitue à γ son estimateur empirique $\hat{\gamma}$, puis on calcule l'algorithme pour la longueur q et on choisit comme estimateur $\hat{b}_i = \theta_{q,i}$.

6.2.3 Méthode de maximum de vraisemblance et de contraste

Dans un premier temps nous allons modéliser la série de données par un modèle ARMA(p,q) gaussien. Nous appliquons la méthode générale du maximum de vraisemblance pour déterminer les paramètres de notre modèle, c'est-à-dire les coefficients des deux polynômes P et Q et la variance σ^2 du bruit. Les degrés des

polynômes ont été choisis préalablement. La méthode consiste à écrire la densité de probabilité correspondant aux données $(X_1, \ldots, X_n)$. Cette densité de probabilité dépend des paramètres du modèle. Nous choisissons les paramètres qui rendent cette fonction la plus grande possible. La densité de probabilité exprimée en fonction des paramètres est appelée vraisemblance, d'où le nom de maximum de vraisemblance pour cette méthode. Dans le cas d'un processus gaussien centré nous avons vu que la vraisemblance a la forme :

$$f(X) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{{}^tX\Sigma^{-1}X}{2}\right)$$

où Σ est la matrice de covariance du vecteur X. Il suffit donc de trouver parmi les processus ARMA(p, q), celui dont la matrice de covariance maximise cette quantité pour $X = (X_1, \ldots, X_n)$. Cette recherche de maximum peut être très coûteuse en temps si le calcul n'est pas correctement préparé. Pour cela on diagonalise la matrice de covariance en utilisant l'algorithme des innovations :

Proposition 15. *Soit un processus gaussien $(X_i)_{i>0}$ de fonction d'autocovariance $\gamma(i, j)$. La densité de probabilité du vecteur $(X_1, \ldots, X_n)$ peut s'écrire*

$$f(X_1, \ldots, X_n) = \frac{1}{(2\pi 2\sigma^2)^{n/2}\sqrt{v_0 \cdots v_{n-1}}} \exp\left(-\frac{2}{\sum}_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{v_{i-1}}\right)$$

Les v_i et les $\hat{X}_i$ sont calculés récursivement grâce à la proposition **??**.

Il ne nous reste qu'à appliquer cette méthode au cas particulier des processus ARMA gaussiens. La vraisemblance s'écrit

$$L_n(\mathbf{a}, \mathbf{b}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}\sqrt{v_0 \cdots v_{n-1}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{v_{i-1}}\right)$$

avec $\mathbf{a}$, $\mathbf{b}$ coefficients respectifs des polynômes P et Q et σ^2, variance du bruit. Les v_i et $\hat{X}_i$ sont calculés par l'algorithme de la proposition **??** :

$$\hat{X}_{i+1} = \sum_{j=1}^{i} \theta_{i,j}(X_{n+1-j} - \hat{X}_{i+1-j}) \qquad \text{si } i < m$$

$$\hat{X}_{i+1} = a_1 X_i + \cdots + a_p X_{i+1-p} + \sum_{j=1}^{q} \theta_{i,j}(X_{i+1-j} - \hat{X}_{i+1-j}) \qquad \text{si } i \geq m$$

Les coefficients v_i et $\theta_{i,j}$ sont indépendants de σ^2. En dérivant par rapport à σ^2, on observe que le maximum est obtenu pour $\hat{\sigma}^2 = n^{-1}S(\mathbf{a}^*, \mathbf{b}^*)$ où $S(\mathbf{a}^*, \mathbf{b}^*)$ est la valeur minimale de

$$S(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{v_{i-1}}.$$

Pour trouver le maximum de $L_n(\mathbf{a}, \mathbf{b}, \hat{\sigma}^2)$ par rapport aux coefficients de $\mathbf{a}$ et $\mathbf{b}$, il suffit de déterminer le minimum de la fonction:

$$l(\mathbf{a}, \mathbf{b}) = \log(n^{-1}S(\mathbf{a}, \mathbf{b})) + n^{-1}\sum_{i=1}^{n} \log(v_{i-1}).$$

La méthode du maximum de vraisemblance consiste à chercher les valeurs de $\mathbf{a}$ et $\mathbf{b}$ qui minimisent cette quantité. Cette recherche est réalisée par un algorithme d'optimisation non linéaire à partir d'une valeur initiale des paramètres $\mathbf{a}$ et $\mathbf{b}$. Cette première valeur proposée doit se trouver près du vrai maximum et correspondre à un modèle causal, car les algorithmes de projections utilisés exploitent cette hypothèse. Les logiciels d'optimisation utilisent une valeur de départ obtenu par l'estimateur de Yule-Walker. La théorie du maximum de vraisemblance donne des conditions pour que cet estimateur soit non seulement consistant mais efficace, c'est-à-dire avec une vitesse de convergence en $\sqrt{n}$ et une variance asymptotique minimale. Cette méthode d'estimation n'est justifiée que lorsque le processus ARMA est gaussien, mais lorsque l'on cherche le prédicteur linéaire, un modèle gaussien et un modèle non gaussien de même fonction d'autocovariance sont parfaitement équivalents, car le prédicteur ne dépend que de la fonction d'autocovariance.

6.2.4 Estimateur de Whittle

Dans le cas d'un processus ARMA(p,q), on pourra préférer une approximation de l'estimateur du maximum de vraisemblance dite approximation de Whittle (voir ci-dessous) qui offre la possibilité de traiter des processus non gaussiens et procure un grand gain en terme de calcul tout en conservant la même vitesse de convergence pour l'estimateur.

Théorème 5. *Soit $X=(X_k)_{k\in\mathbb{Z}}$ un processus ARMA(p,q) stationnaire et de densité spectrale $f_{(\mathbf{a},\mathbf{b})}$. Soit $I_n(\lambda)$ l'estimateur empirique de la densité spectrale défini par*

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

On considère le contraste de Whittle:

$$U_n(\mathbf{a},\mathbf{b})=\frac{1}{4\pi}\int_{-\pi}^{\pi}\log(f_{(\mathbf{a},\mathbf{b})}(\lambda))+\frac{I_n(\lambda)}{f_{(\mathbf{a},\mathbf{b})}(\lambda)}d\lambda.$$

L'estimateur de $(\mathbf{a},\mathbf{b})$ par minimum de contraste est la valeur de $(\mathbf{a},\mathbf{b})$ qui minimise $U_n(\mathbf{a},\mathbf{b})$.

Cet estimateur converge à la même vitesse que l'estimateur du maximum de vraisemblance. De plus, ces propriétés de convergence sont établies même dans le cas où le processus étudié n'est pas gaussien. Il est bien adapté au cas des processus ARMA, car la densité spectrale de ces processus s'exprime directement par rapport aux coefficients des polynômes, ce qui n'est pas le cas pour la fonction d'autocovariance.

6.3 Significativité des paramètres

Une fois calculées les estimations des paramètres, on peut tester si tous les paramètres estimés sont significativement distincts de 0. Les estimateurs de paramètres étant asymptotiquement gaussiens et de variance estimable, on peut construire pour chaque paramètre, un test de l'hypothèse "ce paramètre est nul" et ne conserver dans le modèle que les paramètres rejetés par ce test de nullité. Si les paramètres de grand ordre a_p ou b_q sont acceptés comme nuls, cela veut dire qu'ils ne sont pas nécessaires dans le modèle et qu'il faut baisser l'ordre p ou q correspondant puis réestimer tous les paramètres.

6.4 Test d'adéquation

Il faut également vérifier que le modèle est adéquat, au sens où ses résidus forment bien un bruit blanc. Si ce n'est pas le cas, c'est qu'il y a encore de la dépendance dans les résidus. On cherchera à utiliser un modèle d'ordre p ou q plus grand afin de prendre en compte cettte dépendance. Pour tester l'adéquation à un processus ARMA, on peut utiliser un test dit de Portmanteau (ce qui signifie “fourre-tout” en anglais). Après avoir calculé les résidus estimés $(\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_n)$ obtenus à partir $(X_1,\cdots,X_n)$ et des coefficients estimés par une méthode de type maximum de vraisemblance ou minimum de contraste. On définit

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

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

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

6.5 Sélection de modèles

Supposons que la procédure précédente ait permis de construire un modèle valide pour les données dont on dispose. Il est possible qu'un autre modèle soit également valide, et s'adapte mieux aux données. De fait, plus on va choisir des ordres p et q importants,plus le modèle a des chances de présenter des résidus plus petits ; l'inconvénient est que les paramètres seront plus nombreux et moins correctement estimés. On va introduire un critère de choix entre les modèles qui met en balance l'adéquation du modèle avec le nombre de paramètres. De cette façon, on établira un compromis entre les deux inconvénients. Pour ce faire on utilisera un critère de sélection de modèle, par exemple le critère AICC qui met en balance le nombre de paramètres du modèle $p+q$ avec la vraisemblance obtenue par les estimateurs du maximum de vraisemblance.

$$AICC(p,q) = -2\log L_n(\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\sigma}^2) + 2(p+q+1)\frac{n}{n-p-q-2},$$

On peut également utiliser le critère bayésien BIC. Dans le cas d'un processus ARMA causal inversible celui-ci s'écrit:

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

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

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

6.6 Prévision

Connaissant les estimations des de la tendance f, des paramètres du modèle ARMA et les estimations du bruit ε, on propose d'estimer X_{n+1} par

$$\hat{X}_{n+1} = f(n+1) - \hat{a}_1(X_n - f(n)) - \cdots - \hat{a}_p(X_{n-p+1} - f(n-p+1)) + \hat{b}_1\hat{\varepsilon}_n + \cdots + \hat{b}_q\hat{\varepsilon}_{n-q+1}.$$

Le calcul de l'intervalle de confiance repose sur l'étude des covariances des différents estimateurs utilisés. Ces estimateurs étant très dépendants, le calcul est compliqué, mais il est généralement effectué par le logiciel qui calcule les estimations des paramètres. On observe dans la formule la diminution du nombre de termes correspondant aux valeurs réellement observées. L'information apportée par le modèle ARMA diminue et la prévision se rapproche de la prévision déterministe définie par la fonction f. La prévision utilisant les modèles ARMA n'a d'intérêt qu'à court terme.

7 Présentation d'autres modèles

Les méthodes présentées précédemment sont applicables à des processus stationnaires du second ordre pour lesquels on se satisfait d'un prédicteur linéaire. Les modèles ARMA stationnaires proposés sont représentables sous forme d'un filtre linéaire dont les coefficients forment une série sommable. Dans ce paragraphe nous présentons d'autres modèles de processus pour lesquels l'approche précédente n'est pas suffisante.

7.1 Modèles ARIMA et SARIMA

Pour certaines séries, on observe que les méthodes de régression déterministes pemettent de modéliser le niveau moyen local de la série, mais que les résidus de régression ont une variance qui augmente avec le temps. Nous avons vu que l'augmentation de variance peut provenir du fait que le processus observé est la sommes d'incréments indépendants stationnaires (voir le processus de marche aléatoire). Pour cette raison,

on applique un opérateur de différence au processus X_t de variance croissante pour voir si le processus résultant est stationnaire :

$$Y = (1 - B)X$$

Si la varîance de Y est stable, on modélise Y par un processus ARMA. Sinon on peut faire opérer la différentiation plusieurs fois

$$Z = (1 - B)^k X$$

Si le processus Z obtenu après k différentiations est un processus ARMA(p, q), on dit que X est un processus ARIMA(p, k, q).
On peut par ailleurs observer sur certaines séries que la modélisation déterministe n'est pas suffisante pour traiter la saisonalité de période T présente dans la série. Après régression linéaire, il reste des covariances importantes à distance T, $2T$... Pour traiter ce cas, on applique un filtre de différentiation à distance T :

$$Y = (1 - B^T)X$$

et on regarde si les covariances à distance T, $2T$... ont disparu. On peut également appliquer ce filtre à plusieurs reprises

$$Z = (1 - B^T)^k X$$

Si le processus Z obtenu après k différentiations est un processus ARMA(p, q), on dit que X est un processus SARIMA(p, k, q) de période T.

Remarque : les processus X précédents n'entrent pas dans la famille ARMA stationnaires, car les monômes ajoutés $(1 - B)$ et $(1 - B^T)$ ont des racines de module égale à 1. Les processus résultants sont non stationnaires. Ces définitions permettent d'aggrandir simplement l'ensemble des séries modélisables par les techniques ARMA.

7.2 Modèles ARFIMA

Les modèles ARFIMA sont un cas particulier de modèles linéaires à longue portée. Ces modèles sont caractérisés par une fonction d'autocovariance formant une série non sommable.

Définition 9. *Un processus linéaire à longue portée est un processus linéaire causal dont les coefficients de filtre ne sont pas sommables mais de carrés sommables:*

$$X_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}$$

où ε est un bruit blanc faible, et $\sum_{i=0}^{\infty} |a_i| = \infty$ mais $\sum_{i=0}^{\infty} a_i^2 < \infty$.

Un exemple de tel filtre est donné par l'opérateur inverse de l'opérateur de différentiation fractionnaire $(1 - B)^d$ avec $0 < d < 0,5$.

Définition 10. *L'opérateur de différentiation fractionnaire $(1 - B)^d$ est défini par*

$$(1 - B)^d = \sum_{i=0}^{\infty} \pi_i B^i$$

où

$$\pi_i = \frac{\Gamma(i - d)}{\Gamma(i)\Gamma(-d)} = \frac{(i - 1 - d)(i - 2 - d) \cdots (-d)}{i!}.$$

L'inverse de l'opérateur de différentiation fractionnaire $(1 - B)^d$ est défini par

$$(1 - B)^{-d} = \sum_{i=0}^{\infty} \psi_i B^i$$

où

$$\psi_i = \frac{\Gamma(i + d)}{\Gamma(i)\Gamma(d)} = \frac{(i - 1 + d)(i - 2 + d) \cdots d}{i!}.$$

On montre que $\psi_i \approx C i^{d-1}$ quand i tend vers l'infini, donc $\sum_{i=0}^{\infty} |\psi_i| = \infty$ et $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.

A partir de ce filtre, on génère un processus appelé bruit blanc fractionnaire.

Définition 11. *Le bruit blanc fractionnaire de paramètre d (*$0 < d < 0,5$*) est le processus solution de l'équation :*

$$(1-B)^{-d}X = \varepsilon$$

Il peut donc s'écrire

$$X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

C'est un processus linéaire à longue portée car sa fonction d'autocovariance $\gamma(k) \approx Ck^{2d-1}$ en $+\infty$ et n'est donc pas sommable.
On peut généraliser cette construction aux modèles ARFIMA :

Définition 12. *Soit Y_t un processus causal ARMA(p, q). Le processus ARFIMA(p, d, q) (*$0 < d < 0,5$*) est le processus stationnaire satisfaisant à l'équation :*

$$(1-B)^{-d}X = Y$$

Il peut donc s'écrire

$$X_t = \sum_{i=0}^{\infty} \psi_i Y_{t-i}$$

Le processus X est le résultat de la composition de deux filtres dont les coefficients sont de carré sommable appliqué au bruit blanc d'innovation de Y. Le filtre résultant est de carré sommable, donc le processus X est bien défini dans $\mathbb{L}^2$. On peut également montrer que c'est un processus linéaire à longue portée.
Remarque : la longue portée modifie les propriétés des trajectoires des processus dont les valeurs successives sont plus corrélées que celle d'un processus ARMA : on observe de longues période où le processus est au-dessus de son niveau moyen suivi de longues périodes où il se trouve en-dessous. L'estimation du niveau moyen s'en trouve ralenti, ce que confirme le simple calcul de la variance de l'estimateur empirique de la moyenne : Si $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$\mathrm{var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(|j-i|)$$

Si la somme des covariances est finie et vaut C la variance de $\bar{X}_n$ est équivalente à C/n. Si le processus est un bruit blanc fractionnaire, la somme des covariances est équivalente à Cn^{2d} et la variance est équivalente Cn^{2d-1}. Cette variance tend toujours vers 0 quand le nombre d'observations disponibles augmente, mais beaucoup plus lentement que dans le cas des ARMA.

7.3 Modèles ARCH

Les processus ARCH(p) ont été introduits par Engle (1982) et Bollersev (1986) dans le cadre de données financières où deux comportements particuliers avaient été repérés :

- La série se comporte comme un bruit blanc du second ordre (les corrélations entre données mesurées à des dates différentes sont nulles), mais les carrés des observations sont nettement corrélés. C'est donc bien la structure de dépendance et non simplement les covariances que nous allons vouloir représenter. La théorie des prédicteurs linéaires n'apporte rien pour de telles séries, car ils sont tous égaux à 0.
- Les séries sont globalement stationnaires et centrées mais la variance augmente brutalement pendant de courts instants pour diminuer ensuite

Dans le modèle proposé, la variance dépend conditionnellement du passé.

Définition 13. *Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc fort de variance 1. $(X_t)_{t\in\mathbb{Z}}$ est un processus ARCH(1), s'il existe un couple (a_0, a_1) de constantes réelles positives telles que*

$$\begin{aligned} X_t &= \sigma_t \varepsilon_t \\ \sigma_t &= \sqrt{a_0 + a_1 X_{t-1}^2} \end{aligned}$$

On vérifie facilement que X_t est décorrélé. En effet

$$\mathbb{E}(X_t X_{t+k}) = \mathbb{E}(\varepsilon_t \sigma_t \varepsilon_{t+k} \sigma_{t+k}) = \mathbb{E}(\varepsilon_t)\mathbb{E}(\sigma_t \varepsilon_{t+k} \sigma_{t+k}) = 0,$$

car ε_t est indépendant des autres variables. L'espérance conditionelle de X_t^2 est donné par la deuxième formule :

$$\mathbb{E}(X_t^2 | X_{t-1}) = a_0 + a_1 X_{t-1}^2,$$

ce qui prouve que les carrés sont corrélés. Pour ce modèle, si on veut construire le processsus à partir d'une valeur $\sigma(0) = \sigma_0$, on peut donner une condition nécessaire d'existence d'une solution stationnaire : $\mathbb{E}(X_1^2) = a_0 + a_1 \mathbb{E}(X_0^2) = \sigma_0^2$ ce qui donne une relation entre la variance de la première valeur et les paramètres : $(1 - a_1)\sigma_0^2 = a_0$. Il faut donc que $0 < a_1 < 1$. Cette condition est par ailleurs suffisante. Il est possible comme dans le cas des processus AR de généraliser le modèle en prenant en compte plus de valeurs du passé :

Définition 14. *Soit $(\varepsilon_k)_{k\in\mathbb{Z}}$ un bruit blanc. Alors on dit que $(X_k)_{k\in\mathbb{Z}}$ est un processus ARCH(p), où $p \in \mathbb{N}^*$ s'il existe une famille de constantes réelles positives $(a_0, a_1, \cdots, a_p)$ telles que*

$$\begin{aligned} X_t &= \sigma_t \varepsilon_t \\ \sigma_t &= \sqrt{a_0 + \sum_{j=1}^{p} a_j X_{k-j}^2} \end{aligned}$$

On détermine des conditions d'existence de processus stationnaires vérifiant cette équation de récurrence :

Propriété 13. *Il existe une solution stationnaire si et seulement si $\mathbb{E}\left[\varepsilon_0^2\right] \sum_{k=1}^{p} a_k < 1$.*

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

Un processus ARCH(p) ne pourra pas être identifié à partir de sa densité spectrale qui est celle d'un bruit blanc. Une conséquence de ceci est également qu'un ARCH(p) ne peut pas être gaussien (car sinon ce serait un bruit blanc, ce qu'il n'est pas).
On peut généraliser ces processus en prenant en compte un passé infini (ARCH(∞)) ou en remplaçant l'équation de variance par

$$\sigma_t = \sqrt{a_0 + \sum_{j=1}^{p} a_j X_{k-j}^2 + \sum_{j=1}^{p} b_j \sigma_{k-j}^2}$$

pour des coefficients b_j positifs. Ces derniers modèles sont appelés GARCH(p, q). Pour tous ces modèles des recherches sont actuellement menés pour obtenir tous les outils dont nous disposons dans le cadre des modèles ARMA : conditions d'existence des solutions stationnaires, méthodes d'estimation des paramètres, sélection de modèles, estimations et tests des résidus ; il faut à cela rajouter : estimation de la loi des variables (qui ne peuvent plus être considérées comme gaussiennes) et construction de prédicteurs non linéaires...