Econométrie

Guillaume Chevillon
OFCE & Univ of Oxford
guillaume.chevillon@sciences-po.fr

Majeure Economie
HEC 2005

Table des matières

1 Variables aléatoires et limites **7**

1.1 Qu'est-ce que l'économétrie ? 7

1.2 Notions de probabilités 9

1.2.1 Espaces et axiomes 9

1.2.2 Indépendance 11

1.2.3 Probabilité conditionnelle 11

1.3 Variables aléatoires 13

1.3.1 Fonction de distribution 14

1.3.2 Distribution Normale 15

1.3.3 Autres distributions 16

1.3.4 Distributions multivariées 19

1.3.5 Moments 20

1.3.6 Estimateurs 22

1.4 Approximations asymptotiques 24

1.4.1 Motivations 24

1.4.2 Définitions 25

1.4.3 Autres mesures de convergence 27

1.4.4 Notation de l'ordre 29

2 Inférence **31**

2.1 Motivations 31

2.2 Choix du modèle 33

2.3 Stratégies de test 33

2.3.1 Erreurs de test 34

2.3.2 Fonction de puissance 35

2.3.3 Tests unilatéraux 38

2.4 Test de Student 38

2.4.1 Les autres tests de restriction 39

3 Régression **41**
3.1 Introduction . . . 41
3.1.1 La régression linéaire et ses problèmes potentiels . . . 41
3.1.2 Notation vectorielle et matricielle . . . 43
3.2 Régression . . . 44
3.2.1 Maximum de vraisemblance . . . 44
3.2.2 Moindres carrés (Least squares) . . . 50
3.2.3 Erreurs de spécification . . . 53
3.2.4 Choix du modèle . . . 55

4 Séries temporelles **57**
4.1 Introduction . . . 57
4.1.1 Qu'appelle-t-on série temporelle ? . . . 57
4.1.2 Quels sont les buts de cette analyse ? . . . 59
4.1.3 En quoi cette démarche consiste-t-elle ? . . . 62
4.2 Concepts des séries temporelles . . . 63
4.2.1 Processus stochastiques . . . 63
4.2.2 Stationnarité . . . 66
4.2.3 Ergodicité . . . 67
4.3 La caractérisation des séries temporelles en économie . . . 68
4.3.1 Moyenne de l'échantillon . . . 68
4.3.2 **ACF**, fonction empirique d'autocorrélation . . . 68
4.3.3 **PACF**, fonction empirique d'autocorrélation partielle . . . 69
4.4 Processus intégrés . . . 69
4.5 Quelques processus courants . . . 73

5 Méthodes sans modèle **75**
5.1 Extrapolation déterministe des séries . . . 75
5.1.1 Tendances linéaires . . . 75
5.1.2 Tendances autorégressives . . . 77
5.1.3 Modèles non linéaires . . . 77
5.2 Moyennes mobiles . . . 77
5.3 Lissages . . . 78
5.3.1 Moyennes mobiles . . . 79
5.3.2 Lissage exponentiel . . . 79
5.4 Ajustements saisonniers . . . 80
5.4.1 Méthode multiplicative . . . 80
5.4.2 Méthode additive . . . 81

6 Modèles linéaires de séries temporelles **83**
6.1 Processus linéaires 83
6.1.1 Concepts 83
6.1.2 Théorème de décomposition de Wold 84
6.1.3 Modélisation ARMA 85
6.2 Prédiction des processus ARMA(p, q) 89
6.3 Algorithme de Box-Jenkins 91
6.3.1 Principe de la méthode 91
6.3.2 Travailler sur données stationnaires 91
6.3.3 Etablir une hypothèse 92
6.3.4 Estimation 93
6.3.5 Diagnostic 93
6.4 Estimation des modèles dynamiques 94
6.4.1 Equations de Yule-Walker 94
6.4.2 Fonction de vraisemblance 94
6.4.3 Maximum de vraisemblance d'un ARMA 95

7 Les variables intégrées **99**
7.1 Les tests de racine unitaire 99
7.1.1 Problèmes des processus intégrés 100
7.1.2 Test de Dickey-Fuller 100
7.2 Les différents tests 101
7.3 Les tendances et constantes 101
7.4 Modèles univariés de cointégration 104
7.4.1 Procédure en deux étapes d'Engle et Granger 104
7.4.2 Procédure en une étape de Banerjee, Dolado et Mestre . 106
7.4.3 Références bibliographiques 108
Annexe 7.A Décomposition du MCE 109
Annexe 7.B Neutralité et Homogénéité 109

8 Processus autorégressifs vectoriels **113**
8.1 Processus autorégressifs vectoriels stables 113
8.2 Processus vectoriels et cointégration 115

9 Exercices corrigés **117**

Avant-Propos

Ce cours vise à fournir une introduction à la pratique contemporaine de l'économétrie, principalement en macro. Il ne se propose pas d'être exhaustif mais tâche de conduire le lecteur à travers les étapes principales de la compréhension, en alternant des parties plus mathématiques et d'autres plus explicatives, dans le but de lui fournir les outils de base lui permettant de comprendre les modèles économiques de l'économie appliquée. Ce cours ne saurait être compris sans application à des cas empiriques via l'utilisation de logiciels.

Je renvoie les lecteurs souhaitant davantage d'informations aux ouvrages suivants :

Introductif :

Gujarati, D. N. (1995) *Basic Econometrics*, 3rd ed. MacGraw Hill.

Pindyck, R. S. & D. L. Rubinfeld (1998) *Econometric Models and Economic Forecasts*, 4th ed. McGraw Hill.

Cours :

Greene, W. H. (1993) *Econometric Analysis*, 3rd ed. Prentice Hall.

Hendry & Doornik (2001) *Empirical Econometric Modelling using PcGive 10* : Volume 1, chaps 11-15. London : Timberlanke Consultants Press, 2001.

Johnston, J. & J. DiNardo (1997) *Econometric Methods*, 4th Edition. MacGraw Hill

Lardic, S. & V. Mignon (2002) *Econométrie des séries temporelles macroéconomiques et financières*. Paris : Economica.

Ruud, P. A. (2000) *An introduction to Classical Econometric Theory*. Oxford University Press.

Séries temporelles :

Gourieroux, C et A. Monfort. (1995). *Séries Temporelles et Modèles Dynamiques* (2ème éd.). Paris : Economica.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton : Princeton University Press.

Harvey, A. C. (1993). *Time Series Models* (2nd ed.). Hemel Hempstead : Harvester Wheatsheaf.

Macroéconométrie :

Hendry, D. F. (1995) *Dynamic Econometrics*. Oxford University Press.

Chapitre 1

Variables aléatoires et limites

1.1 Qu'est-ce que l'économétrie ?

La définition du terme économétrie a évolué depuis l'émergence de cette discipline dans les années 1930. A l'origine, elle représentait une voie de formalisation de l'économie par l'usage de mathématiques, probabilités et statistiques. La formalisation présente des avantages et des inconvénients : elle permet d'établir des arguments précis et rapidement compréhensibles grâce à une absence d'ambiguïté. En revanche, elle fait aussi apparaître le domaine plus abstrait et accroît les barrières à l'entrée pour les néophytes. Par ailleurs, elle peut entraîner la théorie dans des directions où des théorèmes peuvent être établis, et ainsi éviter des problèmes économiques importants mais dont la formalisation se révèle plus ardue.

Dans ce sens traditionnel, la quasi-intégralité de la microéconomie et l'essentiel de la macro enseignées appartiennent à l'"économétrie". La revue la plus associée à ce courant est Econometrica, fondée dans les années 30. Une des plus prestigieuses, elle publie essentiellement ce qu'on appelle dorénavant l'économie théorique et la théorie économétrique.

Dans les années 1960, la définition traditionnelle de l'économétrie s'est révélée désuette car la plupart des domaines de l'économie avaient été gagnées par l'approche économétrique, bien que subsistent des débats sur le dégré de formalisation de l'analyse. Une rédéfinition du terme s'ensuivit, et le nouveau sens est plus proche de l'utilisation du suffixe métrie rencontré dans d'autres sciences, comme la biométrie.

L'économétrie moderne concerne le développement de méthodes probabilistes et statistiques dans le contexte d'une compréhension détaillée des données, et des théories économiques, les concernant afin d'obtenir une analyse économique empirique rigoureuse. Elle se situe à l'interface entre l'informa-

tique, les statistiques, les probabilités et la théorie économique. Elle est donc très influencée pas des développements hors du domaine propre de la pensée économique, en particulier informatiques et probabilistes. Divers chercheurs font davantage porter l'accent sur l'un ou l'autre de ces ingrédients, ce qui génère une grande part des conflits dans ce domaine. Les termes importants de la définition ci-dessus sont empiriques et rigoureuses : il s'agit bien d'une discipline qui vise à être appliquée à des problèmes concrets et ainsi ignore une grande part des développements théoriques purs sans possibilité d'application ; par ailleurs, l'économétrie a dans une certaine mesure vocation à rapprocher l'économie des sciences expérimentales : il s'agit de tirer des événements passés et des données le maximum d'informations afin d'utiliser les "expériences historiques", à défaut de pouvoir les réproduire ex abstractum.

L'économétrie appliquée utilise, quant à elle, les dévéloppements théoriques pour analyser des cas concrets afin d'obtenir des recommandations politiques, de tester la théorie économique ou de suggérer de nouvelles manières d'améliorer cette dernière. Au vu de la rapidité des développements, il est essentiel pour toute personne qui s'intéresse aux études économiques de pouvoir en comprendre les forces et faiblesses car *des méthodes économétriques appliquées à mauvais escient entraînent souvent des résultats sans fondements.*

Au cours de la dernière décennie, divers économètres ont reçu le prix Nobel d'économie :

En microéconométrie, Dan McFadden a développé des méthodes permettant d'analyser de manière formelle comment les individus prennent des décisions économiques, par exemple comment les habitants de San Francisco choisissent entre divers modes de transport ; et ce, afin de prévoir l'impact de l'introduction de nouvelles formes de transport et donc de savoir si elle se révélerait bénéfique pour le bien-être global et de quantifier cet aspect. Jim Heckman a pour sa part étudié comment analyser les facteurs influençant les choix individuels en matière de quantité de travail.

En économétrie financière, Rob Engle a analysé la modification de la volatilité au cours du temps. Ceci lui a permis d'étudier comment des agents rationnels devraient equilibrer risque (volatilité) et rentabilité au cours du temps. Une autre application concerne l'évaluation du risque associé à un investissement bancaire (riskmetrics).

En macroéconométrie, Clive Granger a permis une modélisation temporelle de variables instables, comme le revenu et la consommation, qui sont liées par des relations économiques de long terme.

La compréhension de l'économétrie nécessite par conséquent une maîtrise des outils de probabilités et de statistiques.

1.2 Notions de probabilités

1.2.1 Espaces et axiomes

La théorie probabiliste est construite autour de la notion d'ensemble. Les principales définitions concernent les événements possible et leur probabilité.

Définition 1 (Espace d'échantillonnage) *L'ensemble* Ω *est nommé espace d'échantillonnage s'il contient toutes les réalisations possibles considérées, par exemple si un consommateur a la possibilité d'acheter 0,1,2 ou 3 bouteilles de Coca Light :* $\Omega = \{0, 1, 2, 3\}$.

Définition 2 (Evénement) *Un événement est un sous-ensemble de* Ω *(ou* Ω *lui-même) : exemple l'événement* $A = \{0\}$ *correspond à un consommateur n'achetant pas de Coca Light,* $B = \{1, 2\}$ *s'il achète une ou deux bouteilles.*

Les principales notations concernant les ensembles sont :

Union. $A \cup B$, ex. $A \cup B = \{0, 1, 2\}$.
Intersection. $A \cap B$, ex. $A \cap B = \varnothing$.
Complémentarité. A^c ou $\bar{A} = \Omega \backslash A$, ex. $A^c = \{1, 2, 3\}$

La théorie probabiliste est construite autour des développements de la théorie des ensembles. En particulier, on appelle *espace probabilisable* le couple $(\Omega, \mathcal{F})$ où $\mathcal{F}$ est une tribu (sigma algèbre) associée à Ω, il s'agit de l'ensemble des combinaisons d'événements possibles[1]. Pour un espace d'échantillonnage Ω, une fonction de probabilité Pr est une fonction définie sur une tribu associée $\mathcal{F}$ qui satisfait trois axiomes :

1. $\Pr(\omega) \geq 0$ pour tous $\omega \in \mathcal{F}$.
2. $\Pr(\Omega) = 1$.
3. Si les $\{A_i \in \mathcal{F}\}$ sont disjoints alors

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

[1]Une collection de sous-ensembles de S est appelée tribu $\mathcal{F}$ si elle satisfait à trois conditions :

1. $\varnothing \in \mathcal{F}$
2. Si $A \in \mathcal{F}$ alors $A^c \in \mathcal{F}$
3. Si $A_1, A_2, ... \in \mathcal{F}$ alors $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$

i.e. l'ensemble vide est un membre de la tribu, le complémentaire de tout membre est un membre, toute union de membres de la tribu appartient à la tribu.

Remarque 1 – *Les probabilités sont des fonctions s'appliquant à des ensembles*
- *La tribu représente tous les sous-ensembles de* Ω *et fournit ainsi la base sur laquelle les événements valides peuvent être définis.*
- *Tous les événements ont probabilités positives ou nulles.*
- *Au moins un événement de l'espace d'échantillonnage arrive. L'espace d'échantillonnage est l'*Univers *considéré.*
- *Si les événements sont disjoints (pas de superposition) alors la probabilité qu'un d'entre eux (et un seul) arrive est la somme des probabilités que chacun survienne. Par exemple pour deux événements disjoints* $\{A_{1,}A_2\}$, *alors*

$$\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2).$$

Exemple 1 *Divisons* Ω *en deux événements* $\{A, A^c\}$. *Alors* $\Omega = \{A\} \cup \{A^c\}$ *et ces événements sont nécessairement disjoints. Par conséquent*

$$\Pr(A \cup A^c) = \Pr(A) + \Pr(A^c) = \Pr(\Omega) = 1$$

donc

$$\Pr(A^c) = 1 - \Pr(A) = \Pr(\Omega \backslash A).$$

Puisque $\Pr(A^c) \geq 0$, *on en déduit que*

$$\Pr(A) \leq 1.$$

Enfin, les propriétés de la tribu nous permettent d'écrire $\Omega = \Omega \cup \varnothing$ *et donc*

$$\Pr(\varnothing) = 0.$$

Exemple 2 *Soient deux événements* A *et* B *appartenant à* $\mathcal{F}$. *Ces événements ne sont pas nécessairement disjoints et nous ne pouvons utiliser le troisième axiome directement. Cependant*

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Ainsi la probabilité qu'au moins un de A ou de B se produise est la probabilité qu'A arrive plus celle de B moins la probabilité que les deux se produisent.

1.2.2 Indépendance

Considérons deux événements A et B appartenant à $\mathcal{F}$. On s'intéresse ici au concept selon lequel la réalisation d'un événement ne modifie pas la probabilité qu'à un autre événement de se réaliser. Quand ceci est vrai, on parle d'indépendance. Mathématiquement, on note que A et B sont indépendants (dans $\mathcal{F}$) si et seulement si

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B).$$

Noter que deux événements ne peuvent être indépendants s'ils sont disjoints, car alors $\Pr(A \cap B) = 0$. On note parfois l'indépendance entre deux événements :

$$A \perp\!\!\!\perp B.$$

Exemple 3 *Soit A le rendement (géométrique) d'un actif sur un jour donné et B son rendement pour le jour suivant. Beaucoup de modèles en économie financière font l'hypothèse d'indépendance de A et de B. Ceci est pourtant rejeté empiriquement car si A et B sont typiquement presque non-corrélés, il ne sont pas indépendants. Une forte volatilité à tendance à suivre une forte volatilité.*

Exemple 4 *Par définition si on jette un dé deux fois successives, le résultat du premier jet n'influence pas celui du second, et ainsi les deux résultats sont indépendants.*

1.2.3 Probabilité conditionnelle

Il est parfois souhaitable de changer d'espace d'échantillonnage, d'univers pour calculer les probabilités. On peu soit redéfinir Ω à chque fois, par exemple en calculant séparément avec un Ω pour les employés masculins de plus de 45 ans et un pour les employées féminines de moins de 21 ans... Afin d'éviter toute confusion, on utiliser le concept de probabilité conditionnelle, qui vise le même but, mais conserve Ω constant. Si on conditionne sur B, les axiomes de probabilité demeurent les mêmes :

$$\begin{aligned}
\Pr(\omega \in A | \omega \in B) &\geq 0 \\
\Pr(\omega \in B | \omega \in B) &= 1 \\
\Pr\left(\omega \in \bigcup_{i=1}^{\infty} A_i | \omega \in B\right) &= \sum_{i=1}^{\infty} \Pr(\omega \in A_i | \omega \in B)
\end{aligned}$$

si les A_i sont disjoints.

Exemple 5 *Si Ω représente l'ensemble des niveau de salaire de population résidant en France. On peut par exemple s'intéresser à*

$$\Pr(salaire|employé)$$
$$\Pr(salaire|femme)$$
$$\Pr(salaire|femme, employé, français)$$

Remarquer que les événements sont indépendants au sein de Ω. Il s'agit de définir au préalable Ω et toutes les propositions en dépendent.

Il est souvent plus utile de définir la probabilité conditionnelle à l'aide de la distribution conjointe : si nous connaissons la distribution conjointe des salaires et de l'emploi il serait agréable de pouvoir déduire de manière automatique des informations concernant la probabilité conditionnelle des salaires, connaissant le statut d'emploi. On procède de la manière suivante.

Soit un univers comportant deux réalisations A et B. On peut s'intéresser soit à $\Pr(A)$, soit à $\Pr(B)$, soit encore à $\Pr(A \cap B)$. Enfin on peut aussi s'intéresser à $\Pr(A|B)$ pourvu que $\Pr(B) > 0$ (i.e. que B ait une chance de se réaliser). Il s'agit ainsi de contraindre notre monde de sorte que B se réalise et de se demander ce qui se passe alors pour A. Ceci ne peut arriver que si A et B peuvent arriver coinjointement, on définit alors :

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Il est facile de constater que cette définition satisfait aux axiomes de probabilité. La probabilité conditionnelle est un concept vital en économétrie. On peut alors s'intéresser à la probabilité conditionnelle coinjointe de deux événements :

$$\Pr(A \cap B|C).$$

Si $\Pr(A \cap B|C) = \Pr(A|C) \times \Pr(B|C)$, on dit alors que conditionnellement à B, A et C sont indépendents.

Exemple 6 *Supposons que A soit la richesse accumulée au cours de la vie ; B le fait d'étudier à HEC ; C les capacités, l'éducation, la formation. L'indépendance entre A et B sachant C revient à dire que le label HEC n'apporterait aucune valeur pour aucun individu, ce qui est une déclaration plus forte qu'une absence de valeur ajoutée en moyenne.*

En réarrangeant les formules précédentes

$$\Pr(B)\Pr(A|B) = \Pr(A \cap B),$$

et

$$\Pr(A)\Pr(B|A) = \Pr(A \cap B),$$

en réarrangeant on obtient un des théorèmes les plus connus en probabilités, le

Théorème 2 (Théorème de Bayes) *si* $\Pr(A) > 0$ *et* $\Pr(B) > 0$,

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(B)\Pr(A|B)}{\Pr(A)}.$$

Celui-ci revient à dire que pour passer de $\Pr(A|B)$ *à* $\Pr(B|A)$, *il suffit de multiplier par le ratio* $\Pr(B) / \Pr(A)$.

1.3 Variables aléatoires

Nous avons pour l'instant utilsé ω pour représenter chacun des événements associés au triplet $(\Omega, \mathcal{F}, \Pr)$, i.e. $\mathcal{F}$ est généré à partir de Ω, $\omega \in \mathcal{F}$, et Pr est la fonction qui associe une probabilité.

Ces événements ne sont pas nécessairement numériques. En particulier pour diverses applications on peut choisir de s'intéresser à de mulitples facettes d'un même événement et ainsi utiliser des fonctions numériques de cet événement. Si on choisit une fonction $X(\omega)$ qui mène à une valeur numérique (potentiellement un vecteur ou une matrice), on appelle X variable aléatoire. Les distributions sont des familles spécifiques de variables aléatoires.

Exemple 7
- *Soit* Ω *l'univers des nouveau-nés en France en 2005. Pour chaque* ω_i *individuel, i.e. chaque naissance, on peut choisir de s'intéresser à des fonctions diverses : la taille, le poids du nouveau-né, la durée de la grossesse, qui sont des fonctions réelles ; le nombre de frères et soeurs de l'enfant, qui est une fonction entière ; ou une fonction indicatrice qui prend la valeur 1 si l'enfant a des cheveux et 0 sinon.*
- ***Distribution de Bernoulli****. Une personne est employée un non ; on note employé* $\omega = E$, *sans emploi* $\omega = U$. *Soit* $X(\omega = E) = 1$ *et* $X(\omega = U) = 0$ *la variable indicatrice qui renvoie 1 si un individu possède un emploi et 0 sinon. On note* $\Pr(X = 1) = p$ *et* $\Pr(X = 0) = 1 - p$. *La distribution de Bernoulli joue un rôle important en microéconométrie quand des variables prennent les valeurs 0 et 1.*

– ***Distribution Binômiale.*** *Si on réalise n tirages indépendants de la distribution de Bernoulli (par exemple en choisissant n personnes au hasard au sein de la population française et en leur demandant s'ils ont un emploi, 1 si vrai, 0 si faux), et on note le nombre de fois que la réponse est 1. Alors le total est appelé distribution binômiale : soit p la proportion de la population ayant un emploi*

$$Y = \sum_{i=1}^{n} X_i, \quad \Pr(X_i = 1) = p, \quad \Pr(X_i = 0) = 1 - p, \quad X_i \sim iid$$

Alors

$$\Pr(Y = y) = \frac{n!}{y!\,(n-y)!} p^y (1-p)^{n-y}, \quad y = 0, 1, ..., n.$$

1.3.1 Fonction de distribution

La fonction de distribution d'une variable aléatoire $X : \Omega \to \mathbb{R}$ (ou $\mathbb{N}$) est définie par

$$\begin{aligned} F_X &: \mathbb{R}(\text{ou } \mathbb{N}) \to [0,1] \\ &: x \to \Pr(X \leq x). \end{aligned}$$

où X est ici évaluée sur Ω, i.e. pour l'ensemble des événements. La densité de X est (pour les variables aléatoires continues)

$$f_X(x) = \frac{\partial F_X}{\partial x}.$$

On note que pour les fonctions continues

$$\Pr(X = x) = 0$$

pour tout x et qu'on peut noter en revanche

$$\Pr(X \in [x, x + dx]) = f_X(x)\, dx$$

et que si X prend un nombre fini de valeurs

$$f_X(x) = \Pr(X = x).$$

Le lien entre les distribution et densité est donc fourni par

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du.$$

On note que pour les variables aléatoires réelles (définies sur $\mathbb{R}$)

$$F_X(x) \underset{x\to+\infty}{\to} 1, \quad F_X(x) \underset{x\to-\infty}{\to} 0,$$
$$f_X(x) \underset{x\to\pm\infty}{\to} 0.$$

Les **quantiles** d'une distribution sont fournies par la fonction inverse de F_X. Ainsi si on souhaite savoir quelle est la valeur x telle que pour une proportion p de la population X prend une valeur inférieure ou égale à x,

$$p = F_X(x)$$

et donc

$$x = F_X^{-1}(p)$$

est appelée fonction quantile de X. On appelle médiane le quantile 0,5 (50% de la population de parts et d'autres de cette valeur).

Exemple 8 *Les quantiles sont des mesures essentielles de l'inégalité. Ainsi des politiques économiques peuvent par exemple cibler les 10% de la population ayant les revenus les plus faibles.*

Exemple 9 *Une variable aléatoire exponentielle possède la distribution*

$$f_X(x) = \frac{1}{\beta}\exp(-x/\beta), \quad x, \beta \in \mathbb{R}_+.$$

Sa fonction de distribution est

$$F_X = 1 - \exp(-x/\beta),$$

et sa fonction quantile

$$F_X^{-1}(u) = -\beta\log(1-u).$$

1.3.2 Distribution Normale

La distribution Normale ou Gaussienne est la plus couramment utilisée. Elle apparaît naturellement lorsque on s'intéresse à la distribution de la moyenne et présente des propriétés pratiques de translation. Sa forme ne paraît malheureusement pas immédiatement attractive : sa densité est

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad x, \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}_+.$$

Les distributions étant des familles de variable aléatoire, on constate ici que chaque X Gaussienne est paramétrée par sa moyenne μ et sa variance σ^2, ce qu'on note

$$X \sim \mathsf{N}\left(\mu, \sigma^2\right).$$

Mathématiquement, on peut penser à la densité f_X de la manière suivante :

$$\log f_X(x) = c - \frac{1}{2\sigma^2}(x-\mu)^2.$$

La log-densité est quadratique en x, la constante c est determinée de sorte que

$$\Pr(\Omega) = 1 = \int_{-\infty}^{+\infty} f_X(x)\, dx.$$

La densité Normale a $\mathbb{R}$ comme support est centrée autour de μ, σ contrôle sa dispersion. Une propriété importante de la distributon Normale est que si $X \sim \mathsf{N}(\mu, \sigma^2)$ alors

$$\gamma + \lambda X \sim \mathsf{N}\left(\gamma + \lambda\mu, \lambda^2\sigma^2\right),$$

i.e. les transformations affines d'une Normale sont Normales. Ceci entraîne qu'on puisse écrire toute distribution Gaussienne comme

$$X \overset{loi}{=} \mu + \sigma N,$$

où N suit une Normale standard $N \sim \mathsf{N}(0,1)$. Ainsi si X et Y suivent deux Normales indépendantes

$$X + Y \sim \mathsf{N}\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right).$$

1.3.3 Autres distributions

Il existe une multitude de distributions parmi lesquelles on en rencontre fréquemment certaines en économétrie.

Khi-deux

Supposons que $X_i \overset{iid}{\sim} \mathsf{N}(0,1)$, (souvent écrit $\mathsf{NID}(0,1)$ ou $\mathsf{IN}(0,1)$, ce qui signifie que les X_i sont des copies indépendantes et identiquement Normalement distribuées),alors

$$Y = \sum_{i=1}^{v} X_i^2 \sim \chi_\nu^2,$$

une distribution khi-deux avec ν "degrés de liberté". L'espérance et la variance d'une distribution χ_ν^2 sont respectivement ν et 2ν.

Uniforme

On contraint parfois les variables sur de petits intervalles. L'exemple le plus simple est la distribution uniforme standard

$$f_X(x) = 1, \quad x \in [0,1].$$

Cette variable n'a que le segment $[0,1]$ comme support. Cette distribution est souvent utilisée dans les modèles stylisés afin d'introduire une idée ou un concept. Elle est aussi utilisée en simulation. Une Uniforme plus générale se définit comme

$$f_X(x) = \frac{1}{b-a}, \quad x \in [a,b].$$

Poisson

Les modèles de comptage sont souvent utilisés en économie, par exemple le nombre de brevets déposés en un intervalle de temps, le nombre d'échanges sur un marché... La distribution la plus courante est celle de Poisson :

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0,1,2,...$$

Student

Si Z suit une $\mathsf{N}(0,1)$ et X une χ^2_ν et est indépendante de Z, alors le ratio

$$t_\nu = \frac{Z}{\sqrt{X/\nu}}$$

suit une distribution dite de Student avec ν degrés de liberté. Celle-ci a la même forme qu'une distribution Normale, mais avec des bords plus épais. Quand ν augmente, t_ν se comporte de plus en plus comme une Normale, entre $\nu = 30$ et 100 une Normale standard est une bonne approximation, au delà de 100 on ne peut les distinguer.

Fischer

Si X_1 et X_2 sont deux distributions khi-deux indépendantes avec pour degrés de liberté ν_1 et ν_2, alors le ratio

$$F_{\nu_1,\nu_2} = \frac{X_1/\nu_1}{X_2/\nu_2}$$

suit une loi de Fischer avec ν_1 et ν_2 degrés de liberté. Quand ν_2 est nettement plus grand que ν_1, comme est courant en économétrie, X_2/ν_2 tend vers 1 et F_{ν_1,ν_2} se comporte comme un $\chi^2_{\nu_1}$ divisé par ν_1.

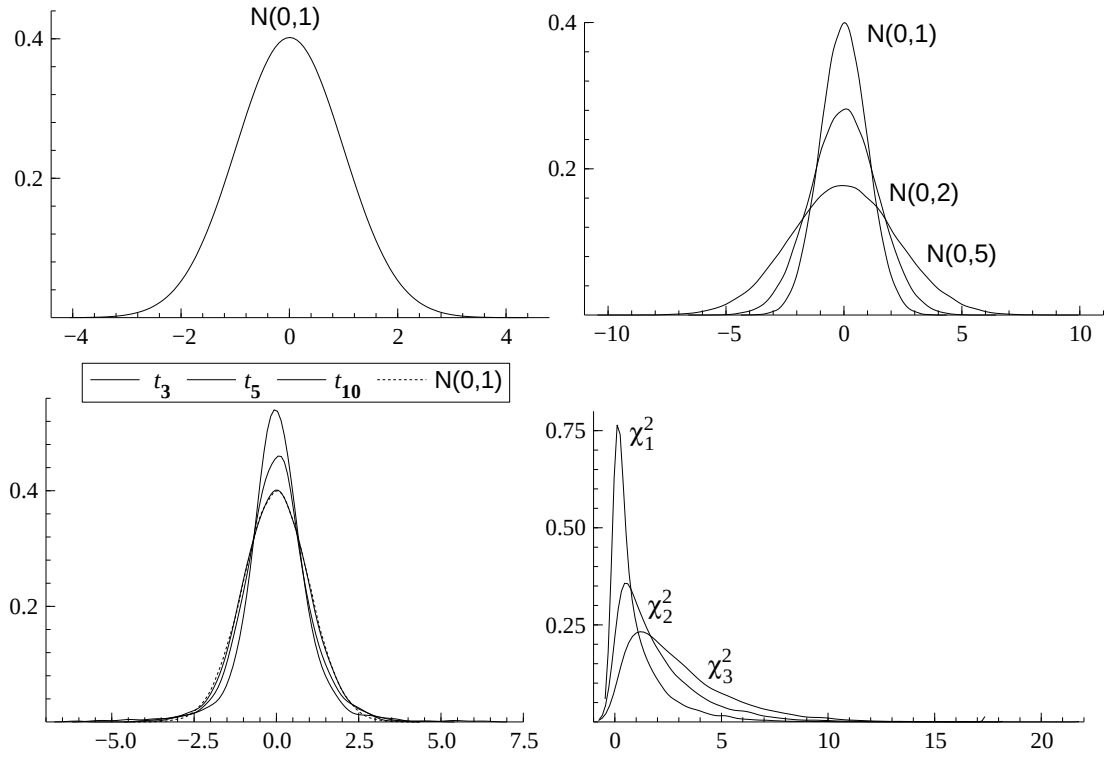


Figure 1.1 – Graphiques des densités des distributions Standard Normale, Normales, de Student et Khi-deux.

1.3.4 Distributions multivariées

Tous les résultats précédents sont aussi valables lorsqu'on s'intéresse au vecteur multivarié de dimension p :

$$X = (X_1, ..., X_p)'.$$

Les éléments de ce vecteur ne sont pas nécessairement indépendants, ils peuvent par exemple représenter une série chronologique ou un panel d'observations économiques. En particulier, si $p = 2$ de sorte que

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

alors

$$F_X(x_1, x_2) = \Pr(X_1 \leq x_1, X_2 \leq x_2)$$

qui, dans le cas continu s'écrit

$$F_X(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_X(u_1, u_2)\, du_1 du_2.$$

Et de manière similaire

$$f_X(x_1, x_2) = \frac{\partial^2 F_X}{\partial x_1 \partial x_2}.$$

Quand X_1 et X_2 sont indépendantes, la densité s'écrit

$$f_X(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2).$$

Dans le cas général en intégrant par rapport à une variable sur son support, on obtient la densité marginale de la seconde :

$$f_{X_2}(x_2) = \int_{-\infty}^{+\infty} f_X(u_1, x_2)\, du_1.$$

La distribution conditionnelle prend la forme

$$F_{X_1|X_2=x_2}(x_1) = \Pr(X_1 \leq x_1 | X_2 = x_2)$$

ce qui donne la densité conditionnelle

$$f_{X_1|X_2=x_2}(x_1) = \frac{\partial \Pr(X_1 \leq x_1 | X_2 = x_2)}{\partial x_1}$$

qui possède toutes les propriétés d'une densité, en particulier on peut montrer que

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_X(x_1, x_2)}{f_{X_2}(x_2)}.$$

1.3.5 Moments

Soit X une variable aléatoire, on définit de manière générale les moments de X comme l'intégrale (lorsqu'elle existe)

$$\mathsf{E}\left[g\left(X\right)\right]=\int g\left(x\right)f_{X}\left(x\right)dx$$

avec des cas spécifiques de fonction polynômiales $g\left(\cdot\right)$. $\mathsf{E}\left[\cdot\right]$ est appelé espérance et est un opérateur linéaire : pour a et b constantes

$$\mathsf{E}\left[a+bg\left(X\right)\right]=a+b\mathsf{E}\left[g\left(X\right)\right].$$

Espérance mathématique

Le cas le plus simple d'espérance est la moyenne ou moment de premier ordre définie par

$$\mu\left(X\right)=\mathsf{E}\left[X\right]=\int xf_{X}\left(x\right)dx.$$

et de manière plus générale on définit le r-ième moment (non centré)

$$\mu_{r}\left(X\right)=\mathsf{E}\left[X^{r}\right]=\int x^{r}f_{X}\left(x\right)dx.$$

Variance

Dans le cas d'une variable univariée, la variance est définie comme second moment de la variable centrée $X-\mathsf{E}\left[X\right]$:

$$\begin{aligned}\mathsf{V}\left[X\right] &= \mathsf{E}\left[\left(X-\mathsf{E}\left[X\right]\right)^{2}\right]\\ &= \int\left(x-\mathsf{E}\left[X\right]\right)^{2}f_{X}\left(x\right)dx\\ &= \mathsf{E}\left[X^{2}\right]-\left(\mathsf{E}\left[X\right]\right)^{2}.\end{aligned}$$

La variance est égale à l'espérance du carré, moins le carré de l'espérance.

Exercice 1 *Prouver que* $\mathsf{V}\left[a+bX\right]=b^{2}\mathsf{V}\left[X\right]$.

Exercice 2 *Montrer que l'espérance et la variance de la distribution Normale :*

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

sont μ *et* σ^{2} *respectivement.*

Exercice 3 *Quelles sont les espérance et variance d'une distribution uniforme standard ?*

Covariance

La covariance de X et Y est définie, lorsqu'elle existe comme

$$\begin{aligned}\mathsf{Cov}[X,Y] &= \mathsf{E}[(X-\mathsf{E}[X])(Y-\mathsf{E}[Y])] \\ &= \int\int (x-\mathsf{E}[X])(y-\mathsf{E}[Y]) f_{X,Y}(x,y)\,dxdy \\ &= \mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y].\end{aligned}$$

Exercice 4 *Prouver que* $\mathsf{Cov}[a+bX, c+dY] = bd\mathsf{Cov}[X,Y]$, *i.e. que la covariance est invariante par translation.*

Exercice 5 *Montrer que* $\mathsf{V}[aX+bY] = a^2\mathsf{V}[X] + 2ab\mathsf{Cov}[X,Y] + b^2\mathsf{V}[Y]$ *comme une identité remarquable.*

Exercice 6 *Montrer que si les* $X_i, ...X_n$ *sont indépendantes alors*

$$\mathsf{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{V}[X_i]$$

et que par conséquent la variance de la moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *des variables aléatoire a pour variance* $\frac{1}{n}\mathsf{V}[X_1]$ *si les* X_i *sont identiquement distribuées.*

L'indépendance implique la non-corrélation lorsque la covariance existe. En effet si X et Y sont indépendantes alors $\mathsf{E}[XY] = \mathsf{E}[X]\mathsf{E}[Y]$ et donc $\mathsf{Cov}[X,Y] = 0$. La réciproque n'est vrai que si X et Y sont Gaussiennes.

Exemple 10 *On suppose que* $X \sim \mathsf{N}(0,1)$, $Y = X^2$ *suit alors une distribution* χ^2_1. *Et*

$$\mathsf{Cov}[X,Y] = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y] = \mathsf{E}\left[X^3\right].$$

or X *étant symmétrique autour de zéro,* X^3 *l'est aussi et* $\mathsf{E}[X^3] = 0$. X *et* Y *ne sont donc pas corrélées mais non-indépendantes.*

Corrélation

On définit la corrélation comme

$$\mathsf{Cor}[X,Y] = \frac{\mathsf{Cov}[X,Y]}{\sqrt{\mathsf{V}[X]\mathsf{V}[Y]}}.$$

L'inégalité de Cauchy-Schwarz implique que

$$\mathsf{Cor}[X,Y] \in [-1,1].$$

Plus la corrélation est proche de ± 1, plus les variables sont liées.

Exercice 7 *Prouver que*

$$\mathsf{Cor}\left[a+bX,c+dY\right]=\mathsf{Cor}\left[X,Y\right].$$

Matrice de Covariance

Si $\mathbf{X}$ est multivariée, alors

$$\mathsf{Cov}\left[\mathbf{X}\right]=\mathsf{E}\left[\left(\mathbf{X}-\mathsf{E}\left[\mathbf{X}\right]\right)\left(\mathbf{X}-\mathsf{E}\left[\mathbf{X}\right]\right)'\right].$$

Cette matrice est symétrique, i.e. $\mathsf{Cov}\left[\mathbf{X}\right]=\mathsf{Cov}\left[\mathbf{X}\right]'$, et définie, i.e. pour tous vecteurs $\mathbf{u}$: $\mathbf{u}\mathsf{Cov}\left[\mathbf{X}\right]\mathbf{u}'\geq 0$. La matrice de covariance de $\mathbf{X}=(X_1,...,X_n)'$ s'écrit

$$\mathsf{Cov}\left[\mathbf{X}\right]=\begin{bmatrix}\mathsf{V}\left[X_1\right] & \mathsf{Cov}\left[X_1,X_2\right] & \cdots & \mathsf{Cov}\left[X_1,X_n\right]\\ \mathsf{Cov}\left[X_2,X_1\right] & \mathsf{V}\left[X_2\right] & \cdots & \mathsf{Cov}\left[X_2,X_n\right]\\ \vdots & \vdots & \ddots & \vdots\\ \mathsf{Cov}\left[X_n,X_1\right] & \mathsf{Cov}\left[X_n,X_2\right] & \cdots & \mathsf{V}\left[X_n\right]\end{bmatrix}.$$

Un résultat important : si $\mathbf{B}$ est une matrice de constantes, $\mathbf{a}$ un vecteur, alors

$$\begin{aligned}\mathsf{E}\left[\mathbf{a}+\mathbf{BX}\right] &= \mathbf{a}+\mathbf{B}\mathsf{E}\left[\mathbf{X}\right]\\ \mathsf{Cov}\left[\mathbf{a}+\mathbf{BX}\right] &= \mathbf{B}\mathsf{Cov}\left[\mathbf{X}\right]\mathbf{B}'.\end{aligned}$$

La matrice de corrélation est définie de manière similaire avec des 1 sur la diagonale principale et les corrélations de parts et d'autres.

Distribution Normale multivariée

La variable aléatoire $\mathbf{X}$ de dimension n est dite suivre une distribution Normale multivariée d'espérance $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$ (symétrique et positive définie, i.e.pour tous $\mathbf{z}\neq\mathbf{0}$, $\mathbf{z}\boldsymbol{\Sigma}\mathbf{z}'>\mathbf{0}$) si

$$f_{\mathbf{X}}\left(\mathbf{x}\right)=\left|2\pi\boldsymbol{\Sigma}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)'\right],\quad \mathbf{x}\in\mathbb{R}^n.$$

Si $\mathbf{a}$ est $q\times 1$ et $\mathbf{B}$ est $q\times n$ alors

$$\mathbf{Y}=\mathbf{a}+\mathbf{BX}\sim\mathsf{N}\left(\mathbf{a}+\mathbf{B}\boldsymbol{\mu},\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\right).$$

1.3.6 Estimateurs

Une statistique $S\left(X\right)$ est une fonction d'une variable aléatoire (vectorielle) X. Quand on utilise cette statistique pour apprendre des propriétés du modèle probabiliste, on dit qu'on *estime* le modèle. La version aléatoire de cette fonction $S\left(X\right)$ est appelée **estimateur**, dans le cas d'un vecteur observé

(une réalisation, ou un échantillon de données) on parle d'**estimation** $S(x)$. L'usage de X et de x est le même que précédemment, X est une variable aléatoire qui possède une certaine distribution, x est une valeur qui provient d'un tirage de X ou d'une réalisation d'un événement.

Exemple 11 *L'example le plus simple est la moyenne arithmétique de variables aléatoires*

$$S(X) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Si les X_i *sont* $\mathsf{NID}(\mu, \sigma^2)$, *alors en utilisant le fait que* $S(X)$ *est une combinaison linéaire de Normales :*

$$S(X) \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Si n *est très grand, l'estimateur a une variance qui tend vers zéro et donc sa distribution tend vers une constante, l'espérance commune à tous les* X_i.

Biais

On suppose qu'on ait défini un estimateur afin d'estimer une certaine quantité θ. On peut souhaiter que $S(X)$ soit en moyenne proche de θ. Une manière de voir ceci est de s'intéresser au biais d'estimation $\mathsf{E}[S(X) - \theta]$.

Exemple 12 *Si* $X_i \sim \mathsf{NID}(\mu, \sigma^2)$ *alors*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

la moyenne sur l'échantillon est un estimateur de biais nul.

Quand le biais est nul on parle d'estimateur non biaisé. Les estimateurs non-biaisés peuvent être très imprécis car ils peuvent présenter une très forte dispersion. Une manière d'évaluer leur imprécision est via le critère de moyenne d'erreur quadratique (Mean Square Error, ou MSE) :

$$\mathsf{E}\left[(S(X) - \theta)^2\right] = \mathsf{V}[S(X)] + (\mathsf{E}[S(X) - \theta])^2.$$

Et ainsi un estimateur plus précis peut se révéler biaisé.

Exercice 8 *Estimer σ^2 à l'aide d'un échantillon aléatoire tiré de $\mathsf{NID}(\mu, \sigma^2)$ en utilisant*

$$S(X) = \frac{1}{n-k}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Montrer que le minimum de MSE est atteint pour $k = -1$ tandis que l'estimateur est non-biaisé pour $k = 1$. Pour ce faire remarquer que

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - \mu\right)^2,$$

et que $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 \sim \chi^2_n$ tandis que $\sqrt{n}\left(\overline{X} - \mu\right)/\sigma \sim \mathsf{N}(0,1)$.

1.4 Approximations asymptotiques

1.4.1 Motivations

Exemple 13 *Convergence Classique*

$$X_n = 3 + \frac{1}{n} \to 3 \quad \textit{quand } n \to \infty.$$

Mais que dire de

$$X_n = 3 + \frac{Y}{n}$$

quand Y est une variable aléatoire ? Il existe diverses mesures de convergences, certaines nécessitant l'existence de moments, d'autres non.

La théorie des distributions peut se révéler très compliquée et parfois inextricable. Par conséquent, nous sommes souvent obligés d'utiliser des approximations. Parmi les nombreuses méthodes, celle qui domine consiste à rechercher l'erreur faite par une approximation consistant à supposer qu'on possède un grand échantillon et qu'on est proche des distributions asymptotiques pour la taille de l'échantillon. Cette idée est particulièrement attractive si on estime un paramètre et qu'on souhaite augmenter la précision avec le nombre d'observations. Deux résultats principaux sont utilisés dans la littérature afférente : la loi des grands nombres et le théorème limite central. Ces approximations sont des exemples de concepts plus généraux de "convergence en probabilité" et de "convergence en distribution".

Formellement, nous observons une suite de variables aléatoires $X_1, ..., X_n$ telles que, lorsque n croît, X_n se comporte comme une autre variable aléatoire ou une constante X.

Exemple 14 *Si on s'intéresse à*

$$X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

les X_i forment une suite

$$X_1 = Y_1, \quad X_2 = \frac{1}{2}(Y_1 + Y_2), \quad X_3 = \frac{1}{3}(Y_1 + Y_2 + Y_3).$$

Comment se comporte $\frac{1}{n}\sum_{i=1}^{n} Y_i$ quand n est grand ? Vers quoi X_n converge–t-elle quand n croît ?

1.4.2 Définitions

Lorsqu'on s'intéresse à une suite de variables aléatoires $\{X_n\}$ et qu'on se demande quelle est la distance entre $\{X_n\}$ et une autre variable X quand n tend vers l'infini, il existe de multiples manières de mesure la petitesse et de nombreuses notions de convergences. Nous en présentons ici trois, la seconde étant la plus importante.

Définition 3 (Convergence en moyenne quadratique) *Soient X et $X_1, ..., X_n$ des variables aléatoires. La série $\{X_n\}$ est dite converger vers X en moyenne quadratique, ce qu'on note*

$$X_n \overset{m.s.}{\to} X$$

si et seulement si

$$\lim_{n\to\infty} \mathsf{E}\left[(X_n - X)^2\right] = 0.$$

Il est nécessaire et suffisant pour que $X_n \overset{m.s.}{\to} X$ que

$$\lim_{n\to\infty} \mathsf{E}[X_n - X] = 0 \quad et \quad \lim_{n\to\infty} \mathsf{V}[X_n - X] = 0.$$

Exemple 15 *Soient $Y_1, ..., Y_n$ des variables aléatoires iid d'espérance μ et de variance σ^2. On définit*

$$X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

telle que

$$\mathsf{E}[X_n] = \mu \quad et \quad \mathsf{V}[X_n] = \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{V}[Y_i] = \frac{1}{n}\sigma^2.$$

X_n est donc un estimateur sans biais de μ et sa variance tend vers zéro. Ainsi

$$X_n \overset{m.s.}{\to} \mu.$$

Définition 4 (Convergence en probabilité) *Si quels que soient ε et $\eta > 0$, il existe un n_0 tel que*

$$\forall n > n_0 \quad \Pr\left(|X_n - X| < \eta\right) > 1 - \varepsilon,$$

alors on dit que la suite $X_1, ..., X_n$ converge en probabilité vers la variable aléatoire X, ce qu'on note

$$X_n \xrightarrow{p} X, \quad ou \quad plim\, X_n = X.$$

La convergence en moyenne quadratique est une notion plus forte que la convergence en probabilité et la première implique la seconde. Il est possible de prouver ceci à l'aide de l'inégalité de Tchébitchev :

$$\Pr\left(|X_n - X| < \eta\right) \leq \frac{1}{\eta^r}\mathsf{E}\left[|X_n - X|^r\right], \quad \text{pour tous } \eta > 0.$$

ainsi

$$X_n \xrightarrow{m.s.} X \Rightarrow X_n \xrightarrow{p} X$$

Définition 5 (Convergence presque sûre) *Si quels que soient ε et $\eta > 0$, il existe un n_0 tel que*

$$\Pr\left(|X_n - X| < \eta, \forall n > n_0\right) > 1 - \varepsilon,$$

alors on dit que la suite $X_1, ..., X_n$ converge presque sûrement vers la variable aléatoire X, ce qu'on note

$$X_n \xrightarrow{a.s.} X.$$

La convergence presque sûre vérifie qu'au delà de n_0 la distribution conjointe de tous les événements (pour tous n) se comporte comme il faut alors que la convergence en probabilité ne s'intéresse qu'aux probabilités à chaque n.

La convergence $a.s.$ est plus forte que la p-convergence :

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$$

mais $X_n \xrightarrow{a.s.} X$ n'implique ni n'est impliqué par $X_n \xrightarrow{m.s.} X$.

Théorème 3 (Loi faible des grands nombres) *Soit $X_i \sim iid$ telle que $\mathsf{E}[X_i]$ et $\mathsf{V}[X_i]$ existent alors*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\xrightarrow{p}} \mathsf{E}\left[X_i\right].$$

Cette loi peut être étendue : si $\overline{A} \xrightarrow{p} a$ et $\overline{B} \xrightarrow{p} b$, alors $g\left(\overline{A}\right) h\left(\overline{B}\right) \xrightarrow{p} g(a) h(b)$. La loi forte des grands nombres ou théorème de Kolmogorov implique que même si la variance n'existe pas la moyenne de l'échantillon converge (presque sûrement) vers l'espérance commune à toutes les variables aléatoires. Ceci implique alors que

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \underset{n\to\infty}{\xrightarrow{p}} \mathsf{E}\left[g(X_i)\right]$$

mais la convergence peut se révéler très lente !

1.4.3 Autres mesures de convergence

Les notions de convergence presque sûre ou en probabilité sont assez grossières, car elles impliquent essentiellement que $X_n - X$ implose vers zéro alors que n augmente. Ceci n'indique pas la vitesse de convergence ni ne fournit aucune information sur la forme de la distribution de $X_n - X$. Pour améliorer notre compréhension, nous devons faire appel au concept de convergence en distribution ou en loi.

Définition 6 (Convergence en loi) *Soient X et $X_1, ..., X_n$ des variables aléatoires. La série $\{X_n\}$ est dite converger vers X en loi ou en distribution, ce qu'on note*

$$X_n \xrightarrow{d} X, \quad ou \quad X_n \xrightarrow{L} X$$

si et seulement si

$$F_{X_n} \to F_X.$$

Dans ce contexte, divers résultats importants sont dus à Slustky :
- Si $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{p} \mu$, alors $X_n Y_n \xrightarrow{d} X\mu$ et $X_n/Y_n \xrightarrow{d} X/\mu$ si $\mu \neq 0$.
- Si $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{p} \mu$, soit φ une fonction continue, alors $\varphi(X_n, Y_n) \xrightarrow{d} \varphi(X, \mu)$.

Exemple 16 *Soient $X_1, ..., X_n$ des variables aléatoires iid Normales, d'espérance μ et de variance σ^2, alors*

$$\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \sim \mathsf{N}(0, 1)$$

et

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - X)^2 \xrightarrow{a.s.} \sigma^2.$$

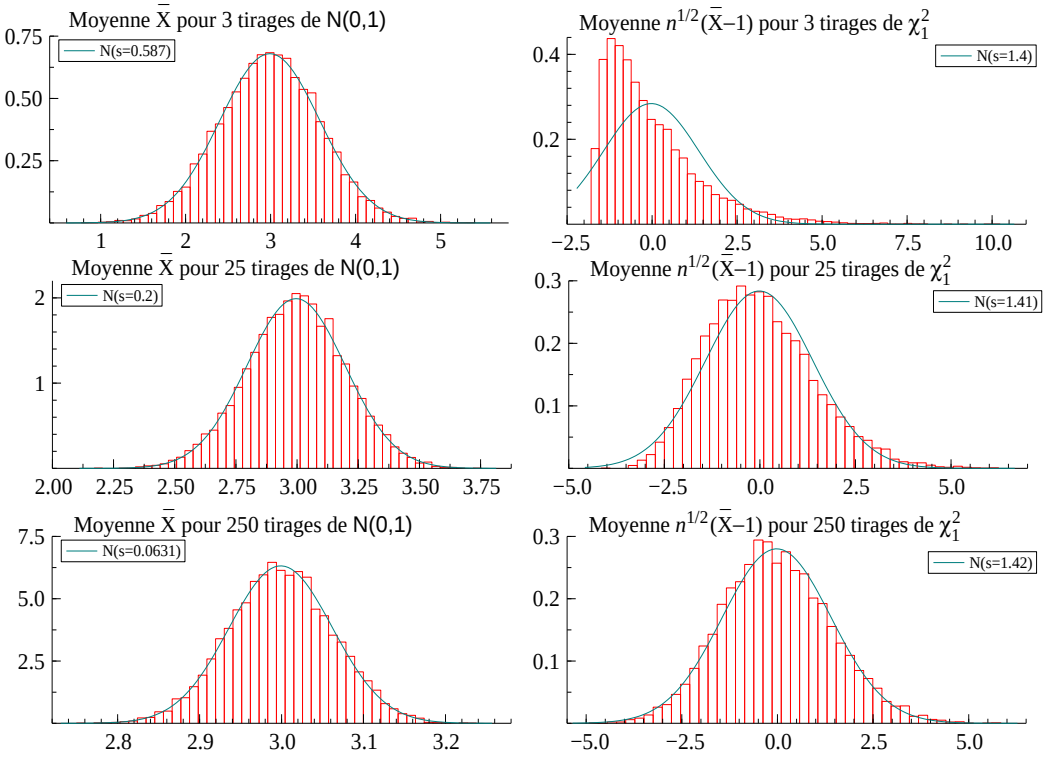


Figure 1.2 – Graphiques des densités de la moyenne d'échantillons de $n = 3, 25$ et 250 observations obtenues grâce à 10 000 simulations. La colonne de gauche représente $\overline{X}$ pour des variables Normalement distribuées, la colonne de droite $\sqrt{n}\left(\overline{X} - 1\right)$ pour des tirages de distribution χ^2_1.

Alors par le théorème de Slutsky

$$\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\widehat{\sigma}} \xrightarrow{d.} \mathsf{N}\left(0, 1\right).$$

Afin de pouvoir utiliser la théorie asymptotique, il faut recourir à des théorèmes limites centraux, le plus connu étant le

Théorème 4 (Théorême de Lindberg-Levy) *Soient $X_1, ..., X_n$ des variables aléatoires identiquement et indépendamment distribuées de sorte que $\mathsf{E}[X_i] = \mu$ et $\mathsf{V}[X_i] = \sigma^2$. On pose $\overline{X} = (X_1 + ... + X_n)/n$. Alors*

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{d.} \mathsf{N}\left(0, \sigma^2\right)$$

1.4.4 Notation de l'ordre

La théorie asymptotique est généralement traitée pour $n \to \infty$. Il peut se révéler intéressant de discuter de l'ordre de magnitude des estimateurs et des termes restants. En calcul, la notation suivante est utilisée :
Notation. Soient deux fonctions $f(x)$ et $g(x)$. Si

$$\frac{f(x)}{g(x)} \to 0 \quad \text{quand} \quad x \to \infty$$

alors f est d'ordre inférieur à g, ce qu'on note

$$f(x) = o(g(x)) \quad \text{“petit o”}.$$

Si

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} \text{ est bornée}$$

alors f est de même ordre que g, et on écrit

$$f(x) = O(g(x)) \quad \text{“grand O”}.$$

Exemple 17 $0 < a < b \Rightarrow n^a = o\left(n^b\right)$.

Exemple 18 $0 < a \Rightarrow \log n = o\left(n^a\right)$.

La notation correspondante en probabilités est
Notation. Soit une suite de variables aléatoires $X_1, ..., X_n$ et f une fonction à valeurs réelles. Si

$$X_n / f(n) \xrightarrow{p} 0,$$

on écrit alors

$$X_n = o_p(f(n)) \quad \text{“petit } o_p\text{”}.$$

Si

$$X_n / f(n) \xrightarrow{d} X$$

alors

$$X_n = O_p(f(n)) \quad \text{“grand } O_p\text{”}.$$

Exemple 19 *Selon la loi des grands nombres* $\sum_{i=1}^{n} X_i/n \xrightarrow{p} \mu$, *ainsi*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} X_i &= O_p(1)\\ \frac{1}{n}\sum_{i=1}^{n} X_i &= \mu + o_p(1).\end{aligned}$$

Exemple 20 *Le théorème limite central fournit, pour* $\mu = 0$ *et* $\sigma^2 = 1$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \xrightarrow{d} \mathsf{N}(0,1)$$

et donc

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i = O_p(1)$$

ou en d'autres termes

$$\sum_{i=1}^{n} X_i = O_p\left(\sqrt{n}\right).$$

Chapitre 2

Inférence

L'inférence statistique utilisée dans le cadre des modèles économétriques présente divers aspects. Une distinction principale s'opère entre l'estimation et les tests ; une autre entre les objets de l'attention : ils sont au nombre de trois :

Moments (méthodes d'estimation des moments). Ceci peut se révéler utile pour tester des contraintes sur les moments qui sont issues de la théorie économique (hypothèse de revenu permanent et d'anticipations rationnelles qui impliquent par exemple que la variation de la consommation est imprévisible).

Distributions (non-paramétrique). On peut souhaiter estimer la distribution du rendement d'un actif sans supposer a priori de famille de distributions au sein de laquelle on procèderait à une estimation.

Paramètres (paramétrique). Paramètres qui sous-tendent une distribution spécifique (et la distribution proprement dite peut être vérifiée par des tests de diagnostic).

Dans le cadre de ce chapitre, nous allons nous intéresser aux statégies de test. Les parties estimation et modélisation forment le sujet des chapitres suivants.

2.1 Motivations

La notation que nous allons suivre ici nécessite de différentier les observations des variables aléatoires (v.a.) qui fournissent leur distribution. Ainsi on observe un tirage $y = (y_1, ..., y_n)'$ issu des v.a. $Y = (Y_1, ..., Y_n)'$ dont la

distribution est donnée par le *modèle* : $F_Y(y;\theta)$, pour $\theta \in \Theta$. où

$$\begin{array}{ccccccc} & & F & {\scriptstyle Y} & (y;\theta); & \theta & \in \Theta \\ & \nearrow & & \uparrow & \uparrow & \nearrow & \nwarrow \\ \text{fonction de distribution} & & & v.a. & \text{réalisation} & \text{paramètre} & \text{Espace des paramètres} \\ \text{(connue)} & & & & & & \text{“admissibles”} \end{array}$$

On suppose que θ est vectoriel de dimension p.

Exemple 21 *$Y_i \sim \mathsf{NID}(\mu, \sigma^2)$, alors $\theta = (\mu, \sigma^2)'$, $\Theta \subset \mathbb{R} \times \mathbb{R}_+$ et par hypothèse d'indépendance*

$$F_Y(y;\theta) = \prod_{i=1}^{n} F_{Y_i}(y_i;\theta).$$

Exemple 22 *Régression linéaire. $Y_i|X_i = y_i \sim \mathsf{NID}(\beta' x_i, \sigma^2)$ avec indépendance entre les x_i. $\theta = (\beta, \sigma^2)'$ et $\Theta \subset \mathbb{R}^{\dim(\beta)} \times \mathbb{R}_+$. Alors*

$$F_{Y|X=x}(y;\theta) = \prod_{i=1}^{n} F_{Y_i|X_i=x_i}(y_i;\theta).$$

Ici le fait de “conditionner” sur les régresseurs x_i apparaît dans le modèle.

Exemple 23 *Y_i est une autorégression de premier ordre (AR) (une série temporelle, qui est le plus simple des modèles utilisés en macroéconométrie) :*

$$Y_i|Y_{i-1} = y_i \sim \mathsf{N}\left(\mu + \beta y_{i-1}, \sigma^2\right),$$

alors $\theta = (\mu, \beta, \sigma^2)'$, $\Theta \subset \mathbb{R}^2 \times \mathbb{R}_+$ et

$$\begin{aligned} F_{Y_2,Y_3|Y_1=y_1}(y_2, y_3;\theta) &= F_{Y_2|Y_1=y_1}(y_2;\theta)\, F_{Y_3|Y_2=y_2,Y_1=y_1}(y_3;\theta) \\ &= F_{Y_2|Y_1=y_1}(y_2;\theta)\, F_{Y_3|Y_2=y_2}(y_3;\theta), \end{aligned}$$

ce qui se généralise à

$$F_{Y_2,Y_3,\ldots,Y_n|Y_1=y_1}(y;\theta) = \prod_{i=1}^{n} F_{Y_i|Y_{i-1}=y_{i-1}}(y_i;\theta).$$

Ainsi dans le cadre des séries temporelles, la distribution conjointe n'est pas le produit des distributions individuelles, mais celui des distributions conditionnées sur le passé, i.e. les distributions des prévisions.

2.2 Choix du modèle

La partie la plus ardue de la modélisation, la sélection du modèle, peut être influencée par de nombreuses sources. Il est plus facile de décrire les propriétés d'un "bon" modèle. Nous reviendrons sur ce point plus tard.

1. Parcimonie. F contient peu de paramètres.
2. Cohérence des données. Un diagnostic permet de vérifier que le modèle ne contredit pas les données.
3. Cohérent avec des connaissances préexistantes, p. ex. une théorie économique bien établie et testée.
4. Stabilité. Représente bien le comportement des données sur différents sous-échantillons.
5. Encompassing. Explique les résultats empiriques de modèles rivaux.
6. Capacité de prévision. Capable de prévoir les observations futures.

Quand F est connue la méthode dominante pour estimer θ consiste en l'utilisation de la fonction de vraisemblance via le maximum de vraisemblance (Maximum Likelihood Estimator ou MLE) ou des méthodes dites Bayésiennes. Les méthodes des moindres carrés sont aussi couramment utilisées et coïncident souvent avec le MLE. Nous les étudierons dans le chapitre suivant. Nous supposons à présent que nous disposons d'estimateurs et d'estimations, il s'agit alors de valider des hypothèses les concernant.

L'inférence consiste à tâcher d'obtenir des informations sur la véritable distribution des variables aléatoires dont on observe un seul tirage, l'échantillon, grâce à des statistiques qui condensent une partie de l'information disponible.

2.3 Stratégies de test

Lorsqu'on dispose d'une estimation $\widehat{\theta}$ d'un paramètre θ obtenue à l'aide d'un échantillon de n observations, il y a fortes chances que la valeur de l'estimation soit légèrement différente si on utilise un échantillon de taille plus faible ou plus grande. Comment savoir alors quelle est la vraie valeur du paramètre ? quel est l'incertitude statistique de notre estimation ? Il s'agit d'établir une stratégie de test.

Lors de l'élaboration d'un test, il s'agit de définir une hypothèse qu'on souhaite confirmer ou infirmer, dite hypothèse nulle

$$H_0 : \theta = \theta_0,$$

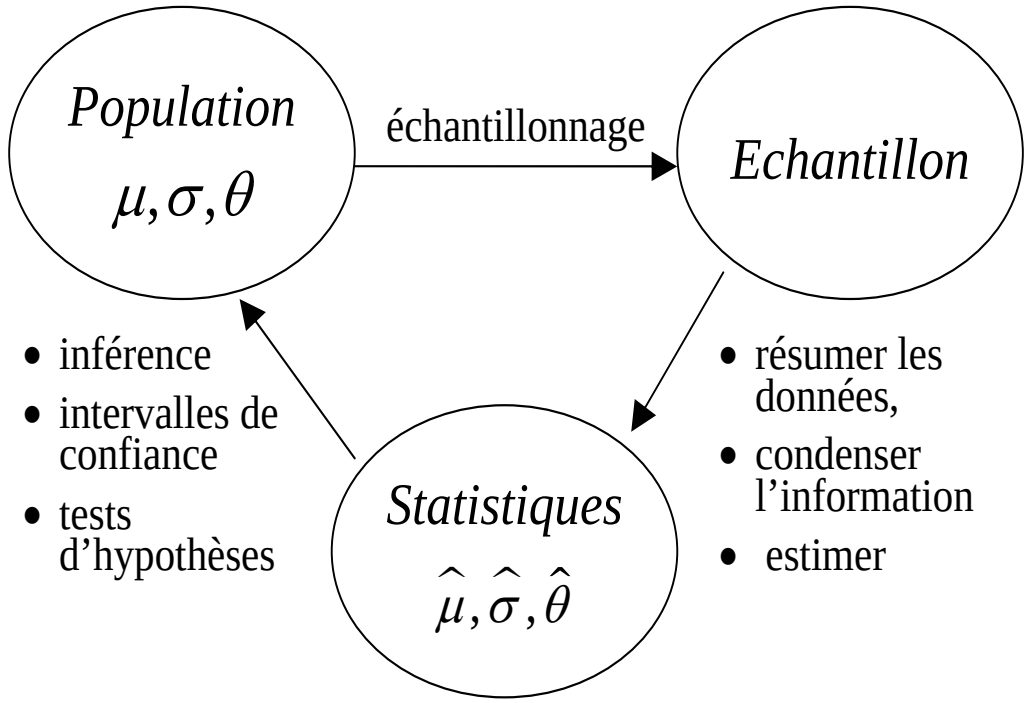


Figure 2.1 – Stratégie de modélisation et d'inférence

et d'établir l'hypothèse alternative, celle que l'on va accepter si on rejette H_0 :

$$H_1 : \theta \neq 0.$$

H_0 est ici une hypothèse simple car elle spécifie complètement la valeur de θ, tandis que H_1 est dite composite. Il existe quatre états possibles concernant ces hypothèses :

- H_0 est vraie et on l'accepte avec raison
- H_0 est vraie, mais on la rejette à tort : on fait une erreur.
- H_0 est fausse et correctement, on la rejette
- H_0 est fausse mais on l'accepte à tort, on fait ici aussi une erreur.

Ce problème présente donc deux décisions correctes et deux types d'erreurs. Le tableau 2.1 présente ces résultats de manière stylisée.

2.3.1 Erreurs de test

Si on fonde un test sur des données présentant un caractère aléatoire, il est inévitable que des erreurs surviennent et on rejette nécessairement l'hypothèse nulle même quand elle est vraie. Ce qui est important, c'est de savoir et de

	Vérité	
	H_0	H_1
Accepte H_0		$\times$ erreur de Type 2
Rejette H_0 (degré de significativité)	$\times$ erreur de Type 1	Puissance du test

Tableau 2.1 – Présentation stylisée d'une stratégie de test. Vérité représente l'état de Nature et Accepte correspond à l'hypothèse retenue.

contrôler le risque de ce type d'erreur. Ce problème est appelé erreur de Type 1. La probabilité de faire ce genre d'erreur est appelée significativité du test (ou taille, size). On la note généralement :

$$\Pr\left(\text{rejeter } H_0 | \ H_0 \text{ est vraie}\right) = \alpha,$$

où α est le degré de significativité. En pratique, on choisit souvent $\alpha = 0,05$ en économétrie, mais ce choix est arbitraire.

Pour un niveau de significativité donné, il serait attractif si l'autre type d'erreur, l'acceptation de H_0 quand elle est en fait fausse, avait une probabilité faible. Cette autre erreur est dite de Type 2.

2.3.2 Fonction de puissance

Une contrepartie de l'erreur de type 2 est le rejet correct de l'hypothèse nulle quand celle-ci est fausse. La probabilité de cette décision est appelée fonction de puissance du test :

$$\textit{Puissance}\left(\theta_1\right) = \Pr\left(\text{rejeter } H_0 | \text{ Vraie valeur de } \theta \text{ est } \theta_1\right).$$

La probabilité d'une erreur de type 2 est

$$\Pr\left(\text{erreur de type 2}\right) = 1 - \textit{Puissance}\left(\theta_1\right).$$

Typiquement, pour une valeur donnée de α, il est souhaitable d'obtenir une puissance la plus élevée possible pour tous les $\theta_1 \neq \theta_0$. Un tel test est dit puissant, et si un test particulier présente une puissance supérieure à tous les autres pour tous les θ_1, ce test est appelé "le plus puissant".

Exemple 24 *On suppose que $Y_i = \mu + \varepsilon_i$, où ε_i est iid mais la distribution n'est pas connue, quoiqu'on suppose que sa variance existe. On souhaite tester*

$$H_0 : \mu_0 = 1 \quad \textit{contre} \quad H_1 : \mu \neq 1.$$

Pour ce faire, on recourt à la statistique

$$Z = \frac{\sqrt{n}\left(\overline{Y} - \mu_0\right)}{Var\left(Y_i\right)}$$

qui, selon le théorème limite central (chapitre précédent) et sous l'hypothèse H_0

$$Z = \frac{\sqrt{n}\left(\overline{Y} - \mu_0\right)}{Var\left(Y_i\right)} \underset{H_0}{\overset{d}{\rightarrow}} \mathsf{N}\left(0,1\right).$$

Ainsi on rejette H_0 *si* Z *est grand. Selon la loi Normale, dans 5% des cas* $|Z|$ *est supérieur à 1,96. Comment ceci fonctionne-t-il en pratique, quelle est la fonction de puissance ?*

- *En réalité il s'agit ici d'un test asymptotique car si les* ε_i *ne sont pas Normalement distribués alors* Z *ne suit pas une Normale standard quand le nombre d'observations* n *est fini.*
- *La variance de* Y_i *n'est pas connue donc il faut se résoudre à utiliser un estimateur, ainsi même si les* ε_i *suivent une loi normale, on ne peut obtenir la distribution exacte de* Z*.*
- *La fonction de puissance de ce test ne peut être connue si on ne spécifie pas la distributions des* ε_i*. En revanche, si on la connaît, il peut se réveler possible de l'obtenir analytiquement.*

Si, à présent, $\varepsilon_i \sim \mathsf{NID}\left(0,\sigma^2\right)$*, alors* $Y_i \sim \mathsf{NID}\left(\mu,\sigma^2\right)$ *et* $\overline{Y} \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$*.*
Ainsi

$$\overline{Y} - \mu_0 \sim \mathsf{N}\left(\mu - \mu_0, \frac{\sigma^2}{n}\right)$$

et donc

$$Z \sim \mathsf{N}\left(\mu - \mu_0, 1\right) \tag{2.1}$$

Le test qui consiste à accepter H_0 *si* $Z \in [-1,96; +1,96]$ *présente un degré de significativité de 5% car sous* H_0 *:*

$$H_0 : Z \sim \mathsf{N}\left(0,1\right)$$

et la probabilité qu'une loi Normale Standard fournisse une valeur supérieure à 1,96 est de 2,5%. Grâce à la fonction de distribution de la loi Normale, on peut calculer, pour chaque μ *et pour* $\mu_0 = 1$*, quelle est la probabilité que la statistique* Z *soit inférieure à 1,96 en valeur absolue car si* $\mu \neq \mu_0$ *on utilise* (2.1)*.*

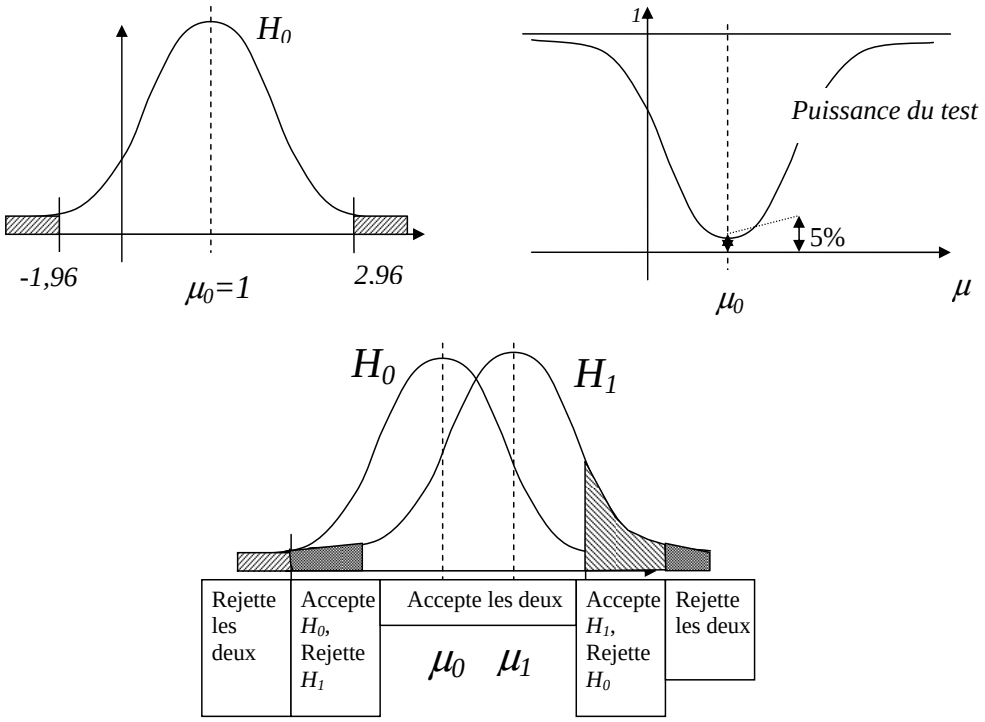


Figure 2.2 – Distributions correspondant à des startégies de tests. A gauche : le test est rejeté si la statistique de test se situe au delà des valeurs critiques $\mu_0 \pm 1,96\sigma$. A droite, la fonction de puissance est égale à la taille du test pour l'hypothèse nulle et tend vers l'unité au fur et à mesure qu'on s'éloigne de μ_0, mais elle est faible à son voisinage. En bas, les deux distributions correspondant à des hypothèses H_0 et H_1 proches ne permettent pas de rejeter ni l'une ni l'autre des hypothèses pour un grand ensemble de valeurs.

De manière générale, la stratégie de test consiste à obtenir une statistique, déduire sa distribution sous l'hypothèse nulle et/ou sous l'alternative ; décider d'un degré de significativité et calculer des valeurs critiques correspondantes pour la distribution de la statistique (des quantiles) ; rejeter ou accepter l'hypothèse nulle. En pratique rejeter un test permet une incertitude plus faible : on se place alors dans une intervalle qui correspond à 5% de probabilité ; accepter une hypothèse revient souvent à accepter aussi des hypothèses proches : voir figure 2.2, graphique du bas.

2.3.3 Tests unilatéraux

Si l'hypothèse alternative n'est pas symétrique $H_1 : \mu \neq \mu_0$ mais

$$H_1 : \mu > \mu_0$$

alors le test devient unilatéral et on ne rejetera plus H_0 lorsque Z est faible car $\mu << \mu_0$ n'est pris en compte ni dans H_0, ni par H_1. Ainsi on ne rejettera l'hypothèse nulle que si la statistique Z est supérieure au quantile à 95%, i.e. au-delà d'une valeur critique de 1,64.

2.4 Test de Student

Dans l'exemple précédent, on a défini une statistique Z et calculé sa loi de distribution asymptotique (et en échantillons de taille finie dans le cas où la distribution des ε_i était Normale). Dans cet exemple, la variance de Y_i était supposée connue ; cette hypothèse n'est pas valable en pratique et il faut se résoudre à l'estimer, ce qui influe sur la distribution de la statistique.

Dans le cas le plus simple de test concernant un paramètre unique :

$$H_0 : \theta = \theta_0 \quad \text{contre} \quad H_1 : \theta \neq \theta_0,$$

dont on possède un estimateur $\widehat{\theta}$, la pratique la plus courante est de calculer la statistique de Student, ou t-statistique : qui dans le cas d'innovations ε_i Normalement distribuées :

$$t = \frac{\widehat{\theta} - \theta_0}{\sqrt{\mathsf{V}\left[\widehat{\theta}\right]}} \sim \mathsf{N}(0,1)$$

et on rejette H_0 si $|t| > 1,96$ par exemple (dans 5% des cas sous hypothèse nulle). La théorème limite central fournit une approximation asymptotique dans le cas de ε_i suivant une autre distribution :

$$t = \frac{\widehat{\theta} - \theta_0}{\sqrt{\mathsf{V}\left[\widehat{\theta}\right]}} \underset{n\to\infty}{\to} \mathsf{N}(0,1)$$

En pratique, il faut recourrir à un estimateur de $\mathsf{V}\left[\widehat{\theta}\right]$ car celle-ci n'est pas connue de manière parfaite. Ainsi si on dispose par ailleurs de $\widehat{\sigma}^2_{\widehat{\theta}}$ un estimateur de $\mathsf{V}\left[\widehat{\theta}\right]$ calculé sur la base d'un échantillon de n observations :

$$\widehat{\sigma}^2_{\widehat{\theta}} \underset{n\to\infty}{\to} \mathsf{V}\left[\widehat{\theta}\right]$$

et

$$t = \frac{\widehat{\theta} - \theta_0}{\sqrt{\widehat{\sigma}^2_{\widehat{\theta}}}} \underset{n\to\infty}{\to} \mathsf{N}(0,1)\,. \tag{2.2}$$

Il est donc possible d'utiliser la loi Normale dans le test asymptotique. En échantillons de taille finie, il faut recourir à un *ajustement* de la loi Normale : la loi de Student, et ainsi pour un échantillon de taille n, t, définie en (2.2), suit une loi de Student avec $n-1$ degrés de liberté :

$$t \sim t_{n-1}.$$

Les distributions de Student présentent des "bords plus épais" que la distribution Normale, i.e. la probabilité de rencontrer de grandes valeurs y est légèrement supérieure : ceci tient au fait qu'une loi t_n correspond à la distribution d'un petit échantillon d'observations issues d'une loi Normale. Ainsi la présence d'une valeur élevée, qui serait expceptionnelle dans les cas d'une normale, reçoit alors une probabilité de $1/n$ soit nettement supérieure à sa probabilité théorique. Les valeurs critiques à 5% pour des tests bilatéraux sont données dans le tableau ci-dessous :

Taille de l'échantillon (n)	Degrés de liberté ($n-1$)	Valeur critique à 5%
3	2	4,30
4	3	3,18
5	4	2,78
10	9	2,26
20	19	2,09
50	49	2,01
∞	$-$	1,96

Dans le cas d'hypothèses impliquant plusieurs estimateurs, on doit recourir à une extension du test t et procéder à un test F (ratio de Fisher). Nous y reviendrons dans le cadre des régressions.

2.4.1 Les autres tests de restriction

Trois tests principaux sont utilisés dans les modèles économétriques afin de tester des restrictions (ou hypothèses) sur des paramètres du modèle. L'un requière d'estimer le modèle sous H_0 (le test du score, ou dit des multiplicateurs de Lagrange), d'autres sous H_1 (test de Wald), ou enfin d'estimer les deux

modèles (test de ratio des vraisemblances, ou test LR, pour likelihood ratio). Ainsi selon la facilité d'estimation et la précision du test dans un cadre donné, (sous H_0 beaucoup de paramètres sont par hypothèse connus et l'estimation peut se révéler plus facile) l'une ou l'autre des stratégies peut être employée.

Chapitre 3

Régression

3.1 Introduction

Le comportement économique est un phénomèe complexe et l'interaction des agents au sein d'un marché ou d'une économie complets forme un système difficile à analyser. Un très grand nombre de facteurs peut potentiellement influencer les décisions. Pour cette raison, toute tentative visant à résumer le comportement d'un système micro ou macro doit explicitement adopter une démarche multivariée ; il est par conséquent important de formaliser les méthodes d'analyse de telles relations. Si seulement deux variables étaient impliquées, il serait possible de décrire leur interaction à l'aide d'un graphique et cette méthode se révèlerait extrêment informatrice. L'économétrie peut être vue comme une approche visant à apporter une réponse à la complexité des interactions et à pallier l'impossibilité de multiplier des expériences grâce à l'usage de méthodes permettant d'isoler les phénomènes économiques dans l'analyse historique : faire de l'économie une science (pseudo-) expérimentale. Au cours du dernier siècle l'économétrie a évolué en tâchant de satisfaire à une certain nombre de demandes :

- La simplicité
- La précision
- L'information
- La robustesse.

3.1.1 La régression linéaire et ses problèmes potentiels

Imaginons que nous ayons à notre disposition un échantillon d'observations contenant $K + 1$ variables :

$$y_i, x_{i,1}, ..., x_{i,K}.$$

Nous nous intéressons à la relation entre les variables aléatoires

$$Y_i, X_{i,1}, ..., X_{i,K}$$

dont l'échantillon fournit une réalisation. Si nous voyons Y_i comme la variable au sujet de laquelle une décision doit intervenir et les X_{ij} comme les variables qui influencent cette décision, on essaie alors souvent de représenter cette relation à l'aide de l'espérance conditionnelle $\mathsf{E}[Y_i|X_{i,1}, ..., X_{i,K}]$. On pourrait aussi s'intéresser à d'autres informations que la simple moyenne conditionnelle, comme la variance de la distribution conjointe de $(Y_i, X_{i,1}, ..., X_{i,K})$. Mais dans le cas de la régression, c'est l'espérance qui forme l'objet de notre analyse. Dans l'exemple décrit ci-dessus, un agent décide de la valeur de Y_i en s'aidant d'informations représentées par les X_{ij} mais peut-être sa décision est elle soumises à des alea extérieurs : ce modèle de décision s'appelle un plan contingent et s'applique par exemple aux décisions de taux d'intérêt des banques centrales. D'autres types de relations peuvent être analysées, comme le résultat pour le taux de change des interactions entre les économies domestiques (taux de croissance, déficit publique, inflation, taux d'intérêt...).

Dès qu'on souhaite estimer une moyenne conditionnelle, les hypothèses commencent à s'accumuler (la première d'entr'elles est que nous connaissions les variables Y et X). En particulier, deux hypothèses sont essentielles concernant $\mathsf{E}[Y_i|X_{i,1}, ..., X_{i,K}]$:

1. Elle est linéaire vis-à-vis des $X_{i,k}$, i.e.

$$\mathsf{E}\left[Y_i|X_{i,1}, ..., X_{i,K}\right] = \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_K X_{i,K}$$

2. Les paramètres qui nous intéressent sont les coefficients des $X_{i,k}$ (ces derniers sont appelés régresseurs).

Afin de convertir ces hypothèses en une relation impliquant Y_i, on définit une variable u_i qui contient toute l'information qui n'est pas comprise dans $X_{i,1}, ..., X_{i,K}$:

$$u_i \equiv Y_i - \mathsf{E}\left[Y_i|X_{i,1}, ..., X_{i,K}\right]$$

où nécessairement

$$\mathsf{E}\left[u_i|X_{i,1}, ..., X_{i,K}\right] = 0$$

et ainsi

$$Y_i = \mathsf{E}\left[Y_i|X_{i,1}, ..., X_{i,K}\right] + u_i.$$

D'après la définition de l'espérance conditionnelle, u_i est souvent appelé terme d'*erreur* ou innovation puisqu'il représente la part de Y_i non prise en compte par les X_{ij}.

La dernière étape sur le chemin d'une modélisation simple consiste à émettre des hypothèses au sujet de u_i (autres qu'une moyenne conditionnelle nulle). Celles-ci sont en général :

3. u_i est identiquement distribué (pour tous i) avec une variance constante σ^2.

4. les u_i sont indépendamment distribués pour $i = 1, ..., n$.

5. u_i suit une distribution Normale.

La troisième des ces hypothèses est ce qu'on appelle une erreur *homoscédastique* (par opposition à *hétéroscédastique* quand σ_i^2 varie), et la quatrième correspond à une erreur *non-autocorrélée* ou ne présentant pas de *corrélation sérielle*. 3. et 4. ensemble fournissent une erreur *identiquement et indépendamment distribuée* :

$$u_i \sim i.i.d. \left(0, \sigma^2\right).$$

La cinquième hypothèse renforce les précédentes et n'est pas toujours imposée. Ce modèle permet donc d'écrire

$$Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_K X_{i,K} + u_i$$

et forme ce qu'on appelle le modèle de régression linéaire. Il est important de conserver en mémoire que ce modèle a été obtenu à la suite d'une série d'étapes de réduction et rien ne nous laisse à penser qu'elles soient toujours vraies. Dans la suite ce chapitre nous allons nous intéresser à l'estimation de ce modèle et allons tâcher de voir comment les hypothèses présentées ci-dessus peuvent se révéler erronées et quelles en sont les conséquences.

3.1.2 Notation vectorielle et matricielle

Les notions présentées ci-dessus sont notées sous forme vectorielle :

$$\underset{K\times i}{x_i} = \begin{pmatrix} 1 \\ x_{i2} \\ . \\ . \\ . \\ x_{iK} \end{pmatrix}, \quad \underset{K\times i}{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_K \end{pmatrix}$$

de sorte que

$$y_i = x_i'\beta + u_i, \quad i = 1, 2, ..., n.$$

On suppose alors

$$\mathsf{E}[Y_i|X_i = x_i] = x_i'\beta \quad et \quad \mathsf{E}[u_i|X_i = x_i] = 0.$$

Il est souvent plus facile d'empiler toutes les n observations dans un vecteur unique

$$\underset{n\times 1}{y} = (y_1, ..., y_n)'$$

et

$$\underset{n\times 1}{u} = (u_1, ..., u_n)'$$

$$\underset{n\times K}{X} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nK} \end{pmatrix}$$

de sorte que

$$y = X\beta + u$$

et les hypothèse (3) et (4) impliquent que

$$\mathsf{Cov}[u|X] = \sigma^2 I.$$

car $\mathsf{Cov}[u_i, u_j|X] = \sigma^2 1_{\{i=j\}}$, où $1_{\{i=j\}} = 1$ si $i = j$ et 0 sinon.

3.2 Régression

3.2.1 Maximum de vraisemblance

Cas général

On note $f(Y; X, \theta)$ la densité conjointe de l'ensemble des observations Y (conditionnée à X) et soumise à un paramètre θ, ici $\theta = (\beta, \sigma^2)$. Alors quand $Y = y$ est observé, la fonction de θ :

$$L(\theta; y, X) = f_{Y|X}(y; \theta, X)$$

est appelé **fonction de vraisemblance**. Pour une densité donnée, θ est fixé quand y varie. Dans le cas d'une fonction de vraisemblance, les rôles sont inversés : on fixe l'échantillon aux valeurs observées et θ varie. Ainsi la fonction de vraisemblance fournit la probabilité d'observer ledit échantillon si le paramètre de la distribution est θ. Une remarque : $\log L(\theta; y)$ est souvent plus facile à utiliser que $L(\theta; y)$ et il existe une bijection entre les deux.

Exemple 25 *Soit* $Y_i \sim \mathsf{NID}(\mu, \sigma^2)$ *et* $\theta = (\mu, \sigma^2)$ *de sorte que*

$$\begin{aligned} f_Y(y;\theta) &= \prod_{i=1}^{n} f_{Y_i}(y_i;\theta) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]. \end{aligned}$$

L'estimateur dit du maximum de vraisemblance (Maximum Likelihood estimator ou MLE) est la valeur de θ qui fournit la plus forte probabilité d'observer $Y = y$:

$$\widehat{\theta} = \underset{\theta}{argmax} L(\theta; y)$$

La vraisemblance est une mesure du caractère plausible, on maximise la "plausibilité".

Quand θ n'a pas d'influence sur le domaine de définition de Y, on s'intéresse plus souvent à la dérivée de $\log L$, ce qu'on appelle le "score" :

$$\frac{\partial \log L(\theta; y)}{\partial \theta}$$

et $\widehat{\theta}$ est défini comme solution de l'équation d'estimation ;

$$\left.\frac{\partial \log L(\theta; y)}{\partial \theta}\right|_{\theta=\widehat{\theta}} = 0$$

qui n'est malheureusement pas nécessairement unique.

Exercice 9 *suite de l'exemple précédent :*

$$\log L(\theta; y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

et donc

$$\begin{aligned} \frac{\partial \log L(\theta; y)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 \\ \frac{\partial \log L(\theta; y)}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \end{aligned}$$

et donc les zéros de ces équations fournissent

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad et \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{\mu})^2$$

on peut vérifier que les dérivées secondes sont négatives en $\widehat{\theta}$.

Il faut bien garder en mémoire que l'estimateur du maximum de vraisemblance présuppose qu'on connaisse la distribution exacte des données, et il permet d'obtenir des valeurs des paramètres sous cette hypothèse, qui n'est d'ailleurs guère plus restrictive que celles nécessaires à d'autres méthodes d'estimation. Il existe des familles de distributions très générales présentant davantage de paramètres et dont la Gaussienne est une sous-famillle qui permettent de travailler dans un cadre très général. L'intérêt du MLE est qu'il est le plus efficient quand le modèle est bien spécifié (i.e. la plus précis car sa variance est minimale), qu'il peut prendre des formes très complexes, et qu'enfin même quand le modèle est faux (mal-spécifié) le MLE (alors appelé quasi-MLE) peut fournir les vrais paramètres !

Dans la régression linéaire

Sous les conditions $(1)-(5)$ du § 3.1.1,.

$$y_i|x_i \sim \mathsf{NID}\left(x_i'\beta, \sigma^2\right)$$

donc la fonction de densité conditionnelle est

$$\begin{aligned} f\left(y|x;\sigma^2,\beta\right) &= \prod_{i=1}^{n} f_{Y_i}\left(y_i, x_i; \sigma^2, \beta\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left(y - X\beta\right)'\left(y - X\beta\right)\right] \end{aligned}$$

et la fonction de log-vraisemblance conditionnelle

$$\log L\left(\beta, \sigma^2\right) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y - X\beta\right)'\left(y - X\beta\right)$$

donc $\widehat{\beta}$ apparaît sous $-\left(y - X\beta\right)'\left(y - X\beta\right)$ et

$$\widehat{\beta} = \underset{\beta}{argmin}\left(y - X\beta\right)'\left(y - X\beta\right)$$

$$\widehat{\sigma}^2 = \underset{\sigma^2}{argmax}\left[-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right].$$

A présent

$$\begin{aligned}
\frac{\partial \log L(\theta; y)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 \\
\frac{\partial \log L(\theta; y)}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log L(\beta, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\
\frac{\partial \log L(\beta, \sigma^2)}{\partial \beta_j} &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ij}(y_i - x_i'\beta), \quad j = 1, 2, ..., K
\end{aligned}$$

donc les zéros de ces équations sont

$$\widehat{\sigma}^2 = \frac{1}{n}\left(y - X\widehat{\beta}\right)'\left(y - X\widehat{\beta}\right)$$

et on constate que

$$\sum_{i=1}^{n} x_{ij}(y_i - x_i'\beta) = 0$$

implique

$$(x_{1j}, x_{2j}, ..., x_{nj})\left(y - X\widehat{\beta}\right) = 0$$

et ce pour tous j, ainsi

$$X'\left(y - X\widehat{\beta}\right) = 0$$

ce qu'on résoud si $X'X$ est inversible par

$$\widehat{\beta} = (X'X)^{-1}X'y. \tag{3.1}$$

On constate que le MLE est linéaire vis-à-vis de y.

Résidus

On appelle résidus les estimateurs de la série d'erreurs u_i, définis par

$$\begin{aligned}
\widehat{u} &= y - \widehat{y} = y - X\widehat{\beta} \\
&= y - X(X'X)^{-1}X'y \\
&= \left(I - X(X'X)^{-1}X'\right)y.
\end{aligned}$$

On pose en général

$$M_X = I - X\left(X'X\right)^{-1}X'$$

une matrice symétrique idempotente, i.e. $M_X^2 = M_X = M_X'$ de sorte que

$$\widehat{u} = M_X y = M_X u.$$

On constate que la matrice $X\left(X'X\right)^{-1}X' = P_X$ est une matrice de projection sur l'espace défini par les combinaisons linéaires des vecteurs de X et que M_X est la projection sur une direction orthogonale à l'espace généré par X : $M_X P_X = 0$ et

$$I = P_X + M_X$$

où

$$\begin{array}{rcl} \widehat{u} & = & M_X y \\ \widehat{y} & = & P_X y \\ \widehat{u}'\widehat{y} & = & 0. \end{array}$$

et surtout les régresseurs sont orthogonaux aux résidus :

$$X'M_X = X'\widehat{u} = 0$$

Ainsi $\widehat{u} = M_X u$ implique que

$$\widehat{u}|X \sim \mathsf{N}\left(0, \sigma^2 M_X M_X'\right) = \mathsf{N}\left(0, \sigma^2 M_X\right).$$

A présent, les équations du MLE précédentes ont donné :

$$\widehat{\sigma}^2 = \frac{1}{n}\widehat{u}'\widehat{u}$$

et donc

$$\widehat{\sigma}^2 = \frac{1}{n}u'M_X'M_X u = \frac{1}{n}u'M_X u$$

où

$$\mathsf{E}\left[\widehat{\sigma}^2|X\right] = \mathsf{E}\left[\frac{1}{n}u'M_X u|X\right]$$

$u'M_X u$ est un scalaire donc il est égal à sa trace $u'M_X u = tr\left[u'M_X u\right]$ et on sait que $tr\left[AB\right] = tr\left[BA\right]$, ainsi :

$$\begin{aligned}
\mathsf{E}\left[\widehat{\sigma}^2|X\right] &= \mathsf{E}\left[\frac{1}{n}tr\left(M_X u'u\right)|X\right] \\
&= \frac{1}{n}tr\left(M_X \mathsf{E}\left[u'u|X\right]\right) \\
&= \frac{1}{n}tr\left(M_X\right)\mathsf{E}\left[u'u|X\right] \\
&= \frac{\sigma^2}{n}tr\left(M_X\right) \\
&= \frac{n-K}{n}\sigma^2
\end{aligned}$$

Cet estimateur est donc biaisé, un estimateur sans biais serait :

$$\widetilde{\sigma}^2 = \frac{1}{n-K}\widehat{u}'\widehat{u}.$$

Propriétés

L'estimateur du maximum de vraisemblance satisfait à beaucoup de propriétés. Puisque

$$\widehat{\beta} = \left(X'X\right)^{-1}X'y. \quad et \quad y|X \sim \mathsf{N}\left(X\beta, \sigma^2 I\right)$$

on peut utiliser les résultats des distributions Normales :

$$\begin{aligned}
\widehat{\beta}|X &= \left(X'X\right)^{-1}X'y|X \sim \mathsf{N}\left(\left(X'X\right)^{-1}X'\left[X\beta\right], \left(X'X\right)^{-1}X'\left[\sigma^2 I\right]X\left(X'X\right)^{-1}\right) \\
\widehat{\beta}|X &\sim \mathsf{N}\left(\beta, \sigma^2\left(X'X\right)^{-1}\right).
\end{aligned}$$

Ceci implique, en utilisant la loi des espérances itérées (qui dit que $\mathsf{E}\left[\mathsf{E}\left[A|B\right]\right] = \mathsf{E}\left[A\right]$) :

$$\mathsf{E}\left[\widehat{\beta}\right] = \beta \quad et \quad \mathsf{Cov}\left[\widehat{\beta}\right] = \sigma^2\mathsf{E}\left[\left(X'X\right)^{-1}\right].$$

On peu aussi montrer que pour tout autre estimateur $\widetilde{\beta}$,

$$\mathsf{Cov}\left[\widetilde{\beta}|X\right] \geq \sigma^2\left(X'X\right)^{-1}$$

et ainsi le MLE atteint la plus faible variance (dite borne de Cramér-Rao), il est dit *efficient*. Enfin, remarquons que nous avons procédé à une maximisation de la fonction de vraisemblance en séparant β et σ^2; si on maximise de manière

multivariée par rapport à $\theta = (\beta, \sigma^2)$, cela nous permet d'obtenir la covariance de $\widehat{\beta}$ et $\widehat{\sigma}^2$:

$$\begin{aligned} \mathsf{E}\left[\widehat{\theta}|X\right] &= \mathsf{E}\begin{bmatrix}\widehat{\beta}|X\\ \widehat{\sigma}^2|X\end{bmatrix} = \begin{bmatrix}\beta\\ \sigma^2\end{bmatrix} \\ \mathsf{Cov}\left[\widehat{\theta}|X\right] &= \begin{bmatrix} \sigma^2\left(X'X\right)^{-1} & 0 \\ 0 & -\dfrac{2\sigma^4}{n} \end{bmatrix}. \end{aligned}$$

Où la distribution de $\widehat{\sigma}^2|X$ suit une loi χ^2. Cependant le théorème limite central nous permet d'affirmer :

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) \underset{n\to\infty}{\to} \mathsf{N}\left(0, \begin{bmatrix} \sigma^2 Q^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}\right),$$

où

$$Q = \lim_{n\to\infty} \frac{1}{n} X'X$$

qui existe en général car la variance des regresseurs X est bornée. Dans le cas contraire, on a affaire à des régresseurs dits non-stationnaires et nous verrons comment traiter ce point dans les chapitres concernant les séries temporelles. Nous constatons que

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) \sim O_p\left(1\right)$$

et donc que

$$\widehat{\theta} = \theta + O_p\left(\frac{1}{\sqrt{n}}\right)$$

l'estimateur tend vers sa cible à un taux de $n^{-1/2}$, il est dit cohérent à l'ordre $\sqrt{n}$ (root n consistent).

3.2.2 Moindres carrés (Least squares)

Définition

On appelle estimateurs des moindres carrés tout estimateur qui est obtenu en minimisant un critère quadratique des résidus : l'exemple le plus simple, appelé moindres carrés ordinaires (ordinary least squares) vise à minimiser le critère :

$$\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 = \left(Y - X\beta\right)'\left(Y - X\beta\right) \tag{3.2}$$

vis-à-vis de β. (On aurait pu choisie de minimiser la valeur absolue de $\sum_{i=1}^{n} |y_i - x_i'\beta|$, mais cet estimateur (moindre distance absolue, ou least absolute deviation, LAD) est moins utilisé car il n'est pas dérivable).

La condition de premier ordre (i.e. la première dérivée de (3.2) est nulle) donne :

$$2\sum_{i=1}^{n} x_i \left(y_i - x_i'\widehat{\beta}\right) = 0$$

et ainsi on reconnaît le MLE :

$$\widehat{\beta} = (X'X)^{-1} X'y.$$

MCO (OLS) et MLE coïncident dans le cas des régressions Gaussiennes. Il est important d'être familier avec de nombreuses propriétés des estimateurs MCO.

Propriétés

Théorème 5 (Gauss-Markov) *L'estimateur des MCO du modèle linéaire où u_i est iid $(0, \sigma^2)$ est le meilleur (i.e. de variance minimale) estimateur au sein de la classe des estimateurs linéaires non-biaisés.*

Pour obtenir des informations sur la distribution de $\widehat{\beta}$, on constate que, conditionnellement à X

$$\widehat{\beta} = \beta + (X'X)^{-1} X'u$$

et donc puisque $u \sim \mathsf{N}(0, \sigma^2 I)$

$$\begin{aligned} \widehat{\beta} - \beta | X &\sim \mathsf{N}\left(0, (X'X)^{-1} X' \left[\sigma^2 I\right] X (X'X)^{-1}\right) \\ &= \mathsf{N}\left(0, \sigma^2 (X'X)^{-1}\right). \end{aligned}$$

Théorème 6

$$(X'X)^{1/2} \left(\widehat{\beta} - \beta\right) \sim \mathsf{N}\left(0, \sigma^2 I_K\right).$$

Théorème 7

$$\sigma^{-2} \left(\widehat{\beta} - \beta\right)' (X'X) \left(\widehat{\beta} - \beta\right) \sim \chi^2_K.$$

Les théorèmes précédents nous permettent de tester des hypothèses concernant β à l'aide de $\widehat{\beta}$, à condition de connaître σ^2. Quand ce dernier est inconnu, il nous faut l'estimer. Mais auparavant, remarquons que si l'un des regresseurs est une constante (i.e. 1) alors la somme des résidus est nulle :

$$\sum_{i=1}^{n} \widehat{u}_i = 0, \tag{3.3}$$

ceci n'est pas une propriété des erreurs, car en général

$$\sum_{i=1}^{n} u_i \neq 0$$

même si

$$\mathsf{E}\left[\sum_{i=1}^{n} u_i\right] = 0.$$

On doit donc faire bien attention lorsqu'on remplace les erreurs par les résidus. Il faut aussi faire attention aux logiciels car la plupart des tests qu'ils calculent font l'hypothèse (3.3) , ce qui n'est pas vrai si aucune constante n'est incluse parmi les régresseurs.

Indicateurs de précision

On définit les résidus

$$\widehat{u} = y - X\widehat{\beta}$$

de sorte que

$$y = X\widehat{\beta} + \widehat{u} = \widehat{y} + \widehat{u}, \quad \widehat{u}'X = 0$$

et donc

$$\begin{aligned} y'y &= (\widehat{y}' + \widehat{u}')(\widehat{y} + \widehat{u}) = \widehat{y}'\widehat{y} + 2\widehat{y}'\widehat{u} + \widehat{u}'\widehat{u} \\ &= \widehat{y}'\widehat{y} + \widehat{u}'\widehat{u} + 2\widehat{\beta}'X'\widehat{u} \\ &= \widehat{y}'\widehat{y} + \widehat{u}'\widehat{u} \end{aligned}$$

La somme des carrés des y_i est égale à la somme des carrés des variables estimées $(\widehat{y}_i)$ plus la somme des carrés des résidus. On appelle les sommes

$$\begin{aligned} \sum_i y_i^2 &= TSS \,(\text{Total sum of squares}) \\ \sum_i \widehat{y}_i^2 &= ESS \,(\text{Explained sum of squares}) \\ \sum_i \widehat{u}_i^2 &= RSS \,(\text{Residual sum of squares}) \end{aligned}$$

de sorte que

$$TSS = ESS + RSS$$

et la mesure habituelle de précision de la régression est

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \in]0, 1[\,,$$

plus le R^2 (R-deux) est proche de l'unité meilleure est la régression (Goodness of fit). Attention tout de même, dans le cadre des variables non-stationnaires, le R^2 peut se révéler très proche de 1 alors même que les variables ne sont pas corrélées !

3.2.3 Erreurs de spécification

Il peut arriver qu'on fasse une erreur sur le modèle mais que les estimateurs convergent toutefois vers leur vraies valeurs. Ou alors, les résultats peuvent n'avoir aucun sens. Afin d'étudier les propriétés des estimateurs de modèles mal spécifiés, la procédure suivie est :

1. Trouver une expression pour les estimateurs dans le cadre du modèle tel que spécifié, en fonction des variables aléatoires le composant.
2. Remplacer les variables aléatoires par leur vraie distribution probabiliste dans le bon modèle (le processus générateur des données ou DGP) et observer la loi de distribution des estimateurs, pour voir s'ils ont biaisés, cohérents...

Pour comprendre les divers cas on étudie :

$$y = X_1\beta_1 + X_2\beta_2 + u$$

où on appelle DGP le modèle correct et M celui utilisé dans la régression.

– Modèles sur-spécifiés :

$$\begin{aligned} \mathsf{DGP} &: y = X_1\beta_1 + u \\ \mathsf{M} &: y = X_1\beta_1 + X_2\beta_2 + u \end{aligned}$$

i.e. $\beta_2 = 0$.

Un modèle est dit sur-spécifié lorsqu'il inclut des régresseurs dont le vrai coefficient est nul et qui ne devraient pas intervenir dans la régression. Les estimateurs sont cohérents et non-biaisés mais leur variance est supérieure à celle qu'on obtiendrait grâce au modèle correctement spécifié.

– Modèles sous-spécifiés

$$\begin{aligned} \mathsf{DGP} &: \; y = X_1\beta_1 + X_2\beta_2 + u \\ \mathsf{M} &: \; y = X_1\beta_1 + u \end{aligned}$$

i.e. $\beta_2 \neq 0$. Alors l'estimateur est biaisé dans des échantillons de taille finie et non cohérent dans de grands échantillons : si $\widehat{\beta}_1 = (X_1'X_1)^{-1} X_1'y$ alors

$$\widehat{\beta}_1 \underset{n\to\infty}{\to} \beta_1 + \left(\lim_{n\to\infty} \mathsf{E}\left[\frac{X_1'X_1}{n}\right]^{-1}\right)\left(\lim_{n\to\infty} \mathsf{E}\left[\frac{X_1'X_2}{n}\right]\right)\beta_2$$

– Hétéroscédasticité

Si la série d'erreur u_i n'est pas identiquement distribuée mais de variance σ_i^2 alors on parle de présence d'hétéroscédasticité. Ceci ne modifie pas le caractère non-biaisé de l'estimateur OLS mais entraîne des problèmes concernant se variance. Ainsi les tests de Student se révèleront faux car ils ne prennent pas en compte la bonne distribution. On utilise alors l'estimateur des moindres carrés généralisés qui consiste a travailler sur des données transformées : $y_i^* = y_i/\sigma_i$, $x_i^* = x_i/\sigma_i$.
Comment se rendre compte de la présence d'hétéroscédasticité ? Soit par des méthodes informalles d'observations des carrés des résidus, soit par divers tests (on estime par exemple un modèle plus général et on teste la présence de coefficients non nuls). Les logiciels en présentent souvent plusieurs.

– Autocorrélation

Souvent il n'est pas possible de faire l'hypothèse que les $y_i|x_i$ sont indépendants, c'est le cas en finance, macroéconométrie et en micro lorsqu'on suit un même individu au cours du temps. Si les données présentent une certaine dépendance temporelle, on ne peut plus supposer que $\mathsf{Cov}[u|X]$ soit diagonale mais on doit poser :

$$\mathsf{Cov}\left[u|X\right] = \Omega.$$

On peut sinon utiliser des modèles du type

$$y_t = \alpha y_{t-1} + x_t'\beta + u_t, \quad u_t|y_{t-1}, x_t \sim \mathsf{NID}\left(0, \sigma^2\right)$$

en utilisant des *retards* (valeurs retardées) de la variable endogène (et/ou des variables exogènes). Ces modèles sont appelés autorégressifs dynamiques.

– Corrélation des régresseurs et des erreurs
Une des hypothèses du modèle classique de régression fait l'hypothèse selon laquelle

$$\mathsf{E}[u|X] = 0$$

i.e.

$$\begin{aligned} \mathsf{Cov}[u_i x_j] &= 0 \quad \forall i,j \\ \mathsf{E}[X'u] &= 0 \end{aligned}$$

qui intervient dans le carctère non-biaisé de $\widehat{\beta}$ car

$$\widehat{\beta} = \beta + (X'X)^{-1} X'u.$$

Si $\mathsf{E}[X'u] \neq 0$ alors l'estimateur n'est pas cohérent et ne tend pas vers sa vraie valeur β. Ce cas apparaît souvent en pratique, par exemple dans le cadre de modèles d'anticipations rationnelles où ce n'est pas x_i qui intervient dans la régression mais sa valeur anticipée x_i^* alors que x_i est inconnue. On observe le même phénomène quand la mesure de x_i est imprécise (p.ex le PIB qui n'est en général connu qu'après deux ans, on travaille avec des estimations dans l'intervalle) et ainsi la mesure $x_i^* = x_i + \eta_i$. On rencontre aussi beaucoup ce problème en microéconomie. On doit alors chercher des variables instrumentales, qui sont corrélées avec X mais non avec u. C'est le travail du modélisateur que de choisir de bonnes variables (il peut s'agir de variables retardées de x_i, p.ex. x_{i-1}, corrélé avec x_i mais non avec u_i). La méthode des variables instrumentales dans le cadre le plus simple consiste à regresser X sur les instruments Z et d'utiliser les variables estimées $\widehat{X}$ pour le calcul de $\widehat{\beta}_{IV}$.

3.2.4 Choix du modèle

Il apparaît donc que le coût de l'omission de régresseurs soit nettement supérieur à celui de l'inclusion de variables inutiles, car dans ce dernier cas, il s'agit alors surtout d'un problème de précision de l'estimation. Ceci est à l'origine de la méthode généralement préconisée de Géneral vers Spécifique, où on commence la modélisation par l'inclusion de l'ensemble des variables théoriquement possibles, et estimation après estimation on ôte celles dont le coefficient n'est pas statistiquement différent de zéro (par des tests de Student). On devrait normalement aboutir au bon modèle si les variables de départ contiennent l'ensemble de celles qui interviennent effectivement et si le modèle est stable sur l'intégralité de l'échantillon. On doit toutefois procéder à la

fin à des tests de spécification afin de vérifier que le modèle est bon (homoscédastique, résidus non autocorrélés...).

Si plusieurs modèles résistent à l'analyse et à des tests de spécification. On peut choisir d'en privilégier un sur la base de critères *ad hoc*. Il en existe divers dont le plus connu est le critère d'information d'Akaike (AIC) qui combine la précision de l'estimation et son caractère parcimonieux (faible nombre de régresseurs) : il s'agit de minimiser

$$AIC = \log\left(\widehat{\sigma}^2\right) + 2\frac{K}{n},$$

où K est le nombre de paramètres et n la taille de l'échantillon.

Chapitre 4

Séries temporelles

4.1 Introduction

4.1.1 Qu'appelle-t-on série temporelle ?

Contrairement à l'économétrie traditionnelle, le but de l'analyse des séries temporelles n'est pas de relier des variables entre elles, mais de s'intéresser à la *dynamique* d'une variable. Cette dernière est en effet essentielle pour deux raisons : les avancées de l'économétrie ont montré qu'on ne peut relier que des variables qui présentent des propriétés similaires, en particulier une même stabilité ou instabilité ; les propriétés mathématiques des modèles permettant d'estimer le lien entre deux variables dépendent de leur dynamique.

Définition 7 (Série Temporelle) *La suite d'observations* $(y_t, t \in \mathbb{T})$ *d'une variable* y *à différentes dates* t *est appelée série temporelle. Habituellement,* $\mathbb{T}$ *est dénombrable, de sorte que* $t = 1, ...T$.

Remarque 8 *En mathématiques, la définition de série temporelle ci-dessus correspond à la définition d'une* suite, $\{u_n\}_{n\in\mathbb{I}}$, *tandis qu'on nomme série la suite définie à partir de la somme des termes de la suite :* $s_n = \sum_{i=0}^{n} u_i$.

Une série temporelle est donc toute suite d'observations correspondant à la même variable : il peut s'agir de données macroéconomiques (le PIB d'un pays, l'inflation, les exportations...), microéconomiques (les ventes d'une entreprise donnée, son nombre d'employés, le revenu d'un individu, le nombre d'enfants d'une femme...), financières (le CAC40, le prix d'une option d'achat ou de vente, le cours d'une action), météorologiques (la pluviosité, le nombre de jours de soleil par an), politiques (le nombre de votants, de voix reçues par un candidat...), démographiques (la taille moyenne des habitants, leur âge...).

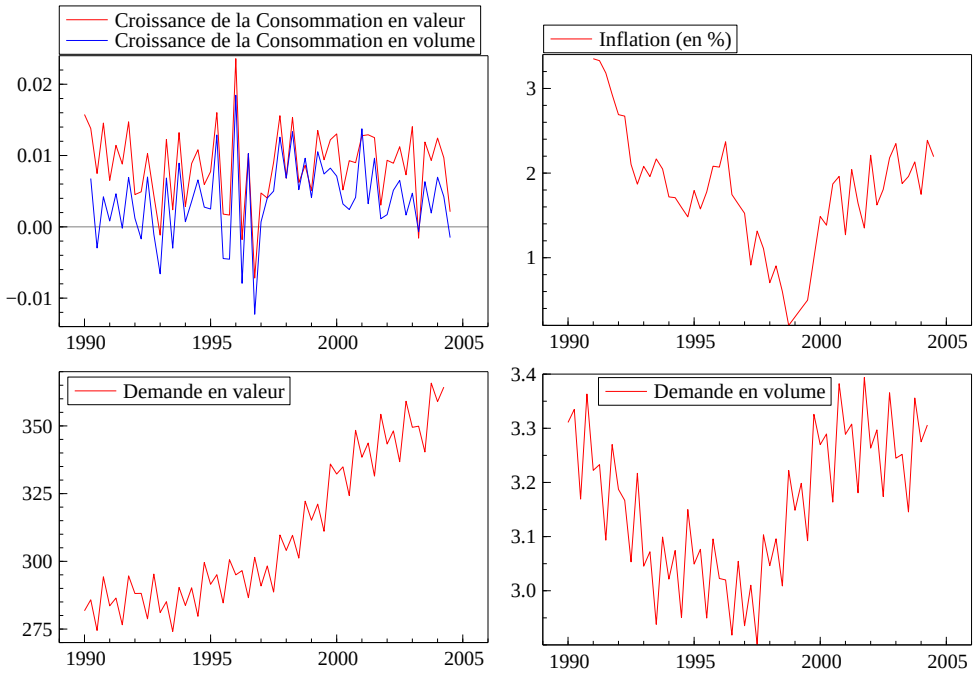


Figure 4.1 – Données françaises trimestrielles de : (a) croissance de la consommation domestique en valeur et en volume (i.e. sans effet d'augmentation des prix) ; (b) Inflation ; (c) Demande en valeur ; (d) Demande en volume.

En pratique, tout ce qui est chiffrable et varie en fonction du temps. La dimension temporelle est ici importante car il s'agit de l'analyse d'une chronique historique : des variations d'une même variable au cours de temps, afin de pouvoir en comprendre la dynamique. Les données de panel s'intéressent pour leur part à la variabilité de caractéristiques entre individus, agents, entreprises. La périodicité de la série n'importe en revanche pas : il peut s'agir de mesures quotidiennes, mensuelles, trimestrielles, annuelles... voire même sans périodicité.

On représente en général les séries temporelles sur des graphiques de valeurs (ordonnées) en fonction du temps (abscisses). Une telle observation constitue un outil essentiel qui permet au modélisateur ayant un peu d'expérience de tout de suite se rendre compte des propriétés dynamiques principales, afin de savoir quel test statistique pratiquer. Sur la figure 4.1, quatre graphiques montrent des séries ayant des propriétés différentes. Le panneau (a) présente deux séries qui oscillent autour d'une valeur comprise entre 0 et 0,01 : elles sont stables autour de leur moyenne. On parle dans ce cas de séries stationnaires. En (b) , l'inflation décroît fortement jusqu'en 1999 pour remonter ensuite : elle n'oscille pas autour d'une moyenne bien qu'elle ne soit jamais très loin de 2% ;

sous réserve de tests statistiques fins, elle semble moins stable que les séries en (a) , elle est donc peut-être non stationnaire. La série (c), quant à elle croît sur l'ensemble de l'échantillon observé ; on parle dans ce cas de tendance et sa moyenne n'est pas constante (sa moyenne entre 1990 et 1995 est radicalement différente de celle mesurée entre 2000 et 2004). Enfin, le panneau (d) reproduit la même série mais hors effet de prix (l'inflation étant toujours positive les prix croissent naturellement sans que le volumes augmentent nécessairement) : la tendance à la hausse provenait uniquement de l'augmentation des prix et la demande en volume décroît au début des années 1990. De plus, cette dernière série présente un comportement assez régulier, le niveau respectif des trimestres d'une année se reproduit tous les ans, il s'agit d'un phénomène saisonnier.

Les caractéristiques de ces graphiques sont toutes modélisables et analysables dans le cadre de l'analyse des séries temporelles. Nous allons introduire plus loin, les concepts de saisonnalité, stationnarité, tendances qui permettront de tester diverses hypothèses sur ces données et de connaître a priori lesquelles peuvent être réliées. Nous verrons aussi que, comme ici pour le passage des valeurs aux volumes, il est possible de trouver des combinaisons entre des séries qui annulent certains effets (ici, la tendance continue à la hausse).

4.1.2 Quels sont les buts de cette analyse ?

Parmi les multiples applications de l'analyse des séries temporelles, il est possible d'en distinguer neuf principales.

Prévoir

La fonction première pour laquelle il est intéressant d'observer l'historique d'une variable vise à en découvrir certaines régularités afin de pouvoir extrapoler et d'établir une prévision. Il s'agit ici de comprendre la dynamique qui relie une observation à celles qui l'ont précédée et de supposer, sous réserve qu'on puisse justifier une telle hypothèse, que les mêmes causes produisent les mêmes effets. Avec une analyse fine, il est même possible d'établir des prévisions "robustes" vis-à-vis de ruptures brusques et de changements non anticipables.

Relier les variables

Il est important de de savoir a priori si certaines relations sont "économétriquement" possibles et d'éviter les équations qui ne présentent aucun sens. Reprenons les séries présentées figure 4.1 : soit la demande en valeur (panneau c) à la

date t, notée D_t, et l'inflation notée i_t. Peut-on faire l'hypothèse que l'inflation influence positivement la demande ? Ce qui reviendrait à dire par qu'en période de forte inflation, les citoyens souhaitent consommer davantage qu'en une période où elle est faible. Ce qu'on peut noter

$$D_t = \alpha + \beta i_t + \varepsilon_t, \tag{4.1}$$

où ε_t représente un écart entre la demande et ce que peut prévoir l'inflation. Si notre modèle représente bien la manière dont est générée la demande, ε_t doit être de moyenne nulle. Dans ce cas, si on note $\mathsf{E}[\cdot]$ l'espérance mathématique, i.e. la moyenne, celle-ci doit satisfaire :

$$\begin{aligned} \mathsf{E}[D_t] &= \mathsf{E}[\alpha + \beta i_t + \varepsilon_t] \\ &= \alpha + \beta \mathsf{E}[i_t] + \mathsf{E}[\varepsilon_t] \\ &= \alpha + \beta \mathsf{E}[i_t]. \end{aligned}$$

car $\mathsf{E}[\varepsilon_t] = 0$. Or nous avons vu sur la graphique que l'inflation avait tendance à ne pas trop s'éloigner de 2%, sa moyenne doit donc être constante et se situer entre 1 et 3%, mettons 2% pour simplifier. Dans ce cas $\mathsf{E}[D_t] = \alpha + \beta \times 2$ est constante, ce qui est contradictoire avec notre observation précédente qui montrait que la demande en valeur était monotone et donc que sa moyenne variait au cours du temps. Une relation comme (4.1) n'a donc aucun sens ; il est revanche statistiquement possible qu'il faille s'intéresser au lien entre l'inflation et le taux de croissance de la demande. L'analyse des séries temporelles permet de savoir quelles équations sont a priori grotesques.

Déterminer la causalité

Un approche dynamique permet aussi de s'intéresser aux relations de causalité. Pour qu'un mouvement en provoque un autre, il est nécessaire qu'il le précède. Une simple concomittance de deux événements révèle davantage une source commune. L'utilisation de *retards* d'une variable, i.e. de ses valeurs aux périodes précédentes, dans les équations autorise la mesure des effets de causalité et permet également de connaître la durée de transmission entre une source et son effet.

Distinguer entre court et long-terme

Certaines lois de comportement ne sont jamais vérifiées en pratique car elles ne s'appliquent que sur les équilibres de long terme. A plus courte échéance, des variations contrarient perpétuellement leur mise en oeuvre. Cependant, des

ajustements transitoires s'opèrent continuellement afin de s'approcher de ces équilibres. On reformule alors le modèle sous la forme d'un mécanisme dit de *correction d'équilibre* (ou d'erreur), selon lequel un écart (une erreur) positif par rapport à l'équilibre de long terme entraîne une variation de court terme négative, afin de réduire cet écart.

Etudier des anticipations des agents

Comment prendre en compte les anticipations des agents ? Dans une décision entre épargne et consommation, ce ne sont pas seulement les revenus actuel et passé qui comptent, mais aussi l'idée qu'on se fait de l'avenir. Il faut donc dans certaines équations faire intervenir des valeurs *avancées* des variables, via leur anticipation en utilisant la manière dont celles-ci ont été formées dans le passé.

Repérer les tendances et cycles

Des méthodes dynamiques repèrent des tendances mouvantes des données. Par différence, l'écart entre le niveau de la variable (localement monotone) et la position de sa tendance est en moyenne nul : il repère la position dans le cycle. Selon le modèle de tendance utilisé, il est possible d'analyser les interactions entre diverses variables afin d'atteindre un équilibre entre méthodes économétriques et purement statistiques.

Corriger des variations saisonnières

Comme constaté figure 4.1, la série de demande présente des variations régulières trimestrielles que nous avons nommées variations saisonnières. Celles-ci peuvent être stables au cours du temps et ainsi l'écart entre les premiers et deuxième trimestre sera le même en 1991 et en 2004. En retirant cet effet habituel et en lissant la série, il est alors possible de comparer le niveau entre ces années. La correction des variations saisonnières (cvs) devient plus complexe quand les comportements évoluent davantage ; l'écart entre deux trimestres consécutifs peut se modifer et la série *cvs* apportera alors une information supplémentaire.

Détecter les chocs structurels

Un choc structurel est défini comme une modification permanente ou temporaire de la façon dont est générée une variable. Ils sont fréquents, souvent non-anticipables et difficiles à mesurer. Il est cependant essentiel de savoir qu'une telle rupture a eu lieu car sa présence change les interactions et

équilibres, souvent radicalement. L'ignorer engendre alors des effets contraires aux buts poursuivis.

Contrôler les processus

Lorsqu'une autorité fixe librement le niveau d'une variable ayant une forte influence sur le reste de l'économie, comme par exemple le taux d'intérêt directeur sur lequel la banque centrale a autorité, il lui faut à la fois quantifier l'ampleur de son impact et mesurer la durée de transmission de son effet dans l'économie. En retour, cette autorité peut prendre en compte son propre comportement afin d'anticiper les évolutions d'une variable cible, comme l'inflation.

4.1.3 En quoi cette démarche consiste-t-elle ?

But

Le but poursuivi est la formulation d'un modèle statistique qui soit une représentation congruente du **processus stochastique** (inconnu) qui a généré la série observée. Tout comme en probabilités/statistiques, il faut bien comprendre la différence entre le processus sous-jacent qui génère des données (data generating process), sa réalisation telle qu'on l'observe sur l'échantillon historique à notre disposition, les futures réalisations et le modèle qu'on construit afin de tâcher de le représenter. Par représentation congruente, on entend un modèle qui soit conforme aux données sous tous les angles mesurables et testables.

Approche

Il est en pratique impossible de connaître la distribution d'une série temporelle $\{y_t\}_{t\geq 0}$, on s'intéresse par conséquent à la modélisation de la **distribution conditionnelle** (a priori constante dans le temps) de $\{y_t\}$ via sa densité :

$$f\left(y_t|Y_{t-1}\right).$$

Conditionnée sur l'historique du processus : $Y_{t-1} = (y_{t-1}, y_{t-2}, ..., y_0)$. Il s'agit donc d'exprimer y_t en fonction de son passé.

Résultat

L'approche conditionnelle fournit une **Décomposition Prévision–Erreur**, selon laquelle :

$$y_t = \mathsf{E}\left[y_t|Y_{t-1}\right] + \epsilon_t,$$

où $\begin{cases} (i) & \mathsf{E}[y_t|Y_{t-1}] \text{ est la composante de } y_t \text{ qui peut donner lieu à une prévision, quand l'historique du processus, } Y_{t-1}, \text{ est connu ; et} \\ (ii) & \epsilon_t \text{ représente les informations imprévisibles.} \end{cases}$

Exemple 26 (Modèles de séries temporelles) *1. Processus autoregressifs d'ordre 1, AR*(1) **:**

$$\begin{aligned} y_t &= \alpha y_{t-1} + \epsilon_t, \\ \epsilon_t &\sim \mathsf{WN}\left(0, \sigma^2\right) \text{ (bruit blanc)} \end{aligned}$$

La valeur y_t ne dépend que de son prédecesseur. Ses propriétés sont fonctions de α qui est un facteur d'inertie : quand $\alpha = 0$, y_t est imprévisible et ne dépend pas de son passé, on parle de bruit blanc ; si $\alpha \in]-1, 1[$, y_t est stable autour de zéro ; si $|\alpha| = 1$, y_t est instable et ses variations $y_t - y_{t-1}$ sont imprévisibles ; enfin si $|\alpha| > 1$, y_t est explosif. Des exemples sont présentés figure 4.2.

2. Séries multivariées :

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\epsilon}_t &\sim \mathsf{WN}\left(\mathbf{0}, \boldsymbol{\Sigma}\right). \end{aligned}$$

3. Pocessus autorégressif vectoriel, VAR(1) *:*

$$\begin{aligned} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}, \\ \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} &\sim \mathsf{WN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right). \end{aligned}$$

4. Modèle autorégressif à retards distribués (autoregressive distributed lags, ADL) : si $\alpha_{12} \neq 0$, ce modèle implique une relation de causalité entre $y_{2,t-1}$ et $y_{1,t}$.

$$\begin{aligned} y_{1,t} &= \alpha_{11} y_{1,t-1} + \alpha_{12} y_{2,t-1} + \epsilon_{1t}, \\ \epsilon_{1t} &\sim \mathsf{WN}\left(0, \sigma_1^2\right). \end{aligned}$$

4.2 Concepts des séries temporelles

4.2.1 Processus stochastiques

Soit $(\Omega, \mathcal{M}, P)$ un espace de probabilité, où Ω est l'espace des événements, $\mathcal{M}$ est une tribu adaptée à Ω (c'est l'ensemble qui contient les combinaisons possibles d'événements) et P est une mesure de probabilité définie sur $\mathcal{M}$.

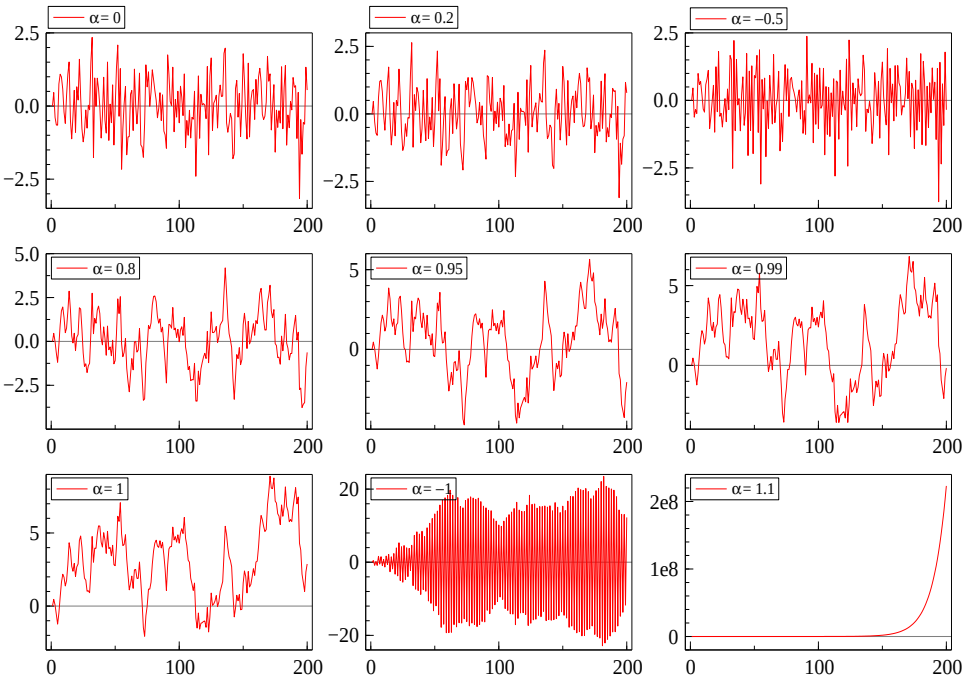


Figure 4.2 – Séries temporelles simulées à partir d'un modèle AR(1) $y_t = \alpha y_{t-1} + \varepsilon_t$ pour diverses valeurs de α. On constate la continuité des observations quand α tend vers l'unité bien que les propriétés statistiques de y_t soient radicalement différentes pour $|\alpha| < 1$ (stationnarité) et $|\alpha| = 1$ (non stationnarité). Remarquer le caractère explosif de la série pour $\alpha > 1$.

Définition 8 *Une* ***variable aléatoire*** *réelle (v.a.r) est une fonction à valeurs réelles* $y : \Omega \to \mathbb{R}$ *telle que pour tout réel* c, $A_c = \{\omega \in \Omega | y(\omega) \leq c\} \in \mathcal{M}$.

En d'autres termes, A_c est un événement dont la probabilité est définie en termes de P. La fonction $F : \mathbb{R} \to [0,1]$ définie par $F(c) = P(A_c)$ est la *fonction de distribution* de y.

Soit $\mathbb{T}$ un ensemble d'indexation dénombrable contenu dans l'ensemble des entiers naturels ou dans celui des entiers relatifs.

Définition 9 *Un* ***processus stochastique*** *(discret) est une fonction à valeurs réelles*

$$y : \mathbb{T} \times \Omega \to \mathbb{R},$$

telle que pour tout $t \in \mathbb{T}$ *donné,* $y_t(\cdot)$ *soit une variable aléatoire.*

En d'autres termes, un processus stochastique est une suite ordonnée de variables aléatoires $\{y_t(\omega), \omega \in \Omega, t \in \mathbb{T}\}$, telle que pour tout $t \in \mathbb{T}$, y_t soit une variable aléatoire sur Ω et que pour tout $\omega \in \Omega$, $y_t(\omega)$ soit une réalisation du processus stochastique sur l'ensemble d'indexation $\mathbb{T}$.

	ω_0	$\cdots$	ω_j	$\cdots$	ω_m
t_0	$y_{t_0}(\omega_0)$	$\cdots$	$y_{t_0}(\omega_j)$	$\cdots$	$y_{t_0}(\omega_m)$
$\vdots$			$\vdots$		
t_i	$y_{t_i}(\omega_0)$	$\cdots$	$y_{t_i}(\omega_j)$	$\cdots$	$y_{t_i}(\omega_m)$
$\vdots$			$\vdots$		
t_n	$y_{t_n}(\omega_0)$	$\cdots$	$y_{t_n}(\omega_j)$	$\cdots$	$y_{t_n}(\omega_m)$

Définition 10 *Une* ***série temporelle*** $\{y_t\}_{t=1}^{T}$ *est (la partie de dimension finie d') une réalisation d'un processus stochastique* $\{y_t\}$.

La réalisation d'un processus stochastique est une fonction $\mathbb{T} \to \mathbb{R}$ où $t \to y_t(\omega)$. Le processus sous-jacent est dit avoir *généré* la série temporelle. La série temporelle $y_1(\omega), ..., y_T(\omega)$ est généralement notée $y_1, ..., y_T$ ou simplement y_t.

Un processus stochastique peut être décrit par la fonction de distribution commune des toutes les sous-collections de dimension finie de y_t, $t \in \mathbb{S} \subset \mathbb{T}$. En pratique le système complet de distributions est souvent inconnu et on se cantonne aux premiers et seconds moments.

La distribution conjointe de $(y_t, y_{t-1}, ..., y_{t-h})$ est généralement caractérisée par sa *fonction d'autocovariance* qui représente le lien entre les valeurs à des

dates différentes :

$$\begin{aligned}\gamma_t(h) &= \mathsf{Cov}(y_t, y_{t-h}) \\ &= \mathsf{E}\left[(y_t-\mu_t)\left(y_{t-h}-\mu_{t-h}\right)\right] \\ &= \int ... \int (y_t-\mu_t)\left(y_{t-h}-\mu_{t-h}\right) f(y_t,...,y_{t-h})\, dy_t ... dy_{t-h},\end{aligned}$$

avec $\mu_t = \mathsf{E}[y_t] = \int y_t f(y_t)\, dy_t$, l'espérance (ou moyenne) inconditionnelle de y_t.

La *fonction d'autocorrelation* est donnée par :

$$\rho_t(h) = \frac{\gamma_t(h)}{\sqrt{\gamma_t(0)\,\gamma_{t-h}(0)}}.$$

4.2.2 Stationnarité

Définition 11 *Le processus $\{y_t\}$ est dit **stationnaire au sens faible**, ou **stationnaire au second ordre** si les premier (moyenne ou espérance mathématique) et second (variance et autocovariances) moments du processus existent et sont indépendants de t :*

$$\begin{aligned}\mathsf{E}[y_t] &= \mu_t < \infty, \quad \textit{pour tout } t \in \mathbb{T}, \\ \mathsf{E}\left[(y_t-\mu_t)\left(y_{t-h}-\mu_{t-h}\right)\right] &= \gamma_t(h), \quad \textit{pour tous } h \textit{ et } t.\end{aligned}$$

Lorsqu'y_t est stationnaire $\gamma_t(h) = \gamma_t(-h) = \gamma(h)$. La stationnarité est une propriété de stabilité, la distribution de y_t est identique à celle de y_{t-1}. la série oscille autour de sa moyenne avec une variance constante ; le lien entre y_t et y_{t-h} ne dépend alors que de l'intervalle h et non de la date t.

Définition 12 *Le processus $\{y_t\}$ est dit **strictement ou fortement stationnaire** si pour tous $h_1, ..., h_n$, la distribution conjointe de $(y_t, y_{t+h_1}, ..., y_{t+h_n})$ dépend uniquement des intervalles $h_1, ..., h_n$ et non de t :*

$$f(y_t, y_{t+h_1}, ..., y_{t+h_n}) = f(y_\tau, y_{\tau+h_1}, ..., y_{\tau+h_n}), \quad \forall (t,\tau).$$

La stationnarité stricte implique que tous les moments soient indépendants du temps.

Définition 13 *Le processus $\{y_t\}$ est appelé **Gaussien** si la distribution de $(y_t, y_{t+h_1}, ..., y_{t+h_n})$, notée*

$$f(y_t, y_{t+h_1}, ..., y_{t+h_n})$$

suit une loi Normale multivariée pour tous $h_1, ..., h_n$.

En pratique, pour les séries suivant une distribution Gaussienne, la stationnarité au sens faible est équivalente à la stationnarité au sens strict.

4.2.3 Ergodicité

Le théorème d'ergodicité statistique concerne l'information qui peut être obtenue à partir d'une moyenne sur le temps concernant la moyenne commune à tout instant. Remarquons que la loi faible des grands nombres ne s'applique pas car la série temporelle observée correspond à *une seule* observation du processus stochastique.

Définition 14 *Soit* $\{y_t(\omega), \omega \in \Omega, t \in \mathbb{T}\}$ *un processus stationnaire au sens faible, tel que* $\mathsf{E}[y_t(\omega)] = \mu < \infty$ *et* $\mathsf{E}\left[(y_t - \mu)^2\right] = \sigma_y^2 < \infty$ *pour tous* t*. Soit* $\overline{y}_t = T^{-1}\sum_{t=1}^{T} y_t$ *la moyenne temporelle. Si* $\overline{y}_t$ *converge en probabilité vers* μ *quand* $T \to \infty$*, alors* $\{y_t(\omega)\}$ *est* ***ergodique*** *pour la moyenne.*

Il faut noter que le concept d'ergodicité repose sur une indépendance asymptotique, alors que la stationnarité concerne l'indépendance par rapport au temps du processus. Pour le type de processus considéré dans ce cours, l'un implique l'autre, mais il convient de noter qu'ils peuvent différer ainsi que dans l'exemple suivant.

Exemple 27 *On considère le processus stochastique* $\{y_t\}$ *défini par :*

$$y_t = \begin{cases} u_0 & \text{à } t = 0, \text{ avec } u_0 \sim \mathsf{N}(0, \sigma^2); \\ y_{t-1} & \text{pour } t > 0. \end{cases}$$

Alors, $\{y_t\}$ *est strictement stationnaire mais non ergodique.*

Démonstration : Clairement, $y_t = u_0$ pour tout $t \geq 0.$, et ainsi :

$$\begin{aligned} \mathsf{E}[y_t] &= \mathsf{E}[u_0] = 0, \\ \mathsf{E}\left[y_t^2\right] &= \mathsf{E}\left[u_0^2\right] = \sigma^2, \\ \mathsf{E}[y_t y_{t-h}] &= \mathsf{E}\left[u_0^2\right] = \sigma^2, \end{aligned}$$

ce qui implique que $\{y_t\}$ soit stationnaire au sens faible, car $\mu = 0$, $\gamma(h) = \sigma^2$ et $\rho(h) = 1$ sont indépendants du temps.

L'ergodicité pour la moyenne nécessite que

$$\overline{y}_T = T^{-1}\sum_{t=1}^{T} y_t \xrightarrow{P} 0,$$

mais il est évident que $\overline{y}_T = T^{-1}\sum_{t=0}^{T-1} y_t = u_0$, qui est, pour la série observée, une réalisation d'une variable aléatoire de distribution Normale et donc ne tend pas vers zéro. ■

Pour être ergodique, la mémoire d'un processus stochastique doit diminuer de façon à ce que la covariance entre des observations de plus en plus distantes converge vers zéro de manière suffisament rapide. Pour les processus stationnaires, il est possible de démontrer que l'absolue sommabilité des covariances (i.e. $\sum_{h=0}^{\infty}|\gamma(h)|<0$) est une condition suffisante pour obtenir l'ergodicité.

De manière similaire, il est possible de définir l'ergodicité pour les seconds moments :

$$\widehat{\gamma}(h)=(T-h)^{-1}\sum_{t=h+1}^{t}(y_t-\mu_t)\left(y_{t-h}-\mu_{t-h}\right)\xrightarrow{P}\gamma(h).$$

4.3 La caractérisation des séries temporelles en économie

Le caractère d'ergodicité autorise la caractérisation des processus stochastiques par leur moments empiriques :

4.3.1 Moyenne de l'échantillon

$$\overline{y}=T^{-1}\sum_{t=0}^{T-1}y_t,$$

en notant que si $y_t\sim\mathsf{i.i.d}(\mu,\sigma^2)$ (identiquement et indépendamment distribuée de moyenne μ et de variance σ^2), le théorème limite central implique que :

$$\sqrt{T}\overline{y}\xrightarrow{\mathcal{L}}\mathsf{N}\left(\mu,\sigma^2\right).$$

4.3.2 ACF, fonction empirique d'autocorrélation

On la définit par $\widehat{\rho}_h=\widehat{\gamma}_h/\widehat{\gamma}_0$, où :

$$\widehat{\gamma}_h=T^{-1}\sum_{t=h+1}^{T}(y_t-\overline{y})(y_{t-h}-\overline{y}),$$

qui est un estimateur de la fonction d'autocovariance. Dans le cas d'une série stationnaire, la fonction d'autocorrélation décroit exponentiellement vers zéro. La décroissance est davantage linéaire pour les séries non stationnaires rencontrées en pratique.

4.3.3 PACF, fonction empirique d'autocorrélation partielle

Il s'agit de la suite de valeurs $\alpha_h^{(h)} = \mathsf{Corr}\left(y_t, y_{t-h}|y_{t-1}, ..., y_{t-h+1}\right)$, qui correspondent au dernier coefficient dans une régression linéaire de y_t sur une constante et ses h dernières valeurs :

$$y_t = \alpha_0^{(h)} + \alpha_1^{(h)} y_{t-1} + \alpha_2^{(h)} y_{t-2} + ... + \alpha_h^{(h)} y_{t-h} + \epsilon_t.$$

Le passage de l'autocorrélation à l'autocorrélation partielle se fait grâce à la relation :

$$\begin{bmatrix} \alpha_1^{(h)} \\ \vdots \\ \alpha_h^{(h)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \cdots & \gamma_{h-1} \\ \vdots & \ddots & \vdots \\ \gamma_{h-1} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_0 \\ \vdots \\ \gamma_h \end{bmatrix}.$$

Exemple 28 *On représente figure 4.3, l'inflation française (π_t) et le taux d'intérêt au jour le jour (i_t) en logarithmes ainsi que leur première différence. On constate que si l'inflation et le taux d'intérêt n'évoluent pas de manière stable autour d'une moyenne, leur première différence sont quant à elles nettement plus stables. Sous réserve d'un test précis indiquant que les différences des variables sont ici stationnaires tandis que les niveaux de celles-ci ne le sont pas, on parle alors de variables intégrées (i.e. y_t non stationnaire avec Δy_t stationnaire), ce qu'on note $\pi_t \sim \mathsf{I}(1)$ (variable intégrée d'ordre 1 et par extension variable intégrée) et $\Delta\pi_t \sim \mathsf{I}(0)$ (variable intégrée d'ordre zéro ou stationnaire).*

Les ACF des variables non-stationnaires (figure 4.4) décroissent lentement tandis que ceux des variables stationnaires oscillent autour de zéro. En revanche, la première valeur d'une PACF d'une variable intégrée d'ordre 1 est très proche de l'unité, tandis que les valeurs suivantes tendent très rapidement vers zéro. Les valeurs des PACF des variables stationnaires sont toujours différentes de 1.

4.4 Processus intégrés

Une classe importante de processus non-stationnaires est celle des processus intégrés. On les retrouve couramment en pratique et ils ont l'avantage de présenter un type de non-stationnarité modélisable.

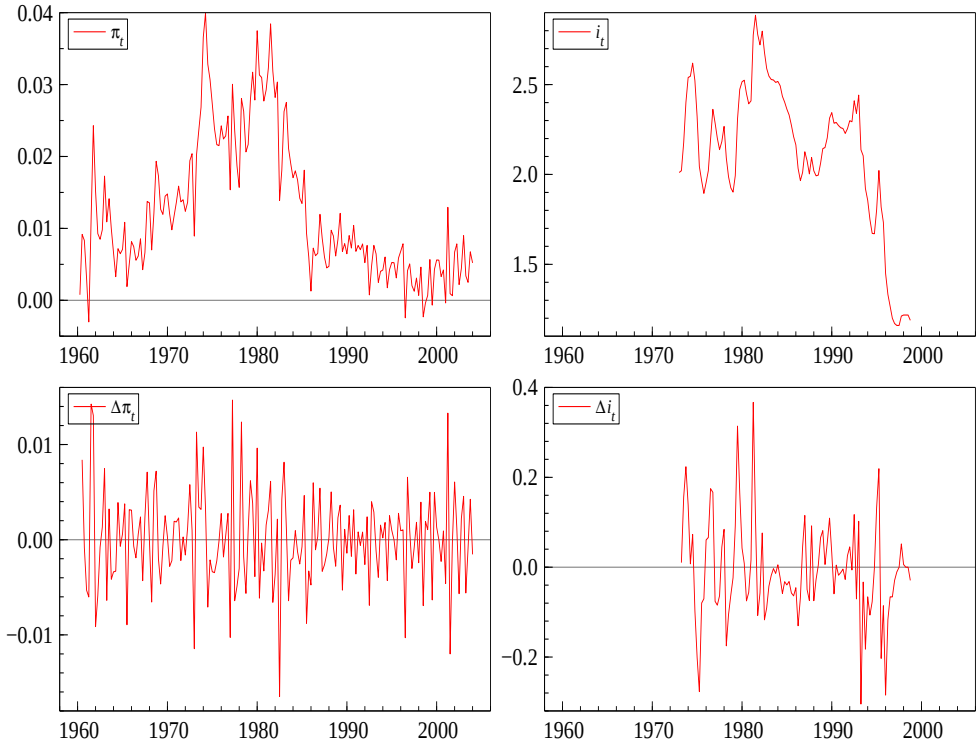


Figure 4.3 – Inflation (π_t) et taux d'intérêt au jour le jour (i_t) français en logarithmes ainsi que leur première différence ($\Delta\pi_t = \pi_t - \pi_{t-1}$). Source DataInsight.

Définition 15 *Un* ***processus intégré*** *est un processus qui peut être rendu stationnaire par* différentiation. *Si un processus stochastique doit être différentié d fois pour atteindre la stationnarité, il est dit être intégré d'ordre d, ou $I(d)$. Une marche aléatoire est intégrée d'ordre 1, ou $I(1)$, les processus stationnaires sont $I(0)$. Par extension on parle de séries intégrées quand leur ordre d'intégration est supérieur ou égal à 1.*

Exemple 29 *Le processus stochastique $\{y_t\}$ est appelé* ***marche aléatoire*** *si*

$$y_t = y_{t-1} + u_t, \text{ pour } t > 0 \text{ et } y_0 = 0,$$

où u_t est indépendamment et identiquement distribué avec une moyenne nulle et une variance $\sigma^2 < \infty$ pour tout t. La marche aléatoire est non-stationnaire et, par conséquent, non-ergodique. La figure 4.5 présente des exemples de marche aléatoire.

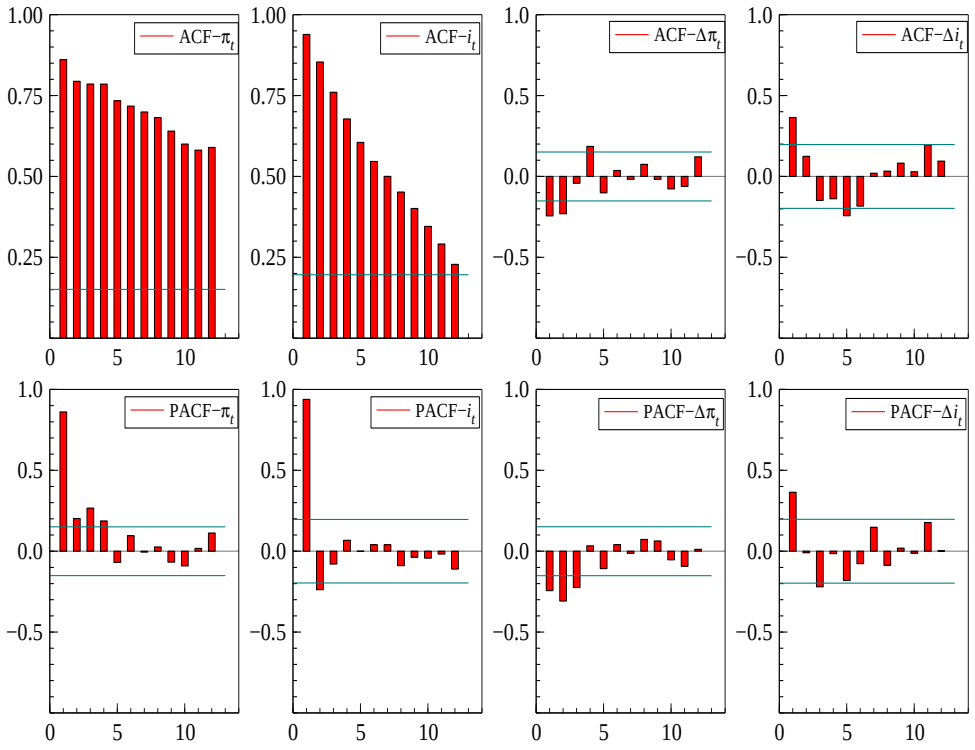


Figure 4.4 – Fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF) de l'inflation (π_t) et des taux d'intérêt au jour le jour (i_t) en logarithmes ainsi que de leur première différence.

Démonstration : Par substitution itérative, on obtient :

$$y_t = u_0 + \sum_{s=1}^{t} u_s \text{ pour tout } t > 0.$$

La moyenne est indépendante du temps :

$$\begin{aligned} \mu &= \mathsf{E}[y_t] = \mathsf{E}\left[y_0 + \sum_{s=1}^{t} u_s\right] \\ &= y_0 + \sum_{s=1}^{t} \mathsf{E}[u_s] = 0. \end{aligned}$$

Mais les moments d'ordre 2 divergent. La variance est donnée par

$$\begin{aligned}
\gamma_t(0) &= \mathsf{E}\left[y_t^2\right] = \mathsf{E}\left[\left(y_0 + \sum_{s=1}^{t} u_s\right)^2\right] = \mathsf{E}\left[\left(\sum_{s=1}^{t} u_s\right)^2\right] \\
&= \mathsf{E}\left[\sum_{s=1}^{t}\sum_{k=1}^{t} u_s u_k\right] = \mathsf{E}\left[\sum_{s=1}^{t} u_s^2 + \sum_{s=1}^{t}\sum_{k\neq s}^{t} u_s u_k\right] \\
&= \sum_{s=1}^{t} \mathsf{E}\left[u_s^2\right] + \sum_{s=1}^{t}\sum_{k\neq s}^{t} \mathsf{E}\left[u_s u_k\right] \\
&= \sum_{s=1}^{t} \sigma^2 = t\sigma^2.
\end{aligned}$$

Les autocovariances sont :

$$\begin{aligned}
\gamma_t(h) &= \mathsf{E}\left[y_t y_{t-h}\right] = \mathsf{E}\left[\left(y_0 + \sum_{s=1}^{t} u_s\right)\left(y_0 + \sum_{k=1}^{t-h} u_k\right)\right] \\
&= \mathsf{E}\left[\sum_{s=1}^{t} u_s\left(y_0 + \sum_{k=1}^{t-h} u_k\right)\right] = \sum_{k=1}^{t-h} \mathsf{E}\left[u_k^2\right] \\
&= \sum_{k=1}^{t-h} \sigma^2 = (t-h)\,\sigma^2, \text{ pour tout } h > 0.
\end{aligned}$$

Et, en conclusion, la fonction d'autocorrélation $\rho_t(h)$, pour $h > 0$, est donnée par :

$$\rho_t^2(h) = \frac{\gamma_t^2(h)}{\gamma_t(0)\,\gamma_{t-h}(0)} = \frac{\left[(t-h)\,\sigma^2\right]^2}{\left[t\sigma^2\right]\left[(t-h)\,\sigma^2\right]} = 1 - \frac{h}{t}, \text{ pour tout } h > 0.$$

■

Les graphiques 4.5 et 4.6 présentent des exemples de séries intégrées respectivement d'ordre 1 et 2. Ces graphiques sont représentatifs de séries de ce type : les variables $\mathsf{I}(1)$ sont en général erratiques et peuvent prendre n'importe quelle valeur, on parle alors de tendance stochastique car elles se comportent comme si pendant de courtes périodes elles suivaient une tendance déterminée, mais cette dernière change elle-même irrégulièrement. Il est en revanche aisé de repérer à l'oeil nu des variables $\mathsf{I}(2)$: elles sont en général très lisses et présentent une réelle direction, soit à la hausse, soit à la baisse mais n'évoluent que très lentement.

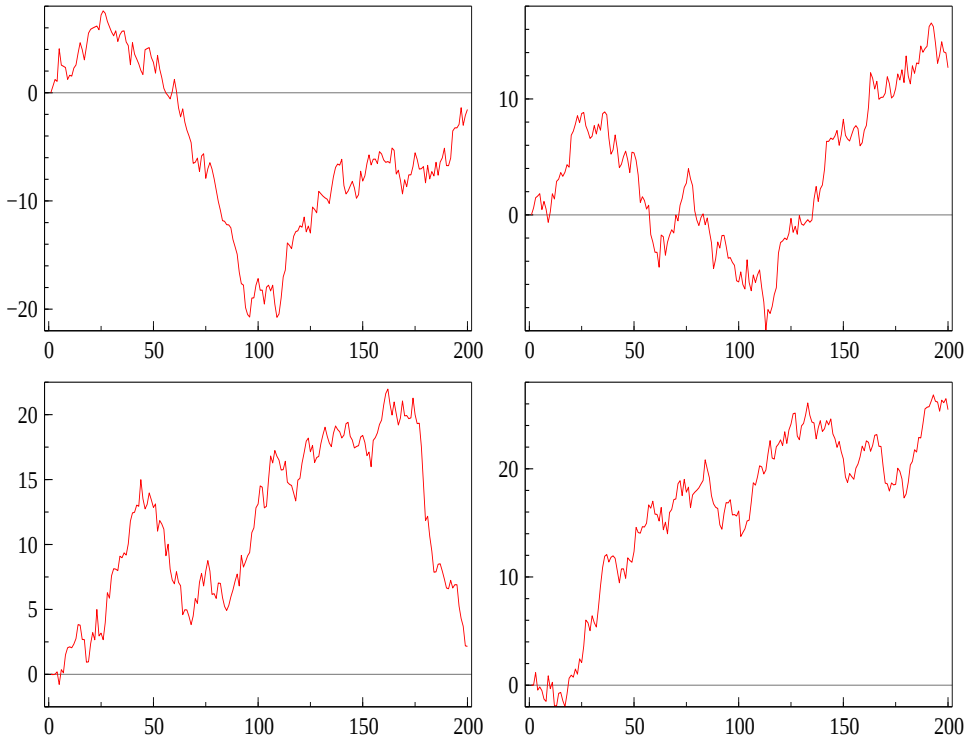


Figure 4.5 – Quatre exemples simulés de variables intégrées d'ordre 1 et de valeur initiale nulle

4.5 Quelques processus courants

Définition 16 *Un* ***bruit blanc*** *(white noise) est un processus stationnaire au sens faible de moyenne zéro et qui est dynamiquement non-corrélé :*

$$u_t \sim \mathsf{WN}\left(0, \sigma^2\right).$$

Ainsi, $\{u_t\}$ *est un bruit blanc si pour tout* $t \in \mathbb{T}$ *:* $\mathsf{E}[u_t] = 0$, $\mathsf{E}[u_t^2] = \sigma^2 < \infty$, *avec* u_t *et* u_{t-h} *indépendants si* $h \neq 0$, t *et* $(t-h) \in \mathbb{T}$.

Définition 17 *Si le bruit blanc* $\{u_t\}$ *est distribué Normalement, on parle de* ***bruit blanc Gaussien :***

$$u_t \sim \mathsf{NID}\left(0, \sigma^2\right).$$

l'hypothèse d'indépendance est alors équivalent à celle de non corrélation : $\mathsf{E}[u_t u_{t-h}] = 0$ *si* $h \neq 0$, t *et* $(t-h) \in \mathbb{T}$.

Noter que l'hypothèse de normalité implique l'indépendance dynamique. Une généralisation des processus NID : les IID avec moments d'ordre supérieur constants mais non précisés.

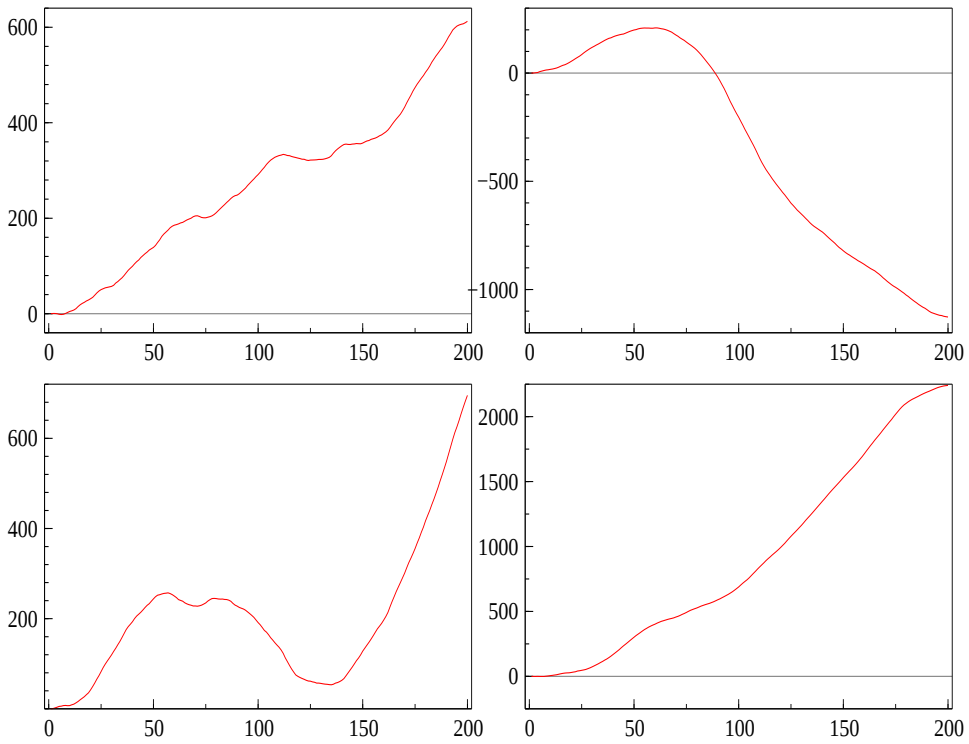


Figure 4.6 – Quatre exemples simulés de variables intégrées d'ordre 2 et de valeur initiale nulle

Définition 18 *Un processus* $\{u_t\}$ *de composantes distribuées indépendamment et identiquement est noté* IID *:*

$$u_t \sim \mathsf{IID}\left(\mu, \sigma^2\right).$$

Tous les u_t *sont issus de la même distribution d'espérance* μ *et de variance* σ^2*, avec* u_t *et* u_{t-h} *indépendants si* $h \neq 0$*,* t *et* $(t-h) \in \mathbb{T}$*.*

Chapitre 5

Méthodes sans modèle

Avant de présenter les modèles stochastiques des séries temporelles, ceux qui sont estimables et dont la précision peut être testée, il en existe d'autres qui sont parfois utilisés pour modifier les données afin de pouvoir les modéliser (voire tâcher de les prévoir).

5.1 Extrapolation déterministe des séries

Les modèles sont dits déterministes lorsque leurs valeurs futures sont connues avec certitude à tout instant. Ainsi ne font-ils aucune référence aux sources d'incertitudes et de hasard des processus stochastiques. Les méthodes purement déterministes apportent une simplicité au détriment de la précision et ne permettent pas d'établir de *quantification* de l'incertitude via, par exemple, **un intervalle de confiance**.

Si on dispose d'un échantillon de T observations d'une série : $y_1, y_2, ..., y_{T-1}, y_T$, il existe un polynôme de degré $n = T - 1$ qui passe par tous les points y_t :

$$f(t) = a_o + a_1 t + a_2 t^2 + ... + a_n t^n. \tag{5.1}$$

Malheureusement, rien ne dit que $f(T+1) = \widehat{y}_{T+1}$ soit proche de y_{T+1}. Ainsi (5.1) ne *décrit* pas y_t, il ne fait que le *reproduire* et ne capture aucune des caractéristisques qui risquent d'apparaître à l'avenir.

5.1.1 Tendances linéaires

Une caractéristique simple de y_t est sa tendance de long terme : si on pense qu'une tendance à la hausse existe et va perdurer, il est possible de construire un modèle simple qui permette de prévoir y_t. Le plus simple consiste en une

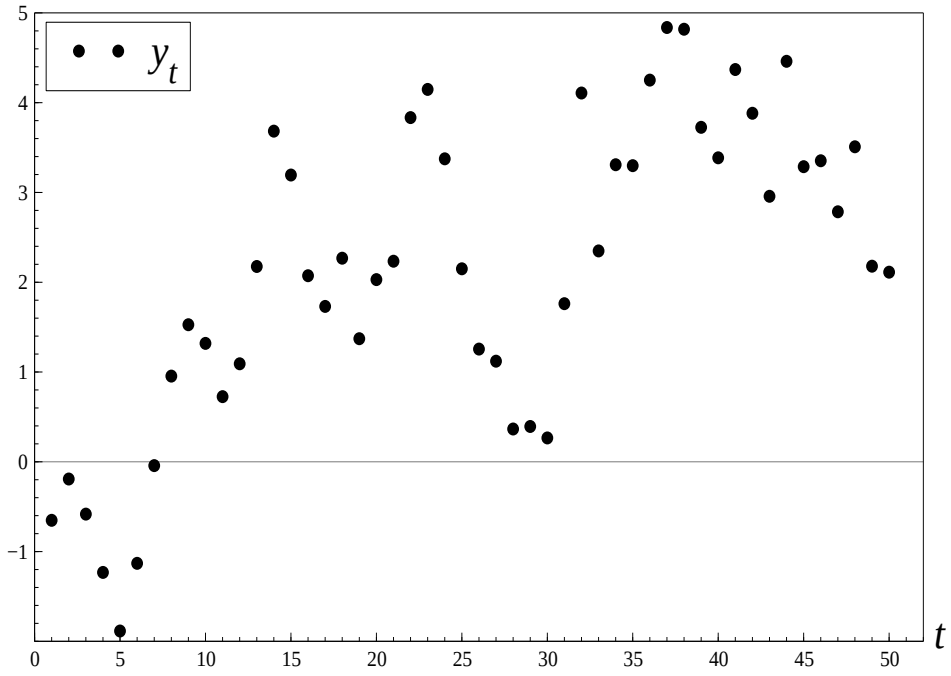


Figure 5.1 – Exemple de série temporelle semblant présenter une tendance à la hausse.

tendance linéaire selon laquelle la série va s'accroître du même montant à chaque période :

$$\begin{aligned} y_t &= a + bt, \\ \Delta y_t &= y_t - y_{t-1} = b. \\ \widehat{y}_{T+h} &= a + b\,(T+h)\,. \end{aligned}$$

Il peut sembler plus réaliste de penser que y_t va s'accroître du même pourcentage à chaque période, auquel cas une tendance exponentielle s'impose :

$$y_t = Ae^{rt}$$

ce qui donne une relation log-linéaire :

$$\log y_t = \log A + rt$$

et le taux de croissance est :

$$\partial/\partial t\,[\log y_t] = \frac{\partial y_t/\partial t}{y_t} = r.$$

5.1.2 Tendances autorégressives

Ici la valeur à t dépend de la valeur précédente :

$$y_t = a + by_{t-1}$$

selon les valeurs de b et a, le comportement de la série diffère. Si $a = 0$ et $|b| \neq 1$, b est le taux de croissance de la série, en revanche si $b = 1$, y_t suit une tendance déterministe.

5.1.3 Modèles non linéaires

Tendance quadratique

$$y_t = a + bt + ct^2$$

Courbe logistique

$$y_t = \frac{1}{k + ab^t}, \quad b > 0$$

5.2 Moyennes mobiles

Il existe deux types de moyenne mobile, l'un sera vu plus loin et correspond au modèle MA, l'autre est davantage une méthode ad hoc permettant de donner une estimation des "alentours" d'une série, on suppose alors que la variable sera proche de sa moyenne récente. Une moyenne mobile est alors simplement une moyenne sur une fenêtre glissante d'observations :

$$\overline{y}_t^{(m)} = \frac{1}{m} \sum_{i=1}^{m} y_{t+k-i},$$

où k est librement fixé selon les besoins du modélisateur, pour une prévision, il est nécessaire que $k \leq 0$.

Exemple 30 *Si les données sont de fréquence mensuelle, et qu'on souhaite prévoir y_t, il est possible d'utiliser la fonction de moyenne mobile*

$$f(t) = \frac{1}{12} \sum_{i=1}^{12} y_{t-i},$$

qui fournit la prévision :

$$\widehat{y}_{T+1} = f(T)$$

qui est la moyenne des 12 dernières observations.

Il peut paraître peu réaliste que la prochaine valeur y_{T+1} puisse être proche d'une simple moyenne des dernières observations. Si on souhaite accorder plus de poids aux observations les plus récentes, on peut utiliser le modèle EWMA (Exponentially Weighted Moving Average) selon lequel :

$$\begin{aligned}\widehat{y}_{T+1} &= \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} .... \\ &= \alpha \sum_{i=0}^{\infty} (1-\alpha)^i y_{T-i},\end{aligned}$$

où α est compris entre 0 et 1 et indique l'importance accordée aux observations les plus récentes, si $\alpha = 1$:

$$\widehat{y}_{T+1} = y_T.$$

Notons qu'il s'agit bien d'une moyenne puisque la somme des coefficients est unitaire :

$$\alpha \sum_{i=0}^{\infty} (1-\alpha)^i = 1$$

Le modèle EWMA se prête mal aux variables présentant une tendance de fond à la hausse ou à la baisse, car il va dans ces cas sous- ou sur-prédire. Il est en revanche possible de l'appliquer à une série dont on a ôté la tendance.

Pour une prévision à horizon $h > 1$, il semble logique d'étendre

$$\widehat{y}_{T+h} = \alpha \sum_{i=1}^{h-1} (1-\alpha)^{i-1} \widehat{y}_{T+h-i} + \alpha \sum_{i=0}^{\infty} (1-\alpha)^{h-1+i} y_{T-i}$$

ce qui donne

$$\widehat{y}_{T+h} = \alpha \sum_{i=0}^{\infty} (1-\alpha)^i y_{T-i}$$

et ainsi le modèle EWMA fournit la même prévision à tous horizons.

5.3 Lissages

Les méthodes de lissage ont pour but de retirer ou de réduire les fluctuations (cycliques ou non) de court terme des séries.

5.3.1 Moyennes mobiles

Les moyennes mobiles présentées précédemment permettent aussi d'obtenir des séries lissées : par exemple en utlisant un moyenne mobile d'ordre n donnée par

$$\widetilde{y}_t = \frac{1}{n}\sum_{i=0}^{n-1} y_{t-i}. \tag{5.2}$$

Plus n est élevé, plus la série sera lissée. Le problème de (5.2) est de n'utiliser que les valeurs passées et présentes. Pour y remédier, on peut faire appel à une moyenne mobile centrée :

$$\widetilde{y}_t = \frac{1}{2k+1}\sum_{i=-k}^{k} y_{t+i}$$

5.3.2 Lissage exponentiel

Le lissage exponentiel fait appel aux modèles EWMA :

$$\widetilde{y}_t = \alpha y_t + \alpha\left(1-\alpha\right) y_{t-1} + \alpha\left(1-\alpha\right)^2 y_{t-2} + ...\alpha\left(1-\alpha\right)^{t-1} y_1. \tag{5.3}$$

En pratique, il est plus facile d'écrire :

$$\left(1-\alpha\right)\widetilde{y}_{t-1} = \alpha\left(1-\alpha\right) y_{t-1} + \alpha\left(1-\alpha\right)^2 y_{t-2} + ...\alpha\left(1-\alpha\right)^{t-1} y_1 \tag{5.4}$$

et en soustrayant (5.4) à (5.3) on obtient la formule de récurrence du lissage exponentiel simple :

$$\widetilde{y}_t = \alpha y_t + \left(1-\alpha\right)\widetilde{y}_{t-1} \tag{5.5}$$

Plus α est proche de zéro, plus la série est lissée. En pratique toutefois, on peut souhaiter effectuer un lissage important mais sans donner trop de poids aux observations lointaines. On applique pour ce faire un lissage exponentiel double, i.e. en réappliquant la formule à $\widetilde{y}_t$ pour obtenir

$$\widetilde{\widetilde{y}}_t = \alpha\widetilde{y}_t + \left(1-\alpha\right)\widetilde{\widetilde{y}}_{t-1}$$

avec une valeur plus élevée de α.

Enfin il est possible d'appliquer (5.5) aux changements moyens de la tendance de long terme de la série en utilisant la formule de lissage exponentiel à deux paramètres de Holt-Winters :

$$\begin{aligned}\widetilde{y}_t &= \alpha y_t + \left(1-\alpha\right)\left(\widetilde{y}_{t-1} + r_{t-1}\right)\\ r_t &= \gamma\left(\widetilde{y}_t - \widetilde{y}_{t-1}\right) + \left(1-\gamma\right) r_{t-1},\end{aligned}$$

où r_t est la série lissée représentant la tendance, i.e. le taux moyen de croissance. Cette tendance est ajoutée lors du lissage afin d'éviter que $\widetilde{y}_t$ ne s'éloigne trop des valeurs récentes de la série originale y_t. Une prévision à horizon h peut être obtenue en posant

$$\widehat{y}_{T+h} = \widetilde{y}_T + hr_T$$

5.4 Ajustements saisonniers

Il existe diverses méthodes de correction des variations saisonnières. Elles fonctionnent pour la plupart sur une décomposition entre tendance sous-jacente et variations saisonnières de la forme (ici multiplicative) :

$$Y_t = L \times S \times C \times I,$$

avec L la valeur de long terme, S le coefficient saisonnier, C le cycle saisonnier, et I une composante irrégulière. Il est aussi possible de décomposer y_t sous forme de comportement saisonnier additif $y_t = L + I + S + C$. Le but des CVS est d'isoler $S \times I$, mais comme ceci n'est pas possible de manière exacte, une méthode de lissage ad hoc doit être utilisée. Quand l'inspection des séries laisse à penser que la variation d'amplitude est constante en valeur, une méthode additive convient, si la variation est constante en pourcentage de la moyenne annuelle, il est préférable de recourir à une désaisonnalisation multiplicative.

5.4.1 Méthode multiplicative

1. Calculer la moyenne mobile centrée de y_t :

$$x_t = \begin{cases} \left(0.5y_{t+6} + y_{t+5} + ... + y_t + ... + y_{t-5} + 0.5y_{t-6}\right)/12 & \text{pour des séries mensuelles} \\ \left(0.5y_{t+2} + y_{t+1} + y_t + y_{t-1} + 0.5y_{t-2}\right)/4 & \text{pour des séries trimestrielles} \end{cases}$$

2. Calculer le ratio $r_t = y_t/x_t$

3. Calculer les indices saisonniers : pour la période m ou q c'est la moyenne des r_t en n'utlisant que le mois m ou le trimestre q (par ex. tous les mois de janvier). Puis ajuster les indices saisonniers pour que leur produit soit égal à 1 :

$$s_{m \text{ ou } q} = \begin{cases} i_m / \sqrt[12]{i_1 i_2 ... i_{12}} & \text{pour des séries mensuelles} \\ i_q / \sqrt[4]{i_1 i_2 i_3 i_4} & \text{pour des séries trimestrielles} \end{cases}$$

La série y_t est ainsi $s_j\%$ supérieure à la série ajustée en période j.

4. Diviser y_t par s_j pour obtenir la série CVS.

5.4.2 Méthode additive

1. Calculer la moyenne mobile centrée de y_t :

$$x_t = \begin{cases} (0.5y_{t+6} + y_{t+5} + ... + y_t + ... + y_{t-5} + 0.5y_{t-6})/12 & \text{pour des séries mensuelles} \\ (0.5y_{t+2} + y_{t+1} + y_t + y_{t-1} + 0.5y_{t-2})/4 & \text{pour des séries trimestrielles} \end{cases}$$

2. Calculer la différence $d_t = y_t - x_t$

3. Calculer les indices saisonniers : pour la période m ou q c'est la moyenne des r_t en n'utlisant que le mois m ou le trimestre q (par ex. tous les mois de janvier). Puis ajuster les indices saisonniers pour que leur somme soit égal à zéro :

$$s_{m \text{ ou } q} = \begin{cases} i_m - \frac{1}{12}\sum i_i & \text{pour des séries mensuelles} \\ i_m - \frac{1}{4}\sum i_i & \text{pour des séries trimestrielles} \end{cases}$$

La série y_t est supérieure à la série ajustée en période j de s_j.

4.La série CVS est donnée par $y_t - s_j$.

Chapitre 6

Modèles linéaires de séries temporelles

6.1 Processus linéaires

6.1.1 Concepts

On appelle processus linéaire toute série temporelle qui puisse être représentée par un modèle linéaire après transformation, par exemple $\log(y_t) = \alpha + \beta t + \varepsilon_t$. Il est toujours surprenant de constater la simplicité de la plupart des modèles linéaires quand on pense à la complexité des modèles dans d'autres disciplines (physique...). En réalité, un modèle linéaire est une approximation (proche de la notion de développement limité) de modèles nettement plus complexes et ils ont la particularité d'être très flexibles et estimables avec un faible nombre d'observations.

Opérateur retard

Soit $\{y_t\}$ un processus stochastique. On définit l'opérateur retard (lag, ou backshift ou backward, operator) L (ou B) tel que

$$\begin{aligned} Ly_t &= y_{t-1}, \\ L^j y_t &= y_{t-j}, \quad \text{pour tout } j \in \mathbb{N}. \end{aligned}$$

et pour c scalaire, $Lc = c$. On peut utiliser l'opérateur L comme un chiffre, il peut multiplier et diviser : si $Ly_t = y_{t-1}$, alors $y_t = L^{-1}y_{t-1}$, ce qu'on note parfois $y_t = L^{-1}y_{t-1} = Fy_{t-1}$ (F forward shift ou opérateur avancé).

Opérateur de différence

Si $\{y_t\}$ est un processus stochastique, les processus suivants existent aussi

$$\begin{aligned}\Delta y_t &= (1-L)\,y_t = y_t - y_{t-1},\\ \Delta^j y_t &= (1-L)^j\,y_t \quad \text{pour tout } j\in\mathbb{N},\\ \Delta_s y_t &= (1-L^s)\,y_t = y_t - y_{t-s} \quad \text{“différence saisonnière” ou glissement.}\end{aligned}$$

Filtre linéaire

Transformation d'une série entrée, $\{x_t\}$, en série sortie $\{y_t\}$ par application du polynôme retard $A(L)$:

$$y_t = A(L)\,x_t = \left(\sum_{j=-n}^{m} a_j L^j\right) x_t = \sum_{j=-n}^{m} a_j x_{t-j} = a_{-n}x_{t+n}+...+a_0x_0+...+a_mx_{t-m}.$$

Exemple 31 *moyenne mobile, moyenne mobile centrée, lissage exponentiel....*

Processus linéaire

Un processus $\{y_t\}$ est dit linéaire s'il existe une série $\{\epsilon_t\}_{t\in\mathbb{R}}$ telle que $\{y_t\}$ puisse être représenté par

$$y_t = A(L)\,\epsilon_t = \left(\sum_{j=-\infty}^{\infty} a_j L^j\right)\epsilon_t = \sum_{j=-\infty}^{\infty} a_j\epsilon_{t-j} \quad où \quad \epsilon_t \sim \mathsf{WN}\left(0,\sigma^2\right).$$

6.1.2 Théorème de décomposition de Wold

Théorème 9 ***(Décomposition de Wold)*** *Tout processus stationnaire au sens faible et de moyenne zéro, $\{y_t\}$, admet la représentation suivante :*

$$y_t = \sum_{j=0}^{\infty}\psi_j\epsilon_{t-j} + \kappa_t,$$

où $\psi_0 = 1$ et $\sum_{j=0}^{\infty}\psi_j^2 < \infty$. Le terme ϵ_t is un bruit blanc qui représente l'erreur faite en prévoyant y_t à partir d'une fonction linéaire de son historique $Y_{t-1} = \{y_{t-j}\}_{j=1}^{\infty}$:

$$\epsilon_t = y_t - \mathsf{E}\left[y_t|Y_{t-1}\right].$$

La variable κ_t est non-corrélée aux ϵ_{t-j}, pour tous $j\in\mathbb{Z}$, bien que κ_t puisse être prévue arbitrairement bien à partir d'une fonction linéaire de Y_{t-1} :

$$\kappa_t = E\left[\kappa_t|Y_{t-1}\right].$$

$\sum_{j=0}^{\infty}\psi_j\epsilon_{t-j}$ est la composante linéaire stochastique et κ_t est la composante linéaire déterministe de y_t.

6.1.3 Modélisation ARMA

Une approximation du polynôme retard d'ordre infini est obtenue à partir d'un ratio de deux polynômes d'ordre fini, $\alpha(L)$ et $\beta(L)$ tels que :

$$\Psi(L) = \sum_{j=0}^{\infty} \psi_j L^j \approx \frac{\beta(L)}{\alpha(L)} = \frac{1 + \beta_1 L + ... + \beta_q L^q}{1 - \alpha_1 L - ... - \alpha_p L^p}.$$

Typologie des modèles linéaires de séries temporelles :

p	q	Modèle	Type	
$p > 0$	$q = 0$	$\alpha(L) y_t = \epsilon_t$	autorégressif (pure) d'ordre p	AR(p)
$p = 0$	$q > 0$	$y_t = \beta(L) \epsilon_t$	moyenne mobile d'ordre q	MA(q)
$p > 0$	$q > 0$	$\alpha(L) y_t = \beta(L) \epsilon_t$	modèle mixte autorégressif–moyenne mobile	ARMA(p, q)

Processus Autorégressifs

Un modèle autorégressif d'ordre p, un modèle AR(p), satisfait l'équation différentielle suivante :

$$y_t = \nu + \sum_{j=1}^{p} \alpha_j y_{t-j} + \epsilon_t, \quad \text{où} \quad \epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right). \tag{6.1}$$

Ou en utilisant l'opérateur retard :

$$\alpha(L) y_t = \nu + \epsilon_t, \quad \text{où } \alpha(L) = 1 - \alpha_1 L - ... - \alpha_p L^p, \quad \alpha_p \neq 0.$$

Stabilité L'hypothèse $\alpha(z) = 0 \Rightarrow |z| > 1$ garantit la stationnarité et l'existence d'une représentation MA(∞) :

$$\begin{aligned} y_t &= \alpha(1)^{-1} \nu + [\alpha(L)]^{-1} \epsilon_t, \\ y_t &= \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \quad o\grave{u} \quad \mu = \frac{\nu}{\alpha(1)} \quad \text{et} \quad \Psi(L) = [\alpha(L)]^{-1} \quad \text{avec} \quad \sum_{j=0}^{\infty} \left|\psi_j\right| < \infty \end{aligned}$$

Analyse de stabilité est fondée sur l'équation différentielle non-homogène d'ordre p :

$$y_t - \mu = \sum_{j=1}^{p} \alpha_j \left(y_{t-j} - \mu\right),$$

$$\text{AR}(1) \quad : \quad \begin{cases} |\alpha_1| < 1 : & \textit{stable} \\ \alpha_1 = 1 : & \text{racine unitaire} \\ |\alpha_1| > 1 : & \textit{instable} \text{ (explosif)} \end{cases},$$

$$\text{AR}(2) \quad : \quad \begin{cases} -1 < \alpha_2 < 1 - |\alpha_1| : & \textit{stabilité} \\ \\ \alpha_1^2 + 4\alpha_2 < 0 : & \text{racines complexes} \end{cases}$$

Fonction d'autocovariance Celle-ci peut s'obtenir analytiquement grâce à l'équation :

$$\begin{aligned}\gamma(h) &= \mathsf{E}[y_t y_{t-h}] = \mathsf{E}[(\alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \epsilon_t) y_{t-h}] \\ &= \alpha_1 \mathsf{E}[y_{t-1} y_{t-h}] + ... + \alpha_p \mathsf{E}[y_{t-p} y_{t-h}] + \mathsf{E}[\epsilon_t y_{t-h}] \\ &= \alpha_1 \gamma(h-1) + ... + \alpha_p \gamma(h-p).\end{aligned}$$

Il suffit alors de résoudre un système. Pour ce qui concerne la fonction d'autocorrélation, il est facile de l'obtenir directement à partir de l'équation différentielle (6.1), on obtient alors un système, nommé **équations de Yule–Walker :**

$$\left.\begin{array}{l}\rho_1 = \alpha_1 + \alpha_2\rho_1 + ... + \alpha_p\rho_{p-1} \\ \rho_2 = \alpha_1\rho_1 + \alpha_2 + ... + \alpha_p\rho_{p-2} \\ \qquad\vdots \\ \rho_p = \alpha_1\rho_{p-1} + \alpha_{p-2} + ... + \alpha_p\end{array}\right\} \Rightarrow \rho_1, ..., \rho_p$$
$$\rho_k = \alpha_1\rho_{k-1} + \alpha_{k-2} + ... + \alpha_{k-p} \quad \text{pour } k > p.$$

Processus de moyenne mobile

Un processus de moyenne mobile d'ordre q, noté MA(q), est caractérisé par :

$$y_t = \mu + \beta(L)\epsilon_t = \mu + \epsilon_t + \sum_{i=1}^{q} \beta_i \epsilon_{t-i},$$

où $\epsilon_t \sim \mathsf{WN}(0, \sigma^2)$ et $\beta(L) = 1 + \beta_1 L + ...\beta_q L^q, \ \beta_q \neq 0.$

Inversibilité Un processus MA est toujours stationnaire, en revanche il existe toujours deux processus MA fournissant les mêmes observations, leurs racines étant inverses les unes des autres. On impose par conséquent une condition dite d'inversibilité, $\beta(z) = 0 \Rightarrow |z| > 1$ qui garantit l'unicité et l'existence d'une représentation AR(∞) :

$$\begin{aligned}\beta(L)^{-1} y_t &= \beta(1)^{-1}\mu + \epsilon_t \\ y_t &= \mu + \sum_{j=1}^{\infty} \phi_j (y_{t-j} - \mu) + \epsilon_t,\end{aligned}$$

où $\phi(L) = 1 - \sum_{j=1}^{\infty} \phi_j L^j = 1 - \phi_1 L - \phi_2 L^2 + ... = \beta(L)^{-1}.$

Fonction d'autocovariance On considère $z_t = y_t - \mu = \sum_{i=0}^{q} \beta_i \epsilon_{t-i}$, $\beta_0 = 1$

$$\begin{aligned}
\gamma_0 &= \left(\sum_{i=0}^{q} \beta_i^2 \right) \sigma^2 \\
\gamma_k &= \left(\sum_{i=0}^{q-k} \beta_i \beta_{i+k} \right) \sigma^2, \quad \text{pour} \quad k = 1, 2, ..., q \\
\gamma_k &= 0, \quad \text{pour } k > q.
\end{aligned}$$

Processus mixtes

Un processus ARMA(p, q) comprend un terme autorégressif et un de moyenne mobile :

$$\begin{aligned}
\alpha(L) y_t &= \nu + \beta(L) \varepsilon_t, \\
y_t &= \nu + \sum_{j=1}^{p} \alpha_j y_{t-j} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i},
\end{aligned}$$

où $\varepsilon_t \sim \mathsf{WN}(0, \sigma^2)$, $\alpha(L) = 1 - \alpha_1 L - ... - \alpha_p L^p$, $\alpha_p \neq 0$, et $\beta(L) = 1 + \beta_1 L + ... + \beta_q L^q$, avec $\beta_q \neq 0$.

Stabilité $\alpha(z) = 0 \Rightarrow |z| > 1$ garantit la stationarité (au sens faible) et l'existence d'une représentation MA(∞) :

$$\begin{aligned}
y_t &= \alpha(1)^{-1} \nu + \alpha(L)^{-1} \beta(L) \varepsilon_t \\
&= \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}
\end{aligned}$$

Inversibilité $\beta(z) = 0 \Rightarrow |z| > 1$ permet une représentation AR(∞) :

$$\begin{aligned}
\beta(L)^{-1} \alpha(L) (y_t - \mu) &= \varepsilon_t \\
y_t &= \mu + \sum_{j=1}^{\infty} \phi_j (y_{t-j} - \mu) + \varepsilon_t
\end{aligned}$$

Unicité Cette propriété requiert l'absence de racines communes entre $\alpha(L)$ et $\beta(L)$:

$$\left.\begin{array}{l} \alpha(L) = \prod_{j=1}^{p} (1 - \lambda_j L) \\ \beta(L) = \prod_{i=1}^{p} (1 - \mu_i L) \end{array}\right\} \Rightarrow \lambda_j \neq \mu_i \text{ pour tous } i \text{ et } j.$$

Dans le cas contraire un ARMA(p,q) : $\alpha(L)\,y_t = \nu + \beta(L)\,\varepsilon_t$, pourrait se réécrire, pour tout polynôme $\delta(L)$ d'ordre r :

$$\begin{aligned}\delta(L)\,\alpha(L)\,y_t &= \delta(L)\,\nu + \delta(L)\,\beta(L)\,\varepsilon_t \\ &= \delta(1)\,\nu + \delta(L)\,\beta(L)\,\varepsilon_t,\end{aligned}$$

et donc

$$y_t \sim \text{ARMA}(p+r, q+r).$$

Fonction d'autocovariance Celle-ci est alors plus difficile à calculer analytiquement, on fait appel à une reparamétrisation utile :

$$\alpha(L)\,y_t = u_t, \text{ avec } u_t = \beta(L)\,\varepsilon_t.$$

On obtient alors la fonction d'autocovariance :

$$\begin{aligned}\gamma(h) &= \mathsf{E}[y_t y_{t-h}] = \mathsf{E}[(\alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + u_t)\,y_{t-h}] \\ &= \alpha_1 \mathsf{E}[y_{t-1}y_{t-h}] + ... + \alpha_p \mathsf{E}[y_{t-p}y_{t-h}] + \mathsf{E}[u_t y_{t-h}] \\ &= \alpha_1\gamma(h-1) + ... + \alpha_p\gamma(h-p) + +\mathsf{E}\left[\left(\varepsilon_t + \beta_1\varepsilon_{t-1} + ... + \beta_q\varepsilon_{t-q}\right)y_{t-h}\right]\end{aligned}$$

et la variance :

$$\gamma(0) = \mathsf{E}\left[y_t^2\right] = \mathsf{E}\left[(\alpha_1 y_{t-1} + ... + \alpha_p y_{t-p})^2\right] + 2\mathsf{E}\left[(\alpha_1 y_{t-1} + ... + \alpha_p y_{t-p})\,u_t\right] + \mathsf{E}\left[u_t^2\right].$$

Exemple 32 *Calcul de la variance d'un ARMA*$(1,1)$:

$$\begin{aligned}\gamma(0) &= \mathsf{E}\left[y_t^2\right] = \mathsf{E}\left[(\alpha_1 y_{t-1})^2\right] + 2\mathsf{E}\left[(\alpha_1 y_{t-1})(\varepsilon_t + \beta_1\varepsilon_{t-1})\right] + \mathsf{E}\left[(\varepsilon_t + \beta_1\varepsilon_{t-1})^2\right] \\ &= \alpha_1^2\mathsf{E}\left[y_{t-1}^2\right] + 2\alpha_1\beta_1\mathsf{E}\left[y_{t-1}\varepsilon_{t-1}\right] + \mathsf{E}\left[\varepsilon_t^2 + \beta_1^2\varepsilon_{t-1}^2\right] \\ &= \alpha_1^2\gamma(0) + \left(1 + 2\alpha_1\beta_1 + \beta_1^2\right)\sigma_\varepsilon^2 \\ &= \left(1-\alpha_1^2\right)^{-1}\left(1 + 2\alpha_1\beta_1 + \beta_1^2\right)\sigma_\varepsilon^2.\end{aligned}$$

Méthode des coefficients indéterminés

Cette méthode peut s'utiliser pour le calcul des coefficients dans la représentation AR(∞) ou MA(∞) d'un processus ARMA d'ordre fini. Par exemple, pour l'obtention des coefficients de la représentation MA(∞) d'un ARMA(p,q) :

$$\begin{aligned}\Psi(L) &= a(L)^{-1}\,\beta(L) \\ &\Leftrightarrow\ a(L)\,\Psi(L) = \beta(L)\end{aligned}$$

Une approche possible consiste à poser $\Psi(L) \equiv \sum_{i=0}^{\infty}\psi_i L^i$ de sorte que

$$(1 - \alpha_1 L - ... - \alpha_p L^p)\left(\psi_0 + \psi_1 L + \psi_2 L^2 + ...\right) = \left(1 + \beta_1 L + ... + \beta_q L^q\right),$$

et à comparer les coeffcients des retards correspondants :

$$\begin{aligned}
L^0 &: \psi_0 = 1 \\
L^1 &: \psi_1 - \alpha_1\psi_0 = \beta_1 \Rightarrow \psi_1 = \beta_1 + \alpha_1\psi_0 \\
L^2 &: \psi_2 - \alpha_1\psi_1 - \alpha_2\psi_0 = \beta_2 \Rightarrow \psi_2 = \beta_2 + \alpha_1\psi_1 + \alpha_2\psi_0 \\
L^h &: -\sum_{i=0}^{h} \alpha_i\psi_{h-i} = \beta_h \Rightarrow \psi_h = \beta_h + \sum_{i=1}^{h} \alpha_i\psi_{h-i},
\end{aligned}$$

avec $\alpha_0 = 1$, $\alpha_h = 0$ pour $h > p$, $\beta_h = 0$ pour $h > q$. Il ne reste plus qu'à résoudre le système.

6.2 Prédiction des processus ARMA(p, q)

En utilisant le critère de moyenne quadratique d'erreur de prévision (mean square forecast (prediction) error, MSFE, MSPE)

$$\min_{\widehat{y}} \mathsf{E}\left[(y_{t+h} - \widehat{y})^2 \middle| \Omega_t\right],$$

le prédicteur optimal de y_{t+h} est donné par l'espérance condtionnelle, étant donné l'ensemble d'information Ω_t :

$$\widehat{y}_{t+h|t} = \mathsf{E}\left[y_{t+h} \middle| \Omega_t\right],$$

où nous considérons ici que l'information disponible est l'historique du processus jusqu'à la date t, $\Omega_t = Y_t = (y_0, ..., y_t)$.

En ce qui concerne les processus ARMA stationnaires, et contrairement à beaucoup de DGP (processus de génération des données) non-linéaires, la moyenne conditionnelle peut être obtenue analytiquement, en utilisant une repésentation AR(∞) :

$$y_{t+h} = \mu + \sum_{j=1}^{\infty} \phi_j \left(y_{t+h-j} - \mu\right) + \varepsilon_{t+h},$$

et en utilisant l'opérateur d'espérance conditionnelle, le prédicteur optimal est donné par :

$$\widehat{y}_{t+h|t} = \mathsf{E}\left[y_{t+h} \middle| Y_t\right] = \mu + \sum_{j=1}^{h-1} \phi_j \left(\mathsf{E}\left[y_{t+h-j} \middle| Y_t\right] - \mu\right) + \sum_{j=0}^{\infty} \phi_{h+j} \left(\mathsf{E}\left[y_{t-j} \middle| Y_t\right] - \mu\right),$$

et sachant que $\mathsf{E}\left[y_s \middle| Y_t\right] = y_s$ pour $s \leq t$, le prédicteur optimal devient :

$$\widehat{y}_{t+h|t} = \mu + \sum_{j=1}^{h-1} \phi_j \left(\widehat{y}_{t+h-j|t} - \mu\right) + \sum_{j=0}^{\infty} \phi_{h+j} \left(y_{t-j} - \mu\right),$$

ce qui peut être calculé de manière récursive à partir du prédicteur une-étape

$$\widehat{y}_{t+1|t} = \mu + \sum_{j=0}^{\infty} \phi_{j+1} \left(y_{t-j} - \mu \right).$$

Ainsi, le prédicteur $\widehat{y}_{t+h|t}$ d'un processus ARMA(p,q) est-il une fonction linéaire des réalisation passées. L'erreur de prédiction qui lui est associée est fournie par :

$$\widehat{e}_{t+h|t} = y_{t+h} - \mathsf{E}\left[y_{t+h} \middle| Y_t \right].$$

Si à présent, on utilise la représentation MA(∞) :

$$y_{t+h} = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t+h-j},$$

et, sachant que $\mathsf{E}\left[\varepsilon_s \middle| Y_t \right] = 0$ pour $s > t$, le prédicteur optimal peut être réécrit :

$$\widehat{y}_{t+h|t} = \mu + \sum_{i=h}^{\infty} \psi_i \varepsilon_{t+h-j},$$

avec, pour erreur correspondante

$$\widehat{e}_{t+h|t} = y_{t+h} - \widehat{y}_{t+h|t} = \sum_{i=0}^{h-1} \psi_i \varepsilon_{t+h-j},$$

dont la variance est donnée par

$$\left(\sum_{i=0}^{h-1} \psi_i^2 \right) \sigma_\varepsilon^2.$$

Et si $\varepsilon_t \sim \mathsf{NID}\left(0, \sigma_\varepsilon^2\right)$

$$y_{t+h} \sim \mathsf{N}\left(\widehat{y}_{t+h|t}, \left[\sum_{i=0}^{h-1} \psi_i^2 \right] \sigma_\varepsilon^2 \right),$$

ce qui permet la construction d'intervalles de confiance.

Remarque 10 *Le caractère optimal du prédicteur est une propriété du processus de génération des donnée, tandis qu'une règle de prévision est toute procédure opérationnelle systématique qui permette d'établir des déclarations concernant l'avenir.*

6.3 Algorithme de Box-Jenkins

L'algorithme de Box-Jenkins vise à formuler un modèle permattant de représenter une série. Son idée principale est le concept de parcimonie, ou de la minimisation du nombre de paramètres. En pratique, ces derniers étant inconnus, ils sont donc remplacés par leur valeur estimée : plus il y a de paramètres, plus nombreuses sont les chances de se tromper.

Les années 1960 virent le développement d'un grand nombre de modèles macro-économiques comportant une forte quantité d'équations et de variables (des centaines). Celles-ci modélisaient très bien l'historique des données mais leur performance en matière de prévision laissait à désirer. D'où l'introduction des modèles ARMA(p,q), avec p et q faibles afin d'améliorer les prédictions :

$$\text{ARMA}\,(p,q) : A\,(L)\,y_t = \nu + B\,(L)\,\varepsilon_t,$$

où $A\,(L) = 1 - a_1 L - ... - a_p L^p$, $B\,(L) = 1 + b_1 L + ... b_q L^q$ et $\varepsilon_t \sim \mathsf{NID}\,(0, \sigma_\varepsilon^2)$.

6.3.1 Principe de la méthode

Il s'agit de procéder en quatre étapes.

(1) Transformer les données de manière à ce que l'hypothèse de stationnarité faible soit raisonnable.

(2) Etablir une hypothèse initiale concernant les paramètres p et q.

(3) Estimer les paramètres de $A\,(L)$ et $B\,(L)$.

(4) Etablir une analyse de diagnostic qui confirme que le modèle est valable.

Les quatres étapes sont détaillées ci-dessous.

6.3.2 Travailler sur données stationnaires

Prenons l'exemple d'un modèle ARIMA(p,d,q), c'est-à-dire d'un modèle ARMA intégré d'ordre d :

$$\text{ARIMA}\,(p,d,q) : A\,(L)\,\Delta^d y_t = \nu + B\,(L)\,\varepsilon_t,$$

il faut ainsi transformer le modèle en utilisant l'opérateur différence (Δ) afin qu'il devienne stationnaire.

Autre exemple : désaisonnalisation

6.3.3 Etablir une hypothèse

Il est essentiel d'établir une hypthèse maximale du nombre de coefficients à utiliser, i.e. des valeurs $p_{\max}$ et $q_{\max}$ à partir desquelles travailler sur le modèle ARMA(p,q). Pour ce faire, on utilise les fonctions ACF (ρ_j) et PACF (α_j) et on les observe visuellement, sachant que leurs propriétés sont les suivantes :

$$\begin{array}{rcl} \mathrm{MA}\,(q) & : & \rho_j = 0 \text{ si } j > q \text{ et non nul pour } j \leq q \\ \mathrm{AR}\,(p) & : & \rho_j \text{ tends graduellement vers } 0 \end{array}$$

et

$$\begin{array}{rcl} \mathrm{AR}\,(p) & : & \alpha_j = 0 \text{ si } j > p \text{ et non nul pour } j \leq p \\ \mathrm{MA}\,(q) & : & \alpha_j \text{ tends graduellement vers } 0 \end{array}$$

Ainsi pour un AR(p) pur, on observe un seuil de PACF pour $j \leq p$, α_j est non nul et il devient nul pour $j > p$. Pour un MA(q) pur, le comportement est le même, mais cette fois-ci en utilisant ACF et q comme valeur de coupure. Pour les ARMA, il faut malheureusement établir un diagnostic en observant séparément les parties ACF et PACF. En pratique, tous les logiciels indiquent par deux lignes en pointillés la bande de valeurs qu'il n'est aps statistiquement possible de différentier de zéro.

Exemple 33 *Soit $\{u_t\}$ un bruit blanc Gaussien de moyenne zéro et de variance constante σ^2. Les processus stochastiques linéaires suivants $\{y_t\}$ présentent un processus d'erreur commun, donné par $\{u_t\}$:*

$$\begin{array}{rrcl} (i) & y_t & = & 0.64y_{t-1} + u_t \\ (ii) & y_t & = & -0.2y_{t-1} + 0.64y_{t-2} + u_t \\ (iii) & y_t & = & u_t + u_{t-1} \\ (iv) & y_t & = & 0.64y_{t-2} + u_t + 0.64u_{t-1} \\ (v) & y_t & = & y_{t-1} + u_t \\ (vi) & y_t & = & y_{t-1} + u_t - 0.9u_{t-1} \end{array}$$

Une réalisation de chacun de ces processus stochastiques est représentée figure 6.1, conjointement avec leur fonctions estimées d'autocorrélation (ACF) et d'autocorrélation partielle (PACF). Identifiez le processus qui correspond à chacune des séries notées de A à F et expliquez votre décision.

6.3.4 Estimation

Pour l'estimation des paramètres de $A(L)$ et $B(L)$, il existe divers logiciels. Pour un AR(p) pur, la méthode la plus simple est les moindres carrés ordinaires ou la résolution des équations de Yule-Walker. En présence de partie MA, il faut utiliser le maximum de vraisemblance (voir section suivante).

6.3.5 Diagnostic

Celui-ci s'opère en plusieurs étapes. Une partie statistique se réfère à divers tests de spécification, pour vérifier que le modèle est congruent, i.e. qu'il ne peut être mis en défaut. Ensuite, si plusieurs modèles résistent à cette batterie de tests, il existe des méthodes ad hoc permettant de choisir entre eux.

Tests statistiques

Il s'agit ici de tester que les residus suivent un bruit blanc, i.e. sont non-corrélés et ne présentent pas d'hétéroscédasticité (i.e. variance constante).

Les tests pour ce faire sont, entre autres.

Test de Breusch-Godfrey pour l'autocorrélation (test LM) Cet test pratique une régression des résidus sur leurs valeurs retardées et vérifie que cette régression n'est pas significative via le R^2 :

$$\frac{R^2}{1-R^2}$$

qui suit une loi de Fischer (ou χ^2 pour les estimations univariées) sous l'hypothèse H_0 d'absence d'autocorrélation. Ce test est utilisable dans le cas d'autorégressions.

Test d'ARCH (AutoRegressive Conditional Heteroscedasticity) Ce test est similaire au précédent mais à présent les carrés des résidus sont régressés sur les carrés de leurs valeurs retardées. De même, sous H_0 : absence d'autocorrélation, la statistique suit une loi χ^2 ou de Fischer.

Test d'hétéroscédasticité de White Ce test utilise une régression des carrés des résidus sur les régresseurs originaux et leurs carrés. De nouveau, sous l'hypothèse nulle d'homoscédasticité, la statistique suit une loi χ^2 ou F.

Test de Normalité Ce test d'hypothèse nulle de Normalité des résidus utilise les propriétés des ratios des troisième et quatrième moments sur la variance dans le cadre des lois Gaussiennes. La statistique suit une loi χ^2 sous H_0.

Critères d'information

Si le choix s'avère difficile entre plusieurs modèles concurrents, il faut utiliser un critère *ad hoc*. Deux sont en général proposés. Selon le critère d'information d'Akaike, le meilleur des modèles est celui qui minimise la statistique :

$$\mathsf{AIC}(p,q) = T\log\left(\sigma^2_{\widehat{\varepsilon}}\right) + 2\left(p+q\right)$$

et la statistique du critère d'information de Schwarz est, quant à elle :

$$\mathsf{SC}\left(p,q\right) = T\log\left(\sigma^2_{\widehat{\varepsilon}}\right) + \left(p+q\right)\log\left(T\right).$$

SC coïncide avec le critère Bayésien d'information (BIC) et est plutôt recommandé pour les modèles ARMA.

6.4 Estimation des modèles dynamiques

6.4.1 Equations de Yule-Walker

En l'absence de composante MA (i.e. $q = 0$ dans ARMA(p,q)) la méthode à utiliser correspond aux moindres carrés ordinaires ou résolution des équations de Yule–Walker :

$$\left.\begin{array}{c}\rho_1 = \alpha_1 + \alpha_2\rho_1 + ... + \alpha_p\rho_{p-1}\\ \rho_2 = \alpha_1\rho_1 + \alpha_2 + ... + \alpha_p\rho_{p-2}\\ \vdots\\ \rho_p = \alpha_1\rho_{p-1} + \alpha_{p-2} + ... + \alpha_p\end{array}\right\} \Rightarrow \rho_1, ..., \rho_p,$$
$$\rho_k = \alpha_1\rho_{k-1} + \alpha_{k-2} + ... + \alpha_{k-p}, \quad \text{pour } k > p,$$

en remplaçant les autocorrélations théoriques par leur estimateurs. En revanche, si $q \neq 0$, il est nécessaire de recourir à la méthode du maximum de vraisemblance (exact ou conditionnel).

6.4.2 Fonction de vraisemblance

La méthode du maximum de vraisemblance part de l'hypothèse que l'échantillon $Y_T = (y_0, ..., y_T)$ observé suit une distribution dont les paramètres $(\boldsymbol{\theta})$ (espérance,

variance, covariances...) sont à estimer. Il faut ici faire l'hypothèse que la distribution est connue (de fonction de densité $f(\cdot)$), seuls ses paramètres ne le sont pas. Il est alors possible d'écrire que, pour $\boldsymbol{\theta}$ donné, la probabilité d'observer l'échantillon Y_T est donnée par

$$P(Y_T) = f(Y_T, \boldsymbol{\theta}).$$

Ainsi la probabilité d'observer Y_T dépend des paramètres $\boldsymbol{\theta}$. Le principe de la méthode est de rechercher quel $\boldsymbol{\theta}$ fournit la probabilité maximale d'observer Y_T. On définit alors la fonction de vraisemblance, qui est une fonction qui dépend de l'échantillon observé et dont le seul paramètre est $\boldsymbol{\theta}$, on la note :

$$L(\boldsymbol{\theta}) = f(Y_T, \boldsymbol{\theta}).$$

Il s'agit alors simplement de rechercher quel $\boldsymbol{\theta}$ maximise $L(\cdot)$ et on obtient l'estimateur :

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}).$$

6.4.3 Maximum de vraisemblance d'un ARMA

La fonction de vraisemblance d'un processus AR(1) Gaussien et stationnaire :

$$y_t = \nu + \alpha_1 y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{NID}\left(0, \sigma^2\right),$$

correspond à la distribution conjointe de $Y_T = (y_1, ..., y_T)'$ qui est Gaussienne $Y_T \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} \gamma_0 & \cdots & \gamma_{T-1} \\ \vdots & \ddots & \vdots \\ \gamma_{T-1} & \cdots & \gamma_0 \end{bmatrix} \right)$$

où $\gamma_h = \alpha_1^h \sigma_y^2$, $\sigma_y^2 = \left(1 - \alpha_1^2\right)^{-1} \sigma^2$ et $\mu = (1 - \alpha_1)^{-1} \nu$.

Ainsi la fonction de densité de l'échantillon $Y_T = (y_1, ..., y_T)$ est-elle donnée par la densité Normale multivariée :

$$f(Y_T) = \left(\frac{1}{\sqrt{2\pi}}\right)^T |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(Y_T - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (Y_T - \boldsymbol{\mu})\right).$$

La méthode de décomposition erreur-prévision utilise le fait que les ε_t sont indépendants, et identiquement distribués, par conséquent

$$f_\varepsilon(\varepsilon_2, ..., \varepsilon_T) = \prod_{t=2}^{T} f_\varepsilon(\varepsilon_t).$$

Et puisque $\varepsilon_t = y_t - (\nu + \alpha_1 y_{t-1})$, on a

$$f(y_t|y_{t-1}) = f_\varepsilon\left(y_t - \left(\nu + \alpha_1 y_{t-1}\right)\right) \quad \text{pour } t = 2, ..., T.$$

Ainsi

$$\begin{aligned} f(y_1, ..., y_T) &= f(y_T|y_{T-1}, ..., y_1) f(y_{T-1}, ..., y_1) \\ &= \left[\prod_{t=2}^{T} f(y_T|y_{T-1})\right] f(y_1). \end{aligned}$$

Pour $\varepsilon_t \sim \mathsf{NID}(0, \sigma^2)$, la fonction de vraisemblance est donnée par :

$$\begin{aligned} L(\boldsymbol{\lambda}) &= f_\varepsilon(\varepsilon_1, ..., \varepsilon_T; \boldsymbol{\lambda}) \\ &= \left[\prod_{t=2}^{T} f(y_t|Y_{t-1}; \boldsymbol{\lambda})\right] f(y_1; \boldsymbol{\lambda}), \end{aligned}$$

où $\boldsymbol{\lambda}$ est le paramètre à estimer (c'est ici un vecteur). Donc

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \left[\prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_t - \left(\nu + \alpha_1 y_{t-1}\right)\right)^2\right]\right] \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma_y^2}(y_1 - \mu)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^T \exp\left[-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(y_t - \left(\nu + \alpha_1 y_{t-1}\right)\right)^2 - \frac{1}{2\sigma^2}(y_1 - \mu)^2\right] \end{aligned}$$

et

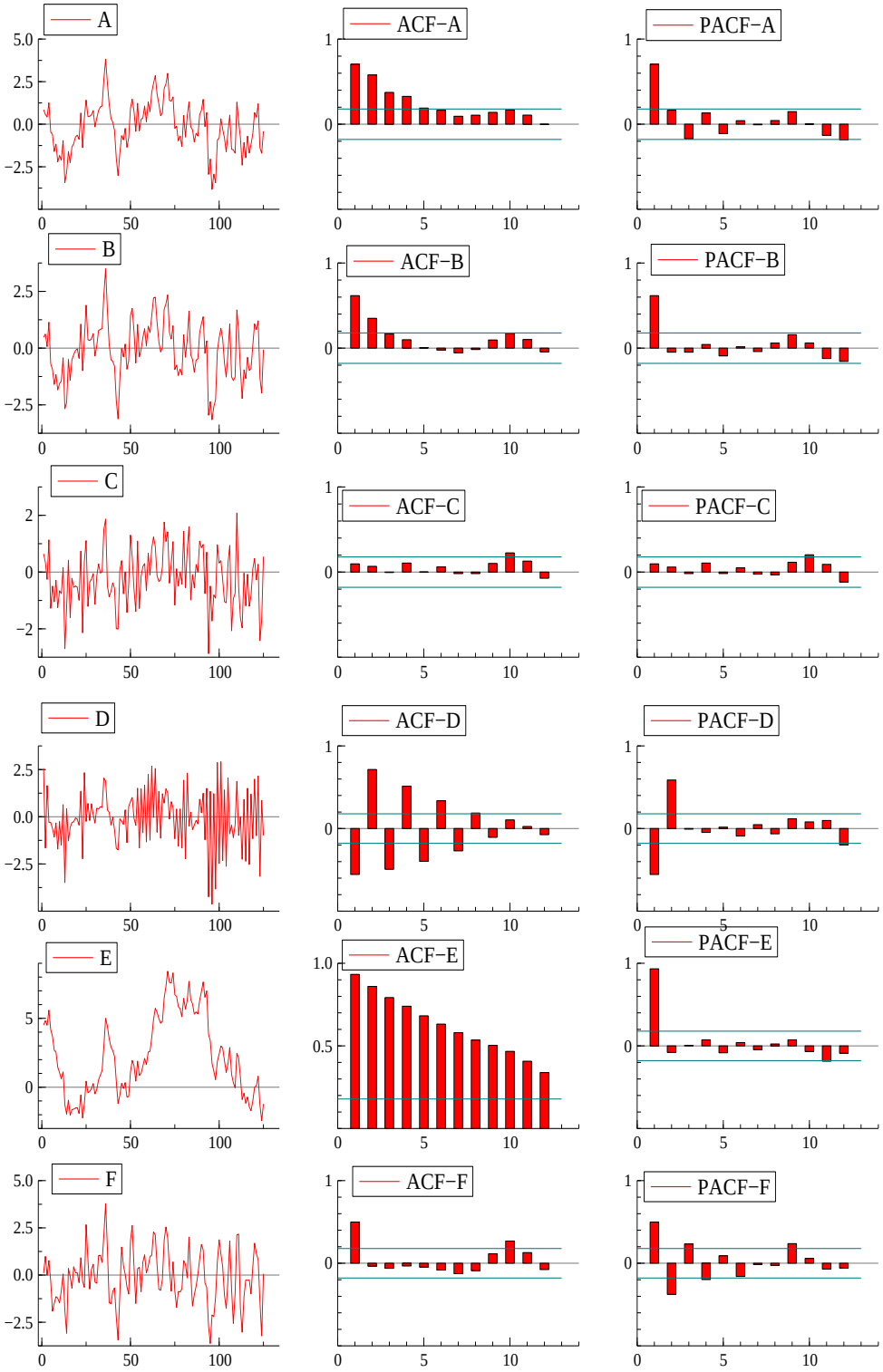
$$l(\boldsymbol{\lambda}) = \log L(\boldsymbol{\lambda}) = -\frac{T}{2}\log(2\pi) - \left(\frac{T-1}{2}\log\left(\sigma^2\right) + \frac{1}{2\sigma^2}\sum_{t=2}^{T}\varepsilon_t^2\right) - \left(\frac{1}{2}\log\left(\sigma_y^2\right) + \frac{1}{2\sigma_y^2}(y_1 - \mu)^2\right)$$

où $y_1 \sim \mathsf{N}\left(\mu, \sigma_y^2\right)$. En général on utilise la fonction conditionnelle de vraisemblance $\left[\prod_{t=2}^{T} f_\varepsilon(\varepsilon_t; \boldsymbol{\lambda})\right]$ (conditionnée à la première observation). Dans le cas des processus ARMA(p, q),

$$\varepsilon_t = y_t - \left[\nu + \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \beta_1 \varepsilon_{t-1} + ...\beta_q \varepsilon_{t-q}\right]$$

et une méthode non-linéaire de maximisation numérique doit être employée.

Figure 6.1 – Séries simulées et leur ACF et PACF. Les DGP sont donnés dans l'exemple 33.



Chapitre 7

Les variables intégrées

La modélisation ARMA repose sur un principe de stationnarité. Il convient donc de s'assurer a priori du dégré d'intégration des séries. Pour ce faire, on utilise un **test de racine unitaire**, i.e. de présence d'une tendance stochastique. Celui-ci s'intègre dans le cadre plus général des modèles ARIMA(p, d, q) pour lesquels la formulation est :

$$\alpha(L)\,\Delta^d y_t = \beta(L)\,\epsilon_t,$$

où $\Delta = (1 - L)$, $\alpha(L)$ est d'ordre p, $\beta(L)$ est d'ordre q, les racines des polynômes sont supérieures à 1 en valeur absolue et ils n'ont pas de racine commune. Il s'agit donc d'un modèle ARMA stationnaire appliqué à une transformation de y_t, sa d-ième différence. Ici un processus $\mathsf{I}(1)$ sera donc une ARIMA$(p, 1, q)$. La marche aléatoire est un processus ARIMA$(0, 1, 0)$ car

$$y_t = y_{t-1} + \epsilon_t$$

se réécrit

$$\Delta y_t = \epsilon_t.$$

7.1 Les tests de racine unitaire

La pratique des tests de racine unitaire repose sur la modélisation d'une série par un processus AR(p) :

$$y_t = \sum_{i=1}^{p} \alpha^i y_{t-i} + \varepsilon_t,$$

le cas le plus simple est celui d'une marche aléatoire :

$$y_t = \rho y_{t-1} + \varepsilon_t. \tag{7.1}$$

Quand $|\rho| < 1$, le processus est stationnaire, explosif si $|\rho| > 1$ et intégré dans le cas $\rho = \pm 1$.

7.1.1 Problèmes des processus intégrés

Les difficultés liées aux racines unitaires sont de trois ordres :

1. Les processus présentant une racine (autorégressive) unitaire sont non-stationnaires (mais toutefois intégrés).
2. Lorsqu'on régresse un processus non-stationnaire sur un autre :

$$y_t = \beta x_t + u_t$$

l'estimateur $\widehat{\beta}$ ne tend pas nécessairement vers 0, même si les deux séries sont en réalité indépendantes, sauf s'il existe une combinaison linéaire de y_t et x_t qui soit elle-même stationnaire. Ce dernier cas se rencontre souvent en pratique lorsque les séries sont intégrées, y_t et x_t sont alors dites co-intégrées.
3. L'estimateur des moindres carrés ordinaires de ρ, noté $\widehat{\rho}$ n'a pas une distribution usuelle :

$$\sqrt{T}\left(\widehat{\rho} - 1\right) \to \mathsf{N}\left(0, \sigma^2_{\widehat{\rho}}\right)$$

mais suit une distribution (non-normale) dite de Dickey–Fuller qui prenne en compte le fait que $\widehat{\rho}$ ait tendance à sous-estimer $\rho = 1$.

7.1.2 Test de Dickey-Fuller

Il est donc essentiel de déterminer *a priori* si les séries présentent une racine unitaire. Pour ce faire, divers tests existent : le plus simple est le test de Dickey-Fuller qui prend pour hypothèses :

$$\mathsf{H}_0 : \rho = 1, \qquad \mathsf{H}_1 : \rho < 1.$$

Sous l'hypothèse H_0, on peut réécrire (7.1)

$$\Delta y_t = y_t - y_{t-1} = \varepsilon_t,$$

et donc en régressant Δy_t sur y_{t-1} dans

$$\Delta y_t = d_t + \alpha y_{t-1} + \varepsilon_t,$$

on doit trouver un estimateur $\widehat{\alpha}$ proche de zéro. d_t contient les termes déterministiques : soit zéro, soit une constante soit une tendance linéaire, i.e. :

$$d_t = \begin{cases} 0 \\ a \\ a + bt \end{cases}.$$

Malheureusement, le test de Student associé à $\widehat{\alpha}$ ne suit pas une distribution habituelle. Il faut donc se reporter aux tables de Fuller, mais les valeurs dépendent des composantes de d_t. Il faut donc élaborer une stratégie.

7.2 Les différents tests

Test de Dickey-Fuller : cas le plus simple de marche aléatoire avec ou sans tendance déterministe.
Dickey–Fuller Augmenté : permet de prendre en compte l'autocorrélation possible de la série différentiée via une correction utilisant les valeurs retardées, sur la base du test :

$$\Delta y_t = d_t + \alpha y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \varepsilon_t,$$

Phillips–Perron : il s'agit d'une procédure de correction *non-paramétrique* (i.e. il n'y a pas de modélisation de l'autocorrélation). Ce test est plus robuste vis-à-vis des erreurs de spécification (i.e. quel que soit le type d'autocorrélation), en revanche il est moins précis que ADF quand le modèle correspond à la réalité.
Schmidt–Phillips : il s'agit de résoudre le problème de la présence, ou non, de tendance déterministe dans le test de D-F. Il consiste en un test qui ôte de manière arbitraire une tendance (détrender). Ainsi travaille-t-il non pas sur la variable x_t mais sur une transformation $S_t = x_t - \widehat{\psi} - \widehat{\xi} t$ où les paramètres sont calculés simplement. Il utilise d'autre distributions que ADF et PP et, comme ces derniers, ne fait pas la différence entre des racines de 0,95 et de 1 dans des échantillons de taille finie.
Eliott-Rothenberg-Stock(ERS) : Utilise le fait que dans de petits échantillons 0,95 et 1 sont indifférentiables. Il procède de ce fait à une quasi-différentiation $\widetilde{x}_t = x_t - \alpha x_{t-1}$, avec $\alpha = 1 - \frac{c}{T}$.
KPSS : prend pour hypothèse, non pas la non-stationnarité mais la stationnarité. Il est malheureusement souvent moins robuste.

ERS et Schmidt–Phillips sont à préférer en général.

7.3 Les tendances et constantes

Les distributions utilisées pour les tests de racine unitaire diffèrent malheureusement selon la présence, ou non, d'une constante et d'une tendance

linéaire. Si on suppose qu'une variable y_t suive :

$$y_t = \alpha + \beta t + \rho y_{t-1} + \varepsilon_t.$$

Parmi les hypothèses potentielles, y_t peut présenter une tendance linéaire mais être stationnaire autour d'elle ($\beta \neq 0$, $|\rho| \neq 1$), de sorte que

$$y_t - \gamma t = \alpha - \gamma + (\beta + \rho\gamma)\, t + \rho\left[y_{t-1} - \gamma\,(t-1)\right] + \varepsilon_t,$$

et ainsi en posant (si $\rho \neq 0$) $\gamma = -\beta/\rho$, et $x_t = y_t + (\beta/\rho)\, t$:

$$x_t = \alpha + \beta/\rho + \rho x_{t-1} + \varepsilon_t$$

ce qui signifie que y_t est stationnaire autour de la tendance linéaire $-(\beta/\rho)\, t$ et la variable"détrendée" x_t est stationnaire. A présent si $|\rho| = 1$, x_t n'est plus stationnaire mais suit une marche aléatoire (avec dérive si $\alpha + \beta/\rho \neq 0$) et ainsi

$$x_t = x_0 + (\alpha + \beta/\rho)\, t + \sum_{i=1}^{t} \varepsilon_i,$$

donc x_t présente à la fois une tendance linéaire déterministe $(\alpha + \beta/\rho)\, t$ et une tendance stochastique $\sum_{i=1}^{t} \varepsilon_i$ (ie. comme une marche aléatoire normale). Ainsi, si $\rho = 1$,

$$\begin{aligned}
y_t &= \alpha + \beta t + y_{t-1} + \varepsilon_t, \\
\Delta y_t &= \alpha + \beta t + \varepsilon_t \\
y_t &= y_0 + \sum_{i=1}^{t} \Delta y_i = y_0 + \sum_{i=1}^{t} (\alpha + \beta i + \varepsilon_i) \\
&= y_0 + \alpha t + \beta \sum_{i=1}^{t} i + \sum_{i=1}^{t} \varepsilon_i \\
&= y_0 + \alpha t + \beta \frac{t\,(t+1)}{2} + \sum_{i=1}^{t} \varepsilon_i \\
&= y_0 + (\alpha + \beta/2)\, t + (\beta/2)\, t^2 + \sum_{i=1}^{t} \varepsilon_i.
\end{aligned}$$

Donc y_t présente à la fois une tendance *quadratique* déterministe $(\beta/2)\, t^2$ et une tendance stochastique $\sum_{i=1}^{t} \varepsilon_i$.. De même si $\beta = 0$, mais $\alpha \neq 0$: quand $\rho = \pm 1$, y_t présente une tendance linéaire déterministe et une tendance stochastique et quand $|\rho| < 1$, y_t est stationnaire de moyenne non nulle. Enfin si $\beta = \alpha = 0$,

Tableau 7.1 – Distribution de F pour le test $(\alpha, \beta, \rho) = (\alpha, 0, 1)$ dans $y_t = \alpha + \beta t + \rho y_{t-1} + \varepsilon_t$.

Taille de	Probabilité d'une valeur inférieure							
l'échantillon	.01	.025	.05	.10	.90	.95	.975	.99
25	.74	.90	1.08	1.33	5.91	7.24	8.65	10.61
20	.76	.93	1.11	1.37	5.61	6.73	7.81	9.31
100	.76	.94	1.12	1.38	5.47	6.49	7.44	8.73
250	.76	.94	1.13	1.39	5.39	6.34	7.25	8.43
500	.76	.94	1.13	1.39	5.36	6.30	7.20	8.34
∞	.77	.94	1.13	1.39	5.34	6.25	7.16	8.27

Source : Dickey & Fuller (1976), Table VI

y_t suit soit une marche aléatoire, soit un processus stationnaire de moyenne nulle. Ce qu'on résume dans le tableau ci-dessous :

(α, β, ρ)	$\lvert\rho\rvert < 1$	$\lvert\rho\rvert = 1$
$\beta \neq 0$	stationnaire autour d'une tendance linéaire	intégré et présentant une tendance quadratique
$\alpha \neq 0, \beta = 0$	stationnaire de moyenne non nulle	intégré et présentant une tendance linéaire
$\alpha = 0, \beta = 0$	stationnaire de moyenne nulle	intégré sans tendance déterministe

Ainsi convient-il de bien spécifier la présence, ou non d'une constante ou d'une tendance dans le modèle. En pratique commencer par le modèle le plus général et vérifier la bonne spécification du modèle, i.e. la présence ou non d'un β ou d'un α. Le test d'hypothèse jointe d'une racine unitaire et d'absence de tendance déterministe se fait théoriquement grâce à la statistique de Fisher

$$F = \frac{ESS_R - ESS_{NR}}{(N-k)\,q},$$

où N est le nombre d'observations, k le nombre de paramètres estimés dans la régression non restreinte (en n'imposant pas l'hypothèse nulle, i.e. en estimant β et ρ), q le nombre de restrictions (ici 2), ESS_R est la somme des carrés des variables modélisées (i.e. $\sum_{i=1}^{T} \widehat{y}_i^2$) sous l'hypothèse $\beta = 0$ (non estimée) et $\rho = 1$ (donc on estime pour $\Delta y_t = \alpha + \varepsilon_t$) et ESS_{NR} est la somme $\sum_{i=1}^{T} \widehat{y}_i^2$ sans restrictions. Se reporter ensuite à la table 1 (pour le test de Dickey-Fuller simple, non augmenté).

7.4 Modèles univariés de cointégration

Pour une équation de comportement, un modèle à correction d'erreurs (MCE), dans le cas simple de deux variables (y) et (x), s'écrit :

$$\Delta y_t = \tau - \alpha\left(y_{t-1} - \beta x_{t-1} - \mu\right) + \sum_{i=1}^{p} \gamma_{y,i}\Delta y_{t-i} + \gamma_{x,0}\Delta x_t + \sum_{i=1}^{q} \gamma_{x,i}\Delta x_{t-i} + \nu_t \tag{7.2}$$

Pour que cette équation soit valable et interprétable, tous les termes de la régression doivent être $\mathsf{I}(0)$. Pour que cela soit le cas, il faut que le terme entre parenthèses soit une relation de cointégration[1] si les variables (y) et (x) sont $\mathsf{I}(1)$. Il est aussi possible qu'une tendance linéaire intervienne dans la relation de cointégration.

Il convient donc de s'assurer, ou de supposer a priori, que (y) et (x) sont intégrées via un test de racine unitaire, par exemple Dickey-Fuller augmenté.

Différentes méthodes permettent d'estimer le modèle (7.2). Nous les présentons ci-dessous et exposons leurs avantages et défauts. Nous terminons par la méthode préconisée.

7.4.1 Procédure en deux étapes d'Engle et Granger

Dans la première étape, on estime la relation de cointégration $y = \beta x + \mu$ et dans la seconde, les coefficients du modèle MCE, en remplaçant l'écart par rapport à l'équilibre (terme entre parenthèses dans l'équation (7.2)) par son estimation.

Détermination de la relation de cointégration

Lors de nos estimations, nous tentons de définir une relation de cointégration pour chacun des grands comportements. Le concept de cointégration permet de définir statistiquement la notion économique d'équilibre (de long terme) entre variables intégrées de même ordre.

Dans notre exemple, la relation de cointégration que l'on estime s'écrit :

$$y_t = \mu + \beta x_t + \varepsilon_t. \tag{7.3}$$

Cette relation est une relation de cointégration si ε_t est stationnaire ($\mathsf{I}(0)$). Le test de cointégration se ramène donc à un test de racine unitaire. La régression

[1] **Définition :** n variables $\mathsf{I}(1)$ sont dites cointégrés s'il existe une combinaison linéaire de celles-ci qui soit $\mathsf{I}(0)$.

qui sert au test est la suivante :

$$\Delta\varepsilon_t = \rho\varepsilon_{t-1} + \sum_{i=1}^{s} \eta_i \Delta\varepsilon_{t-i} + u_t,$$

où on teste $\mathsf{H}_0 : \rho = 0$ contre $\mathsf{H}_1 : \rho < 0$ à partir de la statistique de Student du coefficient ρ. Pour accepter la cointégration, il faut accepter H_1. Mais on ne peut pas utiliser la table de Fuller car ε_t est un résidu d'estimation.

Engle et Granger (1987) ont montré que les coefficients de long terme peuvent être estimés en utilisant la méthode des MCO sur l'équation (7.3). En effet ces estimateurs convergent en probabilité vers leurs vraies valeurs au taux $1/T$ (au lieu de $1/\sqrt{T}$ habituellement). Ils sont qualifiés de "super convergents". Notons que cette convergence a lieu malgré l'oubli de la dynamique de court terme (le résidu ε n'est pas un bruit blanc mais un processus $\mathsf{I}(0)$) et aussi lorsque certaines variables x sont endogènes.

Phillips et Durlauf (1986) ont déterminé la distribution asymptotique des estimateurs des MCO. Elles sont non-standards ainsi que celles des statistiques de student associés. C'est pourquoi on ne peut réaliser de test de significativité sur les coefficients de la relation (7.3) . Ainsi est-il facile d'obtenir des estimateurs convergents des coefficients de long terme mais il est impossible de savoir si les coefficients sont réellement significatifs !

Estimation du MCE

Si on accepte l'hypothèse de cointégration, on passe à la seconde étape de la procédure d'Engle et Granger, c'est-à-dire à l'estimation d'un modèle MCE en remplaçant l'erreur d'équilibre par son estimation.

$$\Delta y_t = \tau - \alpha\varepsilon_{t-1} + \sum\nolimits_{i=1}^{p} \gamma_{y,i}\Delta y_{t-i} + \gamma_{x,0}\Delta x_t + \sum\nolimits_{i=1}^{q} \gamma_{x,i}\Delta x_{t-i} + \nu_t. \quad (7.4)$$

Critique de la procédure en deux étapes

Elle a été développée dans le livre de Banerjee, Dolado, Galbraith, et Hendry[2] et reprise dans le livre de Harris et dans l'article d'Ericsson & McKinnon.

Elle porte d'une part sur l'estimation de la relation de cointégration et d'autre part sur le test de cointégration.

1. Le risque est important de faire intervenir trop de variables dans la relation de cointégration puisqu'on ne dispose pas de test de Student. On force alors la dynamique du modèle vers un équilibre qui n'en est pas un.

[2]Banerjee, Dolado, Galbraith, & Hendry (1993), " Co-integration, Error-correction and the Econometric Analysis of Non-Stationary Data ", Oxford University Press.

2. Lors de l'estimation de la relation de cointégration, on omet (généralement) des variables explicatives (celles de la dynamique de court terme), ce qui entraîne un biais sur les coefficients de la relation. On sait que ce biais disparaît asymptotiquement (estimateurs "super convergents") mais ce biais peut être non négligeable avec des échantillons de taille limitée.
3. Considérons maintenant le test de cointégration. Soit le résidu de la relation de cointégration :

   $$\varepsilon_t = y_t - \beta x_t - \mu.$$

   Prenons la version la plus simple de ce test :

   $$\Delta\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

   avec $\mathsf{H_0} : \rho = 0$ (pas de cointégration) contre $\mathsf{H_1} : \rho < 0$ (cointégration), ce qui peut encore s'écrire

   $$\Delta\left(y_t - \beta x_t - \mu\right) = \rho\left(y_{t-1} - \beta x_{t-1} - \mu\right) + u_t,$$

   soit

   $$\Delta y_t = \beta\Delta x_t + \rho\left(y_{t-1} - \beta x_{t-1} - \mu\right) + u_t.$$

   Ainsi, le test de cointégration se fait-il sur un modèle MCE particulier par rapport au modèle général suivant :

   $$\Delta y_t = \gamma\Delta x_t + \alpha\left(y_{t-1} - \beta x_{t-1} - \mu\right) + \nu_t,$$

   on fait l'hypothèse $\gamma = \beta$.

La méthode en deux étapes contraint donc l'elasticité de long terme (β) à être la même que celle de court terme (γ). Ceci est une restriction très forte non vérifiée en pratique. Ainsi, la cointégration est-elle testée sous cette même hyptothèse, alors que cette restriction n'est en général pas imposée dans la deuxième étape. Ceci entraîne alors une forte incompatibilité des estimations!!

7.4.2 Procédure en une étape de Banerjee, Dolado et Mestre

Cette procédure consiste à estimer par les MCO les coefficients du modèle :

$$\Delta y_t = \delta + \lambda_y y_{t-1} + \lambda_x x_{t-1} + \sum\nolimits_{i=1}^{p} \gamma_{y,i}\Delta y_{t-i} + \gamma_{x,0}\Delta x_t + \sum\nolimits_{i=1}^{q} \gamma_{x,i}\Delta x_{t-i} + \nu_t$$

(7.5)

et d'en déduire la relation de cointégration par division afin de tenir compte de la dynamique de court terme. En effet si l'on omet cette dynamique, on obtient une estimation biaisée du vecteur de cointégration dans les échantillons de taille finie. Ce modèle peut se ré-écrire :

$$\Delta y_t = \delta + \lambda_y \left(y_{t-1} + \frac{\lambda_x}{\lambda_y} x_{t-1} \right) + \sum_{i=1}^{p} \gamma_{y,i} \Delta y_{t-i} + \gamma_{x,0} \Delta x_t + \sum_{i=1}^{q} \gamma_{x,i} \Delta x_{t-i} + \nu_t,$$

ce qui correspond dans (7.2) à

$$\begin{aligned} \alpha &= -\lambda_y, \\ \beta &= -\frac{\lambda_x}{\lambda_y}. \end{aligned}$$

Les statistiques de Student des coefficients de long terme (λ_y et λ_x) ont des lois non standards à la différence de celles des coefficients des variables stationnaires. En conséquence on ne peut pas tester la significativité des coefficients de long terme ni tester l'impact de long terme de x sur y.

Banerjee, Dolado et Mestre (1995) préconisent de tester la cointégration sur la régression (7.5) à partir de la statistique de Student du coefficient λ_y. Pour accepter la cointégration, il faudrait que ce coefficient soit significativement différent de zéro. Cependant, comme nous l'avons précisé précédemment, la statistique de Student de ce coefficient a une distribution non standard.

Test de cointégration

Significativité de λ_y par un test de Student avec les tables de Ericsson et MacKinnon (2002), téléchargeables sous ecmtest.xls sous

`http://qed.econ.queensu.ca/pub/faculty/mackinnon/ecmtest/`

En pratique estimer (7.5) puis se reporter à la feuille excel ecmtest.xls (où on remplit les cellules encadrées sur le graphique 7.1). On doit y reporter le nombre de variables entrant dans la relation de cointégration (y compris l'endogène, soit ici k). Ensuite d est le nombre de composantes déterministes (i.e. soit une constante, plus le cas échéant un trend) toujours compris entre 1 et 2. T est la taille de l'échantillon . Enfin le nombre total de régresseurs, i.e. de variables entrant à droite dans l'équation, y compris les variables déterministes et les retards et dummies.

7.4.3 Références bibliographiques

Ericsson, N & MacKinnon, J (2002). "Distributions of error correction tests for cointegration", *Econometrics Journal*, **5**, *pp.* 285-318.

Harris, R (1995). *Cointegration Analysis in Econometric Models*. Hemel Hempstead : Harvester Wheatsheaf.

Annexe 7.A Décomposition du MCE en court et long termes

MCE :

$$\Delta y_t = \tau - \alpha\left(y_{t-1} - \beta x_{t-1} - \mu\right) + \sum_{i=1}^{p} \gamma_{y,i}\Delta y_{t-i} + \gamma_{x,0}\Delta x_t + \sum_{i=1}^{q} \gamma_{x,i}\Delta x_{t-i} + \nu_t$$

est estimée comme

$$\Delta y_t = \delta + \lambda_y y_{t-1} + \lambda_x x_{t-1} + \sum_{i=1}^{p} \gamma_{y,i}\Delta y_{t-i} + \gamma_{x,0}\Delta x_t + \sum_{i=1}^{q} \gamma_{x,i}\Delta x_{t-i} + \nu_t$$

Long terme :

$$y_t^* = -\frac{\delta}{\lambda_y} - \frac{\lambda_x}{\lambda_y} x_t$$

Puis le court terme, variable yc définie par

$$y_t = y_t^* + yc_t$$

$$yc_t = \left(1 + \lambda_y\right) yc_{t-1} + \left(\gamma_{x,0} + \frac{\lambda_x}{\lambda_y}\right)\Delta x_t + \sum_{i=1}^{p} \gamma_{y,i}\Delta y_{t-1} + \sum_{i=1}^{p} \gamma_{x,i}\Delta x_{t-1} + \nu_t$$

Annexe 7.B Neutralité et Homogénéité du MCE

Reprenons l'équation :

$$\Delta y_t = \tau - \alpha\left(y_{t-1} - \beta x_{t-1} - \mu\right) + \sum_{i=1}^{p} \gamma_{y,i}\Delta y_{t-i} + \gamma_{x,0}\Delta x_t + \sum_{i=1}^{q} \gamma_{x,i}\Delta x_{t-i} + \nu_t.$$

Si à présent on suppose que y et x présentent une tendance linéaire de sorte qu'elles croissent en moyenne de façon constante : $\mathsf{E}[\Delta x_t] = g_x$et $\mathsf{E}[\Delta \mathrm{y}_t] = g_y$. Si on insère un trend dans la relation de cointégration ces taux de croissance peuvent différer car un vecteur de cointégration :

$$c_t = y_t - \beta x_t - \mu - bt$$

implique que

$$g_y - \beta g_x - b = 0.$$

Par conséquent, pour homogénéité, il est nécessaire que

$$g_y = \tau + \sum_{i=1}^{p} \gamma_{y,i} g_y + \gamma_{x,0} g_x + \sum_{i=1}^{q} \gamma_{x,i} g_x,$$

ie

$$\begin{aligned}\left(1-\sum_{i=1}^{p}\gamma_{y,i}\right)g_y &= \tau+\left(\gamma_{x,0}+\sum_{i=1}^{q}\gamma_{x,i}\right)g_x\\ &= \tau+\left(\gamma_{x,0}+\sum_{i=1}^{q}\gamma_{x,i}\right)\left(-\frac{b+g_y}{\beta}\right),\end{aligned}$$

soit

$$\left(1-\sum_{i=1}^{p}\gamma_{y,i}+\frac{1}{\beta}\left(\gamma_{x,0}+\sum_{i=1}^{q}\gamma_{x,i}\right)\right)g_y=\tau-\frac{b}{\beta}\left(\gamma_{x,0}+\sum_{i=1}^{q}\gamma_{x,i}\right).$$

Cette équation à des implications qui dépendent de la modélisation : si on souhaite que la croissance autonome τ soit nulle, cela nécessite que :

$$\begin{aligned}1 &= \sum_{i=1}^{p}\gamma_{y,i},\\ 0 &= \gamma_{x,0}+\sum_{i=1}^{q}\gamma_{x,i}.\end{aligned}$$

Figure 7.1 { chier ecmtest.xls servant a tester l'hypothese de cointegration en utilisant l'estimation a une etape. En rentrant les valeurs correspondantes dans la partie encadree, on obtient la valeur critique de la statistique de Student.

or

$$y_t - \frac{\nu}{1-\alpha} = \alpha\left(y_{t-2} - \frac{\nu}{1-\alpha}\right) + u_t,$$

donc, comme y_{t-2} *et* u_t *sont indépendants :*

$$\begin{aligned}\mathsf{Cov}\left[y_t, y_{t-2}\right] &= \alpha \mathsf{V}\left[y_{t-2} - \frac{\nu}{1-\alpha}\right] \\ &= \frac{\alpha}{1-\alpha^2}\sigma_u^2\end{aligned}$$

(2) *calcul des moments de la série générée par le modèle* $MA(2)$:

$$y_t = \mu + \varepsilon_t + \beta\varepsilon_{t-2}, \quad \varepsilon_t \sim \mathsf{NID}\left(0, \sigma_\varepsilon^2\right).$$

Le processus est stationnaire car il s'agit d'une moyenne mobile.
(*i*) *Espérance :*

$$\mathsf{E}\left[y_t\right] = \mu$$

(*ii*) *Variance :*

$$\mathsf{V}\left[y_t\right] = \left(1+\beta^2\right)\sigma_\varepsilon^2$$

(*iii*) *Covariance entre* y_t *et* y_{t-1} : $\mathsf{Cov}\left[y_t, y_{t-1}\right] = 0$ *car* y_t *et* y_{t-2}*sont fonctions de bruits blanc indépendants enre eux.*
(*iii*) *Covariance entre* y_t *et* y_{t-2} :

$$\begin{aligned}\mathsf{Cov}\left[y_t, y_{t-2}\right] &= \mathsf{Cov}\left[\varepsilon_t + \beta\varepsilon_{t-2}, \varepsilon_{t-2} + \beta\varepsilon_{t-4}\right] \\ &= \beta\sigma_\varepsilon^2.\end{aligned}$$

A présent pour calculer les paramètres à partir des moments, il faut résoudre les équations. Pour ce faire, on utilise le ratio $\frac{\mathsf{Cov}[y_t, y_{t-2}]}{\mathsf{V}[y_t]}$ *et on calcule les coefficients (équation du second degré dans le cas du* $MA(2)$*).*

$$AR\left(2\right) : \begin{cases} \nu = \mathsf{E}\left[y_t\right]\left(1 - \mathsf{Corr}\left[y_t, y_{t-2}\right]\right) \\ \alpha = \mathsf{Corr}\left[y_t, y_{t-2}\right] \\ \sigma_u^2 = \mathsf{V}\left[y_t\right]\left(1 - \left\{\mathsf{Corr}\left[y_t, y_{t-2}\right]\right\}^2\right) \end{cases}$$

$$MA\left(2\right) \begin{cases} \mu = \mathsf{E}\left[y_t\right] \\ \beta = \dfrac{1 \pm \sqrt{1 - 4\left\{\mathsf{Corr}\left[y_t, y_{t-2}\right]\right\}^2}}{2\mathsf{Corr}\left[y_t, y_{t-2}\right]} \\ \sigma_\varepsilon^2 = \dfrac{2\left\{\mathsf{Corr}\left[y_t, y_{t-2}\right]\right\}^2}{1 \pm \sqrt{1 - 4\left\{\mathsf{Corr}\left[y_t, y_{t-2}\right]\right\}^2}} \end{cases}$$

Pour obtenir β, il existe deux solutions de valeurs absolues l'une supérieure à l'unité, l'autre inférieure ; c'est cette dernière racine qu'on utilise afin d'assurer le caractère inversible du MA.

Cas empiriques
(i) on constate que l'autocorrélation du premier ordre mesurée est très forte, tandique que chacun des modèles implique une valeur nulle. Aucun des eux modèles ne peut donc s'appliquer.
(ii) on choisit pour le $MA(2)$, le coefficient implqiuant un processus inversible.

$$AR(2) \quad : \quad \begin{cases} \nu = 0,6 \\ \alpha = 0,4 \\ \sigma_u^2 = 1,68 \end{cases}$$

$$MA(2) \quad : \quad \begin{cases} \mu = 1 \\ \beta = 0,5 \ (ou\ 2) \\ \sigma_\varepsilon^2 = 1,6 \ (ou\ 0,4) \end{cases}$$

Pour pouvoir choisir enter les deux représentations, le corrélogramme se révèlerait par exemple utile.
(iii) Seule la représentation AR est ici possible.

$$AR(2) \quad : \quad \begin{cases} \nu = 0,2 \\ \alpha = 0,8 \\ \sigma_u^2 = 0,72 \end{cases}$$

$$MA(2) \quad : \quad \text{pas de racine réelle pour } \beta$$

Exercice 3 *Soit $\{u_t\}$ un bruit blanc Gaussien, i.e. $u_t \sim \mathsf{NID}(0, \sigma^2)$. Le processus stochastique $\{y_t\}$ dérivé de $\{u_t\}$ est défini par :*

$$\begin{aligned}
(i) \quad y_t &= \alpha y_{t-1} + u_t, \ pour \ \ t > 0, \quad où \ \ \alpha = 1 \ \ et \ \ y_0 = 2; \\
(ii) \quad y_t &= \tau + \alpha y_{t-1} + u_t, \quad pour\ t > 0, \ où\ \alpha = 1, \tau \neq 0, \quad et \ \ y_0 = 0; \\
(iii) \quad y_t &= u_t + \beta u_{t-1}, \quad avec\ \beta = -1/2 \ \ pour \ \ t \ variant \ \ de - \infty \ \ à + \infty; \\
(iv) \quad y_t &= \begin{cases} u_t & pour\ t = 1, 3, 5, ... \\ 2u_t & pour\ t = 2, 4, 6, ... \end{cases}
\end{aligned}$$

Pour chacun des processus $\{y_t\}$ défini ci-dessus :

(a) *Précisez quelles sont la moyenne μ_t et la fonction d'autocovariance $\gamma_t(h)$ de $\{y_t\}$.*
(b) *Déterminez si le processus est (faiblement) stationnaire.*

Correction 3 (i)

$$y_t = y_{t-1} + u_t, \quad avec\ y_0 = 2$$

donc

$$y_t = y_0 + \sum_{i=1}^{t} u_i$$

et

$$\begin{aligned}
\mu_t &= \mathsf{E}[y_t] = y_0 = 2 \\
\gamma_t(0) &= \mathsf{Var}\left[\sum_{i=1}^{t} u_i\right] = \sum_{i=1}^{t} \mathsf{Var}[u_i] = t\sigma_u^2 \\
\gamma_t(h) &= \mathsf{Cov}[y_t, y_{t-h}] = \mathsf{Cov}\left[\sum_{i=1}^{t} u_i, \sum_{i=1}^{t-h} u_i\right] \\
&= \mathsf{Cov}\left[\sum_{i=1}^{t-h} u_i, \sum_{i=1}^{t-h} u_i\right] + \mathsf{Cov}\left[\sum_{i=t-h+1}^{t} u_i, \sum_{i=1}^{t-h} u_i\right]
\end{aligned}$$

or les $(u_{t-h+1}, ..., u_t)$ *sont indépendants des* $(u_1, ..., u_{t-h})$ *donc*

$$\begin{aligned}
\gamma_t(h) &= \mathsf{Cov}\left[\sum_{i=1}^{t-h} u_i, \sum_{i=1}^{t-h} u_i\right] = \mathsf{Var}\left[\sum_{i=1}^{t-h} u_i\right] = (t-h)\,\sigma_u^2 \quad pour\ h \geq 0 \\
\gamma_t(h) &= t\sigma_u^2 \quad pour\ h < 0.
\end{aligned}$$

et le processus n'est pas stationnaire, il s'agit d'une marche aléatoire.

(ii) *ici nous avons affaire à une marche aléatoire avec dérive (non stationnaire)*

$$\begin{aligned}
y_t &= \tau + y_{t-1} + u_t \\
&= \tau t + \sum_{i=1}^{t} u_i
\end{aligned}$$

donc

$$\begin{aligned}
\mu_t &= \mathsf{E}[y_t] = \tau t \\
\gamma_t(0) &= \mathsf{Var}\left[\sum_{i=1}^{t} u_i\right] = \sum_{i=1}^{t} \mathsf{Var}[u_i] = t\sigma_u^2 \\
\gamma_t(h) &= \mathsf{E}\left[\left(\tau t + \sum_{i=1}^{t} u_i\right)\left(\tau(t-h) + \sum_{i=1}^{t-h} u_i\right)\right] - \tau t\,[\tau(t-h)] \\
&= (t-h)\,\sigma_u^2 \quad pour\ h \geq 0 \quad et \quad t\sigma_u^2 \quad pour\ h < 0.
\end{aligned}$$

(*iii*) *A présent le processus* $MA(1)$

$$y_t = u_t - 1/2u_{t-1}$$

nous savons qu'il est stationnaire (propriété des MA) et donc

$$\begin{aligned}
\mu_t &= 0 \\
\gamma_t(0) &= \mathsf{Var}\left[u_t - 1/2u_{t-1}\right] = \left(1 + \frac{1}{4}\right)\sigma_u^2 \\
\gamma_t(1) &= \mathsf{Cov}\left[u_t - 1/2u_{t-1}, u_{t-1} - 1/2u_{t-2}\right] \\
&= -1/2\mathsf{Var}\left[u_{t-1}\right] = -\frac{\sigma_u^2}{2}. \\
\gamma_t(h) &= 0 \quad pour\ h \geq 2.
\end{aligned}$$

(*iv*) *le processus est à présent moins courant*

$$y_t = \begin{cases} u_t & pour\ t\ impair \\ 2u_t & pour\ t\ pair \end{cases}$$

et ainsi

$$\mu_t = 0$$

$$\gamma_t(0) = \begin{cases} \sigma_u^2 & pour\ t\ impair \\ 4\sigma_u^2 & pour\ t\ pair \end{cases}$$

et

$$\gamma_t(h) = 0 \quad pour\ h \geq 1$$

la variance de ce processus dépend t (du moins de sa parité) et donc la série est non stationnaire.

Exercice 4 *On souhaite analyser un modèle ARMA(1,1) donné par :*

$$y_t = \alpha y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1},\ pour t = 1, ..., T \tag{3}$$

avec $\varepsilon_t \sim \mathsf{NID}(0,1)$ *et* $y_0 = 0$. *On définit les polynômes* A *et* B *tels que (3) s'écrive*

$$A(L)\, y_t = B(L)\, \varepsilon_t,$$

où L *représente l'opérateur retard.*
(*i*) *Quels sont les coefficients des polynômes* A *et* B *?*
(*ii*) *Donner les racines de* A *et de* B.

(iii) Quelles conditions les coefficients α et θ doivent-ils satisfaire pour que le processus $\{y_t\}$ soit stationnaire ? inversible ? On suppose ces conditions satisfaites dans les questions suivantes.
(iv) Quel est le rôle de l'hypothèse $y_0 = 0$?
(v) Quelle sont l'espérance $\mathsf{E}[y_t]$, la variance $\mathsf{V}[y_t]$, et l'autocovariance $\mathsf{COV}[y_t, y_{t-h}]$ (pour $h \geq 1$) ?
(vi) Est-il possible que $\mathsf{V}[y_t] < 1$? Commenter les valeurs possibles de $\mathsf{V}[y_t]$, $\frac{\mathsf{COV}[y_t,y_{t-1}]}{\mathsf{V}[y_t]}$ et $\frac{\mathsf{COV}[y_t,y_{t-1}]}{\mathsf{COV}[y_t,y_{t-2}]}$ quand $\alpha = 0$, ou $\theta = 0$.
(vii) Que se passe-t-il quand $\alpha + \theta = 0$? Expliquez.
(viii) Dans chacun des cas suivants, on estime les valeurs des variances et covariances sur un "grand" échantillon, en déduire les paramètres α et θ :

$$\begin{array}{lllll} (a) & \mathsf{V}[y_t] = 1,25 & \mathsf{COV}[y_t, y_{t-1}] = 0,5 & \mathsf{COV}[y_t, y_{t-2}] = 0 \\ (b) & \mathsf{V}[y_t] = 2 & \mathsf{COV}[y_t, y_{t-1}] = \sqrt{2} & \mathsf{COV}[y_t, y_{t-2}] = 1 \end{array}$$

Correction 4 *(i) L'équation (3) peut être réécrite sous la forme :*

$$y_t - \alpha y_{t-1} = \varepsilon_t + \theta \varepsilon_{t-1}$$

soit

$$(1 - \alpha L) y_t = (1 + \theta L) \varepsilon_t$$

ce qui implique que A et B soient de degré 1 et si on pose $A(x) = a_1 x + a_0$ et $B(x) = b_1 x + b_0$, les coefficients sont

$$\begin{aligned} a_1 &= -\alpha \quad et \quad a_0 = 1; \\ b_1 &= \theta \quad et \quad b_0 = 1. \end{aligned}$$

(ii) Les racines des polynômes sont telles que

$$\begin{aligned} A(z_A) &= 0 \Leftrightarrow z_A = \frac{1}{\alpha}, \\ B(z_B) &= 0 \Leftrightarrow z_B = -\frac{1}{\theta}. \end{aligned}$$

(iii) y_t est stationnaire si sa partie autorégressive l'est, i.e, si $|z_A| > 1$, ou $|\alpha| < 1$. La même condition sur la partie de moyenne mobile implique l'inversibilité du processus, i.e. si $|z_B| > 1$, ou $|\theta| < 1$.
(iv). L'hypothèse $y_0 = 0$ permet de s'abstraire de la non-stationarité que la valeur origine entraîne dans les "petits" échantillons.
(v) Calcul de l'espérance : le processus étant stationnaire, $\mathsf{E}[y_t]$ est tel que

$$\begin{aligned} \mathsf{E}[y_t] &= \mathsf{E}[\alpha y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}] \\ &= \alpha \mathsf{E}[y_{t-1}] \end{aligned}$$

et donc $\mathsf{E}[y_t] = 0$.
Calcul de la variance et des covariances : selon les equations de Yule-Walker et puisque l'espérance de y_t *est nulle*

$$\begin{aligned}\mathsf{E}\left[y_t^2\right] &= \gamma_0 = \mathsf{E}\left[\left(\alpha y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}\right) y_t\right] \\ &= \alpha\mathsf{E}\left[y_t y_{t-1} + y_t\varepsilon_t + \theta y_t\varepsilon_{t-1}\right] \\ &= \alpha\mathsf{E}\left[y_t y_{t-1}\right] + \mathsf{E}\left[y_t\varepsilon_t\right] + \theta\mathsf{E}\left[y_t\varepsilon_{t-1}\right] \\ &= \alpha\gamma_1 + \mathsf{E}\left[y_t\varepsilon_t\right] + \theta\mathsf{E}\left[y_t\varepsilon_{t-1}\right].\end{aligned}$$

Or $\mathsf{E}[y_t\varepsilon_t] = \mathsf{E}\left[\left(\alpha y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}\right)\varepsilon_t\right] = \mathsf{E}[\varepsilon_t\varepsilon_t]$ *car* y_{t-1} *et* ε_{t-1} *sont indépendants de* ε_t. *De plus*

$$\begin{aligned}\mathsf{E}[y_t\varepsilon_{t-1}] &= \mathsf{E}\left[\left(\alpha y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}\right)\varepsilon_{t-1}\right] \\ &= \alpha\mathsf{E}[y_t\varepsilon_t] + \theta \\ &= \alpha + \theta\end{aligned}$$

Ainsi

$$\gamma_0 = \alpha\gamma_1 + 1 + \alpha\theta + \theta^2,$$

et de même

$$\begin{aligned}\gamma_1 &= \mathsf{E}[y_t y_{t-1}] = \mathsf{E}\left[\left(\alpha y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}\right) y_{t-1}\right] \\ &= \alpha\gamma_0 + \theta.\end{aligned}$$

Par conséquent

$$\begin{aligned}\gamma_0 &= \alpha\left(\alpha\gamma_0 + \theta\right) + 1 + \alpha\theta + \theta^2 \\ \gamma_0 &= \frac{1 + 2\alpha\theta + \theta^2}{1 - \alpha^2} = \frac{1 - \alpha^2 + (\alpha + \theta)^2}{1 - \alpha^2} \\ &= 1 + \frac{(\alpha + \theta)^2}{1 - \alpha^2}\end{aligned}$$

et

$$\gamma_1 = \frac{(1 + \alpha\theta)(\alpha + \theta)}{1 - \alpha^2}.$$

Enfin pour $h > 1$

$$\begin{aligned}\gamma_h &= \mathsf{E}[y_t y_{t-h}] = \mathsf{E}\left[\left(\alpha y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}\right) y_{t-h}\right] \\ &= \alpha\gamma_{h-1} = \frac{\alpha^{h-1}(\alpha + \theta)}{1 - \alpha^2} = \frac{\alpha^h + \alpha^{h-1}\theta}{1 - \alpha^2}.\end{aligned}$$

(vi) est-il possible que $\mathsf{V}[y_t] < 1$ *?* $\mathsf{V}[y_t] = 1 + \frac{(\alpha+\theta)^2}{1-\alpha^2}$ *donc* $\mathsf{V}[y_t] \geq 1$. *A présent*

$$\frac{\gamma_1}{\gamma_0} = \frac{(1+\alpha\theta)(\alpha+\theta)}{1-\alpha^2+(\alpha+\theta)^2}$$

et

$$\frac{\gamma_1}{\gamma_2} = \frac{1}{\alpha}$$

Donc si $\alpha = 0$, $\frac{\gamma_1}{\gamma_0} = \frac{\theta}{1+\theta^2}$ *est compris entre* $-\frac{1}{2}$ *et* $\frac{1}{2}$. *Par ailleurs* γ_2 *est alors nul. Quand* $\theta = 0$, $\frac{\gamma_1}{\gamma_0} = \alpha = \frac{\gamma_2}{\gamma_1}$.
(vii) Quand $\alpha + \theta = 0$, y_t *ne suit plus un bruit blanc car les polynômes* A *et* B *ont alors une racine commune :*

$$y_t = \varepsilon_t.$$

Ce qu'on perçoit en observant la fonction d'autocovariance : $\gamma_0 = 1$ *et* $\gamma_h = 0$ *pour* $h < 0$.
(viii) (a) On constate que $\gamma_2 = 0$ *mais* γ_1 *non nul, donc* $\alpha = 0$ *et par conséquent* $\gamma_0 = 1 + \theta^2$, *donc* $\theta = 0,5 = \gamma_1$.
(b) A présent $\frac{\gamma_2}{\gamma_1} = \frac{1}{\sqrt{2}} = \alpha$. *Donc* $\gamma_0 = 1 + \frac{\left(1/\sqrt{2}+\theta\right)^2}{1/2} = 2$ *donc* $\left(1/\sqrt{2}+\theta\right)^2 = \frac{1}{2}$ *et* $1/\sqrt{2} + \theta = \frac{1}{\sqrt{2}}$, *i.e.* $\theta = 0$. *On vérifie que* $\frac{\gamma_1}{\gamma_0} = \alpha$.

Exercice 5 *Le but de cet exercice est de comparer les propriétés prévisionnelles de divers modèles. On considère les processus générateurs suivants :*

$$\begin{aligned}
(a) \quad & y_t = \alpha + \beta t + \varepsilon_t \\
(b) \quad & y_t = y_{t-1} + u_t \\
(c) \quad & y_t = \tau + \rho y_{t-1} + v_t \text{ avec } |\rho| < 1 \\
(d) \quad & y_t = \mu + w_t + \theta w_{t-1}
\end{aligned}$$

On suppose que les processus ε_t, u_t, v_t *et* w_t *sont tous des bruits blancs indépendents les uns des autres et de variances toutes égales à 1. On s'intéresse à un échantillon comprenant* T *observations et on suppose que* $y_0 = 0$.
(i) Calculez l'espérance $\mathsf{E}[y_T]$ *et la variance* $\mathsf{V}[y_T]$ *de* y_T *données par chacun des modèles à l'instant* T *(notez qu'elles peuvent dépendre de* T*).*
(ii) Quels sont les processus stationnaires parmi les quatre considérés ? Commentez les propriétés de chacun.
(iii) A la date T, *on souhaite établir une prévision de* y_{T+h}, *où* $h \geq 2$. *Quelle est la valeur* y_{T+h} *qu'implique chacun des modèles, connaissant* y_T *? (vous donnerez une réponse où la valeur* y_{T+h} *dépend de* y_T *et des* ε_i, u_i, v_i *ou* w_i *pour*

$i \geq T$).
(iv) Pour chacun des modèles $(a), (b), (c), (d)$ on souhaite établir une prévision $\widehat{y}_{T+h|T} = \mathsf{E}[y_{T+h}|y_T]$ (l'espérance conditionnelle de y_{T+h} sachant y_T, on suppose w_T inconnu). Calculez ces prévisions.
(v) Pour chacun des modèles quelle est l'erreur de prévision $\widehat{e}_{T+h|T} = y_{T+h} - \widehat{y}_{T+h|T}$ correspondante ?
(vi) Quelles sont pour chacun des modèles, l'espérance $\mathsf{E}\left[\widehat{e}_{T+h|T}|y_T\right]$ et la variance $\mathsf{V}\left[\widehat{e}_{T+h|T}|y_T\right]$ de l'erreur de prévision, connaissant y_T (on suppose w_T inconnu, y_T est vu comme une constante).
(vii) Commenter les différences entre les erreurs de prévision de chacun des modèles, et leur évolution en fonction de T et h.

Correction 5 *(i) Calcul des espérances et variances à l'instant T.*
(a)

$$\begin{aligned}\mathsf{E}[y_T] &= \mathsf{E}[\alpha + \beta T + \varepsilon_T] = \alpha + \beta T \\ \mathsf{V}[y_T] &= \mathsf{V}[\varepsilon_T] = 1\end{aligned}$$

(b)

$$\begin{aligned}\mathsf{E}[y_T] &= \mathsf{E}[y_{T-1}] = ... = \mathsf{E}[y_0] = 0 \\ \mathsf{V}[y_T] &= \mathsf{V}[y_{T-1}] + \mathsf{V}[u_T] + 2\mathsf{Cov}[y_{T-1}, u_T] \\ &= \mathsf{V}[y_{T-1}] + 1 = \mathsf{V}[y_{T-2}] + 2 \\ &= T + \mathsf{V}[y_0] = T\end{aligned}$$

(c) $|\rho| < 1$, *on fait donc l'hypothèse de stationnarité car* $y_0 = 0$:

$$\begin{aligned}\mathsf{E}[y_T] &= \tau + \rho\mathsf{E}[y_{T-1}] = \tau + \rho\mathsf{E}[y_T] \\ &= \frac{\tau}{1-\rho} \\ \mathsf{V}[y_T] &= \rho^2\mathsf{V}[y_{T-1}] + \mathsf{V}[v_T] + 2\rho\mathsf{Cov}[y_{T-1}, v_T] \\ &= \rho^2\mathsf{V}[y_T] + 1 \\ &= \frac{\tau}{1-\rho^2}.\end{aligned}$$

(d) y_t *suit un processus de moyenne mobile, il est donc stationnaire :*

$$\begin{aligned}\mathsf{E}[y_T] &= \mu \\ \mathsf{V}[y_T] &= \mathsf{V}[w_T] + \theta^2\mathsf{V}[w_{T-1}] + 2\theta\mathsf{Cov}[w_T, w_{T-1}] \\ &= 1 + \theta^2.\end{aligned}$$

(ii) Les processus (c) et (d) sont stationnaires ($AR(1)$) et $MA(1)$). (a) est stationnaire autour d'une tendance déterministe linéaire : sa variance est constante

mais son espérance ne l'est pas. (*b*) *suit une marche aléatoire et est intégré (donc non-stationnaire) : sa différence est stationnaire, son espérance est constante et nulle, sa variance croît avec le temps car elle suit une tendance linéaire.*
(*iii*)

$$
\begin{aligned}
(a) \quad &: \quad y_{T+2} = y_T + 2\beta + \varepsilon_{T+2} - \varepsilon_T \\
(b) \quad &: \quad y_{T+2} = y_T + u_{T+2} + u_{T+1} \\
(c) \quad &: \quad y_{T+2} = \tau\left(1+\rho\right) + \rho^2 y_T + v_{T+2} + \rho v_{T+1} \\
(d) \quad &: \quad y_{T+2} = \mu + w_{T+2} + \theta w_{T+1}
\end{aligned}
$$

(*iv*)

$$
\begin{aligned}
(a) \quad &: \quad y_{T+h} = y_T + h\beta + \varepsilon_{T+h} - \varepsilon_T \\
(b) \quad &: \quad y_{T+h} = y_T + \sum_{i=1}^{h} u_{T+i} \\
(c) \quad &: \quad y_{T+h} = \tau\left(\sum_{i=0}^{h-1}\rho^i\right) + \rho^h y_T + \sum_{i=0}^{h-1}\rho^i v_{T+h-i} \\
(d) \quad &: \quad y_{T+h} = \mu + w_{T+h} + \theta w_{T+h-1}
\end{aligned}
$$

donc les prévisions sont données par (on suppose $\mathsf{E}[\varepsilon_T|y_T]$*) :*

$$
\begin{aligned}
(a) \quad &: \quad \widehat{y}_{T+h} = y_T + h\beta \\
(b) \quad &: \quad \widehat{y}_{T+h} = y_T \\
(c) \quad &: \quad \widehat{y}_{T+h} = \tau\left(\sum_{i=0}^{h-1}\rho^i\right) + \rho^h y_T \\
&= \quad \tau\frac{1-\rho^h}{1-\rho} + \rho^h y_T \\
(d) \quad &: \quad \widehat{y}_{T+h} = \mu
\end{aligned}
$$

(*v*) *L'erreur de prévision est obtenue dans chacun des cas par* $\widehat{e}_{T+h} = y_{T+h} - \widehat{y}_{T+h}$

$$
\begin{aligned}
(a) \quad &: \quad \widehat{e}_{T+h} = \varepsilon_{T+h} - \varepsilon_T \\
(b) \quad &: \quad \widehat{e}_{T+h} = \sum_{i=1}^{h} u_{T+i} \\
(c) \quad &: \quad \widehat{e}_{T+h} = \sum_{i=0}^{h-1}\rho^i v_{T+h-i} \\
(d) \quad &: \quad \widehat{e}_{T+h} = w_{T+h} + \theta w_{T+h-1}
\end{aligned}
$$

dont l'espérance est donnée par

$$\begin{aligned}
(a) &: \mathsf{E}\left[\widehat{e}_{T+h}\right] = 0 \\
(b) &: \mathsf{E}\left[\widehat{e}_{T+h}\right] = 0 \\
(c) &: \mathsf{E}\left[\widehat{e}_{T+h}\right] = 0 \\
(d) &: \mathsf{E}\left[\widehat{e}_{T+h}\right] = 0
\end{aligned}$$

et la variance

$$\begin{aligned}
(a) &: \mathsf{V}\left[\widehat{e}_{T+h}\right] = 2 \\
(b) &: \mathsf{V}\left[\widehat{e}_{T+h}\right] = h \\
(c) &: \mathsf{V}\left[\widehat{e}_{T+h}\right] = \sum_{i=0}^{h-1} \rho^{2i} = \frac{1-\rho^{2h}}{1-\rho^2} \\
(d) &: \mathsf{V}\left[\widehat{e}_{T+h}\right] = 1+\theta^2
\end{aligned}$$

(*vii*) *On constate que les divers modèles fournissent des prévisions non-biaisées (l'espérance de l'erreur est nulle) mais qu'en revanche leur comportement est très divers. La prévision selon* (a) *s'accroît de* β *chaque période, elle est stable à la dernière valeur pour* (b), *retourne lentement vers la moyenne* $\frac{\tau}{1-\rho}$*pour* (c) *car* $\rho^h \to 0$ *quand* $h \to \infty$, *et enfin est stable à la moyenne pour* (d).
En ce qui concerne les variances des erreurs (et donc le calcul de l'incertitude autour de la prévision), elle est stable pour (a) *et* (d), *elle croît linéairement pour* (b) *et enfin elle s'accroît progressivement vers la variance de* y_t *pour* (c).