Econométrie de la Finance

Florian Ielpo[1]

24 février 2008

[1]Dexia Group, 7/11 Quai André Citroën, 75015 Paris, Centre d'Economie de la Sorbonne - Antenne de Cachan, Avenue du Président Wilson, 94230 Cachan. E-mail : florian.ielpo@clf-dexia.com

Table des matières

Introductions

"*S*o if you do not accept the Gaussian distribution (i.e. if you have some ethics) AND do not "value" options in a axiomatized top-down fashion (i.e. only price them as some informed and temporary guess), then YOU ARE NOT USING THE BLACK SCHOLES FORMULA, but one of the modifications of the one started by Bachelier (the latest contributer being Ed Thorp's). They did not ground their formula in the Gaussian."
Nassim Nicholas Taleb[1]

0.1 Introduction de la deuxième édition

Voici donc la deuxième année que j'enseigne ce cours, et de nombreuses choses ont changé dans ma compréhension de la finance et de l'économétrie. Ces changements ont mené à un remaniement complet du présent polycopier et à l'apparition de TD associés à ce cours.

- Un chap de rappels, référence.
- Praise to Cochrane et Singleton. Ajout des GMM.
- Chapitre sur les tests d'hypothèse, peut etre
- Chapitre sur les ARMA/GARCH : modelling the first two moments + asymétrie.
- Chapitre spécial GMM sur un modèle d'équilibre tiré du livre sur les GMM ou de Cochrane.
- ACP et multivarié.
- Calibration d'un modèle à vol stochastique ou d'un CIR par fonction caractéristique.

A ceci s'ajoute le mémoire à rendre.

0.2 Introduction de la première édition

Ce cours s'inscrit dans le prolongement de l'U.V. MF 231 [Statistiques I : inférence et estimation]. Il a pour but de présenter certains approfondissements autour des principaux thèmes de l'économétrie financière.

Il s'agit dans un premier temps de revenir sur le modèle linéaire gaussien, dans sa version univariée et multivariée. On présentera quelques questions simples liées à l'inférence statistique de ces modèles ainsi qu'à l'usage qui peut en être fait : expliquer la dynamqiue des séries économiques/financières et permettre la mise en oeuvre de prévisions

[1] `http://www.wilmott.com/blogs/kurtosis/index.cfm/General`

encadrées par des intervalles de confiance.

Il s'agit ensuite de présenter la base de la théorie des séries temporelles : modèle ARMA, GARCH et modèles à facteurs. Là encore, la principale motivation sera l'inférence efficace ainsi que la prévision.

La philosophie de ce cours se veut naturellement pratique : par la compréhension des modélisations et de l'inférence, il s'agit de permettre la mise en oeuvre de ces modèles dans le cadre d'activités de marché sur la base de n'importe quel logiciel de programmation. Une fois la programmation des procédures d'estimation comprise, il est relativement simple de mettre en place des estimations sous n'importe quel environnement. Il sera proposé tout au long de ce cours des exemples de code `R` permettant de réaliser les estimations proposées. `R` est certainement l'un des meilleurs logiciels de statistique disponibles sur le marché actuellement. Il s'agit d'un logiciel open-source : il est gratuitement téléchargeable sur le site de ses développeurs[2]. Le site fourni une séries de manuels permettant une prise en main rapide et efficace du logiciel : il est conseillé de se procurer Paradis (2005) ainsi Faraway (2002) sur le site (section `manual` puis `contributed documentation`).

Ces notes de cours s'appuient sur un certain nombre d'ouvrages de statistiques bien connus ainsi que sur d'autres notes de cours, qui seront citées à chaque fois. Nous y renvoyons un lecteur soucieux de dépasser le niveau de cette introduction. La partie consacrée au modèle linéaire gaussien est grandement inspirée de Greene (2002). La partie consacrée à l'étude des séries temporelles est principalement inspirée de Cochrane (2005).

La lecture de ces notes de cours ce nécessitent pas de connaissance mathématiques étendue : les seules connaissances nécessaires sont des connaissances de base en algèbre matricielle ainsi qu'en analyse (dérivée et formule de Taylor pour la partie consacrée à l'optimimsation). Quand des éléments plus poussés sont nécessaires, ils sont en général rappelé avant utilisation. Dans cette mesure, ce cours tente de se suffire à lui-même et ne requiere pas de lectures annexes. A chaque fois que cela est nécessaire, le lecteur soucieux d'approfondissements qui sont jugés inutiles à ce niveau est renvoyé à un certain nombre de références, citées en annexes. La plupart des références fournies sont par ailleurs des références gratuitement disponibles sur internet : un nombre croissant de professeurs/cherheurs proposent leurs notes de cours sur internet. Les liens sont généralement fournis sur la page web de mes enseignements (`www.mora.ens-cachan.fr/ielpo`). Ces notes de cours ne sont bien entendu pas développées intégralement en cours : le chapitre 1 est notamment laissé de coté lors mes interventions. Il s'agit davantage de rappels que d'éléments développés en cours. Il en va de même de certains passage du chapitre 2 : les élèves sont censés connaitre un certain nombre de résultats tirés de l'économétrie basique (mco). Ces éléments prennent la forme de rapides rappels en cours : il est nécessaire de combler d'éventuelles lacunes par une lecture plus approfondies des passages évoqués.

Enfin, ces notes de cours sont certainement entachées d'inexactitudes ou d'erreurs.

[2] `http://www.r-project.org/`

Celles-ci sont entièrement miennes : tout commentaires/signalement d'erreurs sont bien évidement les bienvenus. La qualité de ce polycopier ira croissante au fil des ans : l'amélioration est naturellement un processus lent, lié aux réactions des élèves ainsi qu'à la croissance de mes propres connaissances en statistiques et en économétrie. J'espère ainsi que ces modestes notes de cours seront un jour suffisament propres et documentées pour fournir *in fine* un manuel de base suffisament rigoureux pour servir de base aux élèves de l'ESILV.

Chapitre 1

Rappels de mathématiques et probabilité

Cette première partie a pour but de revenir sur un certain nombre de concepts et techniques nécessaires pour comprendre et implémenter les différentes méthodes de l'économétrie de la finance. Il s'agit principalement de revenir sur un certain nombre de concepts de probabilités dans un premier temps (définition d'une variable aléatoire, de ses moments et des distributions qu'il est possible de lui affecter). Il sera ensuite question de revenir sur les concepts de convergence (presque sure, en probabilité et en loi), afin d'introduire la Loi des Grands Nombres (LGN hereafter) et le Théorème Central Limite (TCL). Enfin, on finira par quelques éléments de calculs matriciel.

1.1 Des variables aléatoires et des hommes

1.1.1 L'univers... et au dela

Soit $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ un espace fini d'états, représentant les différents états possibles de la nature à un instant donné. On appelle cet espace *l'univers des possibles*. Cet espace est fini : il n'existe qu'un nombre limité d'état atteignable par le cours du monde (du moins dans notre façon de le concevoir). Chaque événement qu'il est possible de voir se réaliser ω_i est appelé événement élémentaire. Ces événements élémentaires sont incompatibles deux à deux. Tout sous-ensemble de Ω est également appelé événement : il s'agit d'un événement composé. On note par exemple $\mathcal{A} = \{\omega_2, \omega_3, \omega_{10}\}$, un sous ensemble d'événement de Ω. Il s'agit d'un événement composé et $\mathcal{A} \subset \Omega$.

1.1.2 A chacun sa tribu

Parmi l'ensemble des sous-ensemble $\mathcal{P}(\Omega)$, on s'interesse seulement à ceux qui sont dotés d'une certaine structure.

Définition 1.1.1 (Notion de tribu)**.** *On dit qu'une partie $\mathcal{A}$ de $\mathcal{P}(\Omega)$ est une tribu si et seulement si elle vérifie les trois propriétés suivantes :*

1. $\Omega \in \mathcal{A}$.

2. *Pour toute partie A de $\mathcal{A}$, $\overline{A} \in \mathcal{A}$.*
3. *Pour toute famille dénombrable $(A_i)_{i\in I}$ de $\mathcal{A}$ alors $\cup_{i\in I} A_i$ est aussi un élément de $\mathcal{A}$.*

En pratique, le concept de tribu est essentiel en finance : il permet de rendre compte de la façon dont l'information s'organise au fur et à mesure que le s'écoule. Il existe d'autres dénominations pour les tribu : σ-algèbre ou filtration (une filtration est une sigma-algèbre). Le concept de filtration est utilisé courament dans le cadre de modèle stochastiques, tel que celui de Black and Scholes (1973). Un développement remarquable sur ce point peut être trouvé dans Münk (2004). On reviendra sur ce point une fois que l'on sera revenu sur les espaces probabilisés.

Les exemples les plus courants de sigma-algèbre sont :

– $\mathcal{A} = \{\oslash, \Omega\}$ est la tribu grossiere.
– $\mathcal{A} = \{\oslash, A, \overline{A}, \Omega\}$ où A est une partie de Ω, est la tribu de Bernouilli.
– $\mathcal{A} = \mathcal{P}(\Omega)$ est la tribu complète ou triviale.

Globalement, deux types de tribu peuvent nous interesser :

– $\mathcal{A}_0$ la tribu engendrée la famille des singletons $\{\omega\}$ de Ω. Cette tribu est utile lors de la détermination d'une loi de probabilité.
– La tribu complète $\mathcal{P}(\Omega)$.

Dans le cas où Ω est fini ou infini dénombrable (ce qui sera toujours le cas dans ce qui suit), alors ces deux tribus sont *identiques*.

Définition 1.1.2 (Espace probabilisable)**.** *Le couple $(\Omega, \mathcal{A})$ est appelé espace probabilisable. Dans le cas où Ω est fini dénombrable, $\mathcal{A} = \mathcal{P}(\Omega)$.*

1.1.3 Probabilités...

Maintenant que l'on a défini la structure de l'univers dans lequel se déroule l'expérience aléatoire qui nous intéresse, reste à donner une forme au hasard. C'est ce qu'on appelle probabiliser l'espace probabilisable. Il s'agit simplement de définir la probabilité qu'un événement $A \in \mathcal{A}$ survienne.

Définition 1.1.3 (Probabilité)**.** *P est une probabilité définie sur l'espace probabilisable $(\Omega, \mathcal{A})$ si et seulement si P est une application de $\mathcal{A}$ vers $\mathbb{R}$ qui vérifie les propriétés suivantes :*

1. *$0 \leq P(A) \leq 1, \forall A \in \mathcal{A}$.*
2. *$P(\Omega) = 1$ (Axiome de normalisation).*
3. *Pour toute famille finie $(A_i)_{0\leq i\leq n}$ d'événements de $\mathcal{A}$, deux à deux incompatibles, on a :*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

(Axiome de simple additivité).

4. *Pour toute famille dénombrable* $(A_i)_{i\in\mathbb{N}}$ *d'événements de* $\mathcal{A}$*, deux à deux incompatibles, on a :*

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

(Axiome de σ*-additivité).*

Le nombre réel du second membre est la somme de la série de terme général $P(A_i)$. Dans la mesure où les termes de cette série sont positifs et que la probabilité de l'union de n A_i est majorée par 1, cette série est toujours convergente.

Définition 1.1.4. *L'espace* $(\Omega, \mathcal{A}, P)$ *est appelé espace probabilisé.*

Ajoutons les deux définitions suivantes :

Définition 1.1.5. *On sait que* $P(\Omega) = 1$*, mais on peut trouver des événements* $A \neq \Omega$ *et tels que* $P(A) = 1$*. On dit que ces événements sont quasi-certains ou presque sûrs. On sait que* $P(\oslash) = 0$*, mais on peut trouver des événements* $A \neq \oslash$ *et tels que* $P(A) = 0$*. On dit que ces événements sont quasi-impossibles ou négligeables.*

Définition 1.1.6. *Deux distributions* P *et* P' *sont dites équivalentes si elles ont les mêmes négligeables.*

Définition 1.1.7 (Espace probabilisé)**.** *Le couple* $(\Omega, \mathcal{A}, P)$ *est appelé espace probabilisé.*

Notons finalement que la donnée d'une probabilité P sur un espace probabilisable est équivalent à la donnée d'une distribution de probabilité sur Ω.

1.1.4 Variables aléatoires

Avec l'ensemble des éléments précédents en tête, il est alors possible de tourner notre attention vers ce qui fera l'objet de ce cours : les variables aléatoires.

Définition 1.1.8. *Toute application* X *telle que :*

$$X : \Omega \to \mathbb{R}$$

est appelée variable aléatoire, ou plus précisement variable aléatoire réelle.

Il est possible de généraliser le concept de variable aléatoire à celui de vecteur aléatoire : il s'agit d'une application quelconque de Ω dans $\mathbb{R}^k$. On définit alors la distribution jointe du vecteur, au lieu de définir la distribution d'une seule variable aléatoire. Notons que l'on note généralement X cette variable aléatoire et $\{x_1, x_2, ..., x_n\}$ n réalisations de cette variable aléatoire.

1.1.5 Les moments

Avant de s'interesser à la distribution d'une variable aléatoire, il existe d'autres quantités utiles à connaitre : les moments.

Définition 1.1.9. *Le moment d'ordre k d'une variable aléatoire X est la quantité*

$$\mathbb{E}\left[X^k\right] = \int_\Omega x^k f_x(x)dx$$

Le moment centré d'ordre k peut se calculer comme suit :

$$\mathbb{E}\left[(X - \mathbb{E}[X])^k\right] = \int_\Omega (x - \mathbb{E}[x])^k f_x(x)dx$$

où f_x est la densité de probabilité de la variable aléatoire X. Cette densité est telle que :

$$P(i^- < X < i^+) = \int_{i^-}^{i^+} f(x)dx \tag{1.1}$$

On reviendra plus tard sur la définition d'une densité. Le moment d'ordre 1 est l'espérance :

$$\mathbb{E}[X] = \int_\Omega x f_x(x)dx \tag{1.2}$$

Le moment centré d'ordre 2 est la variance :

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int_\Omega (x - \mathbb{E}[x])^2 f_x(x)dx \tag{1.3}$$

La variance mesure l'étalement de la distribution autour de l'espérance. En finance, c'est donc un indicateur de risque - risque de perdre autant que risque de gagner.

Le moment d'ordre 3 normé est la skewness ou coefficient d'asymétrie :

$$Sk[X] = \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^3\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^{3/2}} \tag{1.4}$$

Elle mesure l'asymétrie à gauche (négative) ou à droite (positive) d'une distribution. Enfin, le moment centré et normé d'ordre 4 est la kurtosis :

$$Ku[X] = \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^4\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^{4/2}} \tag{1.5}$$

Elle mesure l'épaisseur des queues de distribution, et donc la possibilité de survenance d'événements dits *extrêmes*. Ces quatre moments fournissent une information considérable sur la forme de la distribution d'une variable aléatoire. Il est également possible de calculer des moments entre variables aléatoires, ou d'un vecteur aléatoire.

La covariance est une mesure de dépendance entre deux variables aléatoires. Soit deux variables aléatoires X et Y :

$$\mathbb{C}ov(X, Y) = \int_{X(\omega)} \int_{Y(\omega)} (x - \mathbb{E}[x])\,(y - \mathbb{E}[y])\, f(x, y)dx, dy \tag{1.6}$$

où $f(x, y)$ est la densité de la loi jointe de X et Y. Elle mesure la chance qu'on deux séries d'évoluer de concert. Il est aisé d'interpréter son signe, mais pas son amplitude. En normant la covariance par le produit des écart-types de X et Y, on obtient une mesure dont il est aisé d'interpréter la valeur : le coefficient de corrélation. Il se calcule comme suit :

$$\rho(X, Y) = \frac{\mathbb{C}ov(X, Y)}{\sigma_X \sigma_Y} \tag{1.7}$$

$\rho(X, Y) \in [-1; 1]$, ce qui rend son interprétation aisée.

1.1.6 Distribution, fonction de répartition et densité

Ces moments n'apportent cependant qu'une information partielle sur les distributions des variables aléatoires. Celles ci sont complement définies par la distributions de probabilités. On ne revient pas ici sur les probabilités attachées à des univers fini (cas discret) : il ne sera ici question uniquement des univers infini dénombrables. Les distributions de variables aléatoires dans ce cadre sont approchées par la fonction de répartition et la densité des distributions.

Définition 1.1.10 (Fonction de répartition)**.** *Soit X une variable aléatoire définie sur l'espace probabilisé $(\Omega, \mathcal{A}, P)$. La fonction de répartition notée F de cette variable aléatoire X est la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :*

$$\boxed{\forall a \in \mathbb{R}, F(a) = P(X \leq a)}$$

Une fonction de répartition a les caractéristiques suivantes :

1. F est monotone croissante sur $\mathbb{R}$.
2. F est une fonction continue à droite en tout point de $\mathbb{R}$.
3. $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to \infty} F(x) = 1$

Définition 1.1.11. *Une fonction f est une densité de probabilité si et seulement si elle possède les trois propriétés suivantes :*

1. *f est positive sur $\mathbb{R}$.*
2. *f est continue sur $\mathbb{R}$, sauf peut être sur un ensemble fini de points $\mathcal{D}$.*
3. *$\int_{-\infty}^{\infty} f(x)dx = 1$.*

Notons qu'une densité n'est pas une probabilité : les conditions précédentes ne stipulent par exemple pas que $f(x) \in [0; 1]$, mais que $f(x)$ est positive et que l'intégrale sur l'univers est égale à 1. En revanche, la densité est liée à la distribution par la fonction de répartition, dans la mesure où :

$$P(X \leq a) = F(a) = \int_{-\infty}^{a} f(x)dx,$$

où f est *une* densité de X. En effet, tout autre fonction g de $\mathbb{R}$ dans $\mathbb{R}$, qui coincide avec f sauf sur un ensemble fini de points de $\mathbb{R}$ est aussi une densité de probabilité de X.

On ne propose pas revue des principales distributions, dans la mesure où il est aisé de trouver ces informations sur Wikipédia.

1.1.7 Loi conditionnelle et lemme des espérances itérées

Un aspect particulièrement important des distributions est la différence entre une distirbution non conditionnelle et conditionnelle. Très classiquement, on présente rapidement le cas discret avant de passer au cas continu.

Supposons que l'on ait affaire à un groupe d'individu composé d'hommes et de femmes, de bons et de mauvais élèves. On peut s'interesser à la probabilité de tirer au hasard un bon éléve au sein de cette population, mais on peut aussi s'intéresser au fait de tirer un bon éléve parmi les hommes. Il s'agit ici de la probabilité de tirer un bon éléve, *sachant* que l'on tire parmi les hommes. On parle dans ce cas de probabilité conditionnelle. On note :

$$P(X = \{\text{tirer un bon éléve}\}|\text{il s'agit d'un homme})$$

La règle de Bayes permet de faire le lien entre les probabilités conditionnelles et non conditionnelles :

Définition 1.1.12 (Probabilité conditionnelle)**.** $P(A|B) = \frac{P(A \bigcap B)}{P(B)}$.

Définition 1.1.13 (Règle de Bayes)**.** $P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$

Dans le cas continu, il est possible d'obtenir une densité conditionnelle. Soit un couple de variables aléatoires (X, Y) définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. Alors la densité de X sachant Y s'écrit :

$$f_{X|Y} = \frac{f_{X,Y}}{f_Y}$$

A ceci s'ajoute une propriété importante : la loi des espérances itérées.

Définition 1.1.14 (Espérances itérées)**.** *Soit une variable aléatoire X sur un espace probabilisé. Alors sont espérance peut être calculée comme suit :*

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]] \tag{1.8}$$

où Y est une autre variable aléatoire par rapport à laquelle on conditionne.

Ceci vient du fait que :

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx && (1.9)\\
&= \int_{-\infty}^{\infty} x \int_{\infty}^{\infty} f_{X,Y}(x,y) dy dx && (1.10)\\
&= \int_{-\infty}^{\infty} x \int_{\infty}^{\infty} f_{X|Y}(x) f_Y(y) dy dx && (1.11)\\
&= \int_{-\infty}^{\infty} f_Y(y) \int_{\infty}^{\infty} x f_{X|Y}(x) dx dy && (1.12)\\
&= \mathbb{E}_Y[\mathbb{E}[X|Y]] && (1.13)
\end{aligned}$$

Ajoutons un petit théorème très utile : le théorème de changement de variable.

Théorème 1.1.1. *Soit une variable aléatoire continue X, ayant $f_X(.)$ pour densité et soit le support de X suivant :*

$$\mathcal{X} = \{x | f_X(x) \geq 0\} \tag{1.14}$$

Si $h(.)$ est une fonction dérivable et strictement monotone de domaine $\mathcal{X}$ et d'image $\mathcal{U}$, alors $U = h(X)$ a pour densité :

$$f_U(u) = f_X(h^{-1}(u)) \left| \frac{dx}{du} \right|, \forall u \in \mathcal{U} \tag{1.15}$$

$$= 0 \text{ } sinon \tag{1.16}$$

1.1.8 Fonction génératrice des moments et fonction caractéristique

Une autre façon de caractériser les distributions est d'utiliser la fonction caractéristique et/ou la fonction génératrice des moments.

Définition 1.1.15. *La fonction caractéristique d'une variable aléatoire X est la fonction suivante :*

$$\gamma(t) = \mathbb{E}[e^{itX}] \tag{1.17}$$

$$= \mathbb{E}[cos(tX)] + i\mathbb{E}[sin(tX)] \tag{1.18}$$

Cette fonction est existe toujours. Elle caractérise entièrement la distribution de X, ce qui en fait une contrepartie aux densités. Il sera ainsi possible de travailler à la fois en terme de densité ou de fonction caractéristique. De plus, il existe un lien entre ces deux fonctions :

Proposition 1.1.1. *Soit X une variable aléatoire de densité $f_X(.)$. Alors :*

$$f_X(x) = \int_{-\infty}^{\infty} e^{-itx} \gamma(t) dt \tag{1.19}$$

Notons ensuite que si X et Y sont deux variables aléatoires indépendantes, i.e. telles que $f_{X,Y} = f_X f_Y$, alors $\mathbb{E}[e^{it(Y+X)}] = \gamma_X(t)\gamma_Y(t)$.

Exercice : *déterminer la fonction caractéristique d'une loi normale.*

Il existe enfin une version réelle de cette fonction caractéristique que l'on appelle fonction génératrice des moments.

Définition 1.1.16. *La fonction caractéristique d'une variable aléatoire X est la fonction suivante :*

$$\phi(t) = \mathbb{E}[e^{tX}] \tag{1.20}$$

On l'appelle fonction génératrice des moments du fait de la propriété suivante :

Proposition 1.1.2. *Soit une variable aléatoire X de fonction génératrice des moments $\phi(t)$. Alors on a :*

$$\mathbb{E}[X^k] = \left. \frac{\partial^k \phi(t)}{\partial t^k} \right|_{t=0} \tag{1.21}$$

Exercice : *déterminer les deux premiers moments d'une loi normale à partir de la fonction génératrice des moments..*

1.2 Le petit monde tres fermé des convergences

Cette section a pour but de présenter un certain nombre de rappels (appels ?) concernant les différents types de convergence probabilistes. L'idée est ici de fournir les principales intuitions nécessaire à l'établissement des différentes versions de la loi des grands nombres ainsi que du théorème central limite.

1.2.1 Convergence en probabilité et presque sure

Définition 1.2.1 (Convergence en probabilité). *La variable aléatoire X_n converge en probabilité vers une constante c si*

$$\lim_{n\to\infty} P(|X_n - c| > \epsilon) = 0, \forall \epsilon > 0. \tag{1.22}$$

On note $plim X_n = c$.

Définition 1.2.2 (Convergence presque sure). *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires et X une v.a. définies sur le même espace probabilisé $(\Omega, \mathcal{A}, P)$. On dit que X_n converge presque sûrement vers X si l'ensemble des ω tels que $X_n(\omega)$ converge vers $X(\omega)$ a pour probabilité 1. On note :*

$$Xn \overset{p.s.}{\to} X. \tag{1.23}$$

La différence entre ces deux convergences est que la convergence presque sure implique une convergence ω par ω, sauf pour une poignée de ω qui sont négligeables. La convergence presque sure implique naturellement la convergence en probabilité (correspond au cas où $i = n$).

La convergence en probabilité permet d'établir différentes versions de la loi faible des grands nombres.

Théorème 1.2.1 (Loi faible des grands nombres de Khinchine). *Si $x_1, x_2, ..., x_n$ est un échantillon aléatoire de n réalisations i.i.d. issu d'une distribution de moyenne finie $\mathbb{E}[x_i] = \mu, \forall i$, alors :*

$$plim \frac{1}{n} \sum_{i=1}^{n} x_i = \mu \tag{1.24}$$

Ce théorème est particulièrement important, dans la mesure où il permet l'estimation des moments d'une distribution, pourvu que les conditions d'applicabilité du théorème soient respectées. Une version plus forte de ce théorème existe également, utilisant une convergence p.s. :

Théorème 1.2.2 (Loi forte des grands nombres de Kolmogorov). *Si $x_1, x_2, ..., x_n$ est un échantillon aléatoire de n réalisations indépendantes tel que $\mathbb{E}[X_i] = \mu_i < \infty$ et $\mathbb{V}[X_i] = \sigma_i^2 < \infty$ et $\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty$ lorsque $n \to \infty$ alors*

$$\frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \mu_i \overset{p.s.}{\to} 0 \tag{1.25}$$

Définition 1.2.3 (Estimateur convergent/cohérent). *Un estimateur $\hat{\theta}_n$ d'un paramètre θ est un estimateur convergent ssi*

$$plim \hat{\theta}_n = \theta \tag{1.26}$$

1.2.2 Convergence en distribution et TCL

Définition 1.2.4 (Convergence en distribution)**.** *On dit qu'une suite de variables aléatoires X_n converge en loi vers une variable aléatoire X, si la suite $\{F_n(x)\}$ converge en tout point x où F est continue. On écrit alors :*

$$X_n \xrightarrow{L} X \tag{1.27}$$

Théorème 1.2.3 (Théorème central limite)**.** *Soit $(X_i)_{i\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées, avec $\forall i, \mathbb{E}[X_i] = m, \mathbb{V}[X_i] = \sigma^2$. On a alors :*

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - m}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{L} N(0,1) \tag{1.28}$$

1.3 Vous reprendrez bien un petit peu de calcul matriciel ?

Pour terminer cette section introductive, voici quelques rappels de calcul matriciel. On rappelle qu'une matrice M $\mathcal{M}(n \times p)$ est matrice telle que :

$$M = \begin{pmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,p} \\ m_{2,1} & \dots & \dots & m_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n,1} & m_{n,2} & \dots & m_{n,p} \end{pmatrix} \tag{1.29}$$

Une matrice carrée est telle que $n = p$. Une matrice symétrique est une matrice carrée telle que $m_{i,j} = m_{j,i}, \forall i \neq j$.

Le rang (colonne) d'une matrice $\mathcal{M}(n \times p)$ est le nombre maximum de colonnes qui sont linéairement indépendantes les unes des autres. En notant $r(A)$ le rang d'une matrice A qui soit une $\mathcal{M}(n \times p)$, il vient naturellement que :

$$r(A) \leq min(n, p). \tag{1.30}$$

Une matrice carré d'ordre n est non singulière si son rang est égal à n (par exemple une matrice diagonale).

Définition 1.3.1 (Matrice inverse)**.** *Soit A une matrice carrée d'ordre n. Son inverse, notée A^{-1} si elle existe, est la matrice de même dimension telle que :*

$$AA^{-1} = A^{-1}A = I, \tag{1.31}$$

où I est la matrice identité.

Si la matrice A^{-1} existe, alors on dit que la matrice A est inversible. Cette matrice existe si et seulement si la matrice A est plein rang, autrement dit si la matrice A est non singulière.

Dans ce qui suit, on suppose acquit les éléments suivants : la somme de deux matrices, le produit de deux matrices, la trace d'une matrice ainsi que le déterminant d'une matrice. On rappelle en revanche différentes opérations de différenciation de matrices.

Soit le produit matriciel suivant :

$$y = Ax \tag{1.32}$$

avec A une matrice $n \times p$ et x une matrice colonne $n \times 1$. Si A ne dépend pas de x, alors on a :

$$\frac{\partial y}{\partial x} = A. \tag{1.33}$$

Soit le produit matriciel suivant :

$$y = x^{\mathsf{T}} A \tag{1.34}$$

alors on a :

$$\frac{\partial y}{\partial x} = A^{\mathsf{T}}. \tag{1.35}$$

Soit maintenant :

$$y = x^{\mathsf{T}} A x \tag{1.36}$$

avec A une matrice $p \times p$ et x une matrice colonne de taille p. Alors, on a :

$$\frac{\partial y}{\partial x} = x^{\mathsf{T}}(A + A^{\mathsf{T}}). \tag{1.37}$$

Si A est symétrique, alors :

$$\frac{\partial y}{\partial x} = 2x^{\mathsf{T}} A. \tag{1.38}$$

Chapitre 2

Retour sur le modèle linéaire : cas univarié et multivarié

On présente dans ce qui suit :

- Quelques rappels sur le modèle de régression linéaire multiple : spécification, inférence et test.
- Le modèles à équation multiples : spécification, inférence et tests.

2.1 Le modèle de régression linéaire simple

Le modèle de régression linéraire multiple étudie la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Sa forme est alors :

$$y = f(x_1, x_2, ..., x_n) + \epsilon \tag{2.1}$$
$$= x_1\beta_1 + x_2\beta_2 + ... + x_n\beta_n + \epsilon \tag{2.2}$$

On dit que y est la variable expliquée ou endogène et $\{x_1, x_2, ..., x_n\}$ sont les variables explicatives ou exogènes. ϵ est une perturbation aléatoire : il vient perturber une relation qui, sans lui, resterait stable. Ce terme reçoit de nombreuses dénominations, selon les champs d'application de l'économétrie ainsi que les séries étudiées. Quelques exemples : il est possible de qualifier les ϵ de *bruit* (artefact statistique qui ne comporte pas d'information particulière), d'erreur de mesure (erreur sur la compréhension de y que permet de le modèle), de choc exogène (un choc qui ne transite pas par les variables du modèle)...
Afin d'estimer cette relation, on utilise un échantillon de l'ensemble des variables. On note y_i la $i$ème valeur de l'échantillon des y et $x_{i,j}$ la $i$ème valeur de l'échantillon de la $j$ème variable. La valeur observée de y_i est alors la somme de deux composantes : l'une déterministe (les $x_{i,j}, \forall j$) et l'autre aléatoire, ϵ_i.

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_n x_{i,n} + \epsilon_i \tag{2.3}$$

L'objectif est d'estimer les paramètres inconnus du modèle, d'utiliser les données pour étudier la validité de certaines propositions théoriques, et éventuellement de former une

prévision de y. On s'appuie dans cette démarche sur un corpus statistique bien connu : les estimateurs du maximum de vraisemblance.

2.1.1 Les hypothèses du modèle linéaire simple

Le modèle linéaire simple s'appuie sur un nombre important d'hypothèses que l'on précise ici.

H 1 (Linéarité)**.** *La relation entre y et les x_j est linéaire.*

H 2 (Plein rang)**.** *Il n'existe pas de relation linéaire entre les variables indépendantes. Si X est la matrice des observations, on aura notamment $X'X$ de plein rang, et donc inversible.*

H 3 (Exogéneité des variables indépendantes)**.** $\mathbb{E}[\epsilon_i|x_{i,1}, x_{i,2}, ..., x_{i,n}] = 0$. *L'espérance de la perturbation conditionnellement aux variables exogènes est nulle : les variables exogènes n'apportent plus aucune information sur la perturbation.*

H 4 (Homoscédasticité et absence d'autocorrélation)**.** $\mathbb{V}[\epsilon] = \sigma^2$ *est stable au cours du temps.* $\mathbb{E}[\epsilon_i\epsilon_j] = 0, \forall i \neq j$, *autrement dit la corrélation entre ϵ_i et ϵ_j est nulle.*

H 5 (Données générées de manière exogène)**.** *Les observations de $\{x_1, x_2, ..., x_n\}$ peuvent être un mélange de constantes et de variables aléatoires. Les processus ayant généré ces données sont indépendants de ϵ (il s'agit d'une extension de **H3**.*

H 6 (Distribution normale)**.** *La perturbation suit une loi normale, en générale centrée et variance constante.*

Une fois ces hypothèses mises à jour, on revient sur l'écriture du modèle. On preferera travailler avec des matrices, plutôt qu'avec un indiçage couteux en place et en patience (la mienne). Mieux, la plupart des logiciels de statistique/économétrie ont le bon goût de fonctionner également en matriciel. Cette démarche simplifie grandement les calculs, comme on le verra par la suite.
Soit $x_{:k}$ le vecteur colonne de T observations de la variable x_k, $k = 1, .., n$. Soit X un matrice $\mathcal{M}_{T,n}$ constituée par la concaténation des différents vecteurs colonnes. Dans la plupart des cas, la première colonne de X est constituée par un vecteur colonne unitaire $(1, 1, ..., 1)'$, de façon à ce que β_1 soit la constante du modèle.
Sur la base de ces éléments, il est possible de réécrire 2.2 sous forme matricielle :

$$\underbrace{y}_{\mathcal{M}_{T,1}} = \underbrace{X}_{\mathcal{M}_{T,n}} \underbrace{\beta'}_{\mathcal{M}_{n,1}} + \underbrace{\epsilon}_{\mathcal{M}_{T,1}} \tag{2.4}$$

On notera ici que β est le vecteur ligne des paramètres :

$$\beta = (\beta_1, \beta_2, ..., \beta_n) \tag{2.5}$$

Quelques remarques générales sur les hypothèse citées plus haut :

– L'hypothèse de linéarité (**H1**) implique également l'additivité du terme d'erreur.

– L'hypothèse **H2** rappelle qu'il ne peut exister de relation linéaire entre les variables explicatives. X est une matrice $\mathcal{M}_{T,n}$: elle doit donc être de rang n, i.e. de plein rang colonne. Deux conditions sont à remplir pour cela :
  – une condition d'identification : il est nécessaire de disposer de n observations au moins ;
  – la non-colinéarité entre vecteurs colonne.

  Rajoutons également qu'au moins un régresseur doit varier (et par conséquent être non constant). Dans le cas contraire, la condition de plein rang n'est pas vérfiée (deux colonnes sont identiques à une constante mutliplicative près).

– La nullité de l'espérance conditionnelle des ϵ implique également la nullité de l'espérance non conditionnelle. Ceci se montre très simplement en conditionnant proprement :

$$\begin{aligned}\mathbb{E}[\epsilon] &= \mathbb{E}_x[\mathbb{E}[\epsilon|x_1, x_2, ...x_n]] &(2.6)\\ &= \mathbb{E}_x[0] &(2.7)\\ &= 0 &(2.8)\end{aligned}$$

  L'hypothèse de nullité de l'espérance des erreurs n'est pas une hypothèse contraigante (voir Greene (2002), page 15).

– Hypothèse **H4** : la variance des erreurs est constante. On parle d'homoscédasticité. Dans le cas où elle varie selon les observations, on parle d'hétéroscédasticité. En ajoutant l'hypothèse d'absence d'autocorrélation, on a alors :

$$\begin{aligned}\mathbb{E}[\epsilon'\epsilon|X] &= \begin{pmatrix} \mathbb{E}[\epsilon_1\epsilon_1|X] & \mathbb{E}[\epsilon_1\epsilon_2|X] & ... & \mathbb{E}[\epsilon_1\epsilon_n|X] \\ \mathbb{E}[\epsilon_2\epsilon_1|X] & \mathbb{E}[\epsilon_2\epsilon_2|X] & ... & \mathbb{E}[\epsilon_2\epsilon_n|X] \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{E}[\epsilon_n\epsilon_1|X] & \mathbb{E}[\epsilon_n\epsilon_2|X] & ... & \mathbb{E}[\epsilon_n\epsilon_n|X] \end{pmatrix} &(2.9)\\ &= \begin{pmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & \sigma^2 \end{pmatrix} &(2.10)\\ &= \sigma^2 \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & 1 \end{pmatrix} &(2.11)\\ &= \sigma^2\mathbb{I} &(2.12)\end{aligned}$$

  Ceci résume l'hypothèse **H4** : la matrice de variance/covariance des perturbations est une matrice bloc diagonal, avec $diag(\mathbb{E}[\epsilon'\epsilon|X]) = \sigma^2$. Comme précédement, il est possible d'utiliser l'expression de la variance conditionnelle des erreurs, afin d'en inférer la variance non conditionnelle (formule de décomposition de la variance) :

$$\mathbb{V}[\epsilon] = \mathbb{E}[\mathbb{V}[\epsilon|X]] + \mathbb{V}[\mathbb{E}[\epsilon|X]] \tag{2.13}$$
$$= \sigma^2\mathbb{I} \tag{2.14}$$

Notons finalement que les perturbations satisfaisant à la fois l'hypothèse d'absence d'autocorrélation et d'homoscédasticité sont parfois appelées *perturbations sphériques*.

– Notons enfin que les perturbations sont supposées suivre une loi normale d'espérance nulle et de variance égale à $\sigma^2\mathbb{I}$. Il s'agit d'une hypothèse bien fondée étant donné la structure de ϵ (les perturbations sont formées d'une suite de chocs, de même loi et de mêmes moments 1 et 2 : le théorème central limit s'applique sans restriction).

2.1.2 Les moindres carrés

Les paramètres de la relation $y = X\beta' + \epsilon$ sont à estimer. Les moindres carrés ordinaires forment une méthode simple et très utilisée, même si elle n'est pas toujours la meilleure. Dans ce qui suit, on cherche $\widehat{\beta}$, un estimateur de β, la vraie valeur. Cet estimateur se doit de vérifier un certain nombre de bonnes propriétés que l'on détaillera par la suite. Dans ce qui suit, on appellera $\widehat{\epsilon}$ les erreurs produites par le modèle estimé. La méthode des MCO se propose de minimiser l'erreur quadratique produite par le modèle.
Le programme résolu pour les MCO est alors le suivant :

$$Min\ (Y - X\widehat{\beta}')^2 \tag{2.15}$$

La résolution est simple. Il suffit de dériver l'expression à minimiser par rapport à β et de chercher la valeur de β (unique avec nos conditions) l'annulant :

$$2X'(Y - X\widehat{\beta}') = 0 \tag{2.16}$$
$$\Leftrightarrow X'Y = X'X\widehat{\beta}' \tag{2.17}$$
$$\Leftrightarrow \widehat{\beta}' = (X'X)^{-1}X'Y \tag{2.18}$$

L'estimateur des paramètres du modèle par la méthode MCO est alors :

$$\boxed{\widehat{\beta}' = (X'X)^{-1}X'Y} \tag{2.19}$$

Il est alors possible de montrer que cet estimateur est *sans biais* et de calculer sa variance :

– $\mathbb{E}[\widehat{\beta}'] = \mathbb{E}[(X'X)^{-1}X'Y] = (X'X)^{-1}X'X\beta + \mathbb{E}[(X'X)^{-1}X'\epsilon] = \beta$
– $\mathbb{V}[\widehat{\beta}'] = \mathbb{V}[(X'X)^{-1}X'Y] = \mathbb{V}[(X'X)^{-1}X'\epsilon] = (X'X)^{-1}\sigma$

La distribution de l'estimateur est par conséquent la suivante :

$$\widehat{\beta}' \sim N(\beta', (X'X)^{-1}\sigma^2) \tag{2.20}$$

Le fait de connaitre cette distribution permet d'élaborer un certain nombre de tests.

On ajoute le théorème suivant :

Théorème 2.1.1 (Régression orthogonale)**.** *Si les variables dans une régression mutliple ne sont pas corrélées (autrement dit, si elles sont orthogonales), alors les estimations obtenues sont les mêmes que celles obtenues dans les regressions individuelles simples.*

Ce théorème est d'une importance capitale : lorsque la condition de non colinéarité entre variables explicatives est vérifiée, il est alors identique de procéder à l'estimation des paramètres les uns à la suite des autres ou d'un seul bloc. Ceci mène naturellement au théorème de Frisch-Waugh. On l'illustre comme suit :

Supposons que la régression à mener implique deux sous-ensembles de variables :

$$y = X\beta' + \epsilon = X_1\beta_1' + X_2\beta_2' + \epsilon \tag{2.21}$$

Quelle est alors la solution pour β_2 ? Les équations normales (i.e. l'équation obtenue après dérivation de l'erreur quadratique telle qu'elle est définie par le modèle) sont alors les suivantes :

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \beta_1' \\ \beta_2' \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \tag{2.22}$$

Ce problème peut être résolu de deux façons possibles : soit en utilisant les règles connues sur les matrices partionnées, soit en développant l'expression matricielle.

Rappel 1. Pour la matrice partitionnée de type 2×2, on a l'inverse partitionnée suivante :

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11}^{-1}(I + A_{12}F_2A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}F_2 \\ -F_2A_{21}A_{11}^{-1} & F_2 \end{bmatrix} \tag{2.23}$$

Avec :

$$F_2 = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \tag{2.24}$$
$$F_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} \tag{2.25}$$

Quelle que soit la méthode, on a le résultat suivant :

$$\beta_1' = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2\beta_2 \tag{2.26}$$
$$= (X_1'X_1)^{-1}X_1'(y - X_2\beta_2) \tag{2.27}$$

Ce premier résultat est un début d'illustration du théorème de Firsch-Waugh : l'estimation de β_1 peut donc se faire, non sur y, mais sur y net de l'information sur y contenue dans X_2. En effet, $y - X_2\beta_2$ est le résidus de la regression de y sur X_2. On note y_2 ce résidus. En utilisant ce résultat, et en remplaçant dans la seconde équation, on trouve :

$$X_2'X_1(X_1'X_1)^{-1}X_1'y_2 + X_2'X_2\beta_2' = X_2'y \tag{2.28}$$
$$\Leftrightarrow \beta_2' = (X_2'X_2)^{-1}(X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y_2) \tag{2.29}$$
$$\Leftrightarrow \beta_2' = (X_2'X_2)^{-1}X_2'(y - X_1(X_1'X_1)^{-1}X_1'y_2) \tag{2.30}$$

$X_1(X_1'X_1)^{-1}X_1' = P_1$: il s'agit du projecteur dans l'espace des X_1. Il s'agit d'une matrice symétrique et idempotente [i.e. $P^2 = P$]. On remarque :

$$y = X_1\beta_1' + \epsilon \tag{2.31}$$
$$= X_1(X_1'X_1)^{-1}X_1'y + \epsilon \tag{2.32}$$
$$= P_1y + \epsilon \tag{2.33}$$
$$\Leftrightarrow y - \epsilon = P_1y \tag{2.34}$$

La formule suivante fournit donc l'expression de β_1' par estimation itérative :

$$\beta_2' = (X_2'X_2)^{-1}X_2'(y - P_1y_2) \tag{2.35}$$

L'intuition de ces différents calculs est la suivante : il est identique de procéder à la régression de y sur X, ou de partitionner X entre X_1 et X_2, puis de regresser de façon iterative y sur X_1 et X_2. Ceci est résumé dans les théorème suivant :

Théorème 2.1.2 (Théorème de Frisch-Waugh *modifié*)**.** *Dans la régression linéaire des moindres carrés du vecteur y sur deux ensembles de variables X_1 et X_2, le sous-vecteur des coefficients β_2' est obtenu en régressant $y - P_1y_2$ sur X_2.*

Théorème 2.1.3 (Théorème de Frisch-Waugh)**.** *Dans la régression linéaire des moindres carrés du vecteur y sur deux ensembles de variables X_1 et X_2, le sous-vecteur des coefficients β_2' est obtenu lorsque les résidus de la regression de y sur X_1 sont régressés sur l'ensemble des résidus de la régression de chaque colonne de X_2 sur X_1.*

Ajoutons un dernier théorème, dont on trouvera la démonstration dans Crépon (2005) ainsi qu'un peu plus d'explicitations : le théorème de *Gauss-Markov*. Il s'agit simplement d'un théorème d'optimalité des estimateurs MCO (optimalité du point de vue de la variance des estimateurs).

Théorème 2.1.4 (Gauss Markov)**.** *Sous les hypothèses du modèle linéaire, l'estimateur des moindres carrés ordinaire d'un modèle linéaire est optimal dans la classe des estimateurs sans biais, conditionnellement aux régresseurs.*

2.1.3 Analyse de la variance

Une fois l'estimation accomplie, on dispose de paramètres estimés ainsi que de résidus - les erreurs produites par le modèle, en principe minimales. On est alors en mesure de déterminer par une *analyse de la variance* la part de la variance de Y qui se trouve expliquée par le modèle. On construit ainsi un R^2, ou coefficient de détermination, explicitant l'idée précédente. La formule est la suivante :

$$R^2 = \frac{SCE}{SCT} \tag{2.36}$$
$$= \frac{SCT - SCR}{SCT} \tag{2.37}$$
$$= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \tag{2.38}$$

Dans le précédent calcul, SCT représente la somme des carrés totaux, SCR, la somme des carrés résiduels et SCE, la différence entre les deux, c'est à dire *la somme des carrés expliqués*. $\bar{Y}$ est la moyenne empirique de Y. Le R^2 se définit donc comme le rapport de la somme des carrés expliqués sur la somme des carrés totaux. L'idée est en fait de décomposer la variance de Y en une variance expliquée par le modèle estimé et une variance qui n'a pas pu être expliquée. Naturellement, plus R^2 est grand et plus - en principe - le modèle peut être soupçonné d'être explicatif de la variable endogène. Cet indicateur, par construction est toujours compris entre 0 et 1. Ainsi plus le R^2 est proche de 1 (de 0) et plus (moins) le modèle est explicatif de Y.
Une mise en garde s'impose ici : dans une regression multiple, le R^2 augmente naturellement lorsqu'une variable supplémentaire (vérifiant certaines conditions qui ne sont pas détaillées ici) est ajoutée. Ceci signifie donc que l'intoduction d'un grand nombre de variable peut naturellement conduire à obtenir un R^2 important, quand bien même le pouvoir explicatif du modèle est médiocre.

2.1.4 Quelques tests liés aux MCO

On présente rapidement les tests les plus connus utiles lors de la mise en oeuvre d'estimations basées sur les MCO : test de Fisher, test de Student, test de Durbin et Watson et tests d'adéquation.

2.1.4.1 Test de Fisher

Dans le modèle linéaire simple - et à *fortiori* dans un modèle MCO - le R^2 est utilisé dans le cadre du test de Fisher. Il s'agit d'un test de nullité des paramètres du modèle. L'idée est la suivante : on fait l'hypothèse que l'ensemble des paramètres β ont une valeur égale à 0 et on compare la vraisemblance de cette hypothèse à l'hypothèse alternative dans le cadre de laquelle les paramètres ont la valeur obtenue après estimation. On retient naturellement l'hypothèse la plus vraisemblable - sur la base d'un test statistique.

La statistique de test est la suivante :

$$\boxed{F_{test} = \frac{R^2}{1 - R^2}\frac{N - P - 1}{P} \sim F(P, N - P - 1)} \tag{2.39}$$

Si la valeur de la statistique de test est supérieure au quantile de la loi de Fisher, on rejète l'hypothèse de nullité de l'ensemble des paramètres - autrement dit, le modèle a de bonne chances d'accroitre notre connaissance de Y.

2.1.4.2 Test de Student

Une fois l'ensemble des paramètres testés, il peut être intéressant de tester les paramètres les uns après les autres. Pour cela, on utilise un test de Student. Là encore, on fait l'hypothèse initiale que le paramètre α_i est nul et on compare cette hypothèse à l'alternative de $\alpha_i = \widehat{\alpha_i}$. Dans le cas d'un modèle avec bruits gaussiens, on connait la distribution des estimateurs MCO :

$$\widehat{\beta}' \sim N(\beta, (X'X)^{-1}\sigma^2) \tag{2.40}$$

On en déduit aisément :

$$\boxed{\frac{\widehat{\beta}}{\sqrt{(X'X)^{-1}}\sigma} \sim T_{n-p-1}} \tag{2.41}$$

Pour $n-p-1$ grand (supérieur à 30), on a $T_{n-p-1} \to N(0,1)$. Là encore, lorsque la valeur de la statistique de test est supérieure à la valeur critique pour un niveau de risque défini, on rejette l'hypothèse nulle de nullité du coefficient du paramètre. Ainsi, si le quantile de la loi de Student à 95% et $n-p-1$ degrés de liberté est plus petit que la valeur de la statistique calculée ($\frac{\widehat{\beta}}{\sqrt{(X'X)^{-1}}\sigma}$), on est conduit à rejeter l'hypothèse de nullité du paramètre. Le paramètre est alors significativement différent de 0. En pratique, on compare la valeur de cette statistique à 2 : il s'agit de la valeur la plus courante du quantile d'une loi de Student. La règle est donc : si la valeur de la statistique calculée est inférieure à 2, alors $\alpha_i = 0$; dans le cas contraire, $\alpha_i = \widehat{\alpha_i}$. L'intuition est donc : on conduit un test de Student pour être bien sûr que la valeur estimée par MCO soit bien différente de 0. Il s'agit de vérifier la qualité de l'estimation.

Nota Bene : 1. Intuitivement, plus la variance des résidus σ est importante (autrement dit, moins le modèle semble être explicatif du comportement de Y) et plus l'erreur possible lors de l'estimation des paramètres est potentiellement importante.

2.1.4.3 Test de Durbin et Watson

Une autre problème peut affecter les résidus : la présence d'autocorrélation entre les erreurs. Le résidu spécifié dans le modèle est un bruit blanc : il s'agit d'une innovation pure. Une hypothèse du modèle qui n'apparait pas en première lecture est la suivante : $\mathbb{E}[\epsilon_i \epsilon_{i-1}] = 0$. Dans le cas contraire, la loi des erreurs n'est pas celle spécifiée et les estimations simples par MCO ne sont pas bonnes. Durbin et Watson ont proposé un test astucieux, bâti sur une mesure de distance entre les erreurs en i et en $i-1$:

$$\boxed{d = \frac{\sum_{i=1}^{n} (\widehat{\epsilon_i} - \widehat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n} \widehat{\epsilon_i}^2}} \tag{2.42}$$

Cette statistique peut s'exprimer approximativement en fonction du coefficient d'autocorrélation des résidus ρ :

$$d \sim 2(1 - \rho) \tag{2.43}$$

Cette intuition simplifie grandement la lecture du test :

– Si ρ est nul (pas d'autocorrélation), alors d se situe au voisinage de 2.
– Si ρ est égal à 1 (autocorrélation positive), alors d se situe au voisinage de 0.
– Si ρ est égal à -1 (autocorrélation négative), alors d se situe au voisinage de 4.

Dans les deux derniers cas, l'estimation par les moindres carrés ordinaires n'est pas satisfaisante. Il est alors nécessaire de développer des méthodes plus avancées permettant d'intégrer l'existence de cette autocorrélation.

2.1.4.4 Les tests d'adéquation des résidus

L'hypothèse de normalité des résidus est à confirmer à l'aide de différents tests. On présente ici deux d'entre eux : le test de Jarque et Berra ainsi que les qqplots.

Le test de Jarque et Berra

Le test utilise les estimateurs empiriques de la kurtosis et de la skewness, ainsi que leur distribution afin de juger de la normalité d'une variable aléatoire. L'estimateur empirique du moment centré d'ordre k est le suivant :

$$\mu_k = \frac{1}{T}\sum_{i=1}^{T}(X_i - \bar{X})^k \tag{2.44}$$

La skweness s'estime donc par μ_3/μ_2^3 et la kurtosis par μ_4/μ_2^2. La statistique de Jarque et Berra vaut donc :

$$s = \frac{T}{6}S_k^2 + \frac{T}{24}(K_u - 3)^2 \rightarrow \chi^2(2) \tag{2.45}$$

Le qqplot

Le qqplot compare quantiles empiriques et quantiles théoriques d'une loi donnée. Il permet de se faire une idée de l'adéquation éventuelle de nos données à une loi paramétrique particulière.

Tests d'adéquation paramétrique

Il est également possible de procéder à un test d'adéquation des résidus à une loi paramétrique quelconque. On se reportera au chapitre 1 pour la méthodologie.

2.2 Retour sur le maximum de vraisemblance

Il est possible de retrouver l'ensemble des résultats obtenus jusqu'à maintenant sur la base d'une approche utilisant le maximum de vraisemblance. On rappelle ici les bases de la méthode ainsi que son application au modèle linéaire multivarié gaussien.
La fonction de densité de probabilité d'une variable aléatoire y conditionnellement à un ensemble de paramètres θ est noté $f(y, \theta)$. Cette fonction identifie le processus générant les données qui sous-tend l'échantillon de données, et, en même temps, fournit une description mathématique des données que le processus génère. La densité jointe de n observations indépendantes et distribuées de façon identique (i.i.d.) de ce processus est le produit des densités individuelles :

$$f(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta, \mathrm{y}) \tag{2.46}$$

Cette densité jointe est la fonction de vraisemblance, définie comme une fonction du vecteur de paramètres inconnus (θ), où y indique les données observées (qui ne sont donc pas une inconnue). On remarque que l'on note la densité jointe comme une fonction des données conditionnellement aux paramètres alors que, lorsque l'on forme la fonction de vraisemblance, on note cette fonction en sens inverse, comme une fonction de paramètres conditionnellement aux données observées. Dans ce qui suit, on suppose que les paramètres sont constants et inconnus : l'enjeux de la méthode est d'exploiter l'information disponible dans l'échantillon afin d'en inférer une valeur probable des paramètres.
Il est généralement plus simple de travailler avec le logarithme de la fonction de vraisemblance :

$$lnL(\theta|\mathrm{y}) = \sum_{i=1}^{n} lnf(y_i|\theta) \tag{2.47}$$

On parle dans ce cas de *log-vraisemblance*. Ajoutons qu'il est courant de travailler sur la densité d'un processus *conditionnellement à un autre processus*. C'est du moins ce qui se passe dans le modèle linéraire : les erreurs sont bien i.i.d., ce qui fait que $y|x$ est aussi un processus iid. Soit le modèle linéaire gaussien suivant :

$$y = X\beta' + \epsilon \tag{2.48}$$

$$\Leftrightarrow y_i = \beta_1 + \beta_2 x_{1,i} + ... + \beta_p x_{p-1,i} + \epsilon_i \tag{2.49}$$

On suppose que les perturbations sont gaussiennes : conditionnellement à $x_{:,i}$, y_i est distribué normalement, moyenne $\mu_i = x_{:,i}\beta'$ et de variance σ^2. Cela signifie que les variables aléatoires observées ne sont pas i.i.d. : elles sont de moyenne différentes. Toutefois, les observations sont conditionnelement indépendantes, permettant de travailler sur la vraisemblance conditionnelle. Dans notre cas, elle a la forme suivante :

$$lnL(\theta|\text{y,X}) = \sum_{i=1}^{n} lnf(y_i|X_i,\theta) \tag{2.50}$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\left[ln\sigma^2 + ln(2\pi) + \frac{(y_i - x_{:,i}\beta')^2}{\sigma^2}\right] \tag{2.51}$$

$$= -\frac{1}{2}nln\sigma^2 - nln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - x_{:,i}\beta')^2}{\sigma^2}\right] \tag{2.52}$$

La méthode du maximum de vraisemblance propose de déterminer $\widehat{\theta}$ de façon à ce que la log-vraisemblance soit maximale. Cependant, avant d'exposer la méthode, il est nécessaire de vérifier que cette estimation est réalisable/possible : il s'agit d'étudier ce que l'on appelle *les conditions d'identification*.

Définition 2.2.1 (Identification)**.** *Le vecteur de paramètres θ est identifié (i.e. susceptible d'être estimé) si, pour n'importe quel autre vecteur de paramètre θ^* tel que $\theta \neq \theta^*$, on a pour les données y : $L(\theta^*|y) \neq L(\theta|y)$.*

Il est parfois impossible d'obtenir une valeur unique pour le paramètre θ, rendant toute estimation par maximum de vraisemblance impossible.

2.2.1 Le principe du maximum de vraisemblance

Le principe du maximum de vraisemblance fournit un moyen de choisir un estimateur asymptotiquement efficient (cf. chapitre 1) pour un paramètre ou un ensemble de paramètres. Il est aisé d'illustrer la logique de cette technique dans le cas d'une distribution discrète.
On considère un échantillon aléatoire de 10 observations tirées d'une distribution de Poisson : 5,0,1,1,0,3,2,3,4,1. La densité de chaque obersvation est alors :

$$f(y_i|\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!} \tag{2.53}$$

Puisque les observations sont i.i.d., leur densité jointe, qui est la vraisemblance de cet échantillon, est :

$$f(y_1,...,y_10|\theta) = \prod_{i=1}^{1} 0f(y_i|\theta) = \frac{e^{-10\theta}\theta^{\sum_{i=1}^{1} 0y_i}}{\prod_{i=1}^{1} 0y_i!} = \frac{e^{-10\theta}\theta^{20}}{207,360} \tag{2.54}$$

Ce dernier résultat donne la probabilité d'observer cet échantillon particulier, en supposant qu'une distribution de Poisson de paramètre encore inconnu θ, a généré les données. Quelle est alors la valeur de θ qui rendrait cet échantillon plus probable ? La réponse est fournie par la méthode du maximum de vraisemblance : il s'agit de la valeur qui rend la vraisemblance maximum, i.e. la probabilité jointe la plus importante possible. C'est ce qu'on représent en figure (2.2.1).
Sur la figure, on remarque que la vraisemblance a un mode unique pour $\theta = 2$, qui est l'estimation du maximum de vraisemblance ou EMV de θ.

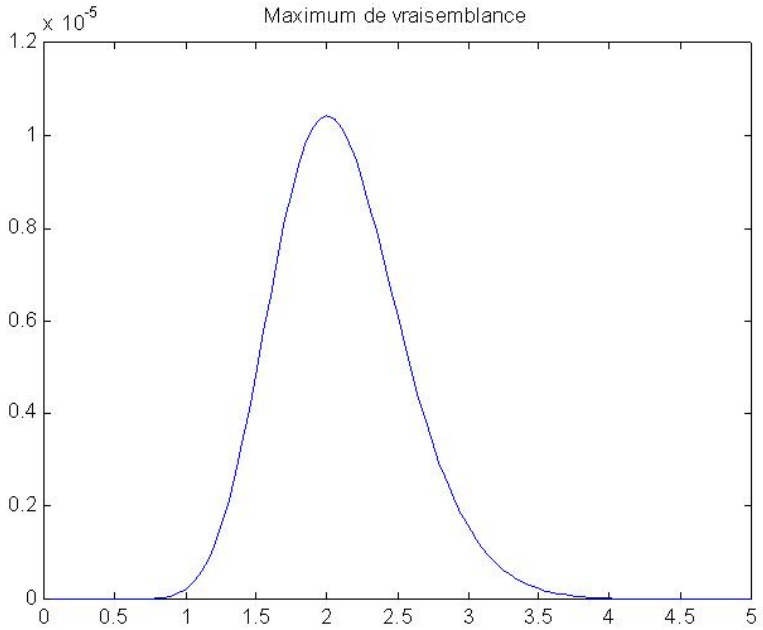

FIG. 2.1 – Représentation graphique du maximum de vraisemblance

On considère la maximisation de $L(\theta|\mathrm{y})$ par rapport à θ. Puisque la fonction logarithme croît de manière monotone et qu'elle est plus simple à utiliser, on maximise généralement $lnL(\theta|\mathrm{y})$ à la place. Dans notre exemple :

$$lnL(\theta|\mathrm{y}) = -n\theta + ln\theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} ln(y_i!) \tag{2.55}$$

$$\frac{\partial lnL(\theta|\mathrm{y})}{\partial \theta} = -n + \frac{1}{\theta} \sum_{i=1}^{n} y_i = 0 \Leftrightarrow \hat{\theta}_{EMV} = \bar{y}_n \tag{2.56}$$

Ainsi, afin de déterminer la maximum de vraisemblance, il "suffit" de dériver la log-vraisemblance par rapport à θ et de l'annuler (comme on le fait très classiquement pour un fonction à une variable). Dans le cas d'une loi de Poisson, on trouve un EMV pour θ égal à la moyenne empirique. Ceci n'est pas vraiment surprenant, dans la mesure où si $X \sim P(\lambda)$, alors $\mathbb{E}[X] = \lambda$.
Annuler la dérivé première ne suffit cependant pas à s'assurer qu'il s'agit d'un maximum : encore faut il prouver qu'en ce point (le prétendant au titre de maximum), la dérivée seconde est négative (fonction concave). En général, la vraisemblance est naturellement strictement concave, ce qui fait que la solution de la dérivée première est toujours un maximum.
La référence à la probabilité d'observer un échantillon donné n'est pas exacte dans une distribution continue, puisqu'un échantillon particulier à un probabilité d'être observé nulle. La principe reste néanmoins le même. Les valeurs des paramètres qui maximisent $L(\theta|\mathrm{y})$ ou son logarithme sont les estimations du maximum de vraisemblance, notées $\hat{\theta}$. Puisque le logarithme est une fonction monotone, les valeurs qui maximisent $L(\theta|\mathrm{y})$ sont

les mêmes que celles qui maximisent $lnL(\theta|\mathrm{y})$. La condition nécessaire pour maximiser $lnL(\theta|\mathrm{y})$ est :

$$\frac{\partial lnL(\theta|\mathrm{y})}{\partial \theta} = 0 \tag{2.57}$$

Il s'agit de l'équation de vraisemblance. Le résultat général est que l'EMV est une racine de l'équation de vraisemblance. L'application aux paramètres du processus générant les données d'une variable aléatoire discrète suggèrent que le maximum de vraisemblance est une bonne utilisation des données. Cette intuition reste à généraliser dans ce qui suit.

2.2.2 Propriétés du maximum de vraisemblance

On introduit dans ce qui suit quelques défintions et propriétés qui ne sont pas démontrées. On lira les preuves avec profit dans Greene (2002).

Définition 2.2.2 (Efficience asymptotique)**.** *Un estimateur est asymptotiquement efficient s'il est convergent, distribué normalement asymptotiquement et s'il a une matrice de covariances asymptotiques qui n'est pas plus grande que celle de n'importe quel autre estimateur convergent distribué normalement asymptotiquement.*

Proposition 2.2.1 (Propriétés d'un EMV)**.**

1. *Convergence :* $\theta_{EMV} \rightarrow \theta_0$ *(convergence en probabilité), où* θ_0 *est la vraie valeur du paramètre.*
2. *Normalité asymptotique :* $\theta_{EMV} \sim N\left[\theta_0, \{I(\theta_0)\}^{-1}\right]$*, où* $I(\theta_0) = -\mathbb{E}[\frac{\partial^2 lnL}{\partial \theta_0 \partial \theta_0'}]$.
3. *Efficience asymptotique :* θ_{EMV} *est asymptotiquement efficient et atteint la borne inférieure de Frechet-Darmois-Cramer-Rao des estimateurs convergents.*
4. *Invariance : l'estimateur du maximum de vraisemblance de* $\gamma_0 = c(\theta_0)$ *est* $c(\theta_{EMV})$ *si* $c(\theta_0)$ *est une fonction continue et continuellement différenciable.*

Ces propriétés reposent principalement sur les *conditions de régularités*, principalement au nombre de trois, que le lecteur trouvera (avec la démonstration des propriétés du maximum de vraisemblance) dans Greene (2002) (page 457 et suivantes). Il s'agit principalement d'être en présence d'une vraisemblance triplement continument dérivables, avec les dérivées d'ordre un et deux appartenant à L^1 et celle d'ordre majorable par une fonction appartenant à L^1. Il s'agit ici de simples rappels de résultats statistiques de base, nous laisserons le soin au lecteur d'aller faire les lectures adéquates pour remédier à la maigre présentation qui en est faite.

2.2.3 EMV du modèle gaussien standard

On rappelle que le modèle de régression linéaire standard est :

$$y_i = x_i \beta' + \epsilon_i \tag{2.58}$$

La vraisemblance d'un processus gaussien x de n observations s'écrit naturellement :

$$lnL(\theta|x) = -nln(\sigma) - \frac{n}{2}ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 \tag{2.59}$$

Dans le cas de ϵ, on a :

$$lnL(\theta|x) = -nln(\sigma) - \frac{n}{2}ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\epsilon_i}{\sigma}\right)^2 \tag{2.60}$$

Il est alors nécessaire de savoir passer de la loi de ϵ à celle de y, lorsque l'on a spécifié le modèle. Pour cela, on utilise un changement de variable qui est rappelé ici :

Proposition 2.2.2 (Changement de variable)**.** *Si x est une variable aléatoire de fonction de répartition $f_x(.)$ et si $y = g(x)$, alors la densité de y s'exprime comme suit :*

$$f_y(y) = f_x(g^{-1}(y))|g^{-1\prime}(y)| \tag{2.61}$$

Cette proposition sera particulièrement importante pour l'analyse des séries temporelles (troisième partie de cet opus). Ici, la transformation de ϵ à y est $\epsilon = y - X\beta'$, donc la jacobienne (matrice des dérivées premières) est égale à l'unité ($\frac{\partial \epsilon}{\partial y} = 1$). On en déduit naturellement la vraisemblance associée à y :

$$lnL(\theta|x) = -nln(\sigma) - \frac{n}{2}ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - x_i\beta'}{\sigma}\right)^2 \tag{2.62}$$

Les conditions nécessaires (équations normales ou de la vraisemblance) sont alors :

$$\begin{bmatrix} \frac{\partial lnL}{\partial \beta} \\ \frac{\partial lnL}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(y - x_i\beta'\right) \\ -\frac{n}{2\sigma^2} + \frac{1}{2}\frac{\sum_{i=1}^{n}(y_i - x_i\beta')^2}{\sigma^4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.63}$$

On en déduit aisément, en passant en forme matricielle :

$$\beta_{EMV} = (X'X)^{-1}X'y \tag{2.64}$$

$$\sigma^2 = \frac{\epsilon'\epsilon}{n} \tag{2.65}$$

On retrouve l'estimateur MCO du modèle de régression linéaire. Dernier calcul, la borne de Cramer-Rao : on calcule tout d'abord pour cela la matrice des dérivées secondes du maximum de vraisemblance, puis en on prend l'espérance. Les calculs sont les suivants :

$$\begin{bmatrix} \frac{\partial^2 lnL}{\partial\beta\partial\beta'} & \frac{\partial^2 lnL}{\partial\beta\partial\sigma} \\ \frac{\partial^2 lnL}{\partial\sigma\partial\beta'} & \frac{\partial^2 lnL}{\partial\sigma\partial\sigma} \end{bmatrix} = \begin{bmatrix} -\frac{\sum_{i=1}^{n} x_i^2}{\sigma^2} & -\frac{\sum_{i=1}^{n} x_i\epsilon_i}{\sigma^4} \\ -\frac{\sum_{i=1}^{n} x_i\epsilon_i}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{\sum_{i=1}^{n}\epsilon_i^2}{\sigma^6} \end{bmatrix} \tag{2.66}$$

L'espérance de la précédente matrice fournit alors la matrice d'information de Fisher :

$$I(\theta) = \begin{bmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \tag{2.67}$$

D'où la matrice de variance/covariance des estimateurs :

$$\mathbb{V}[\theta] = I(\theta)^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix} \tag{2.68}$$

Estimateurs du maximum de vraisemblance et des moindres carrés coincident en tout point[1]. Par conséquent, l'estimateur MCO hérite de toutes les bonnes propriétés asymptotiques souhaitables des estimateurs du maximum de vraisemblance.

2.2.4 Les tests liés à la vraisemblance

L'un des avantages de procéder à des estimations par maximum de vraisemblance est que l'on peut par la suite en procéder à un certain nombre de tests simples. Trois de ces tests sont généralement développés : *le test de Wald*, *le test du ratiode vraisemblance* et enfin *le test du multiplicateur de Lagrange*. On ne présentera ici que celui dont l'implémentation est immédiate une fois les estimations EMV produites : le test du ratio de vraisemblance. Pour les autres, voir Greene (2002), pages 479 et suivantes.
Lors de l'estimation d'un modèle donné, il est possible de devoir opérer un choix sur le nombre de variables à utiliser pour un modèle donné. Soit θ les paramètres du modèle incluant le maximum de paramètres (dit modèle non contraint) et θ^- les paramètres d'un modèle dit contraint, i.e. excluant volontairement certaines variables. Le test du ratio de vraisemblance propose une statistique simple afin de discriminer entre modèle contraint et non contraint. Soit $lnL(\theta|y)$ la log-vraisemblance associée à θ et $lnL(\theta^-|y)$ celle associée θ^-. La statistique est alors égale à :

$$LR = -2[lnL(\theta^-) - lnL(\theta)] \tag{2.69}$$

On parle de rapport de vraisemblance car il s'agit en fait de calculer le rapport de vraisemblance, dont on prend ensuite le logarithme : ceci explique que l'on ait une différence. Cette statistique a sous H_0 (θ^- est une paramétrisation plus raisonnable que θ)la distribution asymptotique suivante :

$$LR \to \chi^2(J) \tag{2.70}$$

Où J est le nombre de contraite pesant sur les paramètres θ.

[1] A peu de choses pret en fait, dans la mesure où l'estimateur des moindres carrés de la variance des erreurs est corrigés dans sa version MCO des degrés de libertés. Un estimateur du maximum de vraisemblance n'est jamais corrigé. Là encore, voir Greene (2002), chapitre 17.

2.3 Prévision à partir du modèle linéraire multiple

Une fois un modèle robuste dégagé des précédentes estimations, il est possible de passer à la prédiction des valeurs de y à partir de celle de x. On suppose que l'on souhaite prédire la valeur de y^0 associée au regresseur x^0. Cette valeur est :

$$y^0 = x^0\beta' + \epsilon^0 \tag{2.71}$$

Par le théorème de Gauss Markov, on a :

$$\mathbb{E}[y^0|x^0] = \hat{y}^0 = x^0\beta' \tag{2.72}$$

est l'estimateur sans biais de variance minimale de $\mathbb{E}[y^0|x^0]$. L'erreur de prévision est alors :

$$\begin{aligned} e^0 &= y^0 - \hat{y}^0 && (2.73)\\ &= x^0(\beta' - \hat{\beta}') + \epsilon^0 && (2.74) \end{aligned}$$

Où β est la vraie valeur du paramètre et $\hat{\beta}$ son estimateur. On en déduit alors aisément la variance de la prédiction :

$$\begin{aligned} \mathbb{V}[e^0|X, x^0] &= \mathbb{V}[x^0(\beta' - \hat{\beta}')|X, x^0] + \sigma^2 && (2.75)\\ &= x^0\mathbb{V}[\hat{\beta}'|X, x^0]x^{0\prime} + \sigma^2 && (2.76)\\ &= x^0\sigma^2(X'X)^{-1}x^{0\prime} + \sigma^2 && (2.77)\\ &= \sigma^2\left[1 + \frac{1}{n} + \tilde{x}^0(\tilde{X}'\tilde{X})^{-1}\tilde{x}^{0\prime}\right] && (2.78) \end{aligned}$$

La dernière équation est obtenue dans le cas où le modèle contient un terme constante. On note alors à l'aide d'un tilde la matrice des données dont on a supprimé la première colonne contenant le terme constant. Ce résultat montre que l'intervalle dépend du rapport entre une version approchée de la variance de x^0 et X. Plus x^0 a une variance approchée importante, et plus la prévision sera incertaine.
Il est alors possible d'inférer un intervalle de confiance pour la prédiction (voir le chapitre 1 sur la construction d'un intervalle de confiance pour la moyenne) :

$$IC_\alpha = \left[\hat{y}^0 \pm t_{\alpha/2}\sqrt{\widehat{\mathbb{V}}[e^0|X, x^0]}\right] \tag{2.79}$$

Afin de déterminer la qualité d'une prévision, on utilise en général les deux statistiques suivantes :

$$RMSE = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2} \tag{2.80}$$

$$MAE = \frac{1}{n^0}\sum_i |y_i - \hat{y}_i| \tag{2.81}$$

On parle de *Root Mean Square Error* et de *Mean Absolute Error*. n^0 désigne le nombre de périodes de prévisions.

2.4 Une calibration simple du CAPM

Suite à l'ensemble de ces éléments théoriques, on propose dans ce qui suit un exemple simple de calibration du CAPM. Cette méthode repose, comme nous le verrons, sur l'hypothèse principale selon laquelle les prix sont des martingales. On montre que le beta coincide exactement avec l'estimateur des moindres carrés ordinaires. On procède de même pour estimer le alpha, après avoir estimer la Security Market Line (SML). Enfin, on montre qu'il est aisé d'obtenir le R^2 du modèle, sur la base du calcul du beta.

2.4.1 L'estimation de la relation du MEDAF par MCO

On cherche à calibrer un modèle de la forme :

$$r_i = r_f + \beta_i(r_m - r_f) \tag{2.82}$$

Une approche classique en économétrie pour résoudre ce type de problème revient à se donner ϵ, une erreur d'estimation, puis à chercher à minimiser cette erreur. On fait apparaitre ϵ comme suit dans la précédente relation :

$$r_i = r_f + \beta_i(r_m - r_f) + \epsilon \tag{2.83}$$

Il s'agit bien d'une erreur : si $r_f + \beta_i(r_m - r_f)$ correspond à l'approximation (ou *estimation*) de la rentabilité du titre i, celle ci n'est pas parfaite. ϵ est précisement la différence entre le vrai r_i et son estimation, que l'on peut noter $\hat{r}_i$. On a donc :

$$\epsilon = r_i - \hat{r}_i \tag{2.84}$$

$$= r_i - r_f + \beta_i(r_m - r_f) \tag{2.85}$$

Ceci est vrai d'après la relation précédente (équation 2.83). La méthode des moindres carrés ordinaires vise à déterminer une valeur pour β (que l'on ne connait pour l'instant pas), qui rende la somme des erreurs au carrés la plus petite possible. L'idée est donc de trouver un beta qui permette de rendre l'erreur la plus petite possible, autrement dit, qui rende le modèle le meilleur possible.

On cherche donc à minimiser la somme des erreurs commises pour chacune des observations de la rentabilité du titre i, élevées au carré. On cherche donc :

$$Min \sum \epsilon_i^2 \tag{2.86}$$

$$\Leftrightarrow \sum (r_i - r_f - \beta(r_m - r_f))^2 \tag{2.87}$$

Pour trouver un minimum sur une fonction convexe, il suffit d'égaler la dérivée à 0. Ici, on :

$$\frac{\partial SSR(\beta)}{\partial \beta} = \sum (r_m - r_f)(r_i - r_f - \beta_i(r_m - r_f)) \tag{2.88}$$

$$= \sum \beta(r_m - r_f)^2 - (r_i - r_f)(r_m - r_f) \tag{2.89}$$

$$= \sum \beta(r_m - r_f)^2 - \sum (r_i - r_f)(r_m - r_f) \tag{2.90}$$

En égalant la dérivée à 0, il vient :

$$\sum \beta(r_m - r_f)^2 - \sum (r_i - r_f)(r_m - r_f) = 0 \tag{2.91}$$

$$\Leftrightarrow \sum \beta(r_m - r_f)^2 = \sum (r_i - r_f)(r_m - r_f) \tag{2.92}$$

$$\Leftrightarrow \beta = \frac{\sum (r_i - r_f)(r_m - r_f)}{\sum (r_m - r_f)^2} \tag{2.93}$$

En notant $\tilde{r}_i$ et $\tilde{r}_m$ les rentabilités du titre i ainsi que du marché dont on a retrancher le taux sans risque, il vient une écriture simple de l'estimateur des MCO, qui peut se réécrire de façon matricielle aisément :

$$\beta = \frac{\sum \tilde{r}_i \tilde{r}_m}{\sum (\tilde{r}_m)^2} \tag{2.94}$$

En notant R_i la matrice $n \times 1$ contenant les n observations de rentabilités du titre (dont on a retranché le taux sans risque) i et R_m de même taille, celle contenant celles du marché, il est alors possible de réécrire ces sommes de façon matricielle (faire les calculs pour s'en convaincre) :

$$\beta = (R_m^T R_m)^{-1} R m^T R_i \tag{2.95}$$

Cet estimateur est l'estimateur MCO de β. On admettra qu'il possède la distribution suivante :

$$\widehat{\beta} \sim N(\beta, (R_m^T R_m)^{-1} \sigma_\epsilon^2) \tag{2.96}$$

σ_ϵ^2 est la variance du terme d'erreur et β la vraie valeur du paramètre à estimer. On est ainsi en mesure de construire un test de Student permettant de tester sur l'estimation de β est différente de 0 avec une probabilité importante. On se borne ici à fournir la méthodologie générale permettant de construire ce test : il s'agit pour nous d'une recette de cuisine financière dont les fondements ne sont pas démontrés.

Pour conduire ce test, il suffit de comparer $\frac{\hat{\beta}}{\mathbb{V}[\hat{\beta}]}$ à 1,96 (le quantile à 95% d'une loi normale). Si la valeur est supérieure, alors $\hat{\beta} \neq 0$.

2.4.2 Lien de l'estimateur MCO avec le beta financier

On a vu que le beta financier pouvait être estimer par :

$$\beta = \frac{cov(r_i, r_m)}{\sigma_m^2} \tag{2.97}$$

au terme de la démonstration de Sharpe. Il est facile de démontrer que l'estimateur MCO et cette expression du beta coincident exactement, à la condition que les prix soient martingale. Quelle est la contrepartie empirique de la covariance et de la variance ? En remplaçant dans l'expression précedentes les moments par leur estimation, il vient :

$$\beta = \frac{\frac{1}{n}\sum (r_i - \mathbb{E}[r_i])(r_m - \mathbb{E}[r_m])}{\frac{1}{n}\sum (r_m - \mathbb{E}[r_m])^2} \tag{2.98}$$

$$= \frac{\sum (r_i - \mathbb{E}[r_i])(r_m - \mathbb{E}[r_m])}{\sum (r_m - \mathbb{E}[r_m])^2} \tag{2.99}$$

Dans le cas où les prix sont martingale, on sait que :

$$\mathbb{E}[P_t] = P_{t-1} \tag{2.100}$$
$$\Leftrightarrow \mathbb{E}[P_t - P_{t-1}] = 0 \tag{2.101}$$
$$\Leftrightarrow \mathbb{E}\left[\frac{P_t - P_{t-1}}{P_{t-1}}\right] = 0 \tag{2.102}$$
$$\Leftrightarrow \mathbb{E}[r_t] = 0 \tag{2.103}$$

Ainsi, l'ensemble des actifs doivent avoir une espérance de rentabilité nulle si le marché est martingale. L'introduction de la nullité des espérances de rendement permet de retrouver (à r_f pret), la formule des MCO :

$$\beta = \frac{\sum r_i r_m}{\sum r_m^2} \tag{2.104}$$

2.4.3 Estimation de la SML

Une fois l'ensemble des β des titres estimés, il est possible d'estimer la SML. Là encore le recours aux MCO permet d'obtenir des résultats simples. On cherche à estimer une relation du type :

$$\mathbb{E}[r_i] = \alpha\beta_i \tag{2.105}$$

L'estimateur MCO est alors le suivant :

$$\hat{\alpha} = \frac{\sum \beta_i \mu_i}{\sum \beta_i^2} \tag{2.106}$$

On propose à la fin de ce chapitre une fonction `R` permettant de réaliser ces estimations. La figure 2.2 est le résultat de cette fonction.

2.4.4 Calcul des alpha

Il est possible d'estimer les surrentabilités eventuelles dégagées par le marché : les alpha. Il s'agit simplement d'introduire une constante dans le modèle du CAPM et de l'estimer par MCO. On note R_i la matrice consituée de deux colonnes : l'une comportant exclusivement des 1 et l'autre l'ensemble des observations des r_i. On fait de même pour R_m. On est alors en mesure d'utiliser là encore la formule des MCO pour estimer le beta et le alpha en un seul coup.

2.4.5 Le R^2

Le R^2 ou coefficient de détermination est une mesure de la qualité globale du modèle proposé. Il s'agit du rapport entre la variance de l'estimation de l'on donne du modèle et la variance de la variable expliquée. Il s'agit donc du rapport :

$$R^2 = \frac{\mathbb{V}[\beta_i r_m]}{\mathbb{V}[r_i]} \tag{2.107}$$

Dans le cas du MEDAF, il est aisé de calculer ce R^2 à partir du β :

$$R^2 = \frac{\beta_i^2 \sigma_m^2}{\sigma_i^2} = \rho_{i,m} \tag{2.108}$$

On aboutit donc à une expression simple de ce R^2.

2.4.6 Code pour le CAPM

```
##############################################################################
#                          Estimation et test du CAPM                        #
##############################################################################

# Dans ce programme, on estime les beta du CAPM par MCO. On teste l'existence de
# alpha. Enfin, on estime la SML par MCO.

capm<-function(x,Rf){
# Formattage de la base de données
x=as.matrix(x)
x.beta=x
x=x-Rf
SP=cbind(matrix(1,nrow(x),1),x[,1])
titres=x[,2:ncol(x)]
xx=solve(t(SP)%*%SP)
theta=xx%*%(t(SP)%*%titres)
res=titres-SP%*%theta
var.res=apply(res,2,var)
test=cbind(diag(xx)[1]*as.matrix(var.res),diag(xx)[2]*as.matrix(var.res))
test=sqrt(test)
test=t(theta)/test
ptest=pnorm(test)

#Calcul de la SML

beta.capm=t(theta)[,2]
renta.moy=as.matrix(apply(x.beta[,2:ncol(x.beta)],2,mean))
pente=sum(renta.moy*beta.capm)/sum(beta.capm^2)
beta.sort=(beta.capm)
renta.est=beta.sort*pente

par(bg="lightyellow")
plot(beta.capm,renta.est,type="l",col="red",ylab="Rentabilité",xlab="Beta",
main="Estimation de la SML")
lines(beta.capm,renta.moy,type="p",col="blue")

# Calcul des R2

R2=as.matrix(beta.capm^2)*var(SP[,2])/apply(titres,2,var)
return(list(theta=theta, R2=R2, test=test, ptest=ptest))
}
```

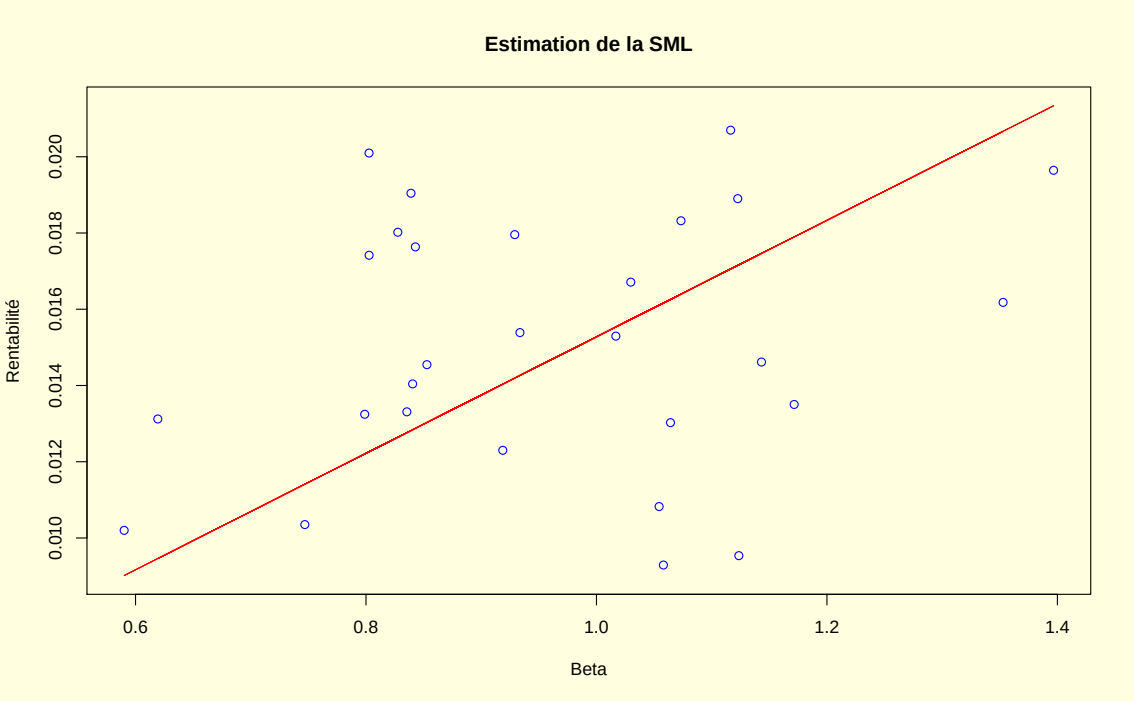

FIG. 2.2 – Security Market Line

Chapitre 3

Extensions du modèle de base

On propose dans ce qui suit quelques extensions du modèle linéraire standard : les modèles non linéaires et les modèles à équations multiples. L'exposé s'appuie à la fois sur Greene (2002) et Davidson and MacKinnon (1993). Il est à noter qu'il existe une version française de ce dernier ouvrage librement accessible sur le site de Russel Davidson (voir l'url fournie en bibliographie). Il s'agit d'une introduction succinte : tout lecteur soucieux de dépasser ce stade se reportera avec profit aux deux références qui sont fournies plus haut.

3.1 Modèle de régression non linéaire

On se contente ici de présenter le modèle de régression non linéaire dans le cas univarié : il est aisé de généraliser les résultats présentés ici au cas multivarié. Classiquement, un modèle non linéaire s'exprime sous forme fonctionnelle comme suit :

$$y = x(\beta) + \epsilon \tag{3.1}$$

avec $x(\beta)$ une forme fonctionnelle liant les variables exogènes x et le vecteur des paramètres β. ϵ est une perturbation i.i.d. de moyenne nulle et de variance égale à σ^2. La fonction scalaire $x(\beta)$ est une fonction de regression (non linéaire) qui détermine l'espérance de y conditionnellement à β et à x. Ici encore, un modèle de régression non linéaire doit être identifié si l'on désire obtenir des estimations uniques des paramètres. Classiquement, l'estimation peut se faire par moindres carrés. La fonction objectif est alors :

$$Min \sum_{i=1}^{n} (y_i - x_i(\beta))^2 \tag{3.2}$$

Dans le cas multivarié, ceci s'écrit matriciellement comme suit :

$$Min \ (y - x(\beta))'(y - x(\beta)) \tag{3.3}$$

où y désigne une matrice $\mathcal{M}(n \times 1)$ composé des différentes valeurs de y_i et $x(\beta)$ une matrice $\mathcal{M}(n \times p)$ composé des n fonctions de regression $x_i(\beta), i = 1, ...p$. Cette somme des carrés peut être explicité comme suit :

$$Min\ y'y - 2y'x(\beta) + x(\beta)'x(\beta) \tag{3.4}$$

En dérivant cette expression par rapport à β et en l'annulant, il vient :

$$-2y\frac{\partial x(\beta)}{\partial \beta}' + 2\frac{\partial x(\beta)}{\partial \beta}' x(\beta) = 0 \tag{3.5}$$

Ceci est équivalent à :

$$\frac{\partial x(\beta)}{\partial \beta}' (y - x(\beta)) = 0 \tag{3.6}$$

On retrouve ici la condition d'orthogonalité entre les résidus et la dérivée des régresseurs, une version modifiée de la condition d'orthogonalité entre regresseurs et résidus dans le cas MCO. Le problème est qu'il n'existe pas formule analytique permettant d'obtenir l'expression des estimateurs comme dans le cas MCO : il est alors nécessaire d'utiliser des algorithmes d'optimisation permettant de trouver le minimum du programme mentionné plus haut. Nous reviendrons sur ces algorithme plus loin : il seront également utiles pour l'estimation des modèles de séries temporelles.
Un dernier point reste à noter dans ce bref exposé des moindres carrés non linéaires : les conditions de premier ordre sont nécessaire, mais pas suffisante pour garantir le fait que $\hat{\beta}$ soit un minimum. Il peut exister plusieurs valeurs de β qui vérifient les conditions de premier ordre, mais qui soient des minima lcaux, des points stationnaires ou même des maxima locaux. En définitive, rien ne garantit que la fonction à minimiser soit globalement convexe, i.e. :

$$H_{ij}(\beta) = \frac{\partial^2 SSR(\beta)}{\partial \beta_i \beta_j} \tag{3.7}$$

soit définie positive en β.

Rappel 2. Une matrice M $\mathcal{M}(n \times n)$ est dite définie positive si $\forall x$, une matrice colonne composée de n éléments réels on a :

$$x'Mx > 0 \tag{3.8}$$

Une autre façon de prouver le caractère défini positif d'une matrice est de montrer que l'ensemble de ses valeurs propres est strictement positif. Enfin, si l'ensemble des sous matrices d'une matrice carré admet un déterminant positif, alors cette matrice est définie positive. Il s'agit de la généralisation matricielle de la postivité d'un scalaire. Une matrice définie négative est une matrice dont l'opposé est définie positive.

Cette condition n'est assurée que pour quelques cas particuliers, dont les MCO :

$$H_{ij}(\beta) = \frac{\partial^2 SSR(\beta)}{\partial \beta_i \beta_j} = -x'x \tag{3.9}$$

Il est aisé de prouver d'après ce qui vient d'être rappelé que cette matrice est bien définie positive.
Terminon enfin par une courte discussion des conditions d'identification de ce type de modèle. On distingue deux srotes d'identification, l'identification locale et l'identification globale. Les estimation des moindres carrés non linéaire ne seront identifiés localement qu'à la condition que pour tout modification infinitésimale de $\hat{\beta}$, la valeur de la fonction objectif s'élève. Ceci revient à supposer que la matrice hessienne est strictement convexe en β, de sorte que :

$$SSR(\hat{\beta}) < SSR(\hat{\beta} + \delta) \tag{3.10}$$

Pour une petite variation δ. Cette condition est analogue au caractère défini positif de la matrice hessienne. La stricte convexité implique que $SSR(\beta)$ soit incurvée dans toutes les directions ; aucun plat n'est autorisé quelle que soit la direction. Si $SSR(\beta)$ était plate dans une direction donnée au voisinage de $\hat{\beta}$, il serait possible de s'éloigner de $\hat{\beta}$ dans cette direction, sans jamais modifier la valeur de la somme des résidus au carré (du fait des conditions du premier ordre). Par conséquent, $\hat{\beta}$ ne sera pas l'unique estimateur des moindres carrés non linéaires.
L'identification locale n'est cependant pas suffisante. Une condition plus générale est l'identification globale :

$$SSR(\hat{\beta}) < SSR(\beta*), \forall \hat{\beta} \neq \beta* \tag{3.11}$$

Il s'agit d'une simple reformulation de la condition d'identification établit au précédent chapitre : l'estimateur doit être unique. Remarquons que même si un modèle est identifié localement, il est toujours possible qu'ily ait deux (ou davantage) estimations distinctes garantissant une même valeur minimale de la fonction objectif. A titre d'exemple :

$$y = \beta\gamma + \gamma^2 x_t + \epsilon \tag{3.12}$$

Il apparait clairement que si $(\hat{beta}, \hat{\gamma})$ minimise la fonction objectif des MCO pour ce modèle, $(-\hat{beta}, -\hat{\gamma})$ en fera autant. Donc le modèle est globalement non identifié par quelque ensemble de données que ce soit, bien que les conditions du premier et second ordre soient remplies.
Dans la pratique, notons qu'il est également possible qu'un modèle non linéaire puisse être linéarisé au moyen d'un passage au logarithme, notamment dans le cas de modèles multiplicatifs. L'exemple le plus courant est la fonction de production de type *Cobb-Douglas* :

$$y = AK^{\beta_1} L^{\beta_2} \epsilon \tag{3.13}$$

qui redevient un modèle linéaire dans le cas où l'on passe au log :

$$y = ln(A) + \beta_1 ln(K) + \beta_2 ln(L) + ln(\epsilon) \tag{3.14}$$

En paramétrisant ϵ de façon à ce qu'il suive une loi log-normale, on retrouve $\epsilon \sim N(\sigma^2)$. Ceci évite d'avoir recours à des procédures d'estimation complexes et consommatrices de temps.

3.2 Les modèles à système d'équations

On introduit brièvement une méthode économétrique utile pour l'un des principaux modèles financiers (le MEDAF ou *CAPM*) : les systèmes d'équations de regression. Là encore, le lecteur se reportera avec profit à
Les modèles décrits dans les chapitres précédents peuvent s'appliquer à des groupes de variables. Dans ce cas, on examine les modèles de manière jointe. L'un des principaux exemple pour l'économétrie de la finance est le **MEDAF**. Ce type de modélisation procède comme suit :

$$y_1 = X_1\beta_1' + \epsilon_1 \tag{3.15}$$
$$y_2 = X_2\beta_2' + \epsilon_2 \tag{3.16}$$
$$... \tag{3.17}$$
$$y_m = X_m\beta_m' + \epsilon_m \tag{3.18}$$

lorsque l'on dispose de m équations et de n observations. On se bornera ici à l'étude d'un cas particulier de ces régressions : le modèle *SUR* (*seemingly unrelated regressions* ou *modèle de régressions apparament indépendantes*). Le modèle se présente comme suit :

$$y_i = X_i\beta_i' + \epsilon_i, i = 1, ..., m \tag{3.19}$$
$$\epsilon = [\epsilon_1, ..., \epsilon_m] \tag{3.20}$$
$$\mathbb{E}[\epsilon | X_1, ..., X_m] = 0 \tag{3.21}$$
$$\mathbb{E}[\epsilon\epsilon' | X_1, ..., X_m] = \Sigma \tag{3.22}$$

On suppose que n observations sont utilisées pour l'estimation des paramètres des m équations. Chaque équation a K_m régresseurs, pour un total de $K = \sum_{i=1}^{m} K_j$. On pose $T > K_i$. On suppose que les perturbations ne sont pas corrélées entre observations. En conséquence,

$$\mathbb{E}[\epsilon_{it}\epsilon_{js} | X_1, ..., X_m] = \sigma_{ij}, \text{ si } t = s, 0 \text{ sinon} \tag{3.23}$$

La structure de perturbation est donc :

$$\mathbb{E}[\epsilon_i'\epsilon_j | X_1, ..., X_m] = \sigma_{ij} I_n \tag{3.24}$$

D'où on en déduit naturellement :

$$\mathbb{E}[\epsilon'\epsilon|X_1,...,X_m] = \Omega = \begin{bmatrix} \sigma_{11}I_n & \sigma_{12}I_n & ... & \sigma_{1m}I_n \\ \sigma_{21}I_n & \sigma_{22}I_n & ... & \sigma_{2m}I_n \\ & \vdots & & \\ \sigma_{m1}I_n & \sigma_{m2}I_n & ... & \sigma_{mm}I_n \end{bmatrix} = \Sigma \otimes I_n \tag{3.25}$$

Chaque équation est une régression classique. Les paramètres peuvent donc être estimés de manière convergent par la méthode MCO. La régression généralisée s'applique aux données dites *empilées* :

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \\ \vdots \\ y_{mn} \end{bmatrix} = \begin{bmatrix} X_1 & 0_{n,k_2} & ... & 0_{n,k_m} \\ 0_{n,k_1} & X_2 & ... & 0_{n,k_m} \\ \vdots & \vdots & ... & \vdots \\ 0_{n,k_1} & 0_{n,k_2} & ... & X_m \end{bmatrix} \begin{bmatrix} \beta_{1,1} \\ \beta_{1,2} \\ \vdots \\ \beta_{1,k_1} \\ \beta_{2,1} \\ \vdots \\ \beta_{m,k_m} \end{bmatrix} \tag{3.26}$$

$$+ \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \vdots \\ \epsilon_{1,n} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{m,n} \end{bmatrix} \tag{3.27}$$

Il est alors possible de réécrire ce modèle de façon matricielle :

$$Y = X\beta' + \epsilon \tag{3.28}$$

Où Y est une matrice $\mathcal{M}(n\times m, 1)$, X une matrice $\mathcal{M}(n\times m, K)$, β' une matrice $\mathcal{M}(K, 1)$ et ϵ une matrice $\mathcal{M}(n \times m, 1)$. Une fois ce travail préliminaire accompli, tournons nous vers l'estimation de ce type de modèle.
A ce stade plusieurs stratégies d'inférence sont envisageables. On montre dans ce qui suit que l'estimation par maximum de vraisemblance ou par Moindres Carrés Généralisés (MCG), comme dans le cas simple des MCO, coincident exactement. On s'intéresse finalement au cas où Σ est bloc diagonal.

3.2.1 Estimation par moindres carrés généralisés et quasi-généralisés

Quelques rappels sur les MCG et MCQG sont ici nécessaires. L'estimation efficace de β dans le modèle de régression généralisé requiert Σ. On suppose pour commencer que Σ est une matrice connues, symétrique et définie positive. Il arrive que ce soit le cas,

mais le plus souvent il est nécessaire de procéder à l'estimation de Σ avant de mettre en oeuvre la méthode (on parle alors de *moindres carrés quasi généralisés*).
L'idée de base consiste à minimiser la somme des carrés des résidus, pondérés par la variance des résidus. Cette somme est appelée mesure de Mahalanobis (utile pour l'analyse en terme de vraisemblance). Les MCG se propose de déterminer un estimateur qui minimise cette distance.

$$Min(Y - X\beta')'\Omega^{-1}(Y - X\beta') \tag{3.29}$$

Où $\Omega^{-1} = \Sigma^{-1} \otimes I_n$. En développant l'expression, il vient :

$$(Y' - \beta X')\Sigma^{-1} \otimes I_n(Y - X\beta') \tag{3.30}$$
$$= Y'\Sigma^{-1} \otimes I_n Y - \beta X'\Sigma^{-1} \otimes I_n Y - Y'\Sigma^{-1} \otimes I_n X\beta' + \beta X'\Sigma^{-1} \otimes I_n X\beta' \tag{3.31}$$

En dérivant la précédente expression, on obtient :

$$-2X'\Sigma^{-1} \otimes I_n Y + X'\Sigma^{-1} \otimes I_n\beta' = 0 \tag{3.32}$$
$$\Leftrightarrow (X'\Sigma^{-1} \otimes I_n X)\beta' = X'\Sigma^{-1} \otimes I_n Y \tag{3.33}$$
$$\Leftrightarrow \beta' = (X'\Sigma^{-1} \otimes I_n X)^{-1} X'\Sigma^{-1} \otimes I_n Y \tag{3.34}$$

On obtient ainsi l'expression de l'estimateur des MCG. On obtient bien un β' de bonne dimension : $\mathcal{M}(K, 1)$. Il est à noté que $\otimes$ est le produit Kronecker. On en rappelle brièvement les propriétés :

Définition 3.2.1 (Produit de Kronecker)**.** *Soit A une matrice $\mathcal{M}(m \times n)$ constituée d'éléments $[a_{ij}]$. Soit B une matrice $\mathcal{M}(p \times q)$. Alors on a le produit de Kronecker suivant :*

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & ... \\ a_{21}B & ... & ... \\ ... & ... & a_{mn}B \end{bmatrix} \tag{3.35}$$

Il s'agit donc d'une matrice $\mathcal{M}((m \times p) \times (n \times q))$

Proposition 3.2.1 (Produit de Kronecker)**.** *On a les propriétés suivantes :*
- $(A \otimes B)(C \otimes D) = AC \otimes BD$
- $(A \otimes B)' = A' \otimes B'$
- $A \otimes (B + C) = A \otimes B + A \otimes C$
- $(B + C) \otimes A = B \otimes A + C \otimes A$
- $A \otimes (B \otimes C) = (A \otimes B) \otimes C$

La matrice de covariance asymptotique de l'estimateur des MCQ est la matrice inverse dans la précédente équation. On rappelle quelques propriétés des estimateurs MCG. On détail le précédent résultat de façon s'en convaincre pleinement. En notant σ^{ij} la contrepartie de σ_{ij} dans Σ^{-1}, il vient :

$$\beta' = \begin{bmatrix} \sigma^{11}X_1'X_1 & \sigma^{12}X_1'X_2 & ... & \sigma^{1m}X_1'X_m \\ \sigma^{21}X_2'X_1 & \sigma^{22}X_2'X_2 & ... & \sigma^{2m}X_2'X_m \\ \vdots & \vdots & \vdots & \vdots \\ \sigma^{m1}X_m'X_1 & \sigma^{m2}X_m'X_2 & ... & \sigma^{mm}X_m'X_m \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^m \sigma^{1j}X_1'y_j \\ \vdots \\ \sum_{j=1}^m \sigma^{mj}X_m'y_j \end{bmatrix} \tag{3.36}$$

Il est aisé de s'en convaincre sur un exemple de faible dimensions, par exemple pour $m = 2, K = 2$. Autre remarque : en spécifiant $\Sigma = \sigma I_m$, on retrouve exactement l'estimateur des moindres carrés. On détaille quelques propriétés des estimateurs MCG :

Proposition 3.2.2 (Estimateur sans biais)**.** *L'estimateur MCG est un estimateur sans biais de β.*

Proposition 3.2.3 (Estimateur efficace)**.** *L'estimateur MCG est un estimateur efficace de β.*

Proposition 3.2.4 (Gauss Markov)**.** *L'estimateur MCG $\hat{\beta}$ est l'estimateur linéaire sans biais de variance minimale pour la régression généralisée. Il s'agit du théorème de* Aitken*[1935], et d'une généralisation de Gauss Markov.*

Pour finir, on a supposé que jusque ici que Σ était connue. Ceci n'est généralement pas le cas. On procède donc à une estimation préliminaire de Σ : les résidus des moindres carrés peuvent être utilisées pour estimer la matrice de variance-covariance des résidus. En note $\hat{\epsilon}$ les résidus des MCO, l'estimation de Σ est alors :

$$\hat{\Sigma} = \frac{1}{n}\hat{\epsilon}'\hat{\epsilon} \tag{3.37}$$

On parle alors d'estimateur des moindres carrés quasi-généralisés pour $\hat{\beta}$ estimé par MCG, une fois l'estimation de Σ accomplie. Les propriétés asymptotiques de l'estimateur MCQG ne proviennent pas de l'estimateur sans biais de Σ ; seule la convergence est nécessaire. L'estimateur a les mêmes propriétés que l'estimateur MCG.

3.2.2 MCO contre MCG et MCQG

L'estimateur MCG diffère bien évidemment de l'estimateur MCO. Pour l'instant, les équations sont seulement liées par leurs perturbations, d'où le terme de régressions apparament indépendantes. Il est alors nécessaire de s'interroger sur le gain d'efficacité provenant de l'estimateur MCO à la place de l'estimateur MCG. L'estimateur MCO est ici un estimateur équation par équation, laissant de coté l'écriture fastidieuse du modèle MCG. On détaille quelques cas particuliers :

– Si les équations sont indépendantes - i.e. si $\sigma_{ij} = 0, i \neq j$ - alors il n'y a aucun avantage à utiliser les MCG pour estimer le système d'équation. En effet, les MCG reviennent aux MCO équation par équation.

– Si les équations ont les mêmes variables explicatives, alors MCO et MCG sont identiques.

– Si les regresseurs dans un bloc d'équations sont un sous-ensemble des regresseurs dans d'un autre , alors MCG n'apporte aucun gain d'efficacité par rapport aux MCO dans l'estimation de l'ensemble de l'ensemble plus petit d'équation ; ainsi MCO et MCG sont de nouveau identiques.

3.2.3 Estimation de systèmes d'équation par maximum de vraisemblance

Il est possible, comme pour le cas des MCO, d'estimer le modèle par maximum de vraisemblance. On rappelle qu'une loi gaussienne multivariée se note comme suit :

$$f(X_1, ..., X_m) = (2\pi)^{-\frac{m}{2}} |\Omega|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(Y - X\beta')'\Omega^{-1}(Y - X\beta')\right\} \tag{3.38}$$

Sur la base des notations introduites plus haut, la vraisemblance concentrée du modèle SURE s'écrit :

$$lnL = -\frac{n}{2}ln(|\Omega|) - \frac{1}{2}(Y - X\beta')'\Omega^{-1}(Y - X\beta') \tag{3.39}$$

En dérivant par rapport à β', il vient :

$$X'\Omega^{-1}(Y - X\beta') = 0 \tag{3.40}$$
$$\Leftrightarrow X'\Omega^{-1}Y = X'\Omega^{-1}X\beta' \tag{3.41}$$
$$\Leftrightarrow \beta' = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \tag{3.42}$$

On retrouve bien l'estimateur des MCG : celui profite donc de l'ensemble des propriétés des estimateurs du maximum, developpées plus haut. L'estimateur de la variance se déduit des propriétés suivantes :

Proposition 3.2.5 (Dérivation matricielle 1)**.** *Soit A une matrice carrée, inversible et de déterminant positif. Alors :*

$$\frac{\partial ln(|A|)}{A} = (A')^{-1} \tag{3.43}$$

Proposition 3.2.6 (Dérivation matricielle 2)**.**

$$-ln(|A|) = -ln(|A^{-1}|)^{-1} = log(|A^{-1}|) \tag{3.44}$$

Une fois que l'on connait ces propriétés, il est aisé de voir que :

$$\frac{\partial lnL}{\partial \Omega^{-1}} = \frac{n}{2}\Omega - \frac{1}{2}(Y - X\beta')(Y - X\beta')' \tag{3.45}$$

En égalisant la dernière dérivée à 0, il vient :

$$\Omega = \frac{1}{n}\epsilon\epsilon' \tag{3.46}$$

où $\epsilon = Y - X\beta'$. On retrouve bien un estimateur habituel de la variance.

3.2.4 Retour sur l'estimation du MEDAF : implémentation des MCQG

On se propose de calibrer le MEDAF à l'aide des *Moindres Carrés Quasi Généralisés*. Comme exposé plus haut, il est tout d'abord nécessaire d'obtenir une estimation de la matrice de covariance des résidus. On utilise pour cela la matrice des résidus issus des estimations produites par les moindres carrés dans la fonction `capm.R`, fournies

précédement. Il est ensuite nécessaire de réécrire le système d'équation sous la forme présentée plus haut : la nouvelle matrice des variables explicatives doit être de la forme $\mathcal{M}(n \times m, m \times 2)$, dans la mesure où l'on a ici uniquement deux variables explicatives pour l'ensemble des équations du systèmes (la constante et l'index de marché).

On commence donc par réécrire l'ensemble des matrices de façon appropriée. Avec l'estimation de Σ, on fournit une estimation de Ω :

$$\Omega = \Sigma \otimes I_n \tag{3.47}$$

Il ne reste alors plus qu'à estimer la matrice β, qui comporte (si tout se passe bien) $n \times m$ éléments en colonne. L'estimateur $\widehat{\beta}$ est alors :

$$\widehat{\beta} = (X^T \Omega^{-1} X)^{-1} (X^T \Omega^{-1} Y) \tag{3.48}$$

La fonction `capm.R` a été complétée pour prendre en compte l'estimation par MCQG : le code est fournit ci-après. Une mise en garde s'impose : le calcul de Ω conduit à construire une matrice aux dimensions imposantes. Il est par conséquent possible que de nombreux PC ne puisse pas permettre l'estimation par MCQG d'un système d'équations où m est *grand*. Dans notre cas, pour 30 titres et 250 dates, la matrice Ω est de taille 7500×7500 : elle comporte donc... 56 250 000 éléments ! Il est par conséquent nécessaire de disposer d'une mémoire vive... très importante.

Les resultats retournés par la procédure sont fournis dans la table 3.2.4 et comparés à ceux obtenus par MCO. Le constat principal est la suivant : les MCQG n'apporte rien en comparaison des MCO lors de l'estimation du CAPM. Ceci tient à la matrice de variance-covariance des erreurs MCO qui est une matrice diagonale presque surement. Pour tester la valeur d'une corrélation ρ, on rappelle que la statistique de test est :

$$T_\rho = \sqrt{n-2} \frac{\rho}{\sqrt{1-\rho^2}} \tag{3.49}$$

Cette statistique suit, sous $H_0 : \rho = 0$, une loi de student à $n-2$ degrés de liberté (où n est le nombre de données disponibles). On compare donc T_ρ à 1,96, comme on l'a fait pour les tests de Student.

Finalement, ceci ne sert qu'à montrer que dans de nombreux cas pour lesquels la structure de corrélation se réduit aux simples variances, il n'est pas nécessairement utile de recourir aux MCG/MCQG : les MCO suffisent amplement. Pour refaire ces estimations vous-mêmes, il suffit d'ajouter à la fonction `capm.R` les quelques lignes qui suivent, ainsi que de modifier `return`, comme c'est ici le cas :

```
# Calcul de l'estimateur des MCQG
# Mise en forme des données
titres.new=matrix(titres,length(titres),1)
n=nrow(titres)
X.new=matrix(0,length(titres),2*ncol(titres))
```

```
for (i in 1:ncol(titres)){
X.new[(n*(i-1)+1):(i*n),(2*(i-1)+1):(2*i)]=SP
}
# Calcul de la matrice de variance covariance des résidus MCO
var.res=var(res)
omega=solve(kronecker(var.res,diag(1,nrow(titres),nrow(titres))))
# Estimation des paramètres
beta.mcqg=t(solve(t(X.new)%*%omega%*%X.new)%*%(t(X.new)%*%omega%*%titres.new))
return(list(theta=theta, R2=R2, test=test, ptest=ptest, beta.mcqg=beta.mcqg))
```

		Alcoa	AT.T	Boeing	Caterpilar	Chevron	Coca.Cola	Disney	DuPont	Eastman.Kodak
MCO	**alpha**	0.003024439	-0.01330993	0.004927275	-0.001995679	-0.006321155	-0.001599203	0.0084873	0.001150471	-0.01899075
	beta	1.029791342	0.74692300	1.143201992	1.054370309	0.840468354	0.802655308	1.1226112	1.016821780	0.59013381
MCQG	**alpha**	0.003024439	-0.01330993	0.004927275	-0.001995679	-0.006321155	-0.001599203	0.0084873	0.001150471	-0.01899075
	beta	1.029791	0.746923	1.143202	1.054370	0.8404684	0.8026553	1.122611	1.016822	0.5901338

TAB. 3.1 – Estimations du CAPM par MCQG

Chapitre 4

Optimisation de fonctions à plusieurs variables par algorithme

On présente dans ce qui suit quelques bases nécessaires à l'approximation numérique de fonctions de plusieurs variables. Comme on l'a vu dans la partie consacrée aux modèles non linéaires, il arrive (assez souvent) lorsque l'on cherche à maximiser une vraisemblance ou des SSR qu'il n'existe pas de forme analytique pour les estimateurs. Il est alors nécessaire d'approximer ce maximum de façon numérique, i.e. déterminer une valeur approchée des paramètres assurant que la vraisemblance soit maximale.

Ne le cachons pas, proposer ce type de chapitre n'est pas monnaie courante : rares sont les cours de statistique à insister sur de tels aspects computationnels. Je suis intimement convaincu que de trop nombreux économètres ne programment pas leurs propres procédures d'estimation, et ceci est extrêment dommageable. Des générations d'élèves sont formés dans l'idée que l'économétrie se fait simplement en lançant des procédures SAS ou pire : Eviews. Il existe un certain illétrisme économétrique parmis beaucoup d'élèves : il est possible de programmer ses propres procédures d'estimation (notamment en `R`) et ce sans faire d'efforts incroyables. Je propose dans ce qui suit les principales intuitions nécessaires à la programmation de fonction permettant d'optimiser un fonction de p variables. Bien évidemment, on ne propose ici pas l'ombre d'une preuve des méthodes proposées : il s'agit bien plutôt d'une chapitre de cuisine économétrique. Il existe de nombreuses références proposant les preuves de ce qui va suivre, mais je doute qu'un élève de M1 en tire un quelconque profit.

Les différentes recettes proposées ont été tirées pelle-melle de : Harvey (1990), Deschamps (2004) Quinn (2001) et Greene (2002). Notez que ce chapitre du Green est librement téléchargeable sur le site de Pearson Education (et en Français !)[1]. On propose trois types de méthodes : une première approche intuitive vise à déterminer graphiquement (méthode de recherche par quadrillage) les fonctions de un ou deux variables. Ce type de méthode laisse ensuite la place à l'ensemble des méthodes basées sur le gradient : plus grande pente, Newton Raphson, méthode du score et BHHH. Ces méthodes se heurtent cependant aux problèmes liées aux multimodalités du maximum de vraisemblance : les méthodes de type recuit simulé sont brièvement présentées.

[1] `www.pearsoneducation.fr`

4.1 Pour commencer...

Dans ce qui va suivre, on s'intéresse à un problème très simple : soit $\{x_1, x_2, ..., x_n\}$ un échantillon que l'on suppose tiré d'une loi normale de paramètres inconnus. On souhait déterminer les estimateurs du maximum de vraisemblance de ces paramètres. [Ce problème se traite aisément sans tout ce qui va suivre, mais autant prendre un problème simple pour introduire des éléments plus complexes.] La vraisemblance associée à l'échantillon est :

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right\} \tag{4.1}$$

La log-vraisemblance associée est naturellement :

$$lnL = -\frac{n}{2}ln(2\pi) - nln(\sigma) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2 \tag{4.2}$$

On cherche à présent à déterminer les paramètres $\theta = (m, \sigma)$ qui maximisent cette expression. La première idée que l'on peut proposer est de représenter cette vraisemblance en deux dimensions en fonction d'un *grid* ou quadrillage. On détermine une valeur minimum ainsi qu'une valeur maximum pour l'ensemble des valeurs que m peut prendre. On fait de même pour σ. Ceci nous fournit alors un damier composé des deux grid ainsi formés (les deux supports de nos paramètres). L'idée est alors de calculer la log-vraisemblance associée à l'ensemble des points d'intersection des lignes composant le damier. L'algorithme à coder pour déterminer le maximum de vraisemblance est alors :

1. Déterminer le support des paramètres
2. Pour chaque valeur de ce support (une boucle *for* par paramètre), on calcule la log-vraisemblance
3. On entamme ensuite une nouvelle boucle (possibilité de la combiner avec la précédente) pour déterminer le max sur les logvraisemblance pour le support.
4. On récupère finalement la valeur des paramètres garantissant cette valeur maximale de la log-vraisemblance.

C'est ce qui est fait par le code suivant (à ceci pret que la fonction proposée représente également la logvraisemblance en 3D).

```
grid<-function(x,n){
mu<-seq(1,3,length=n)
sigma<-seq(1,3,length=n)
T=nrow(x);
lnL=matrix(0,n,n);
for (i in 1:n){
for (j in 1:n){
mm2=sum((x-mu[i])^2);
```

```
lnL[i,j]=-T*log(sigma[j])-1/2*(mm2/(sigma[j]^2));
}}
persp(mu,sigma,lnL, theta = 50, phi = 30, expand = 0.5, col = "lightblue");
max=min(lnL);
for (i in 1:n){
for (j in 1:n){
if(max<lnL[i,j]){max=lnL[i,j]; indexi=i;indexj=j}
}}
mumax=mu[indexi];
sigmamax=sigma[indexj];
return(list(mu=mu,sigma=sigma,lnL=lnL,sol=cbind(max,mumax,sigmamax)))
}
```

On obtient alors les graphiques présentés en figure 4.1 et 4.2, en utilisant les fonction `persp` et `contour` de `R`. On en déduit alors l'estimation de nos paramètres. L'idée est essentiellement de se construire une représentation mentale de ce qu'est une vraisemblance (ça existe !), avant de passer à des méthodes plus avancées.

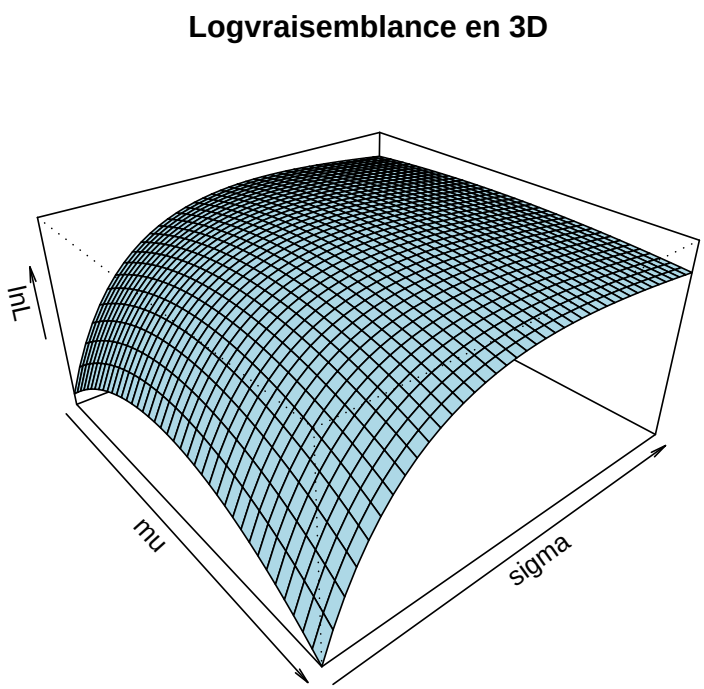

FIG. 4.1 – Log-vraisemblance en 3D

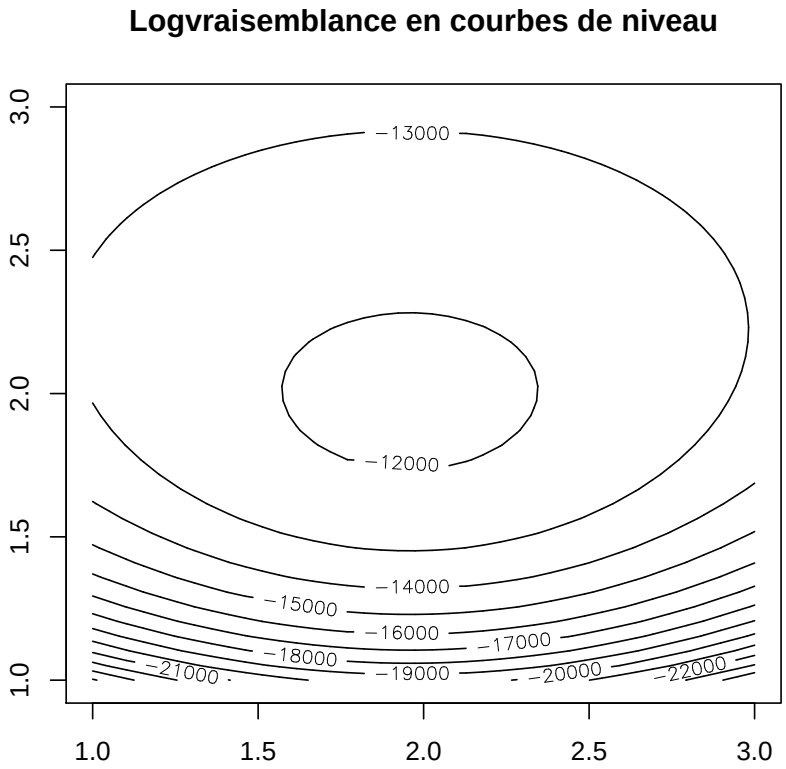

FIG. 4.2 – Log-vraisemblance en 2D

4.2 Les méthodes du gradient

Les méthodes les plus utilisées en économétrie/statistique sont généralement basées sur un calcul approché du gradient [matrice des dérivées premières de la fonction à maximiser]. Il existe différentes méthodes, qui ont chacune leurs qualités et défauts. On les présente ici de façon très pratique : il s'agit de permettre une programmation aisée de ces méthodes de maximisation.

4.2.1 Quelques généralités pour commencer...

L'ensemble des méthodes ci-après partent toutes plus ou moins du même point de départ. On cherche à déterminer le θ optimal, i.e. l'ensemble des paramètres qui maximise une fonction $f(\theta)$. On part d'un point initial noté θ_0, si possible le plus proche des vraies valeurs de θ^* (les paramètres optimaux). On cherche ensuite à modifier la valeur des ces paramètres selon une certain pas λ_t et une certaine direction Δ_t, de façon à former une suite de valeur θ_t qui converge vers la vraie valeur de θ^*. La chaîne des θ_t est alors formée de la façon suivante :

$$\theta_{t+1} = \theta_t + \lambda_t \Delta_t \tag{4.3}$$

Tout l'art de la maximisation vient alors du fait de spécifier λ_t et Δ_t de façon optimale, i.e. :

– de façon à ce que l'algorithme aboutisse effectivement à la valeur optimale (une valeur approchée du moins),
– et de façon à ce qu'on l'atteigne ce point le plus vite possible, c'est à dire :
  1. que l'algorithme nécessite le moins de calculs complexes possibles,
  2. que l'algorithme consomme le moins de ressource machine possible.

On développe ici quelques méthodes bien connues : méthode de la plus grande pente, méthode de Newton Raphson et méthode du score par BHHH. L'ensemble de ces méthodes sont dires méthodes du gradient dans la mesure où Δ_t est systématiquement de la forme :

$$\Delta_t = W_t G_t \tag{4.4}$$

où W_t est une matrice définie positive et G le gradient de f :

$$G = \left[\frac{\partial f}{\partial \theta}\right] \tag{4.5}$$

4.2.2 La méthode de la plus grande pente

Il s'agit de la méthode la plus simple. On utilise G, le gradient de f, comme direction : on a donc $W_t = I$ et $\Delta_t = G_t$. On dirige ainsi l'estimateur du côté de la plus grande pente, i.e. de la plus grande dérivée. On montre que le pas optimal est alors de la forme :

$$\lambda_t = \frac{-G'G}{G_t' H_t G_t} \tag{4.6}$$

où H est la matrice hessienne, i.e. la matrice des dérivées secondes de f. On a donc :

$$H = \frac{\partial^2 f}{\partial \theta \partial \theta'} \tag{4.7}$$

Ainsi l'itération de la plus grande pente est alors :

$$\theta_{t+1} = \theta_t - \frac{G'G}{G_t' H_t G_t} G_t \tag{4.8}$$

En supposant que l'on a p paramètres à estimer, θ est alors une $\mathcal{M}(p, 1)$, G est également une $\mathcal{M}(p, 1)$ et H est une $\mathcal{M}(p, p)$. On vérifie donc que les dimensions des matrices correspondent bien.

L'algorithme est des plus simples à programmer :

1. Etablir un θ_0
2. Déterminer G_0 et H_0
3. Déterminer Δ_t et λ_t

4. Déterminer θ_1
5. Tester $\sum G^2 < \epsilon$. Si oui, stop. Si non, on continue la boucle...

Le code `R` est fourni ici :

```
steep<-function(theta,x,iter){
H=matrix(0,2,2)
G=matrix(1,2,1)
n=nrow(x)
check=matrix(0,2,1)
i=1;
while(sum(G^2)>0.0000001){
mm=sum(x-theta[1,1])
mm2=sum((x-theta[1,1])^2)
H[1,1]=-n/(theta[2,1]^2);
H[2,2]=H[1,1]-3*(mm2)/(theta[2,1]^4);
H[1,2]=-(2*mm)/(theta[2,1]^3);
H[2,1]=H[1,2];
G[1,1]=mm/(theta[2,1]^2);
G[2,1]=-n/theta[2,1]+(mm2)/(theta[2,1]^3);
cat(i,"\n");
check=cbind(check,theta-solve(H)%*%G);
theta=theta-as.numeric((t(G)%*%G)/(t(G)%*%H%*%G))*G;
i=i+1
}
return(list(theta=theta,check=check))
}
```

Ici, on a fixé ϵ à 0.0000001, de façon arbitraire. On illustre les résultats de cette méthode sur notre problème initial de vraisemblance gaussienne en figure 4.3. On a fixé volontairement θ_0 loin de la vraie valeur des paramètres : on observe une certain nombre de sauts, même si on trouve au final la bonne valeur approchée des paramètres.

Deux mises en gardes cependant :

– Le calcul des dérivées secondes peut être long et pénible.
– D'autre part, si θ_t est loin de la vraie valeur du maximum, il est possible que H_t ne soit pas définie négative, et que l'algorithme diverge.

4.2.3 La méthode de Newton-Raphson

Il s'agit d'une méthode classique d'optimisation numérique. L'origine de la méthode est la suivante : supposons que l'on veuille trouver $\hat{x} \in \mathbb{R}^k$ qui maximise la fonction $f : \mathbb{R}^k \to \mathbb{R}$, qui est deux fois continuement dérivable. Il est possible de donner le développement de Taylor suivant de la fonction f :

$$f(x+h) \approx f(x) + G'h + \frac{1}{2}h'Hh \tag{4.9}$$

FIG. 4.3 – Convergence de la méthode de la plus grande pente

avec les notations précédentes pour G et H. Ceci implique naturellement que :

$$\frac{\partial f(x+h)}{\partial h} = G + Hh \tag{4.10}$$

La condition de premier ordre pour un maximum est alors :

$$0 = G + Hh \tag{4.11}$$

$$\Leftrightarrow h = -H^{-1}G \tag{4.12}$$

En d'autres termes, la direction optimale pour optimiser f (retour sur notre problème) est de choisir θ_{t+1} tel que :

$$\theta_{t+1} = \theta_t - H_{-1}G \tag{4.13}$$

On notera qu'avec Newton-Raphson, on a : $\lambda_t = 1 \forall t$. Là encore, la programmation de ce type d'algorithme est très simple :

1. Etablir un θ_0
2. Déterminer G_0 et H_0
3. Déterminer Δ_t et λ_t
4. Déterminer θ_1
5. Tester $\sum G^2 < \epsilon$. Si oui, stop. Si non, on continue la boucle...

```
newton<-function(theta,x,iter){
H=matrix(0,2,2)
G=matrix(1,2,1)
n=nrow(x)
```

```
check=theta
i=1;
while(sum(G^2)>0.0001){
mm=sum(x-theta[1,1])
mm2=sum((x-theta[1,1])^2)
H[1,1]=-n/(theta[2,1]^2);
H[2,2]=H[1,1]-3*(mm2)/(theta[2,1]^4);
H[1,2]=-(2*mm)/(theta[2,1]^3);
H[2,1]=H[1,2];
G[1,1]=mm/(theta[2,1]^2);
G[2,1]=-n/theta[2,1]+(mm2)/(theta[2,1]^3);
cat(i,"\n");
check=cbind(check,theta-solve(H)%*%G);
theta=theta-solve(H)%*%G;
i=i+1
}
return(list(theta=theta,check=check))
}
```

On applique également cette méthode à notre problème, en partant encore d'un point éloigné du véritable maximum. Là, l'algorithme de Newton Raphson diverge et aboutit à la solution $m = -1182571474$ et $\sigma = -168165462$, soit des valeurs abhérantes. Ceci est représenté sur la figure 4.4.

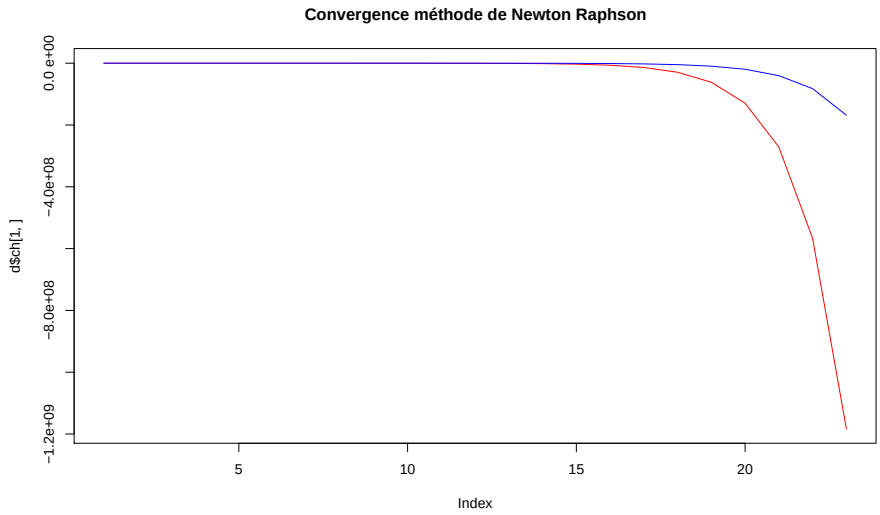

FIG. 4.4 – Convergence de la méthode de Newton Raphson

Cette méthode converge cependant dans de nombreux cas. Si la fonction est quadratique, elle atteint l'optimum en une itération depuis n'importe quel point de départ. Si la fonction est globalement concave, cette méthode reste probablement la meilleure. Elle est particulièrement adaptée à l'estimation du maximum de vraisemblance.

4.2.4 Méthode du score et matrice BHHH

Il n'est parfois pas possible ou simplement trop couteux de calculer la matrice hessienne directement. On la remplace alors par un estimateur, en s'appuyant sur la propriété bien connue :

$$-\mathbb{E}\left[\frac{\partial^2 lnL}{\partial\theta\partial\theta'}\right] = \mathbb{E}\left[\frac{\partial lnL}{\partial\theta}\frac{\partial lnL}{\partial\theta'}\right] \tag{4.14}$$

Là encore, il peut être parfois plus simple d'approximer cette espérance à l'aide de l'estimateur BHHH (Berndt et al. (1974)) :

$$-H = \sum_{i=1}^{n} G'G \tag{4.15}$$

Le calcul de θ_{t+1} se fait alors comme suit :

$$\theta_{t+1} = \theta_t + BHHH_t^{-1}G_t \tag{4.16}$$

L'algorithme est alors :

1. Etablir un θ_0
2. Déterminer G_0 et $BHHH_0$
3. Déterminer Δ_t et λ_t
4. Déterminer θ_1
5. Tester $\sum G^2 < \epsilon$. Si oui, stop. Si non, on continue la boucle...

Dans notre cas, le code `R` peut être écrit comme suit :

```
HHH<-function(theta,x){
G=matrix(1,2,1)
n=nrow(x)
check=theta
i=1;
while(sum(G^2)>0.0001){
mm=sum(x-theta[1,1])
mm2=sum((x-theta[1,1])^2)
BHHH=cbind((x-theta[1,1])/(theta[2,1]^2),-1/theta[2,1]+((x-theta[1,1])^2)/(theta[2,1]^3
H=(t(BHHH)%*%BHHH);
G[1,1]=mm/(theta[2,1]^2);
G[2,1]=-n/theta[2,1]+(mm2)/(theta[2,1]^3);
cat(theta,"\n");
check=cbind(check,theta+solve(H)%*%G);
theta=theta+solve(H)%*%G;
i=i+1
}
return(list(theta=theta,check=check))
}
```

On présente en figure 4.5 une application de ce code à notre problème. La convergence est rapide, et l'algorithme fonction (quasiment) toujours (du moins pour les essais que j'ai fait). Mieux, la matrice $BHHH^{-1}$ fournit à l'optimum une estimation de la variance des estimateurs (i.e. de l'inverse de la matrice d'information de Fisher).

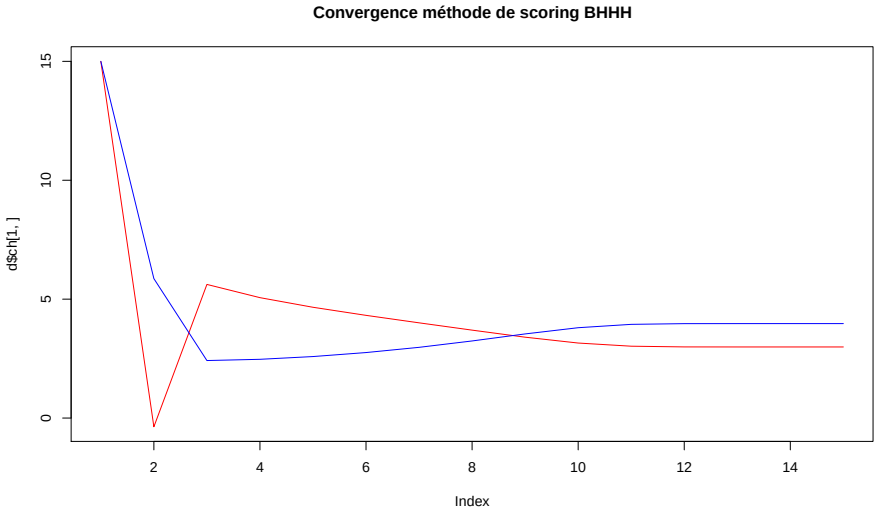

FIG. 4.5 – Convergence de la méthode de scoring par BHHH

4.3 Estimations par algorithme aléatoire

Une autre façon de voir les choses est de compter sur le hasard : lorsque le nombre de paramètres est grand, il n'est pas possible d'utiliser la méthode du quadrillage. Pire, dans un tel cas, il est rarement aisé de s'assurer de la globale concavité de f. L'existence de maximum locaux met en danger les estimations. Une solution peut être de lancer des algorithmes de type gradient en partant de différents points de départ. Il est également possible d'utiliser des méthodes de type recuit simulé, telles que l'algorithme de Métropolis-Hastings. Encore une fois, on en fournit ici une présentation de type "cuisine économétrique". Le lecteur soucieux d'aller plus loin est renvoyé aux références cités dans la partie consacrée à ces méthodes.

4.3.1 Faire jouer le hasard

L'idée de ce type de méthode est de remplacer un grid search par une recherche d'optimum aléatoire. Au lieu de construire une suite de paramètre de façon déterministe (type méthode du gradient), on rend cette suite aléatoire. On a donc :

$$\theta^*_{t+1} = \theta_t + \epsilon_t \tag{4.17}$$

où ϵ_t est tiré d'une loi particulière (uniforme, normale...). A ce stade, θ^*_{t+1} n'est pas encore le paramètre retenu à l'itération $t+1$. A chaque itération, on compare la log-vraisemblance associée à θ_t et à θ^*_{t+1}. On conserve le paramètre qui rend la log-vraisemblance maximale. Il s'agit donc d'un algorithme itératif : il est nécessaire de

déterminer un critère d'arrêt pour l'algorithme. Deux possibilités : il est possible de fixer un nombre d'itérations *ex ante*, ou de d'utiliser un critère de type gradient, comme il en a été question dans les méthodes précédentes.

L'algorithme est donc :

1. Choisir θ_0
2. Tirer autant de ϵ que de paramètres
3. Déterminer θ_1
4. Comparer $lnL(\theta_0)$ et $lnL(\theta_1)$
5. Conserver le θ_i rendant la log-vraisemblance maximale
6. Recommencer la boucle

Le code `R` dans le cadre de notre problème est le suivant :

```
random<-function(theta,x,iter){
nb=length(theta)
logv<-function(x,theta){
mm2=sum((x-theta[1,1])^2);
lnL=-nrow(x)*log(theta[2,1])-1/2*(mm2/theta[2,1]^2)
return(list(lnL=lnL))
}
thetachain=theta
thetanew=matrix(0,2,1)
for(i in 1:iter){
for (j in 1:nb){thetanew[j,1]=theta[j,1]+rnorm(1)/2}
if (logv(x,thetanew)$lnL>logv(x,theta)$lnL){theta=thetanew}
cat(i,"\n")
thetachain=cbind(thetachain,theta)
}
return(list(theta=theta,lnL=logv(x,theta)$lnL,thetachain=thetachain))
}
```

Appliqué à notre problème, ce code fournit une estimation correcte de nos paramètres. En partant d'un point éloigné de maximum, l'algorithme converge cependant lentement. C'est ce qu'on observe sur la figure 4.6.
Il est possible de converger plus rapidement, en effectuant des saut dans les paramètres plus importants. Cependant, ce faisant, on diminue également la probabilité de trouver un point qui maximise la vraisemblance. Notons que l'algorithme s'alourdit considérablement en présence d'un grand nombre de paramètres. En effet, on travaille à chaque itération autour du voisinage du dernier maximum trouvé par l'algorithme : un voisinage à n paramètres est bien plus grand qu'un voisinage à deux paramètres. Pour finir, il est possible de coupler les méthodes basées sur le gradient (définissant la direction) avec un algorithme construit pour générer des nombres positifs aléatoirement : on génère alors une taille de pas aléatoire. L'idéal est de générer au début de l'algorithme des taille de pas importantes, et de diminuer ces tailles au fur et à mesure de l'algorithme. Ceci est d'ailleurs proche de ce qu'on appelle la température dans l'Algorithme de Métropolis Hastings.

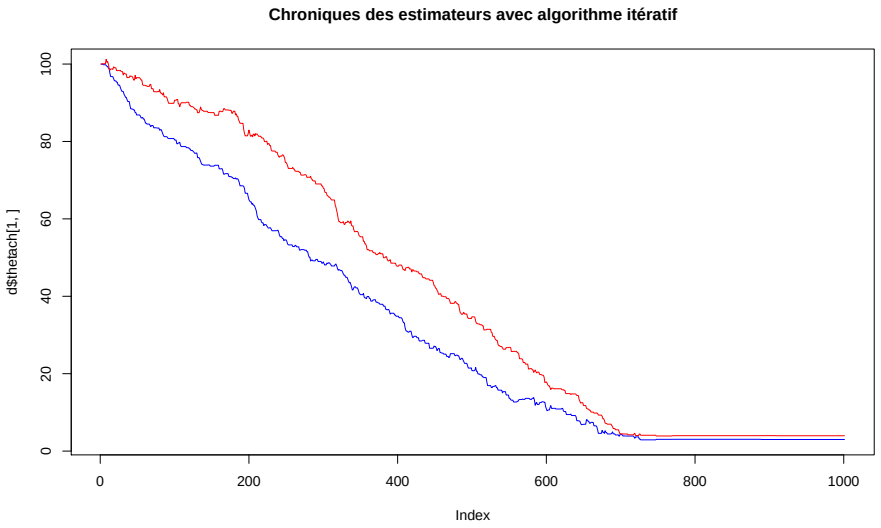

FIG. 4.6 – Convergence de la méthode aléatoire

4.3.2 Moduler le hasard : Metropolis Hastings et le recuit simulé

On présente ici rapidement une façon de résoudre des problèmes de maximisation en présence de maxima locaux et d'un unique maximum global. On utilise pour cela l'algorithme de Métropolis Hastings. Cet algorithme a été introduit par Metropolis et al. (1953) pour minimiser un critère sur un espace d'états fini de grande taille. Cette présentation est inspirée de Robert (1996) et Duflo (1996).

Soit une fonction f à maximiser en θ. La fonction f est souvent appelée fonction d'énergie, dans le cadre de cette méthode. On part de θ_0 et on fixe un paramètre T dit de température : ce paramètre va déterminer la probabilité d'échapper à un maximum local au fur et à mesure des itération. On détermine θ_{t+1} de la même façon que précédement :

$$\theta^*_{t+1} = \theta_t + \epsilon_t \tag{4.18}$$

Ce qui change par rapport à la précédente méthode est le mode de selection entre θ^*_{t+1} et θ_t. On compare la log-vraisemblance associée à chacun d'eux :

- si $lnL(\theta^*_{t+1}) > lnL(\theta_t)$, alors on retient $\theta_{t+1} = \theta^*_{t+1}$.
- sinon, on ne rejette pas nécessairement θ^*_{t+1}. On l'accepte avec une probabilité de la forme : $min(exp(-\frac{lnL(\theta^*_1)-lnL(\theta_0)}{T}), 1)$. Dans la pratique, on procède à un tirage d'une loi uniforme, puis on compare ce tirage avec le précédent calcul. Si le tirage est inférieur, on choisit θ^*_{t+1}, sinon, on conserve θ_t. La température T du système permet d'augmenter cette probabilité d'accepter θ^*_{t+1}. En général, il s'agit d'une fonction du nombre d'itérations de l'algorithme.

L'algorithme est alors le suivant :

1. Déterminer θ_0
2. Tirer ϵ et déterminer θ_1^*
3. Comparaison de la lnL associée aux deux paramètres
4. Si $lnL(\theta_1^*) > lnL(\theta_0) \Rightarrow \theta_1 = \theta_1^*$
5. Sinon : on tire u selon une loi uniforme.
6. On calcule $min(exp(-\frac{lnL(\theta_1^*)-lnL(\theta_0)}{T}), 1)$ et on compare à u
7. Si $u < min(exp(-\frac{lnL(\theta_1^*)-lnL(\theta_0)}{T}), 1)$, alors $\theta_1 = \theta_1^*$.
8. Sinon $\theta_1 = \theta_0$
9. On recommence la boucle...

Le code `R` peut être le suivant :

```
randommh<-function(theta,x,iter){
nb=length(theta)
logv<-function(x,theta){
mm2=sum((x-theta[1,1])^2);
lnL=-nrow(x)*log(theta[2,1])-1/2*(mm2/theta[2,1]^2)
return(list(lnL=lnL))
}
thetachain=theta
thetanew=matrix(0,2,1)
for(i in 1:iter){thetanew[2,1]=-1;
while(thetanew[2,1]<0.1){for (j in 1:nb){thetanew[j,1]=theta[j,1]+rnorm(1)}}
if (logv(x,thetanew)$lnL>logv(x,theta)$lnL){theta=thetanew; cat("-","\n")}
if (logv(x,thetanew)$lnL<logv(x,theta)$lnL){u=runif(1);
if(u<min(exp((logv(x,thetanew)$lnL-logv(x,theta)$lnL))*exp(1+i^(1/100)),1)){theta=theta
}
cat(i,"\n")
thetachain=cbind(thetachain,theta);
plot(thetachain[1,],type="l",col="blue")
}
return(list(theta=theta,lnL=logv(x,theta)$lnL,thetachain=thetachain))
}
```

Il existe des façons bien plus complexe de contrôler la température du système que celle proposée ici qui est déterministe. Graphiquement, sur un problème aussi simple que le notre, il n'y a pas grande différence entre la précédente méthode et celle-ci (cf. figure 4.7).

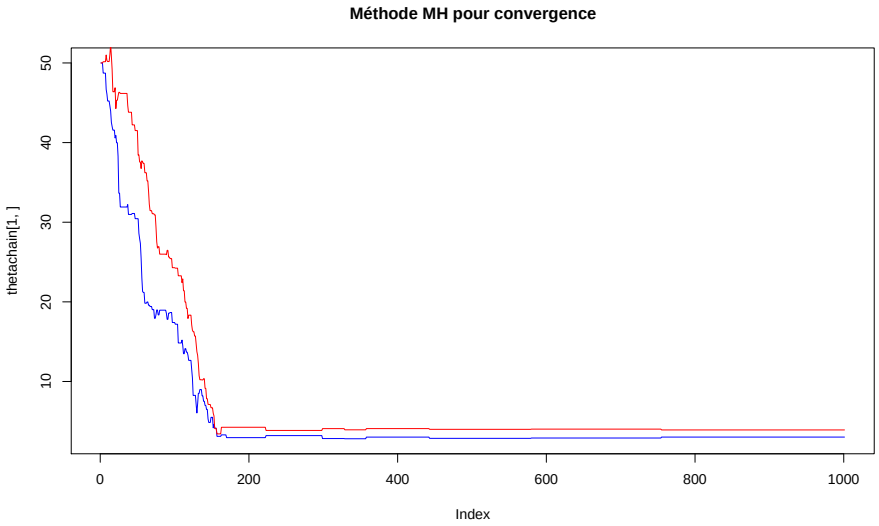

FIG. 4.7 – Convergence de la méthode MH

Chapitre 5

Introduction aux modèles de séries temporelles

5.1 Qu'est-ce qu'une série temporelle ?

La plupart des donnés macroéconomiques et financières prennent la forme de *séries temporelles*, un jeu d'observations répétées d'une même variable, telle que le PIB ou le rendement d'un titre donné. Dans ce qui suit, on note un série temporelle de la façon suivante :

$$\{x_1, x_2, ..., x_T\} \text{ ou } \{x_t\},\ t = 1, 2, ..., T \tag{5.1}$$

x_t est appelé variable aléatoire. En principe, il n'existe pas ou peu de différence entre les séries temporelles et l'économétrie, sinon que les variables sont indicées par t plutôt que par i. Par exemple, si y_t est généré par :

$$y_t = \beta x_t + \epsilon_t,\ \mathbb{E}[\epsilon_t | x_t] = 0 \tag{5.2}$$

alors une estimation par MCO permet d'obtenir des estimateurs consistants, de la même façon que dans le cas classique (i.e. indexé par i).

Le terme de *séries temporelles* est utilisé de façon interchangeable pour désigner à la fois un échantillon de données $\{x_t\}$ et un modèle probabiliste pour cet échantillon. Un exemple de modèle probabiliste peut être le suivant :

$$x_t = \epsilon_t,\ \epsilon_t \sim i.i.d.N(0, \sigma) \tag{5.3}$$

Heureusement pour nous, il est très rare que des séries temporelles soient i.i.d. (indépendantes et identiquement distribuées) : il s'agit précisement de ce qui les rend intéressantes. Le PIB en donnée trimestrielle constitue un bon exemple de cette dépendance temporelle : en général, quand le PIB est anormalement haut en t (i.e. au-dessus de sa moyenne historique ou non-conditionnelle), il y a de très fortes chances pour que la prochaine valeur de ce même PIB soit elle aussi anormalement haute.

L'idée est donc de parvenir à caractériser la loi jointe de $\{..., x_{t-1}, x_t, x_{t+1}, ...\}$. Evidement, il serait interessant de mettre en oeuvre des méthodes non paramétriques

(histogrammes, kernels...) afin d'embrasser toute la dépendance existante. Le problème est que les séries temporelles - du moins les séries économiques - sont souvent réduites à (au plus) 300 à 400 points. Il n'en va pas de même des séries financières, qui peuvent être plus conséquente (centaines de milliers de points) : cependant, la forme de la dépendance sur ce type d'actif est tout sauf stable. D'où le recours à des procédures paramétriques, cadre infiniment plus souple, aisé et agréable à manier. Là encore, le marketing est essentiel.

Dans ce qui suit, on présente deux classes de modèles : l'une s'attachant à modéliser la moyenne conditionnelle des processus (les modèles $ARMA$) et l'autre s'intéressant à la modélisation de la variance conditionnelle (les modèles $ARCH - GARCH$).

5.2 Les modèles ARMA

5.2.1 Au commencement : le bruit blanc

Le fondement de l'ensemble des méthodes de séries temporelles est le *bruit blanc* (*white noise*), que l'on note dans ce qui suit ϵ_t. Dans un cas général, on a :

$$\epsilon_t \sim i.i.d.N(0, \sigma_\epsilon) \tag{5.4}$$

Cette formulation a trois implications :
- $\mathbb{E}[\epsilon_t] = \mathbb{E}[\epsilon_t|\epsilon_{t-1}, \epsilon_{t-2}, ...] = \mathbb{E}[\epsilon_t|$toute l'information disponible en date $t-1]$
- $\mathbb{E}[\epsilon_t\epsilon_{t-j}] = Cov[\epsilon_t\epsilon_{t-j}] = 0$
- $var[\epsilon_t] = var[\epsilon_t|\epsilon_{t-1}, \epsilon_{t-2}, ...] = var[\epsilon_t|$toute l'information disponible en date $t-1] = \sigma_\epsilon^2$

Les premières et deuxième propriétés supposent l'absence d'une corrélation sérielle ou d'une prédictibilité quelconques. La troisième propriété stipule l'homoscédasticité conditionnelle du processus bruit blanc. Ces propriétés seront peu à peu relachées au fur et à mesure des modèles évoqués ci-après.

En lui-même, ϵ_t est un processus plutôt ennuyeux : si ϵ_t atteint une valeur surprenament haute, ϵ_{t+1} ne sera pas nécessairement plus élevé. Il ne s'agit donc pas d'un processus persistant, alors que la plupart des séries chronologiques présentent ce type de persistance.

Notons néanmoins que la théorie financière repose néanmoins sur une hypothèse proche du bruit blanc. Le modèle de Black-Scholes repose sur l'hypothèse d'une diffusion suivant un brownien géométrique pour le prix des actifs. On rappelle que si S_t est le cours du sous-jacent dans le modèle de Black-Scholes, alors sa diffusion est de la forme :

$$dS_t = \mu S_t dt + \sigma S_t dW_t \tag{5.5}$$

où W_t est un mouvement brownien standard. En applicant la formule d'Îto, on parvient à déterminer une expression intégrale permettant d'obtenir la dynamique des prix :

$$dlog(S_t) = (\mu - \frac{\sigma^2}{2})dt + \sigma dW_t \tag{5.6}$$

On en déduit aisément :

$$\int_t^{t+1} dlog(S_s) = \int_t^{t+1} (\mu - \frac{\sigma^2}{2})ds + \int_t^{t+1} \sigma dW_s \tag{5.7}$$

$$\Leftrightarrow log\left(\frac{S_{t+1}}{S_t}\right) = \mu - \frac{\sigma^2}{2} + \sigma(W_{t+1} - W_t) \tag{5.8}$$

Sachant que $(W_{t+1} - W_t) \sim N(0,1)$ (l'intervalle de temps est l'unité), on retrouve donc bien un processus pour les rendements qui est à peu de choses pret un bruit blanc. La seule différence entre les deux processus tient au drift obtenu après application de la formule d'Îto. Malheureusement, le cours des actifs est rarement bruit blanc, ce qui ne va pas sans poser quelques problèmes lors de l'utilisation de la formule de Black et Scholes. Elle constitue néanmoins un benchmark intéressant pour l'évaluation des options.

La discrétisation de la diffusion a été accomplie ici rapidement et sans précautions : nous reviendrons plus loin sur la discrétisation des diffusions, en rappellant la méthode d'Euler. Cette maigre digression a uniquement pour but de mettre à jour quelques liens évidents entre l'analyse des séries temporelles ... et la finance. Tournons nous tout d'abord vers l'analyse des premiers modèles de séries temporelles : les processus ARMA.

5.2.2 Les modèles ARMA de base

La plupart du temps, on étudie une classe de modèles créés par combinaisons linéaires de bruits blancs. Ces modèles sont les suivants :

$$\text{AR(1)} : x_t = \phi x_{t-1} + \epsilon_t \tag{5.9}$$

$$\text{MA(1)} : x_t = \theta \epsilon_{t-1} + \epsilon_t \tag{5.10}$$

$$\text{AR(p)} : x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \epsilon_t \tag{5.11}$$

$$\text{MA(q)} : x_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} + \epsilon_t \tag{5.12}$$

$$\text{ARMA(p,q)} : x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} + \epsilon_t \tag{5.13}$$

Comme on peut le constater, il s'agit à chaque fois d'une recette à base de bruits blancs passés et de valeurs de départ pour x_t. L'ensemble de ces modèles ont une moyenne nulle, et sont utilisés pour représenter l'écart des séries à leur moyenne. Par exemple, si une série $\{x_t\}$ a une moyenne égale à $\bar{x}$ et suit un processus AR(1), alors :

$$(x_t - \bar{x}) = \phi(x_{t-1} - \bar{x}) + \epsilon_t \tag{5.14}$$

est équivalent à :

$$x_t = (1 - \phi)\bar{x} + \phi x_{t-1} + \epsilon_t \tag{5.15}$$

$$\Leftrightarrow x_t = c + \phi x_{t-1} + \epsilon_t \tag{5.16}$$

Ainsi, la constante c absorbe l'effet moyen. On travaille dans ce qui suit principalement à l'aide de modèle excluant la constante : elle est aisément estimable.

5.2.3 L'opérateur retard

Il est aisé de représenter et de manipuler les processus ARMA en utilisant l'opérateur retard L. Cet opérateur retarde les données d'une unité de temps :

$$Lx_t = x_{t-1} \tag{5.17}$$

De façon plus formelle, l'opérateur retard est un opérateur qui produit une nouvelle série de donnée (retardée) à partir d'une série $\{x_t\}$. A partir de cette définition sommaire, il est aisé de voir que :

$$L^2 x_t = LLx_t = Lx_{t-1} = x_{t-2} \tag{5.18}$$

On a donc :

$$L^j x_t = x_{t-j} \qquad\qquad L^{-j} x_t = x_{t+j} \tag{5.19}$$

Il est également possible de définir des polynomes de l'opérateur retard. On a alors :

$$a(L)x_t = (\alpha_0 L^0 + \alpha_1 L^1 + \alpha_2 L^2 + ... + \alpha_p L^p)x_t = \alpha_0 x_t + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + ... + \alpha_p x_{t-p} \tag{5.20}$$

En utilisant ces notations, il est alors possible de réécrire des modèles ARMA comme suit :

$$\text{AR(1)} : (1 - \phi L)x_t = \epsilon_t \tag{5.21}$$
$$\text{MA(1)} : x_t = (1 + \theta L)\epsilon_t \tag{5.22}$$
$$\text{AR(p)} : (1 - \phi_1 L^1 - \phi_2 L^2 - ... - \phi_p L^p)x_t = \epsilon_t \tag{5.23}$$
$$\text{MA(q)} : x_t = (1 + \theta_1 L^1 + \theta_2 L^2 + ... + \theta_p L^p)\epsilon_t \tag{5.24}$$
$$\text{ARMA(p,q)} : (1 - \phi_1 L^1 - \phi_2 L^2 - ... - \phi_p L^p)x_t = (1 + \theta_1 L^1 + \theta_2 L^2 + ... + \theta_p L^p)\epsilon_t \tag{5.25}$$

ou plus simplement :

$$\text{AR} : a(L)x_t = \epsilon_t \tag{5.26}$$
$$\text{MA} : x_t = b(L)\epsilon_t \tag{5.27}$$
$$\text{ARMA(p,q)} : a(L)x_t = b(L)\epsilon_t \tag{5.28}$$

5.2.4 Manipulation les processus ARMA avec L

Un modèle ARMA n'est pas unique. La loi jointe de $\{x_0, x_1, ..., x_n\}$ peut être modélisée par différents processus ARMA. Il est cependant important d'avoir toujours en tête :

- une représentation à l'aide d'un polynome retard le plus petit possible est toujours plus aisé ;
- les modèles AR sont les plus aisés à estimer par MCO ;
- les MA représentent x_t en fonction de variables indépendantes : dans de nombreux cas, ceci facilitera les calculs de variance et de covariance, comme on le verra plus loin.

5.2.5 AR(1) et MA(∞) par recursion

Il est possible dans de nombreux cas de fournir une représentation MA(∞) à partir d'un AR(1). Ceci se montre aisément comme suit :

$$x_t = \phi x_{t-1} + \epsilon_t \tag{5.29}$$
$$\Leftrightarrow x_t = \phi(\phi x_{t-2} + \epsilon_{t-1}) + \epsilon_t = \phi^2 x_{t-2} + \phi\epsilon_{t-1} + \epsilon_t \tag{5.30}$$
$$\Leftrightarrow x_t = \phi^k x_{t-k} + \phi^{k-1}\epsilon_{t-k+1} + ... + \phi^2\epsilon_{t-2} + \phi\epsilon_{t-1} + \epsilon_t \tag{5.31}$$

Ainsi, à la condition que $|\phi| < 1$, on peut écrire :

$$x_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} \tag{5.32}$$

Ainsi un AR(1) peut être exprimé comme MA(∞).

5.2.6 AR(1) et MA(∞) avec L

Les manipulations proposées plus haut sont plus aisées en utilisant le polynome retard :

$$(1 - \phi L)x_t = \epsilon_t \tag{5.33}$$
$$\Leftrightarrow x_t = (1 - \phi L)^{-1}\epsilon_t \tag{5.34}$$

Quel sens peut on donner à $(1-\phi L)^{-1}$? Une façon d'expliciter les choses est la suivante :

$$(1 - z)^{-1} = 1 + z + z^2 + z^3 + ..., \text{ pour } |z| < 1 \tag{5.35}$$

Ceci peut être prouvé en utilisant un développement de Taylor. Ce développement, en supposant que $|\phi| < 1$ implique $|\phi L| < 1$ suggère la chose suivante :

$$x_t = (1 - \phi L)^{-1}\epsilon_t = (1 + \phi L + \phi^2 L^2 + ...)\epsilon_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} \tag{5.36}$$

On retrouve donc le résultat précédent. On a supposé que $|\phi L| < 1$. Tous les processus ARMA n'ont pas de représentation inversible (on parle d'inversibilité des processus ou de processus inversible) de x_t en fonction du passé de ϵ_t.

Il est possible d'étendre les résultats, en montrant sous quelle condition un AR(p) peut admettre une représentation MA(∞). Les calculs sont cependant un peu plus longs et fastidieux. Un lecteur intéressé lira avec intérêt Cochrane (2005), page 14 et suivantes.

5.2.7 Résumé des manipulations possibles de l'opérateur retard

Les règles de calcul pour L sont les suivantes :
– $a(L)b(L) = (a_0 + a_1 L + ...)(b_0 + b_1 L + ...) = a_0 b_0 + (a_0 b_1 + b_0 a_1)L + ...$
– $a(L)b(L) = b(L)a(L)$
– $a(L)^2 = a(L)a(L)$
Il existe d'autres règles de calculs, développée dans Cochrane (2005), page 17. [Les ajouter un jour ?]

5.2.8 La fonction d'autocorrélation

5.2.8.1 Définitions

La fonction d'autocovariance d'une série x_t est définie par :

$$\gamma_j = cov(x_t, x_{t-j}) = \mathbb{E}[x_t - \mathbb{E}[x_t]][x_{t-j} - \mathbb{E}[x_{t-j}]] \tag{5.37}$$

Dans un cas où $\mathbb{E}[x_t] = 0, \forall t$, alors $\gamma_j = cov(x_t, x_{t-j}) = \mathbb{E}[x_t x_{t-j}]$. On a de plus $\gamma_0 = \mathbb{V}[x_t]$.
On en déduit aisément l'expression du coefficient de corrélation :

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \frac{\gamma_j}{\mathbb{V}[x_t]} \tag{5.38}$$

Remarque 1 (Autocorrélation et ARMA)**.** Les processus ARMA sont construits de façon à fournir une modélisation de la loi jointe de $\{x_1, ...x_n\}$. Les fonctions d'autocorrélation et d'autocovariance sont un moyen intéressant de caractériser cette loi jointe : la correlation entre x_t et x_{t-j} est un moyen intéressant (mais imparfait) de mesurer notamment la persistance d'une série. Si on observe une valeur importante pour x_{t-j}, on sera capable de dire si la valeur en x_t sera plus ou moins importante elle aussi.

5.2.8.2 ACF des modèles MA(q)

5.2.8.2.1 Bruit blanc Un bruit blanc $\epsilon_t \sim iidN(0, \sigma_\epsilon^2)$ a les caractéristiques suivantes :

$$\gamma_0 = \sigma_\epsilon^2 \tag{5.39}$$
$$\gamma_j = 0, \forall j \neq 0 \tag{5.40}$$
$$\rho_0 = 1 \tag{5.41}$$
$$\rho_j = 0, \forall j \neq 0 \tag{5.42}$$

5.2.8.2.2 MA(1) Le modèle s'écrit : $x_t = \theta\epsilon_{t-1} + \epsilon_t$. On a les propriétés suivantes :

$$\gamma_0 = \mathbb{V}[x_t] = \mathbb{V}[\epsilon_t + \theta\epsilon_{t-1}] = (1 + \theta^2)\sigma_\epsilon^2 \tag{5.43}$$
$$\gamma_1 = \mathbb{E}[x_t x_{t-1}] = \mathbb{E}[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})] = \theta\sigma_\epsilon^2 \tag{5.44}$$
$$\gamma_2 = \mathbb{E}[x_t x_{t-2}] = \mathbb{E}[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-2} + \theta\epsilon_{t-3})] = 0 \tag{5.45}$$
$$\gamma_j = 0, \forall j \geq 2 \tag{5.46}$$

L'autocorrélation se déduit des précédents calculs :

$$\rho_1 = \frac{\theta}{1 + \theta^2} \tag{5.47}$$
$$\rho_i = 0, \forall i > 1 \tag{5.48}$$

On observe ainsi que les autocorrélations d'un MA(1) s'annulent à partir de l'ordre 1. C'est ce qu'on observe sur les figure 5.2. Notons que la figure 5.1 présente l'allure d'un processus bruit blanc, MA(1), AR(1) et ARMA(1,1).

Le cas d'un MA(2) se traite aisément de la même façon. Le modèle s'écrit :

$$x_t = \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \epsilon_t \tag{5.49}$$

Comme précédement, les covariances se déterminent comme suit :

$$\gamma_0 = \mathbb{V}[x_t] = \mathbb{V}[\theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \epsilon_t] = (1 + \theta_1^2 + \theta_2^2)\sigma_\epsilon^2 \tag{5.50}$$

$$\gamma_1 = \mathbb{E}[x_t x_{t-1}] = \mathbb{E}[(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2})(\epsilon_{t-1} + \theta_1\epsilon_{t-2} + \theta_2\epsilon_{t-3})] = (\theta_1 + \theta_1\theta_2)\sigma_\epsilon^2 \tag{5.51}$$

$$\gamma_2 = \mathbb{E}[x_t x_{t-2}] = \mathbb{E}[(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2})(\epsilon_{t-2} + \theta_1\epsilon_{t-3} + \theta_2\epsilon_{t-4})] = \theta_2\sigma_\epsilon^2 \tag{5.52}$$

$$\gamma_i = 0, \forall i > 3 \tag{5.53}$$

On en déduit la fonction d'autocorrélation suivante :

$$\rho_0 = 1 \tag{5.54}$$

$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2} \tag{5.55}$$

$$\rho_2 = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} \tag{5.56}$$

$$\rho_i = 0, \forall i > 2 \tag{5.57}$$

On déduit aisément les autocorrélations d'un processus MA(q). Le modèle s'écrit :

$$x_t = \theta(L)\epsilon_t = \sum_{i=0}^{q}(\theta_i L^i)\epsilon_t \tag{5.58}$$

On a alors :

$$\gamma_0 = \mathbb{V}[x_t] = \mathbb{V}\left[\sum_{i=0}^{q}(\theta_i L^i)\epsilon_t\right] = \left(\sum_{i=0}^{q}\theta_j^2\right)\sigma_\epsilon^2 \tag{5.59}$$

$$\gamma_k = \mathbb{E}[x_t x_{t-k}] = \mathbb{E}\left[\sum_{i=0}^{q}(\theta_i L^i)\epsilon_t \sum_{i=0}^{q}(\theta_i L^i)\epsilon_{t-k}\right] = \sum_{i=0}^{q}\theta_i\theta_{i+k}\sigma_\epsilon^2, \forall k \leq q \tag{5.60}$$

$$\gamma_k = 0, \forall k > q \tag{5.61}$$

Remarque 2. Il y a une leçon importante à retenir de tout ceci : les calculs de covariance pour les processus MA est simple dans la mesure où les termes en $\mathbb{E}[\epsilon_j\epsilon_k]$ deviennent rapidement nuls quand j et k sont éloignés. Il n'en va pas de même des processus AR.

5.2.9 ACF des modèles AR(p)

Il existe deux façons de calculer l'ACF d'un processus AR(p). La première d'entre elles est de travailler sur la représentation MA(∞) d'un processus AR(p). En utilisant les formules obtenues plus haut, il est aisé de montrer qu'avec un modèle de la forme :

$$(1 - \phi L)x_t = \epsilon_t \Rightarrow x_t = (1 - \phi L)^{-1}\epsilon_t = \sum_{i=0}^{\infty}\phi^i\epsilon_{t-i} \tag{5.62}$$

On obtient des covariances de la forme :

$$\gamma_0 = \left(\sum_{i=0}^{\infty} \phi^{2i}\right) \sigma_\epsilon^2 = \frac{1}{1-\phi^2}\sigma_x^2; \rho_0 = 1 \tag{5.63}$$

$$\gamma_1 = \left(\sum_{i=0}^{\infty} \phi^i \phi^{i+1}\right) \sigma_\epsilon^2 = \phi \left(\sum_{i=0}^{\infty} \phi^i \phi^i\right) \sigma_\epsilon^2 = \frac{\phi}{1-\phi^2}\sigma_x^2; \rho_1 = \phi \tag{5.64}$$

En continuant ainsi, on trouve :

$$\gamma_k = \frac{\phi^k}{1-\phi^2}\sigma_\epsilon^2, \rho_k = \phi^k \tag{5.65}$$

L'autre façon de retrouver ces résultats est de travailler directement sur x_t, sans utiliser l'astuce de l'inversion. Pour le même modèle AR(1) que précédement, on a :

$$\gamma_1 = \mathbb{E}[x_t x_{t-1}] = \mathbb{E}[(\phi x_{t-1} + \epsilon_t)x_{t-1}] = \phi\sigma_x^2, \rho = \phi \tag{5.66}$$

$$\gamma_2 = \mathbb{E}[x_t x_{t-2}] = \mathbb{E}[(\phi^2 x_{t-2} + \phi\epsilon_{t-1} + \epsilon_t)x_{t-1}] = \phi^2\sigma_x^2, \rho = \phi^2 \tag{5.67}$$

$$\ldots \tag{5.68}$$

$$\gamma_k = \mathbb{E}[x_t x_{t-k}] = \mathbb{E}[(\phi^k x_{t-k} + \epsilon_t + \ldots)x_{t-k}] = \phi^k\sigma_x^2, \rho = \phi^k \tag{5.69}$$

$$\tag{5.70}$$

Ainsi l'ACF d'un modèle AR est général un mélange entre une sinusoide et une exponentielle, selon les signes de coefficient du modèle AR. Si les signes sont négatifs, les autocorrélations paires seront positives, et celles impaires seront négatives. Quoi qu'il arrive, $|\phi| < 1$, d'où $\lim_{k\to 0} \phi^k = 0$. Une décroissance relativement lente vers zéro se traduit par un convergence en forme d'exponentiele de l'ACF vers 0.

5.2.10 La fonction d'autocorrélation partielle

Définition 5.2.1 (Autocorrélation partielle)**.** *L'autocorrélation partielle d'ordre k désigne la corrélation entre x_t et x_{t-k} obtenue lorsque l'influence des variables $x_{t-k-i}, i < k$ a été retirée.*

Une définition plus formelle est la suivante :

Définition 5.2.2 (Définition plus formelle)**.** *L'autocorrélation partielle d'ordre k d'un processus $(x_t)_{t\in\mathbb{Z}}$, de moyenne m, notée p(k) est définie par le dernier coefficient de la projection linéaire de x_{t+1} sur ces k précédentes valeurs. $\forall k \in \mathbb{Z}$:*

$$x_{t+1} - m = c_1(x_t - m) + c_2(x_{t-1} - m) + ... + c_{k-1}(x_{t-k} - m) + p(k)(x_{t-k+1} - m) \tag{5.71}$$

ou de façon équivalente par :

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ p_k \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{k-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{k-2} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{k-1} & \gamma_{k-2} & \cdots & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{pmatrix} \tag{5.72}$$

Ceci tient donc pour $m = 0$. Dans un tel cas, $p(k) \in [-1; 1]$. On ajoute la propriété suivante :

Proposition 5.2.1. *De façon générale, la fonction d'autocorrélation partielle d'un processus $(x_t)_{t\in\mathbb{Z}}$ satisfait la relation :*

$$p(k) = \frac{|P_k^*|}{|P_k|} \tag{5.73}$$

avec

$$P_k = \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{k-1} \\ \rho_1 & 1 & \dots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \dots & \dots & 1 \end{pmatrix} \tag{5.74}$$

et

$$P_k = \begin{pmatrix} 1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & 1 & \dots & \rho_2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \dots & \dots & \rho_k \end{pmatrix} \tag{5.75}$$

L'idée est donc de caractériser la dépendance entre les différents retards en éliminant l'impact à chaque fois des retards intermédiaires. La solution la plus simple consiste a utilisée la définition avancée plus haut utilisant les moindres carrés de façon itérative, pour les différents ordres de l'autocorrélation partielle. On en déduit (intuitivement) que la PACF d'un AR(p) devrait être égale à 0 pour les autocorrélations partielles d'un ordre supérieur à p.

Sans en faire la preuve formelle, on retrouve (théoriquement) dans le cas d'un processus MA(q) un fonction d'autocorrélation partielle qui est un mélange d'une fonction exponentielle et d'une sinusoïde. Il s'agit donc d'un cas symétrique de l'ACF pour les AR(p).

5.2.11 Estimation et test des ACF et PACF

Il est essentiel de pouvoir disposer d'une ACF et d'une PACF estimées correctement. On fournit ici l'estimateur empirique ainsi que la région de confiance pour une hypothèse nulle de nullité de la corrélation et de la corrélation partielle.

5.2.11.1 Fonction d'Autocorrélation

L'estimation d'une fonction d'autocorrélation par la biais de son estimateur empirique, grace à la loi des grands nombres. Pour un processus non centré, de moyenne m, on a :

$$\mathbb{C}ov[x_t, x_{t-k}] = \mathbb{E}[(x_t - m)(x_{t-k} - m)] \tag{5.76}$$

Son estimateur est alors :

$$\widehat{\mathbb{C}ov}[x_t, x_{t-k}] = \frac{1}{N} \sum_{t=k+1}^{N} (x_t - \bar{x})(x_{t-k} - \bar{x}) \tag{5.77}$$

où $\bar{x}$ est la moyenne empirique. L'estimateur de l'autocorrélation ρ_k est alors :

$$\hat{\rho}_k = \frac{\widehat{\mathbb{C}ov}[x_t, x_{t-k}]}{\widehat{\mathbb{C}ov}[x_t, x_t]} \tag{5.78}$$

D'après le théorème central limite, la variable centrée t_{ρ_k} suit une loi normale centrée réduite :

$$t_{\rho_k} = \frac{\hat{\rho}_k - \rho_k}{\sqrt{\mathbb{V}[\hat{\rho}_k]}} \xrightarrow{\mathcal{L}} N(O, 1) \tag{5.79}$$

où $\mathbb{V}[\hat{\rho}_k]$ désigne la variance de l'estimateur. Elle est égale à :

$$\mathbb{V}[\hat{\rho}_k] = \frac{1}{N-k} \sum_{i=-K}^{K} \hat{\rho}_i^2, \text{ avec } K < k \tag{5.80}$$

Par symétrie de la fonction d'autocorrélation, on a :

$$\mathbb{V}[\hat{\rho}_k] = \frac{1}{N-k} \left(1 + 2 \sum_{i=1}^{K} \hat{\rho}_i^2 \right), \text{ avec } K < k \tag{5.81}$$

On en déduit aisément la région de confiance pour une hypothèse nulle $\rho_k = 0$:

$$IC = [\pm 1.96 \times \sqrt{\mathbb{V}[\hat{\rho}_k]}] \tag{5.82}$$

On fournit un exemple de code `R` permettant de calculer cette fonction d'autocorrélation ainsi que les bornes du test :

```
autocorrel<-function(x,lag,series){
# This function computes the ACF for any series of data.
#Dimension setting
n=nrow(x);
N=n-lag;
# Scaling of the data
x=scale(x)
#Creation of the matrix containing the lagged series
data=x[(lag+1):n,1];
for (i in 1:lag){
data=cbind(data,x[(lag-i+1):(n-i),1])
}
#Computation of the correlations from lag 1 to lag "lag"
correl=t(data[,1])%*%data[,1]
for (i in 1:lag){
correl=cbind(correl,t(data[,1])%*%data[,(i+1)])
}
```

```
correl=t(as.matrix(correl))
#Normalization to obtain the ACF
correl=correl/correl[1,1]
#Computation of the variance of the estimates
variance=matrix(1/N,nrow(correl),1)
for (i in 1:lag){variance[(i+1),1]=1/N*(1+2*sum(correl[1:i,1]^2))}
#Computation of the intervall around the 0 null hypothesis
intervalleup=1.96*sqrt(variance);
intervalledown=-1.96*sqrt(variance);
#Plotting the results : ACF and tests
#Definition of the plot windows
mindata=min(intervalleup,intervalledown,correl);
maxdata=max(intervalleup,intervalledown,correl);
#Plot
par(bg="lightyellow")
plot(correl,type="h",col="blue",ylim=c(mindata,maxdata),main="Autocorrelation
Function",xlab="Lag",ylab=series,bg="red")
lines(matrix(0,nrow(correl),1),col="black")
lines(intervalleup,type="l",col="red");
lines(intervalledown,type="l",col="red");
return(list(correl=correl, interval=cbind(intervalleup,intervalledown)))
}
```

5.2.11.2 Fonction d'autocorrélation partielle

On a suffisament dévelopé d'éléments relatifs à l'estimation des PACF. Pour déterminer la variance de l'estimateur, plusieurs stratégies sont possibles. Dans le cas d'un AR(p), les coefficients $\hat{p}(k), k > p$ sont distribués selon une loi normale de moyenne nulle et de variance :

$$\mathbb{V}[\hat{p}(k)] = \frac{1}{T}, \forall k > p \tag{5.83}$$

Une autre façon de traiter le problème consiste à déterminer de façon itérative les variances des estimateurs MCO permettant d'obtenir les coefficients de corrélation partielle. La méthode fonctionne quel que soit le modèle sous-jacent. C'est qui est proposé dans le code suivant :

```
pautocorrel<-function(x,lag,series){
# This function computes the PACF for any series of data.
#Dimension setting
n=nrow(x);
N=n-lag;
# Scaling of the data
x=scale(x)
#Creation of the matrix containing the lagged series
data=x[(lag+1):n,1];
for (i in 1:lag){
data=cbind(data,x[(lag-i+1):(n-i),1])
}
```

```
#Computation of the partial autocorrelations from lag 1 to lag "lag"
correl=matrix(0,lag,1)
var1=matrix(0,lag,1)
for (i in 1:lag){
X=data[,2:(i+1)]; Y=data[,1];
coeff=solve(t(X)%*%X)%*%(t(X)%*%Y)
res=sqrt(var(Y-X%*%coeff));
correl[i,1]=coeff[nrow(coeff),1];
var2=solve(t(X)%*%X);
var1[i,1]=res*sqrt(var2[nrow(var2),ncol(var2)])
}
correl=t(as.matrix(correl))
#Computation of the intervall around the 0 null hypothesis
intervalleup=1.96*(var1);
intervalledown=-1.96*(var1);
#Plotting the results : ACF and tests
#Definition of the plot windows
mindata=min(intervalleup,intervalledown,correl);
maxdata=max(intervalleup,intervalledown,correl);
#Plot
correl=t(correl)
par(bg="lightyellow")
plot(correl,type="h",col="blue",ylim=c(mindata,maxdata),main="Autocorrelation
Function",xlab="Lag",ylab=series,bg="red")
lines(matrix(0,nrow(correl),1),col="black")
lines(intervalleup,type="l",col="red");
lines(intervalledown,type="l",col="red");
return(list(correl=correl, interval=cbind(intervalleup,intervalledown)))
}
```

5.2.12 Stationnarité des processus et théorème de Wold

Avant de passer à l'estimation des processus ARMA, il est important de revenir sur certains aspects de ces processus. Il n'a été question jusqu'ici que d'une présentation relativement intuitive des processus ARMA, sans avoir souligné le fait que ces processus sont stationnaires par hypothèse. On distingue généralement deux formes de stationnarité : forte (stricte) ou faible. Soit x_t un processus temporel aléatoire :

Définition 5.2.3 (Stationnarité stricte)**.** *Le processus x_t est dit strictement stationnaire si quel que soit t_i et t_i+h la suite $\{x_{t_1}, ..., x_{t_n}\}$ a la même loi que $\{x_{t_1+h}, ..., x_{t_n+h}\}, \forall h$.*

Dans la pratique, on se limite généralement à supposer la stationnarité faible, qui se définit comme suit :

Définition 5.2.4 (Stationnarité faible)**.** *Le processus x_t est dit stationnaire au second ordre si les trois conditions suivantes sont satisfaites :*

$$- \forall t \in \mathbb{N}, \mathbb{E}[x_t^2] < \infty \tag{5.84}$$

$$- \forall t \in \mathbb{N}, \mathbb{E}[x_t] = m \tag{5.85}$$

$$- \forall (t,h) \in \mathbb{Z}^2, \mathbb{C}ov[(x_{t+h} - m)(x_t - m)] = \gamma_h, \text{ indépendant de } t \tag{5.86}$$

L'idée derrière tout ceci est de travailler sur des processus dont la moyenne et la variance sont constante au cours du temps. Notez que si la loi du processus est gaussienne, il y a équivalence entre les deux définitions de la stationnarité. Détaillons les conditions : la première de ces conditions suppose l'existence ou la convergence du moment d'ordre 2 (la variance pour un processus centré) ; les deux conditions suivantes supposent que l'espérance et la covariance du processus sont constantes au cours du temps. Il n'y a donc ni rupture dans la moyenne, ni rupture dans la structure de dépendance au cours du temps.

Dernier point dans cet errata théorique, le théorème de Wold est le théorème fondamental de l'analyse des séries temporelles stationnaires.

Théorème 5.2.1 (Théorème de Wold)**.** *Tout processus stationnaire d'ordre deux* (x_t) *peut être représenté sous la forme :*

$$x_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} + \kappa_t \tag{5.87}$$

où les paramètres ψ_j *satisfont* $\psi_0 = 1$, $\psi_j \in \mathbb{R} \forall j \in \mathbb{N}^*$ *et* $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ *et où* ϵ_t *est un bruit blanc i.i.d..*

On dit que la somme des chocs passés correspond à la composante linéaire stochastique de x_t. Le terme κ_t désigne la composante linéaire déterministe telle que $\mathbb{C}ov[\kappa_t, \epsilon_t] = 0, \forall j \in \mathbb{Z}$. Il s'agit par conséquent d'une formalisation de ce qui a été déjà été dit sur la transformation d'un processus AR en un MA(∞). On n'en fournit pas la preuve.

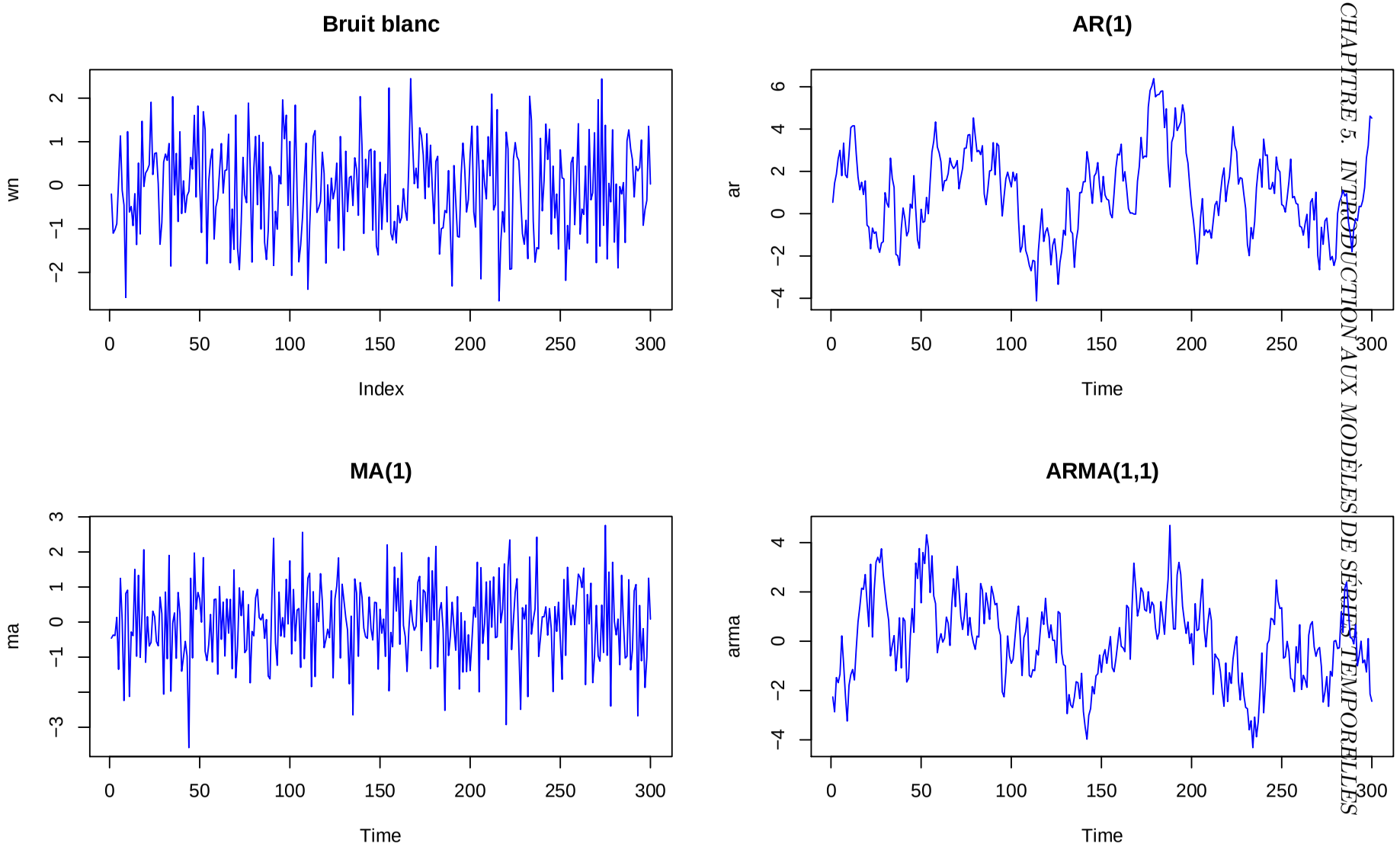

FIG. 5.1 – Chroniques de processus

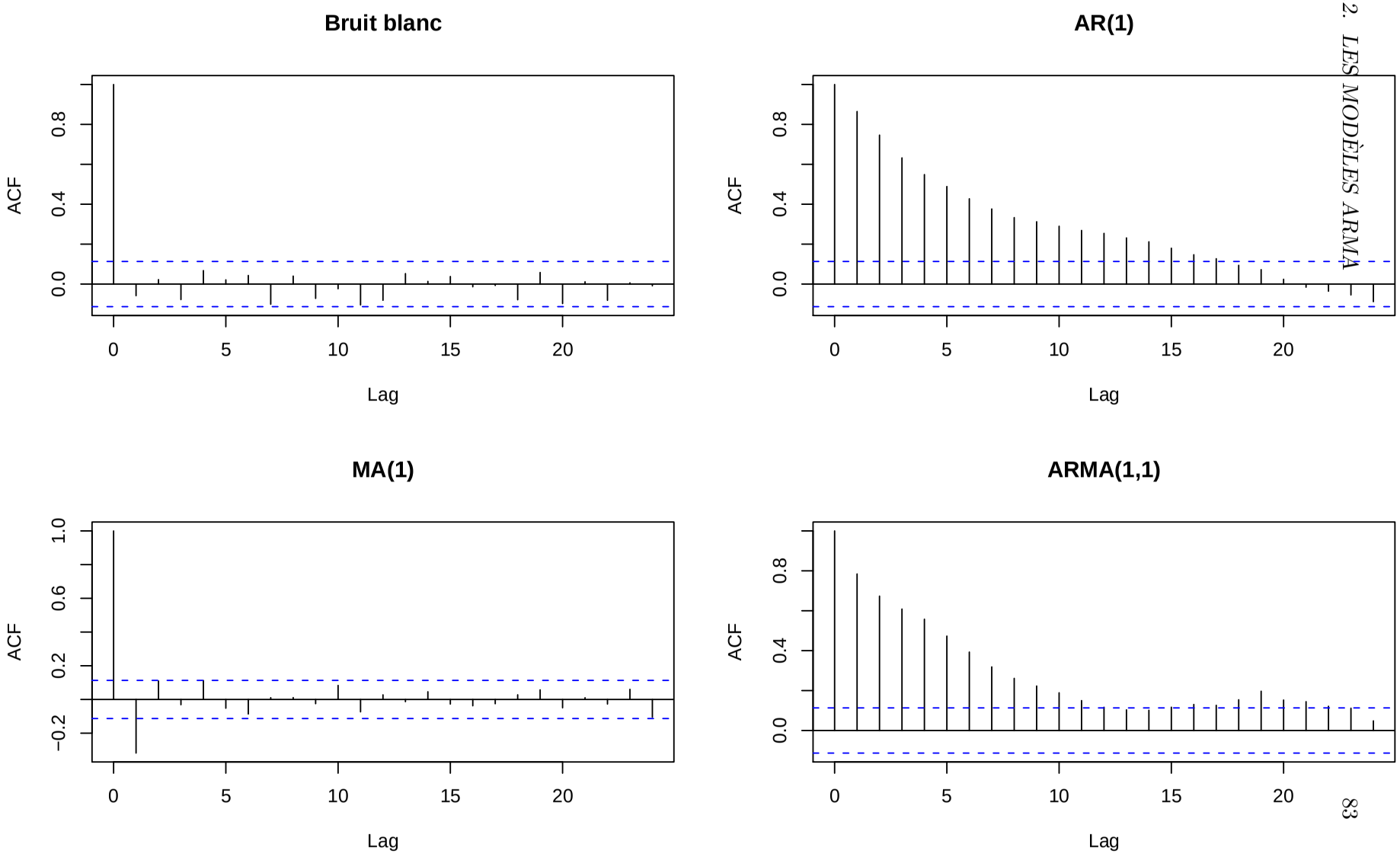

FIG. 5.2 – Chroniques de processus

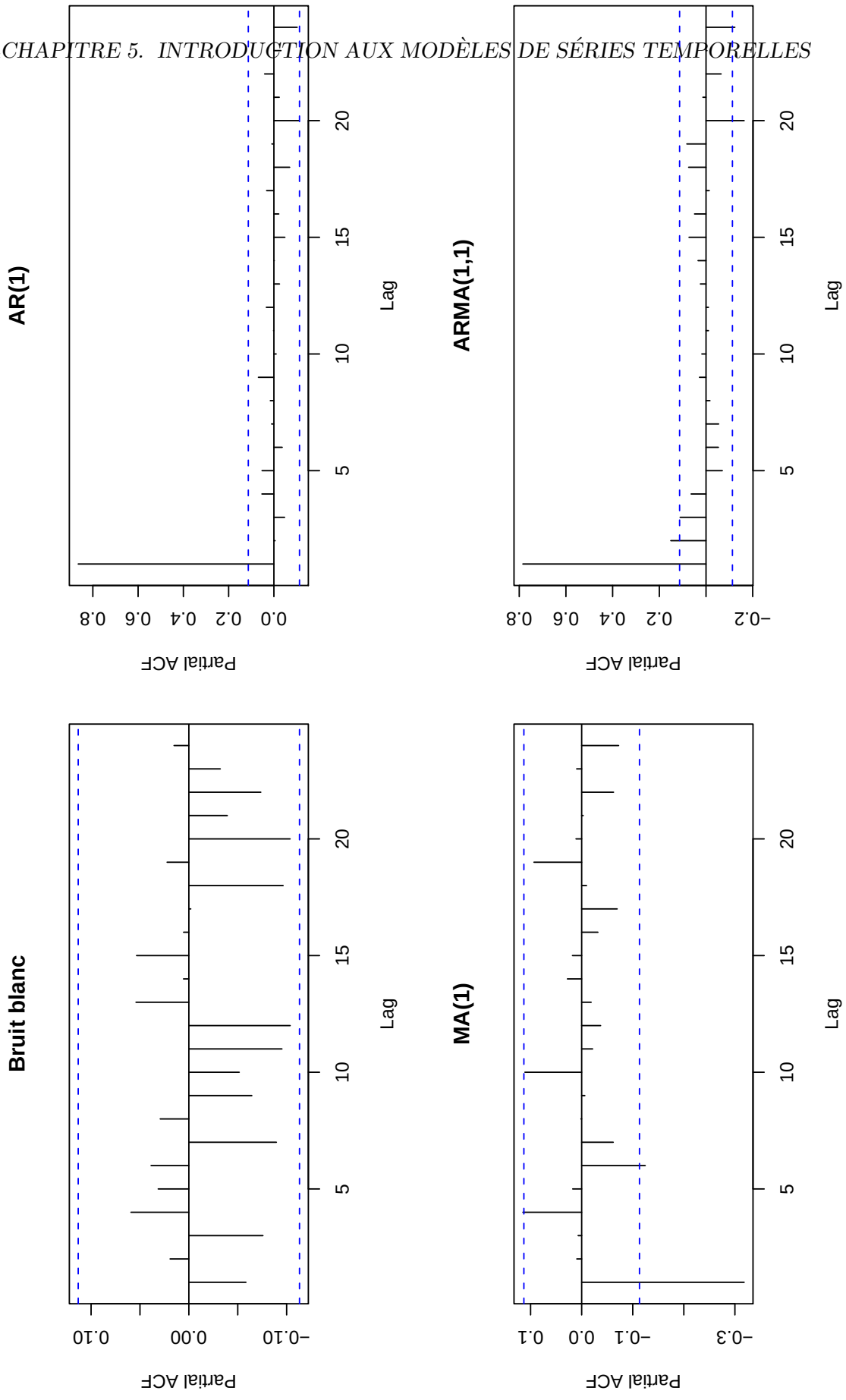

FIG. 5.3 – Chroniques de processus

5.2.13 Estimation des processus ARMA

Il existe différentes méthodes pour estimer les processus ARMA. Les processus AR s'estiment notamment par la méthode de Yule Walker, qui n'est pas présentée ici : la méthode devient très vite complexe lorsque l'on ajoute une composante MA dans le processus. On se bornera à présenter l'estimation par maximum de vraisemblance (conditionnelle) des processus AR,MA et enfin ARMA.

L'estimation par maximum de vraisemblance requiert un background théorique qui n'est pas présenté ici : la preuve de l'efficience des estimateurs du maximum de vraisemblance est notamment développé dans Harvey (1990)[chapitre 3]. Une autre référence bien connue sur la question est Hamilton (1994). Le lecteur soucieux de revenir sur le détail de ces preuves s'y reportera.

On présente rapidement les principes généraux de l'estimation par maximum de vraisemblance des modèles MA, AR et ARMA.

5.2.13.1 Estimation d'un AR(1)

L'estimation par maximum de vraisemblance dans le cas d'un AR(1) est légèrement plus complexe que dans le cas d'un modèle linéaire gaussien, tel qu'il en a été question dans le chapitre 2 de cet opus. Ceci tient au fait que le processus AR(1) non conditionnel n'est pas i.i.d. On rappelle qu'un modèle AR(1) est de la forme :

$$x_t = \phi x_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon) \tag{5.88}$$

Il est aisé de montrer que les moments conditionnels et non-conditionnels ne coincident pas, notamment lorsque le processus x_t n'est pas centré. Soit le modèle AR(1) suivant :

$$x_t = \mu + \phi x_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon) \tag{5.89}$$

Il suffit pour montrer cette divergence de calculer l'espérance conditionnelle et non conditionnelle et de faire de même pour la variance. Pour déterminer l'espérance non conditionnelle, il est utile de recourir à la représentation MA(∞) du processus AR(1). Ceci est possible si $|\phi| < 1$. On a alors :

$$\begin{aligned} x_t &= (1 - \phi L)^{-1}(\mu + \epsilon_t) && (5.90)\\ &= \mu + \epsilon_t + \phi(\mu + \epsilon_{t-1}) + \phi^2(\mu + \epsilon_{t-2}) + ... && (5.91)\\ &= \mu \sum_{i=0}^{\infty} \phi^i + \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i} && (5.92)\\ &= \frac{\mu}{1-\phi} + \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i} && (5.93) \end{aligned}$$

En prenant l'espérance de la précédente expression, on obtient :

$$\mathbb{E}[x_t] = \frac{\mu}{1-\phi} \tag{5.94}$$

Dans le cas où le processus est centré, on obtient :

$$\mathbb{E}[x_t] = 0 \tag{5.95}$$

ce qui n'est bien sur pas surprenant. L'espérance conditionnelle à x_{t-1} est différente de cette dernière expression :

$$\mathbb{E}[x_t|x_{t-1}] = \phi x_{t-1} \neq 0 \tag{5.96}$$

Ainsi conditionellement à x_{t-1}, l'espérance d'un AR(1) est différente de l'espérance non conditionnelle. Il est possible de dérouler les mêmes calculs pour la variance conditionnelle et non conditionnelle :

$$\mathbb{V}[x_t] = \mathbb{V}[\frac{\mu}{1-\phi} + \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}] \tag{5.97}$$

$$= \frac{\sigma_\epsilon^2}{1-\phi^2} \tag{5.98}$$

$$\mathbb{V}[x_t|x_{t-1}] = \mathbb{V}[\mu + \phi x_{t-1} + \epsilon_t] \tag{5.99}$$

$$= \sigma_\epsilon^2 \tag{5.100}$$

Là encore, les moments conditionnels et non conditionnels ne coincident pas. L'estimation de ce type de processus nécessite de travailler non plus sur la vraisemblance mais sur la vraisemblance non-conditionnelle. Harvey (1990) en rappelle le principe général : dans le cas de la vraisemblance non conditionnelle avec observations i.i.d., il est possible d'écrire la loi jointe du processus comme le produit des lois pour chacune des observations, du fait de la propriété d'indépendance des observations. Ici, on travaille en relachant cette hypothèse. On utilise alors le fait que conditionnellement à l'observation du passé, les observations sont i.i.d.. Pour se faire, on applique la règle de Bayes rappelée dans le chapitre 1, de façon à reproduire la décomposition proposée dans le cas de la vraisemblance. Dans le cas où l'on dispose de trois observations $\{x_1, x_2, x_3\}$ il est alors possible d'écrire :

$$f(x_1, x_2, x_3) = f(x_1)f(x_2|x_1)f(x_3|x_2, x_1) \tag{5.101}$$

On en déduit alors la logvraisemblance :

$$lnL = ln(f(x_1)) + ln(f(x_2|x_1)) + ln(f(x_3|x_2, x_1)) \tag{5.102}$$

Il est alors possible d'estimer les paramètres en utilisant les méthodes proposées au chapitre 4. On présente la méthode dans le cadre d'un modèle AR(1). Si $\epsilon_t \sim N(0, \sigma^2)$, alors, en utilisant les calculs précédents des moments conditionnels, le processus $x_t = x_{t-1} + \epsilon_t$ suit une loi normale d'espérance conditionnelle ϕx_{t-1} et de variance σ_ϵ^2. Sa log-vraisemblance s'écrit alors comme suit :

$$lnL(x, \phi, \sigma_\epsilon^2) = -\frac{n-1}{2} ln(2\pi) - (n-1) ln(\sigma) - \frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{n} (x_t - \phi x_{t-1})^2 + ln(f(x_1)) \tag{5.103}$$

Dans la plupart des cas, on néglige le terme $ln(f(x_1))$: dans le cas d'échantillons importants, son influence est minime. Ceci peut se réécrire sous forme concentrée et matricielle :

$$lnL(x, \phi, \sigma_\epsilon^2) = -(n-1)ln(\sigma_\epsilon) - \frac{1}{2\sigma_\epsilon^2}(X_t - \phi X_{t-1})'(X_t - \phi X_{t-1}) \tag{5.104}$$

où X_t et X_{t-1} sont des matrices $\mathcal{M}(n-1 \times 1)$ contenant les observations du processus $(x_t)_{t\in\mathbb{Z}}$. Les équations normales sont alors :

$$\frac{\partial lnL}{\partial \phi} = \frac{1}{\sigma_\epsilon^2} X'_{t-1}(X_t - \phi X_{t-1}) = 0 \tag{5.105}$$

$$\frac{\partial lnL}{\partial \sigma_\epsilon} = -\frac{n-1}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon^3}(X_t - \phi X_{t-1})'(X_t - \phi X_{t-1}) = 0 \tag{5.106}$$

$$\tag{5.107}$$

Ce système admet la solution suivante :

$$\theta = (X'_{t-1}X_{t-1})^{-1}X'_{t-1}X_t \tag{5.108}$$

$$\sigma_\epsilon = \sqrt{\frac{1}{n-1}\epsilon'_t\epsilon_t} \tag{5.109}$$

On retrouve donc exactement la même solution que dans le cas des MCO. Tout ce qui a été dit précédement sur ces estimateurs est donc valable : on n'y reviendra pas.

5.2.13.2 Estimation d'un AR(p)

La généralisation au cas d'un AR(p) se fait aisément une fois la précédente étape en tête. Soit le processus AR(p) suivant :

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t \tag{5.110}$$

utilisant les mêmes conditions que celles évoquées précédement. Conditionnellement à l'information passée, la loi de ce processus est la suivante :

$$x_t | x_{t-1}, x_{t-2}, ..., x_{t-p} \sim N(\sum_{i=1}^{p} \phi_i x_{t-i}, \sigma_\epsilon^2) \tag{5.111}$$

Tout ceci se réécrit naturellement sous forme matricielle :

$$x_t = \underline{X_{t-1:t-p}}\Phi' + \epsilon_t \tag{5.112}$$

$$x_t | \underline{X_{t-1:t-p}} \sim N(\underline{X_{t-1:t-p}}\Phi', \sigma_\epsilon^2) \tag{5.113}$$

où Φ est la matrice $\mathcal{M}(1 \times p)$ des p coefficients à estimer et $\underline{X_{t-1:t-p}}$ est la matrice $\mathcal{M}(n-p \times p)$ des séries de données retardées.
Il est alors aisé de déterminer la log-vraisemblance conditionnelle et de retrouver les estimateurs du maximum de vraisemblance, qui, là encore, coincident avec ceux des MCO :

$$lnL(x, \Phi, \sigma_\epsilon^2) = -(n-1)ln(\sigma_\epsilon) - \frac{1}{2\sigma_\epsilon^2}(X_t - \underline{X_{t-1:t-p}}\Phi')'(X_t - \underline{X_{t-1:t-p}}\Phi') \tag{5.114}$$

Les équations normales sont alors :

$$\frac{\partial lnL}{\partial \phi} = \frac{1}{\sigma_\epsilon^2}\underline{X_{t-1:t-p}}'(X_t - \underline{X_{t-1:t-p}}\Phi') = 0 \tag{5.115}$$

$$\frac{\partial lnL}{\partial \sigma_\epsilon} = -\frac{n-1}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon^3}(X_t - \underline{X_{t-1:t-p}}\Phi')'(X_t - \underline{X_{t-1:t-p}}\Phi') = 0 \tag{5.116}$$

$$\tag{5.117}$$

Les estimateurs sont alors :

$$\theta = (\underline{X_{t-1:t-p}}'\underline{X_{t-1:t-p}})^{-1}\underline{X_{t-1:t-p}}'X_t \tag{5.118}$$

$$\sigma_\epsilon = \sqrt{\frac{1}{n-1}\epsilon_t'\epsilon_t} \tag{5.119}$$

On retrouve donc bien les estimateurs MCO du modèle. Les processus AR sont ainsi très simple à estimer : il "suffit" de regresser x_t sur son passé par MCO pour obtenir des estimateurs sans biais et efficaces. L'estimation des MA est bien plus complexe comme nous allons le voir.

5.2.13.3 Estimation d'un MA(1)

La difficulté de l'estimation d'un MA(1) tient au fait que l'on observe pas directement les résidus passés : il n'est pas possible de regresser x_t sur ϵ_{t-1} car ce dernier n'est pas directement observé. Il est alors nécessaire de procéder de façon itérative pour parvenir à écrire la vraisemblance. Soit le processus MA(1) suivant :

$$x_t = \theta x_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{5.120}$$

On commence comme précédement par constater que les moments d'ordre 1 et 2 conditionnels et non conditionnels ne sont pas les mêmes :

$$\mathbb{E}[x_t|\epsilon_{t-1}] = \theta\epsilon_{t-1} \tag{5.121}$$

$$\mathbb{E}[x_t] = 0 \tag{5.122}$$

$$\mathbb{V}[x_t|\epsilon_{t-1}] = \sigma_\epsilon^2 \tag{5.123}$$

$$\mathbb{V}[x_t] = \sigma_\epsilon^2(1+\theta^2) \tag{5.124}$$

Là encore les moments conditionnels et non conditionnels ne coincident pas. On raisonnera là encore en terme de vraisemblance conditionnelle. On sait que :

$$x_t|\epsilon_{t-1} \sim N(\theta\epsilon_{t-1}, \sigma_\epsilon^2) \tag{5.125}$$

La vraisemblance conditionnelle s'écrit donc :

$$lnL(x, \theta, \sigma_\epsilon^2) = -\frac{n-1}{2}ln(2\pi) - (n-1)ln(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\sum_{t=2}^{n}(x_t - \theta\epsilon_{t-1})^2 + ln(f(x_1)) \tag{5.126}$$

La encore, on ne tient pas compte de $ln(f(x_1))$. Le problème est alors le suivant : il n'est pas possible de calculer la vraisemblance directement. Il est nécessaire pour un $theta^*$ donné de calculer les résidus de façon récursive. L'algorithme serait par exemple :

1. On connait x_2. On est alors en mesure de calculer $\epsilon_2 = x_1 - \theta\epsilon_1$. Il suffit de donner une valeur de départ à ϵ_1 pour pouvoir débuter l'algorithme. La solution généralement considérée est de choisir $\epsilon_1 = \mathbb{E}[\epsilon_t] = 0 \Leftrightarrow \hat{\epsilon}_2 = x_2$.
2. Une fois que l'on connait $epsilon_2$, il est possible de calculer $\epsilon_3 = x_3 - \theta\epsilon_2$...
3. Et ainsi de suite : on parvient à déterminer l'ensemble des valeurs de ϵ_t pour θ fixé ex ante.

On comprend bien qu'il n'est pas possible d'appliquer une méthode standard (i.e. MCO) d'estimation. Il est nécessaire d'avoir recours aux méthodes présentées au chapitre 4. On présente à titre d'exemple un code permettant de simuler un processus MA(1), ainsi qu'un code permettant d'estimer les paramètres de ce modèle.

```
simul.ma<-function(n,theta){
epsilon=as.matrix(rnorm(n));
x=matrix(0,n,1);
for (i in 2:n){
x[i,1]=theta*epsilon[(i-1),1]+epsilon[i,1]
}
x[1,1]=epsilon[1,1]
return(list(x=x))
}
```

```
estim.ma<-function(theta,x){
G=matrix(1,2,1)
n=nrow(x)
check=theta
i=1;
epsilon=matrix(0,n,1);
while(sum(G^2)>0.0001){
for (i in 2:n){epsilon[i,1]=x[i,1]-theta[1,1]*epsilon[i-1,1]}
BHHH=cbind((1/theta[2,1])*epsilon[1:(n-1),1]*(x[2:n,1]-theta[1,1]*epsilon[1:(n-1),1])
,-1/theta[2,1]+(1/theta[2,1]^3)*(x[2:n,1]-theta[1,1]*epsilon[1:(n-1),1])^2);
H=(t(BHHH)%*%BHHH);
G[1,1]=sum(BHHH[,1]);
G[2,1]=sum(BHHH[,2]);
cat(theta,"\n");
check=cbind(check,theta+solve(H)%*%G);
theta=theta+solve(H)%*%G;
i=i+1
}
plot(check[1,],type="l",col="blue",ylim=c(min(check),max(check)))
lines(check[2,],col="red")
return(list(theta=theta,check=check))
}
```

L'utilisation de ce code permet d'obtenir des estimations de processus MA(1). Le graphique 5.2.13.3 présente la chronique d'un processus MA(1) simulé. Le graphe 5.2.13.3 représente les ACF et PACF de ce processus, en utilisant le code fournit plus haut.

Enfin, la figure 5.2.13.3 fournit la trajectoire pour différents points de départ des paramètres estimés.

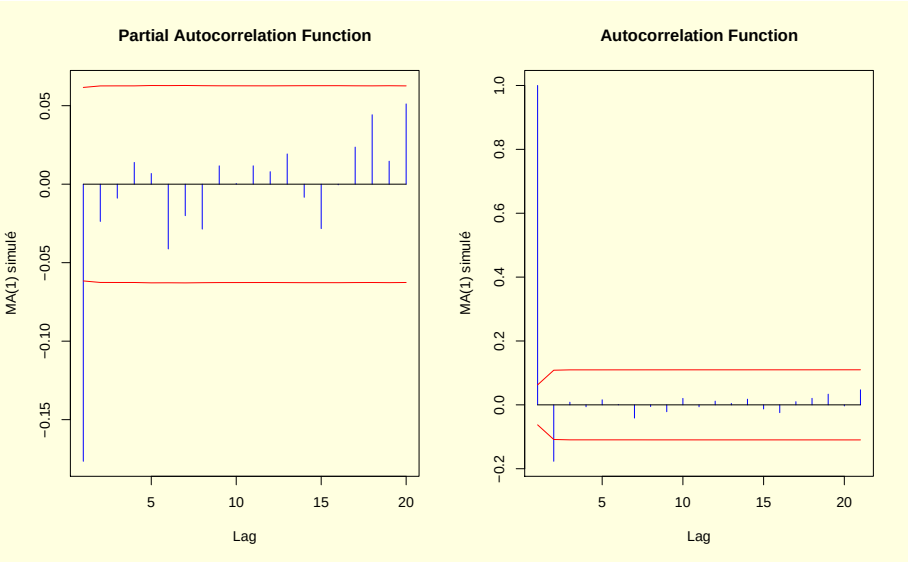

FIG. 5.4 – ACF et PACF du processus MA(1) simulé

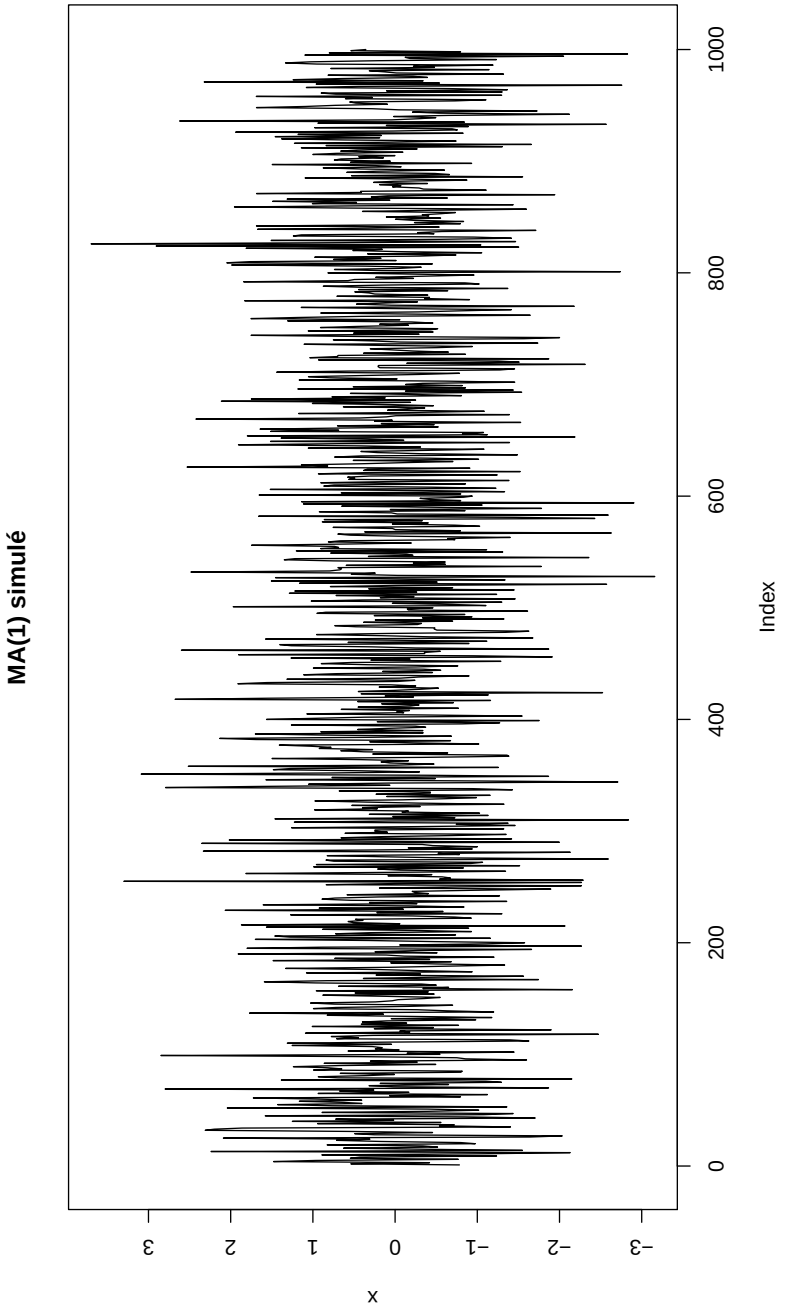

FIG. 5.5 – Trajectoire d'un processus MA(1)

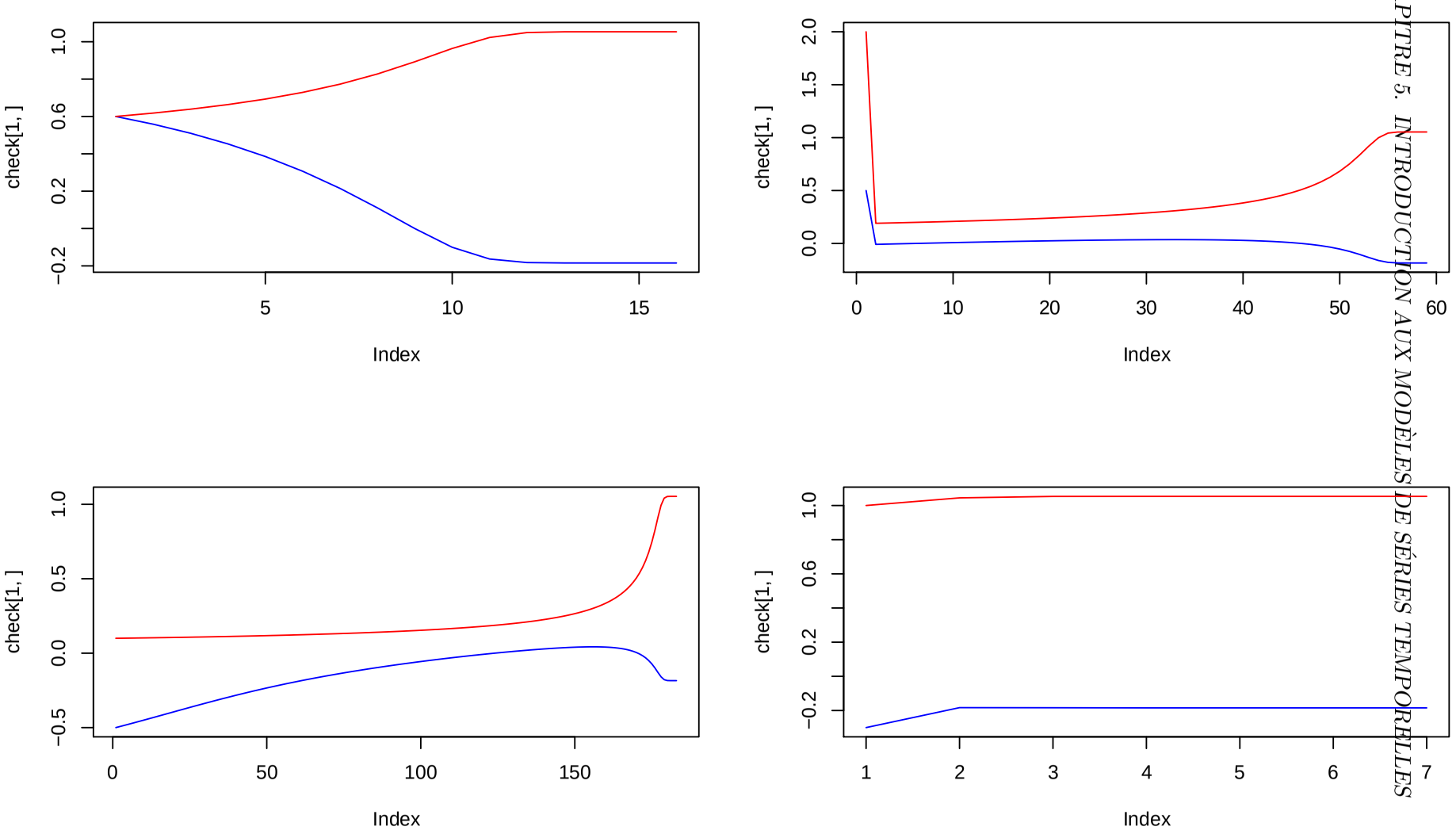

FIG. 5.6 – trajectoire des estimateurs pour différentes valeurs de départ

5.2.13.4 Estimation d'un MA(q)

On généralise ce qui vient d'être dit dans le cas d'un MA(1) au cas d'un MA(q), comme on l'ai fait pour les AR(p). Le modèle s'écrit :

$$x_t = \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{5.127}$$

Là encore les ϵ_t ne sont pas observables : il est nécessaire de procéder à leur estimation à $\Theta = \{\theta_1, ..., \theta_p\}$ fixé. Il est alors possible d'écrire la vraisemblance associée à Θ ainsi qu'aux observations. On détermine les ϵ_t comme suit :

1. On suppose que les ϵ_i pour $i = \{1, ..., q\}$ sont nuls (espérance du processus bruit blanc).
2. On détermine alors $\epsilon_{q+1} = x_{q+1}$.
3. On peut alors déterminer $\epsilon_{q+2} = x_{q+2} - \theta_1 \epsilon_{q+1}$.
4. Puis $\epsilon_{q+3} = x_{q+3} - \theta_1 \epsilon_{q+2} + \theta_2 \epsilon_{q+1}$.
5. On poursuit ainsi jusqu'à $\epsilon_{2q+1} = x_{2q+1} - \sum_{i=1}^{q} \theta_i \epsilon 2q - i$.
6. On poursuit ensuit l'algorithme jusqu'à ce qu'on l'on obtenue l'intégralité de la chronique des ϵ_t en utilisant la formule : $\epsilon_t = x_t - \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$.

Une fois ceci fait, il est aisé de calculer la vraisemblance puis de la maximiser en utilisant les méthodes proposées au chapitre 4. La loi conditionnelle de x_t est :

$$x_t | \epsilon_{t-1}, ..., \epsilon_{t-q} \sim N(\sum_{i=1}^{q} \theta_i \epsilon_{t-i}, \sigma_\epsilon^2) \tag{5.128}$$

On en déduit la vraisemblance du modèle :

$$L(\Theta, x) = \prod_{i=q+1}^{n} \frac{1}{\sqrt{2\pi}\sigma_\epsilon} exp\{-\frac{1}{2} \frac{(x_t - \sum_{i=1}^{q} \theta_i \epsilon_{t-i})^2}{\sigma_\epsilon^2}\} \tag{5.129}$$

La log-vraisemblance est alors :

$$lnL(\Theta, x) = -\frac{n-q}{2} ln(2\pi) - (n-q) ln(\sigma) - \frac{1}{2} \sum_{i=q+1}^{n} \left(\frac{(x_t - \sum_{i=1}^{q} \theta_i \epsilon_{t-i})^2}{\sigma_\epsilon^2} \right) \tag{5.130}$$

On en déduit alors les équations normales :

$$\frac{\partial lnL}{\partial \theta_k} = \sum_{i=q+1}^{n} \frac{\epsilon_{t-k}(x_t - \sum_{i=1}^{q} \theta_i \epsilon_{t-i})}{\sigma_\epsilon^2} \tag{5.131}$$

$$\frac{\partial lnL}{\partial \sigma} = -\frac{n-q}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon^3} \sum_{i=q+1}^{n} \left(x_t - \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \right)^2 \tag{5.132}$$

Là encore, pour maximiser la vraisemblance, il suffit de mettre en oeuvre l'une des méthodes présentées au cours du chapitre 4. On préfèrera naturellement utiliser des méthodes de type scoring qui ont le bon gout de converger à tous les coups (en théorie). L'utilisation de la matrice BHHH simplifie considérablement le calcul de la matrice d'information de Fisher.

Remarque 3 (Espérance et variances conditionnelles)**.** On remarque qu'au cours des calculs des espérances et des variances conditionnelles qui ont été menés on observe que pour un processus AR ou MA, l'espérance conditionnelle varie au cours du temps, alors que la variance conditionnelle, elle, ne change pas. C'est en relachant cette dernière hypothèse que l'on passera au modèles GARCH dans la section suivante. Quoi qu'il en soit, les espérances et variances non conditionnelles sont toujours constantes au cours du temps : les processus MA et AR sont des processus stationnaires au second ordre. Ceci n'est donc pas surprenant.

5.2.13.5 Estimation d'un ARMA(p,q)

On termine cette section consacrée à l'estimation des ARMA par une méthode d'estimation pour les modèles ARMA(p,q). On présente les moments conditionnels et non conditionnels, la loi conditionnelle ainsi que la vraisemblance d'un tel modèle, accompagnée de ses équations normales.

Un modèle ARMA(p,q) s'écrit comme suit :

$$x_t = \sum_{i=1}^{p} x_{t-i} + \sum_{i=1}^{q} \epsilon_{t-i} + \epsilon_t \tag{5.133}$$

Il est aisé de retrouver l'ensemble des moments :

$$\mathbb{E}[x_t|x_{t-1}, ..., x_{t-p}, \epsilon_{t-1}, ..., \epsilon_{t-q}] = \sum_{i=1}^{p} x_{t-i} + \sum_{i=1}^{q} \epsilon_{t-i} \tag{5.134}$$

$$\mathbb{E}[x_t] = 0, \text{ il existe plusieurs façons de le montrer !} \tag{5.135}$$

$$\mathbb{V}[x_t|x_{t-1}, ..., x_{t-p}, \epsilon_{t-1}, ..., \epsilon_{t-q}] = \sigma_\epsilon^2 \tag{5.136}$$

$$\mathbb{V}[x_t] = \frac{\sigma_\epsilon^2 \sum_{i=1}^{q} \theta_i^2}{1 - \sum_{i=1}^{p} \phi_i^2} \tag{5.137}$$

On en déduit la loi de x_t conditionnellement au passé des observations :

$$x_t|x_{t-1}, ..., x_{t-p}, \epsilon_{t-1}, ..., \epsilon_{t-q} \sim N(\sum_{i=1}^{p} x_{t-i} + \sum_{i=1}^{q} \epsilon_{t-i}, \sigma_\epsilon^2) \tag{5.138}$$

Là encore, la difficulté tient au calcul des ϵ_{t-i} liés à l'introduction d'une composante MA(q). L'algorithme peut être de la forme :

1. On fixe $\epsilon_t = 0, t \in (1, max(p, q))$.
2. Puis, à partir de $t = max(p, q) + 1$ et à $\{\theta_1, ..., \theta_q, \phi_1, ..., \phi_p\}$ fixés, on détermine ϵ_t de la façon suivante :

$$\epsilon_t = x_t - \sum_{i=1}^{p} x_{t-i} + \sum_{i=1}^{q} \epsilon_{t-i} \tag{5.139}$$

3. on répète la procédure jusqu'à l'obtention de l'ensemble de la chronique des ϵ_t.

La log vraisemblance concentrée du processus est alors :

$$lnL = -(n - max(p,q))ln(\sigma_\epsilon) - \frac{1}{2}\sum_{t=max(p,q)+1}^{n} \frac{(x_t - \sum_{i=1}^{p}\phi_i x_{t-i} + \sum_{i=1}^{q}\theta_i\epsilon_{t-i})^2}{\sigma_\epsilon^2} \tag{5.140}$$

Les équations normales sont alors :

$$\frac{\partial lnL}{\partial \theta_k} = \sum_{i=max(p,q)+1}^{n} \frac{\epsilon_{t-k}(x_t - \sum_{i=1}^{p}\phi_i x_{t-i} - \sum_{i=1}^{q}\theta_i\epsilon_{t-i})}{\sigma_\epsilon^2} \tag{5.141}$$

$$\frac{\partial lnL}{\partial \sigma} = -\frac{n - max(p,q)}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon^3}\sum_{i=max(p,q)+1}^{n}\left(x_t - \sum_{i=1}^{p}\phi_i x_{t-i} - \sum_{i=1}^{q}\theta_i\epsilon_{t-i}\right)^2 \tag{5.142}$$

Il suffit alors d'appliquer les méthodes présentées au chapitre 4 pour obtenir une estimation des paramètres. Le code à developper est bien entendu plus complexe que ce qui a été développé jusqu'ici. Les concepteurs de `R` proposent par défaut une fonction permettant d'estimer les paramètres d'un modèle ARMA. Il s'agit de la fonction `arima` qui utilise la synthaxe suivante[1] :

```
arima(x, order = c(0, 0, 0),
           seasonal = list(order = c(0, 0, 0), period = NA),
           xreg = NULL, include.mean = TRUE, transform.pars = TRUE,
           fixed = NULL, init = NULL, method = c("CSS-ML", "ML", "CSS"),
           n.cond, optim.control = list(), kappa = 1e6)
```

[Ajouter un code généraliste qui permet d'estimer un ARMA(p,q), un jour... du courage...]

5.2.14 Critères de sélection de l'ordre des processus ARMA

Une fois les estimations conduites, deux questions restent à traiter : 1. les estimations violent-elles les hypothèses du modèle ? 2. quel ordre retenir pour p et q ? Il est à noter que la méthode de Box et Jenkins visant à retenir p et q au vu des ACF et PACF n'est plus utilisée. Au mieux, il s'agit d'un guide dans la sélection des p et q maximaux à tester.

5.2.14.1 Tests sur les résidus

Le premier élément à vérifier est que les résidus ne violent pas les hypothèses du modèle. De façon générale, on vérifie la valeur moyenne des résidus, leur autocorrélation, la stabilité de la variance ainsi que leur normalité. Il est à noter que si les résidus ne sont pas normaux, mais que l'estimation a été conduite en utilisant une hypothèse de normalité des résidus, les estimations des paramètres du modèle reste bonne. Il s'agit en fait d'une estimation par pseudo-maximum de vraisemblance, introduite pas Gourieroux, Monfort et Trognon dans deux articles fameux d'*Econometrica*. L'estimation par BHHH

[1] Il est toujours possible d'obtenir de l'aide sur une fonction, en tapant `help(nom de la fontion)"`

de la matrice d'information de Fisher est particulièrement appropriée à ce cas. Cette approche sera très utile pour l'estimation des processus GARCH.

Test de nullité des résidus

Lorsque le processus est bien estimé, les résidus entre valeurs estimées et réelles par le modèle doivent se comporter comme un bruit blanc. L'une des hypothèses de bruit blanc est que l'espérance des résidus (et la moyenne empirique par conséquent) est nulle. Si le processus (ϵ_t) est i.i.d., on doit alors :

$$\bar{\epsilon_t} = \frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon_t} \to 0 \tag{5.143}$$

On se réfère au chapitre 1 où un test de nullité de la moyenne est développé. La statistique de test est alors (sous $H_0 : \bar{\epsilon_t} = 0$) :

$$\frac{\bar{\epsilon_t}}{\hat{\sigma_\epsilon}/\sqrt{n}} \sim T_{n-1} \tag{5.144}$$

La loi de Student convergent rapidement vers une loi normale centrée réduite. Ainsi, il est possible de calculer la statistique de test et de la comparer à 2 pour obtenir un test à 95% de la nullité de la moyenne. Il est également possible de construire un intervalle de confiance autour de la moyenne en utilisant la loi de cette statistique :

$$IC_{95\%} = [\bar{\epsilon_t} \pm t_{n-1}\frac{\hat{\sigma_{\epsilon_t}}}{\sqrt{n}}] \tag{5.145}$$

Test d'autocorrélation des résidus

Les processus ARMA peuvent être vu comme une façon de stationnariser les séries, abus de langage pour désigner le fait que l'on se débarrasse de la corrélation existant dans les séries. Il est donc nécessaire de tester l'existence d'autocorrélation dans les résidus. Si ce test conduit au constat que les résidus sont corrélés, c'est certainement que le modèle est mal spécifié. Il existe différents tests d'autocorrélation :

- Test de Durbin et Watson : ce test a déjà été brievement présenté au chapitre 2. On n'y revient pas ici.
- Etude des ACF et PACF : si le modèle est bien spécifié, l'ensemble des autocorrélations simples et partielles doivent être nulles.
- Cette dernière étude est complétée par l'étude de la statistique dite du "porte-manteau". Ce test repose sur l'idée que la FAC d'un bruit blanc ne doit pas réveler d'autocorrélations non nulles. En pratique ce test présente deux variantes :

  - *Test de Box et Pierce* : l'idée de ce test est simple. Il est basé sur une somme des carrés de autocorrélations pour un horizon allant de 1 à K. La statistique est la suivante :

$$Q_{BP} = n\sum_{k=1}^{K}\rho_k^2 \to \chi^2_{K-p-q} \tag{5.146}$$

L'hypothèse H_0 est ici : $\rho_1 = ... = \rho_K = 0$ contre $H_1 : \exists j \in [1, K], \rho_j \neq 0$. Il suffit de comparer la valeur de la statistique de test au quantile de la loi du χ^2.

– *Test de Ljung-Box* : l'hypothèse nulle du test est $H_0 : \rho_j = 0, \forall j < K$. On construit la statistique de test suivante :

$$Q_K = n(n-2)\sum_{k=1}^{K}\frac{\rho_k^2}{n-k} \rightarrow \chi^2_{K-p-q} \tag{5.147}$$

Là encore, il suffit de déterminer la valeur de la statistique et de la comparer au quantile d'une loi du χ^2.

5.2.15 Tests sur les résidus ARMA

Dans la mesure où l'estimation suppose que les résidus sont gaussiens, il est souvent conseillé de mener un certain nombre de tests afin de vérifier leur gaussiannité dans les faits. On utilise en général des tests simples (tests d'adéquation ou test de Kolmogorov) afin de s'en assurer.

On rappelle ici le test de Jarque et Berra explicité plus haut. Celui-ci repose sur deux statistiques connues pour la loi normale. La première est la *skewness* ou coefficient d'asymétrie. Son expression est :

$$S_k = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\sigma_X^3} \tag{5.148}$$

Cette statistique est une mesure de l'asymétrie de la distribution, i.e. de la façon dont la densité s'étale de part et d'autre de son espérance. Dans le cas d'une loi normale, cette statistique vaut 0 : la loi est parfaitement centrée.

La seconde statistique sur laquelle s'appuie le test de Jarque et Berra est la *kurtosis* ou coefficient d'applatissement. Cette statistique mesure l'applatissement des queues de distribution : plus celles-ci sont épaisses et moins le processus a de chances d'être gaussien. Un loi normale a théoriquement une kurtosis égale à 3. On mesure cet index à l'aide de la statistique suivante :

$$K_u = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\sigma_X^4} \tag{5.149}$$

La statistique de Jarque et Berra est en fait une mesure de la distance de chacune de ces statistiques aux résultats théoriques connus pour la loi normale. Elle est néanmoins pondérée par les écarts types des lois asymptotiques des estimateurs de la skewness et de la kurtosis :

$$JB = \frac{T}{6}(S_k - 0)^2 + \frac{T}{24}(K_u - 3)^3 \tag{5.150}$$

où T est le nombre d'observation. Cette statistique a pour distribution asymptotique une loi du chi-deux à 2 degrés de liberté.

Attention cependant à la philosophie de la démarche : on ne vérifie pas l'adéquation à une loi gaussienne afin de s'assurer de la justesse des estimations. Depuis Gourieroux et alii (1984), on *sait* qu'à partir du moment où l'estimation est conduite en supposant que les innovations $((\epsilon_t)_{t\in\mathbb{Z}})$ suivent une loi appartenant à la famille des lois exponentielles (dont la loi normale fait partie), les estimateurs utilisés dans le cadre d'une démarche basée sur le maximum de vraisemblance sont des estimateurs consistents ! Finalement, peut importe la loi, pourvu qu'il ne subsiste ni autocorrélation des erreurs, ni hétéroscédasticité (la variance n'est pas constante au cours du temps), alors les estimateurs du maximum de vraisemblance utilisant la gaussiannité des innovations sont convergents.

La méthode d'estimation utilisant une loi des erreurs possiblement différente de la vraie loi de celle-ci, mais appartenant à la famille des lois exponentielles est appelée *Pseudo Maximum de Vraisemblance*. L'estimation de la matrice de variance covariance des estimateurs peut naturellement se faire en utilisant la matrice BHHH présentée au chapitre précédent. Il s'agit d'une méthode particulièrement utile pour l'estimation des processus ARCH/GARCH que l'on verra par la suite.

5.2.15.1 Tests sur les résidus

Avant de passer à la prévision, il convient néanmoins de s'interroger sur la méthodologie permettant de sélectionner l'ordre p et q des processus. Plusieurs méthodologies sont applicables, et reste plus complémentaires que substituables.

1. La première démarche est celle exposée plus haut et fut la démarche fondatrice de ces modèles, telles qu'elle fut proposée par Box et Jenkins. Il s'agit simplement de sélectionner les ordres pour le processus AR et le processus MA au vu des autocorrélogrammes du processus utilisé pour l'estimation. Il s'agit simplement du prolongement naturel de ce qui a été dit lors de l'introduction du présent chapitre. L'étude de l'ACF permet de sélectionner l'ordre du processus MA et l'étude la PACF permet de sélectionner l'ordre du processus AR. Cependant, comme il l'a été mentionné plus haut, un processus AR(1) présente une ACF avec une forte persistance, ce qui n'est dû qu'à la "contagion" de l'autoregressivité sur les erreurs du processus. Autrement dit, un AR(1) présente certes une ACF qui décroit lentement, mais il s'agit simplement d'une persistance née de l'unique retard de l'AR(1), perceptible sur la PACF. Moralité : si cette méthode fournie un point de départ, elle est loin d'être suffisante.

2. Une seconde approche possible, dans le prolongement de la précédente, consiste à utiliser les statistiques de student des estimations par maximum de vraisemblance (estimées généralement par approximation BHHH ou forme analytique) pour juger de la significativité des paramètres associés à chacun des retards. Pour un paramètre ϕ_i donné, le test est de la forme :

$$\frac{\hat{\phi}_i - \phi}{\sigma_\phi} \sim N(0,1) \tag{5.151}$$

Ce type de test a déjà fait l'objet de développements lors de l'exposé des méthodes du maximum de vraisemblance.

3. La dernière méthode s'appuie sur des statistiques composés à partir de la log-vraisemblance, permettant de juger de la distance entre la loi du modèle estimé et celle du processus. On parle de critères d'information pour manifester ce dernier trait associé à ces statistiques. On en détaille quelques unes :

   – Le critère d'Akaike (AIC) : le meilleur des modèles ARMA est celui qui minimise la statistique :

   $$AIC(p,q) = Tlog(\sigma_\epsilon^2) + 2(p+q) \tag{5.152}$$

   – Le critère bayésien (BIC) : le meilleur des modèles ARMA est celui qui minimise la statistique :

   $$BIC(p,q) = Tlog(\sigma_\epsilon^2) - (n-p-q)log\left[1 - \frac{p+q}{T}\right] \tag{5.153}$$

   $$+ (p+q)log(T) + log\left[(p+q)^{-1}\frac{\sigma_x^2}{\sigma_\epsilon^2 - 1}\right] \tag{5.154}$$

   – Le critère de Hanan et Quinn : le meilleur des modèles ARMA est celui qui minimise la statistique :

   $$HQ(p,q) = Tlog(\sigma_\epsilon^2) + (p+q)log\left[\frac{log(T)}{T}\right] \tag{5.155}$$

Au final, lors de l'estimation de processus ARMA, l'ensemble de ces critères sont à utiliser pour l'estimation des ordres p et q : on commence en général par l'étude de l'ACF et de la PACF pour se faire un ordre d'idée sur le processus latent. On définit en général un ordre maximum pour p et q, puis on procède de façon descendante : on élimine progressivement (on parle d'approche *stepwise*) les paramètres non significatifs, tout en conservant un oeil sur les critères d'information.

5.2.16 La prévision à l'aide des modèles ARMA

Le principal avantage des modèles ARMA tient au fait qu'il permettent de fournir des prévisions pour des échéances éloignées (du moins pour une échéance plus éloignée dans le temps que la prochaine date). Comme dans le cas des modèles linéaires présentés plus haut, la prévision se fait en utilisant l'espérance : en supposant que l'on se situe à une date t, la prévision du processus $(x_t)_{t\in\mathbb{Z}}$ est obtenue comme sa projection dans l'espace engendré par le passé de ce processus et de ses erreurs.

Plus simplement, dans le cas d'un processus AR(1) :

$$x_t = \phi x_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{5.156}$$

sa prévision à la date $t+1$ sera :

$$\hat{x}_{t+1} = \mathbb{E}[x_{t+1}|x_t] \tag{5.157}$$

$$= \phi x_t \tag{5.158}$$

Les prévisions suivantes s'obtiennent de façon récursives. Par exemple pour $t+2$:

$$\hat{x}_{t+2} = \mathbb{E}[x_{t+2}|x_t] \tag{5.159}$$
$$= \phi\mathbb{E}[x_{t+1}|x_t] \tag{5.160}$$
$$= \phi^2 x_t \tag{5.161}$$

D'une façon plus générale, une prévision à l'ordre k pour un processus AR(1) s'obtient de la façon suivante :

$$\hat{x}_{t+k} = \phi^k x_t \tag{5.162}$$

On remarque évidement que pour un k suffisament grand, on a :

$$\hat{x}_{t+k} \to 0 \tag{5.163}$$

qui est la moyenne non conditionnelle du processus. Autrement dit, pour un ordre k élevé, le modèle AR(1) se contente de fournir comme prévision la moyenne (historique) du processus.

La simplicité de ces processus les a rendu très attrayants : il est possible de prévoir n'importe quelle série autoregressive pour un ordre important, à partir de la seule connaissance de son passé. Dans le cas d'un AR, l'estimation se fait très simplement, par MCO. On montre encore qu'il est aisé d'obtenir un intervalle de confiance pour la prévision, en utilisant le théorème de Wold.

Si le processus étudié est stationnaire, alors il admet une représentation MA(∞) de la forme :

$$x_t = \sum_{i=0}^{\infty} \phi_i \epsilon_{t-i}, \text{ avec } \phi_0 = 1 \tag{5.164}$$

en faisant abstraction de la composante déterministe. La prévision précédente peut aussi être obtenue de la façon suivante :

$$\hat{x}_{t+1} = \mathbb{E}[x_{t+1}|x_t, ..., x_0] \tag{5.165}$$
$$= \mathbb{E}[x_{t+1}|\epsilon_t, ..., \epsilon_0] \tag{5.166}$$
$$= \sum_{i=1}^{\infty} \phi_i \epsilon_{t+1-i} \tag{5.167}$$

La précédente égalité est obtenue en remarquant que :

$$x_{t+1} = \sum_{i=0}^{\infty} \phi_i \epsilon_{t+1_i} \tag{5.168}$$
$$= \epsilon_{t+1} + \sum_{i=1}^{\infty} \phi_i \epsilon_{t+1_i} \tag{5.169}$$

D'une façon plus générale, on a :

$$\hat{x}_{t+k} = \sum_{i=k}^{\infty} \phi_i \epsilon_{t-i} \tag{5.170}$$

On en déduis l'erreur de prévision :

$$x_{t+k} - \hat{x}_{t+k} = \sum_{i=0}^{\infty} \phi_i \epsilon_{t+k-i} - \sum_{i=k}^{\infty} \phi_i \epsilon_{t+k-i} \tag{5.171}$$

$$= \sum_{i=0}^{k-1} \phi_i \epsilon_{t+k-i} \tag{5.172}$$

On en déduis sans problème la variance des erreurs de prévision :

$$\mathbb{V}[x_{t+k} - \hat{x}_{t+k}] = \sum_{i=0}^{k-1} \phi_i^2 \sigma_\epsilon^2 \tag{5.173}$$

Première remarque : non, le théorème de Wold ne sert pas à rien. Il permet de déterminer l'intervalle de confiance de la prévision simplement en s'appuyant sur la représentation MA des processus stationnaires. Deuxième remarque : l'erreur de prévision va grandissante au fur et à mesure que l'on s'éloigne de la date t. Troisème remarque : il est toujours possible d'obtenir cet intervalle de confiance en estimant un modèle MA avec un ordre important afin d'obtenir les ϕ_i nécessaire à l'estimation de l'intervalle de confiance.

D'après ce qui vient d'être dit, il est évident qu'il est possible de construire un intervalle de confiance de la forme :

$$IC_\alpha = \left[\hat{x}_{t+k} \pm t_{\alpha/2} \sum_{i=0}^{k-1} \phi_i^2 \sigma_\epsilon^2 \right] \tag{5.174}$$

Ceci tient naturellement au fait que la loi asymptotique de l'estimateur du maximum de vraisemblance est gaussienne (cf. chapitres précédents). On a donc naturellement :

$$\frac{\hat{x}_{t+k} - x_{t+k}}{\sqrt{\sum_{i=0}^{k-1} \phi_i^2 \sigma_\epsilon^2}} \rightarrow N(0,1) \tag{5.175}$$

où x_{t+k} est la vraie valeur du processus à la date $t+k$.

5.2.17 A vrai dire...

Arrivé à ce stade, il est évident que la clef du succès des modèles de séries temporelles univariés et linéaires réside dans leur simplicité et l'interprétation des résultats fournis.

Malheureusement, qui dit simplicité dit également pauvreté de la prévision. Les modèles ARMA sont incapables de percevoir un retournement de tendance, puisqu'on ne fait que multiplier la tendance actuelle. Pour remédier à ceci, des modèles à seuil ont été introduits. Il en sera question plus tard lors d'une simple application. Pire, les mouvements de prix, de rentabilité ou bien encore les évolutions des chiffres de l'économie sont le résultats d'une mutlitude de chocs, de dépendances conditionnnelles entre séries... Bref, x_t ne contient certainement pas assez d'information pour parvenir à construire une prévision robuste de son avenir ! Là encore, pour remédier à ceci, des modèles multivariés ont été introduits. On traitera de ces modèles dans la section consacrée à la prise en compte des liens entre variables macroéconomiques.

5.2.18 Quelques applications de modèles ARMA

5.2.18.1 Modélisation de l'inflation

L'inflation est l'une des variables macroéconomiques les plus suivies par les marchés, et notamment les marchés de taux. La raison à cela est simple : il s'agit d'une variable suivie par les Banques Centrales afin d'établir leur politique. On parle d'instrument de la politique monétaire : les Banques Centrales sont censées réagir par une montée des taux à tout accroissement de l'inflation, de façon mécanique. En réalité, la véritable variable cible des banquiers centraux est l'inflation qu'il anticipent à moyen terme. Celle-ci n'est malheureusement pas mesurable : on a recourt en France et en Europe à L'Indice des Prix à la Consommation ou IPC pour mesurer cette inflation.

La mesure d'inflation suivie par les Banques Centrales est en réalité le glissement mensuel de l'indice des prix à la consommation. En notant P_t l'indice des prix à la consommation en date t, le glissement mensuel se note comme suit :

$$\pi_t = \frac{P_t - P_{t-12}}{P_{t-12}} \tag{5.176}$$

La chronique de l'inflation est présentée en figure 5.7. Sa trajectoire ressemble furieusement à une marche aléatoire : on s'attend à trouver un modèle de type AR, avec un premier paramètre proche de 1.

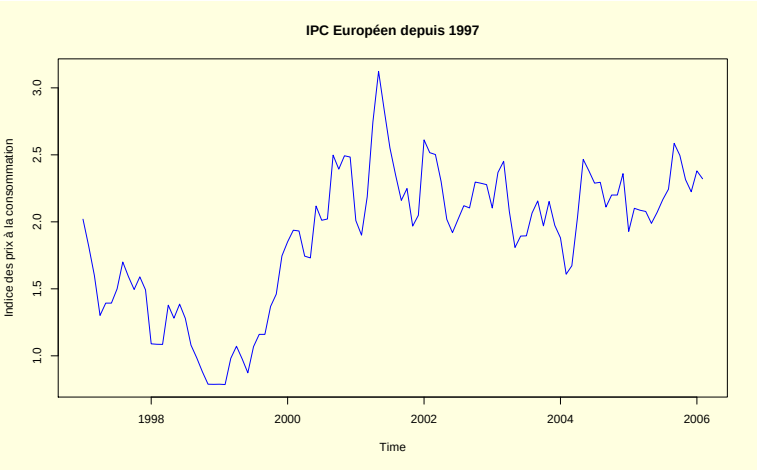

FIG. 5.7 – IPC européen depuis 1997

Cette intuition est confirmée par l'étude des ACF/PACF qui exhibent naturellement une persistance importante, pour des ordres de retard eux-mêmes importants. Ces graphiques sont représentés en figure 5.8.

Finalement, l'estimation du modèle ARMA optimal conduit aux paramètres suivants :

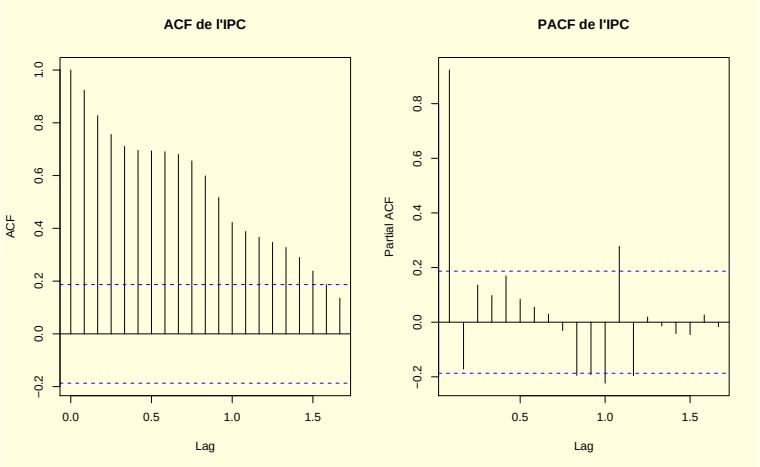

FIG. 5.8 – ACF et PACF de l'IPC européen

	ar1	ma1	intercept
Estimation	0,89	0,23	1,92
Ecart type	0,04	0,11	0,19
T-stat	19,96	2,15	10,06

Conclusion de cette courte étude : l'indice des prix à la consommation semble se comporter comme une marche aléatoire. La meilleure prévision de l'inflation de demain devrait ainsi être celle d'aujourd'hui. En oubliant la composante ma1, et en posant $\phi = 1$, on a alors :

$$\pi_t = \pi_{t+1} + \epsilon_t \tag{5.177}$$

On a donc très naturellement :

$$\mathbb{E}[\pi_{t+h}|I_t] = \pi_t \tag{5.178}$$

où I_t est l'information disponible à la date t. Il s'agit d'un fait qui commence à être bien connu dans le monde de l'économie monétaire.

Dernier point de remarque : il est possible d'interpréter le paramètre de la composante MA comme un paramètre de retour à la moyenne/autoexcitation de l'inflation. On remarque tout d'abord que l'on a approximativement :

$$\epsilon_{t-1} = \pi_{t-1} - (\mu + \phi\pi_{t-2}) \tag{5.179}$$

Le paramètre associé à la composante MA est positif : l'inflation a ainsi tendance à se surexciter. Si l'inflation s'écarte de sa moyenne de long terme (ici 1,92%), elle aura tendance à rester au dessus de cette moyenne pour quelques périodes.

Notons enfin que la constante du modèle correspond à la cible d'inflation de la BCE. La figure 5.9 présente une prévision tirée du modèle ARMA estimé.

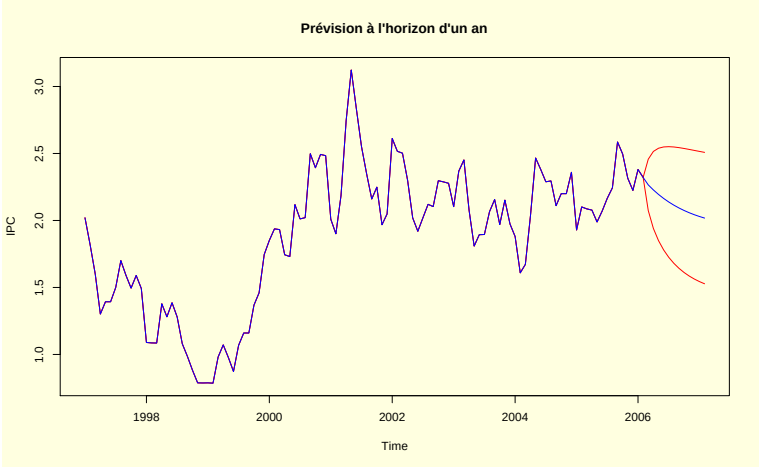

FIG. 5.9 – Prévision de l'IPC européen

Quoiqu'il en soit, il n'est pas certain que l'inflation soit tout à fait une marche aléatoire pour ce qui est de la zone euro depuis 2000. En effet, on remarque visuellement que la moyenne pour la période 1997-2000 ne semble pas être la même que celle pour la période 2000-2006 (cf figure 5.7). Confirmons cette intuition : étudions la structure de la série de l'IPC depuis 2000. La série est présentée en figure 5.10.

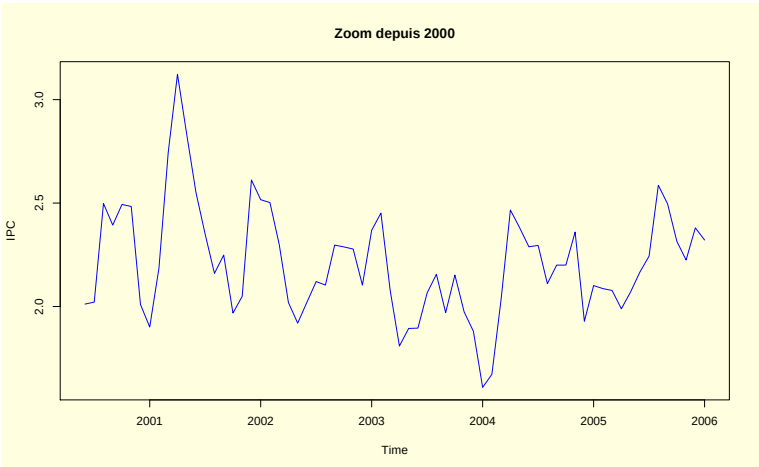

FIG. 5.10 – Zoom sur l'IPC depuis 2000

Etudions les ACF/PACF de la série sur la figure 5.11.

Elle semble assez différente de celle étudiée pour le même processus, avec une fenetre de temps plus large. La série semble se rapprocher d'une série nettement moins proche de la marche aléatoire : il s'agit simplement d'un AR(2), générant dans l'ACF un comportement de mélange entre sinusoide et exponentielle. L'estimation d'un AR(2) sur la série confirme les résultats :

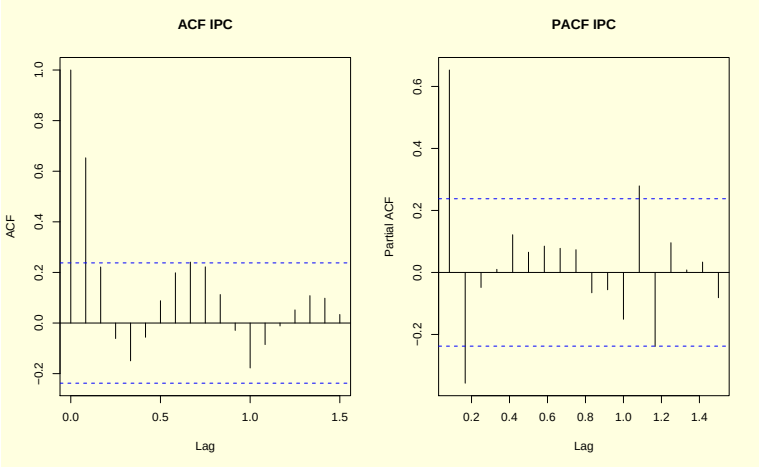

FIG. 5.11 – ACF et PACF de l'IPC depuis 2000

	ar1	ma1	intercept
Estimation	0,8808	-0,3499	2,2105
Ecart type	0,1123	0,1121	0,048
T-stat	7,84	-3,12	46,05

Finalement, les résultats obtenus sont réellements différents des précédents : l'inflation moyenne est plus importante que dans le cas précédent et le modèle est globalement un modèle autoregressif qui n'est plus marche aléatoire (le terme en AR(2) vient "corriger" l'importance du 0,88. Au final, notre petite estimation souligne que la zone euro est actuellement dans un régime inflationniste plus important que son régime de long terme. Ceci est certainement dû à l'effet Balassa-Samuelson : lors de la formation d'unions monétaires, un effet rattrapage par le haut du niveau des prix se produit. L'intégration de nouveaux pays à bas niveau de vie conduit à chaque fois à l'accroissement du niveau général des prix en Europe (il s'agit d'un index prix en glissement annuel).

La figure 5.12 présente la prévision associée au modèle.

5.2.18.2 Modélisation du taux cible de la BCE

Le taux directeur de la Banque Centrale Européenne est défini chaque mois lors de la réunion du conseil des gouverneurs. Ce taux constitue une cible permettant de guider le taux court (journalier) sur les marchés obligataires européens. La plupart des modèles de taux font de la politique monétaire le premier facteur de la courbe des taux : la compréhension du processus latent au taux cible est essentiel pour la gestion obligataire. Est-il possible d'utiliser les modèles ARMA à cette fin ?

La figure 5.13 présente l'évolution (sur une base mensuelle) de ces taux depuis 2000. Premier constat : il existe un nombre important de dates pour lesquelles le taux ne change pas. En notant r_t le taux cible à la date t, il existe de nombreuses dates pour lesquelles on a :

$$r_t = r_{t-1} \tag{5.180}$$

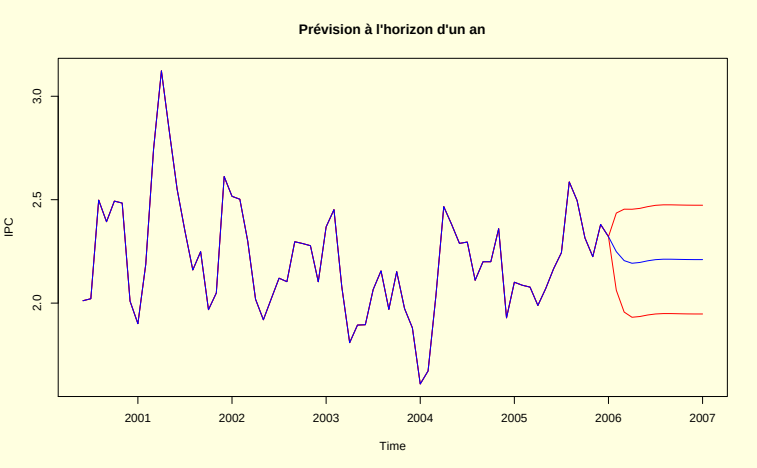

FIG. 5.12 – Prévision de l'IPC

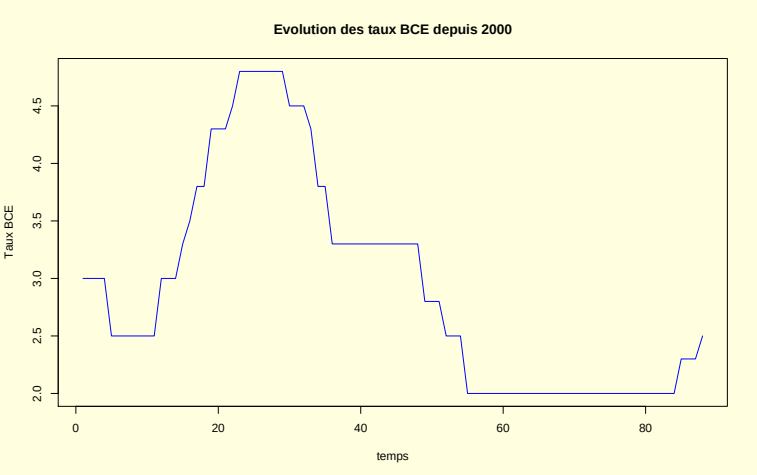

FIG. 5.13 – Chronique des taux BCE depuis 2000

Ceci risque bel-et-bien de faire apparaitre une correlation persistante dans l'ACF du processus. On observe ceci en figure 5.14. En effet, on observe tout d'abord un phénomène de contagion évident entre les différentes dates sur l'ACF qui décroit lentement vers 0. Au contraire, la PACF admet un unique pic, proche de 1, suggérant que l'on peut écrire pour le taux BCE un modèle de la forme :

$$r_t = \phi r_{t-1} + \epsilon_t \tag{5.181}$$

avec ϕ proche de 1. Dans le cas où $\phi = 1$, on retrouverai naturellement une marche aléatoire pour le taux BCE : le processus ne serait ainsi pas stationnaire. Faisons l'hypothèse que tel n'est pas le cas. Deux stratégies sont alors implémentables :

– Il est tout d'abord possible d'estimer un modèle MA avec un lag important. Ceci signifierai que chaque taux BCE peut s'écrire comme la somme des chocs passés sur ces taux (une sorte de représentation MA(∞) d'un processus AR(1). Après estimation, un modèle MA(15) semble convenir : il minimise le critère BIC. Les coefficients estimés sont les suivantes :

	Estimations	**T-stats**
ma1	0,9115	6,80731889
ma2	0,8645	3,92954545
ma3	1,0792	5,96902655
ma4	1,4888	7,61145194
ma5	1,5917	6,49408405
ma6	1,783	4,97766611
ma7	1,5166	5,9732178
ma8	1,5969	6,67600334
ma9	1,6871	5,59568823
ma10	1,3082	4,19698428
ma11	0,9441	4,30506156
ma12	0,6917	3,56546392
ma13	1,0329	4,61116071
ma14	0,7159	3,71124935
ma15	0,2105	1,72258592
Constante	2,9544	12,8788143

– La seconde stratégie, qui est la plus naturelle, consiste à estimer un AR(1), tout en se doutant que les taux suivent probablement un processus proche de la marche aléatoire. L'estimation d'un simple AR(1) donne des résultats similaires au MA(15). Les estimations sont :

	AR(1)	**Constante**
Estimations	0,9772	2,8363
T-stats	59,5853659	5,11229272

5.2.18.3 Modélisation de la volatilité implicite d'options sur DAX

Parmi les processus connus pour être autoregressifs, les processus de volatilité présentent des comportements modélisables à l'aide de processus ARMA. Nous verrons plus loin

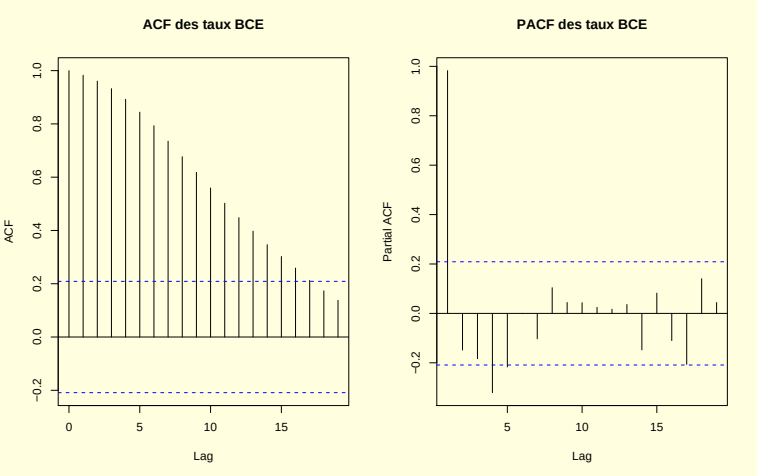

FIG. 5.14 – ACF et PACF des taux BCE depuis 2000

une raison à cette situation.

On présente en figure 5.15 les prix black scholes calibrés sur volatilité historique, implicite globale et implicite locale.

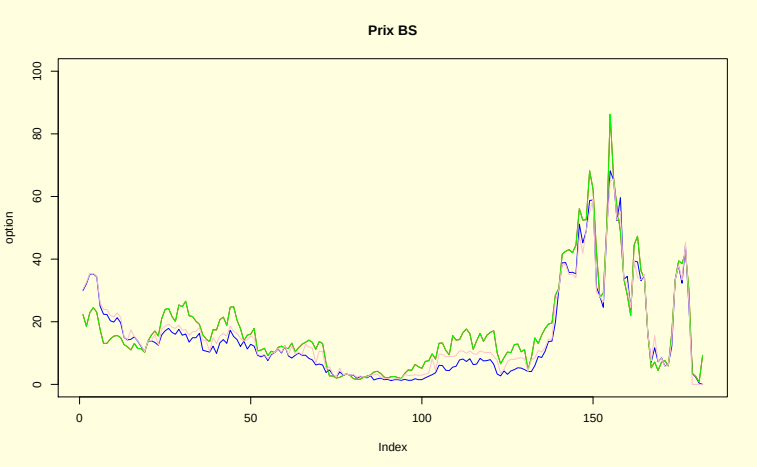

FIG. 5.15 – Prix BS d'une option sur DAX de strike 5000

La volatitilé implicite extraite en inversant numériquement la formule de BS présente la dynamique présentée en figure 5.16.

L'étude de l'ACF/PACF de la série des volatilités implicites ainsi obtenues permet de se convaincre de l'existence de pattern ARMA dans la série (cf. figure 5.17).

L'estimation d'un modèle ARMA permet de conclure à un modèle ARMA(1,1) dont les estimations sont fournies dans le tableau suivant :

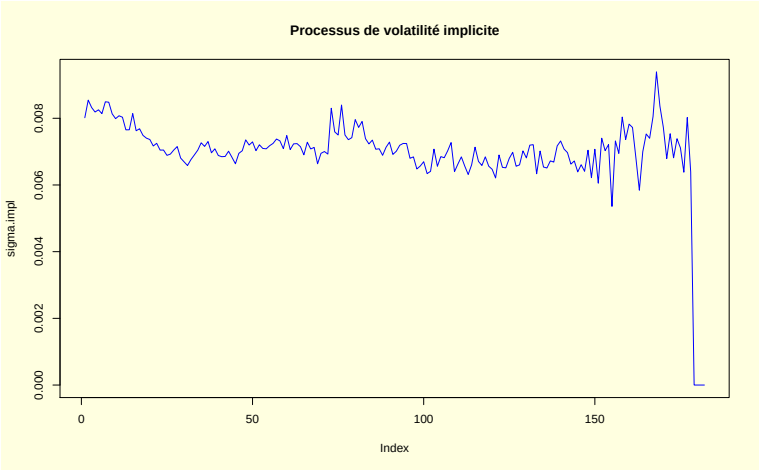

FIG. 5.16 – Volatilité implicite

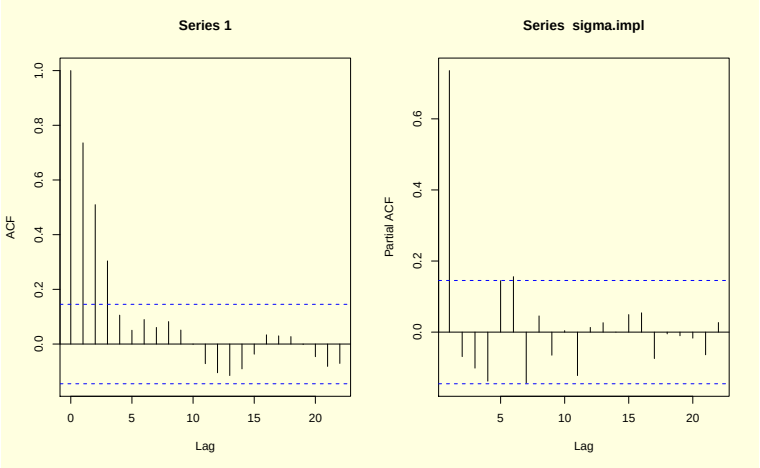

FIG. 5.17 – ACF de la volatilité implicite

	ar1	ma1	intercept
Estimation	0,9169	-0,5431	0,0072
Ecart type	0,0483	0,1042	0,0002
T-stat	18,9834	-5,2121	36,0000

La non constance de la volatilité a des implications particulières pour la gestion de portefeuille d'option, au nombre desquelles la courverture. La simple couverture d'un portefeuille de sous-jacent à l'aide du vega traditionnel pose problème : la couverture est statique alors que la volatilité, elle, est dynamique. On proposera l'utilisation des modèles ARCH/GARCH pour permettre de prendre cet élément en compte.

Poussons tout de même l'analyse un peu plus loin. Les prix d'options génèrent en général ce que l'on appelle un smile : la volatilité est une fonction non linéaire du strike. On représente cette surface de volatilité en figure 5.18.

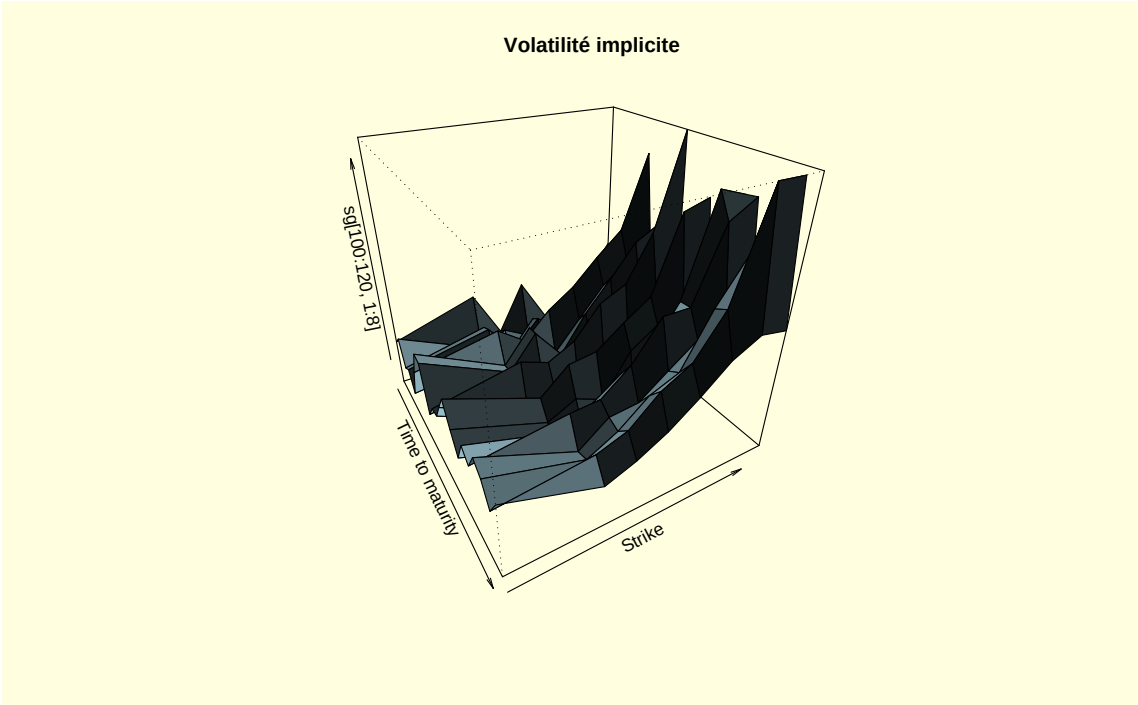

FIG. 5.18 – Surface de volatilité implicite

Une piste d'explication possible pour rendre compte de cette relation est d'étudier les composantes AR et MA de chacune des volatilité implicites. C'est ce qu'on représente en figures 5.19 et figures 5.20 : on retrouve une relation non linéaire entre strike et composante.

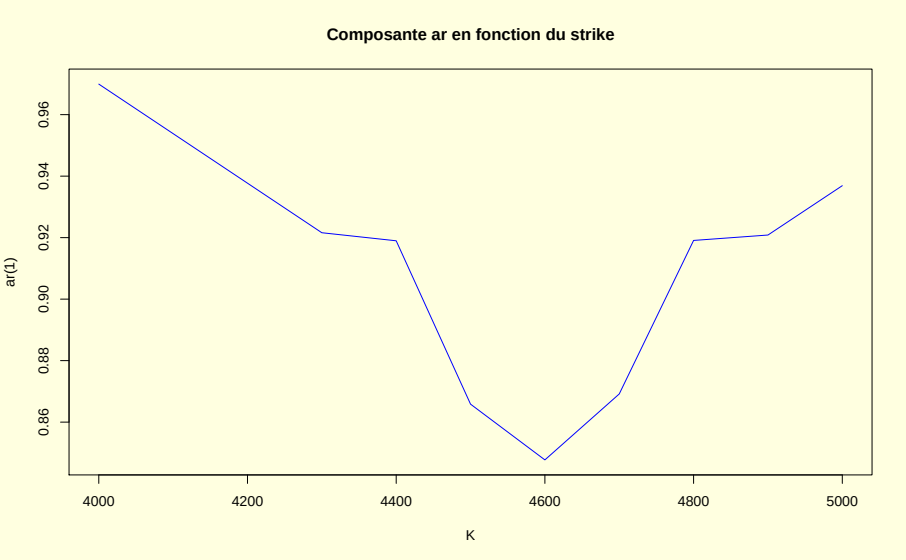

FIG. 5.19 – Relation entre strike et composante AR

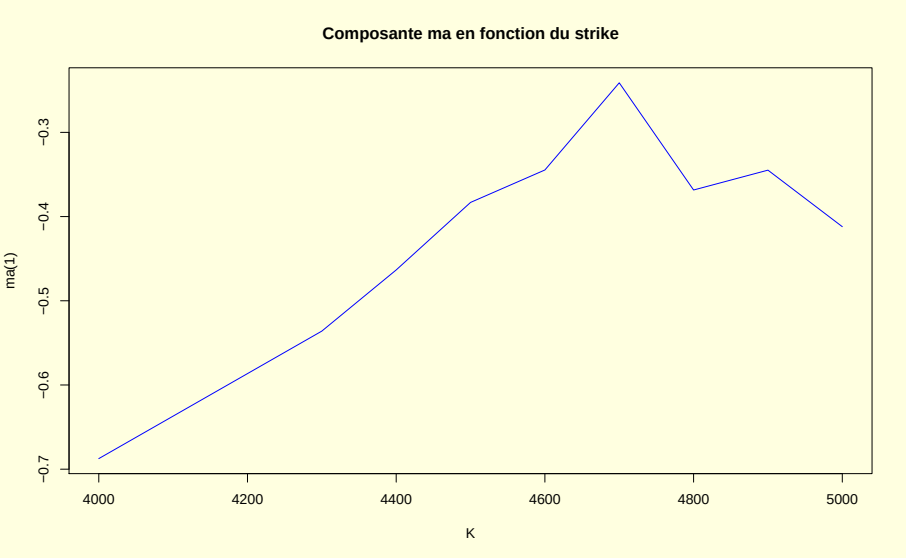

FIG. 5.20 – Relation entre strike et composante MA

5.3 Les modèles ARCH-GARCH

L'ensemble des notes qui suivent visent à introduire un certain nombre de concepts simples relatifs à la modélisation de la volatilité en finance. Il s'agit d'une problématique essentielle, pour de nombreux champs d'étude de la finance contemporaine. Le simple fait que la gestion de portefeuille s'appuie le plus souvent sur des algorithme espérance-variance souligne l'importance de la construction d'indicateurs de variance correctement spécifiés. La gestion des options s'appuie également sur une étude rigoureuse de l'évolution de la variance.

Ces dix dernières années furent l'occasion de développements importants, conduisant à une appréhension plus claire et dans un même temps plus complexe de ce qu'est réellement la variance du rendement d'un actif. Le lecteur soucieux d'accroitre sa culture dans ce domaine (au delà de ces maigres notes) lira avec intérêt Poon (2005), Gourieroux (1992), Gourieroux and Jasiak (2001) et Wang (2003). Une revue de littérature en français tournant autour des applications des modèles de volatilité appliqués aux options est disponible dans Aboura (2005).

La suite du cours est constuite de la façon suivante : on s'attarde dans un premier temps sur quelques faits stylisés en finance ainsi que sur la mesure de la volatilité des rendements. Puis, dans un second temps, on présente simplement les modèles dédiés à la volatilité historique des titres (modèles ARCH-GARCH). Ceci est suivi d'un retour sur l'inférence et la prédiction de la variance sur la base d'un modèle GARCH, avant de conclure cette section par l'étude des modèles de Duan et de Heston et Nandi visant à évaluer des actifs contingent sur la base d'un processus de variance GARCH.

5.3.1 Présentation des faits stylisés en finance

Les séries financières présentent un certain nombre de traits caractéristiques qu'il est nécessaire de souligner avant de passer à leur modélisation. Ces principaux faits sont les suivants (on suit ici ce qui en est dit dans Poon (2005)[page 7-8] :

1. Les rendements ne sont pas autocorrélés, ce qui semble confirmer l'hypothèse faible d'efficience des marchés financiers.
2. Les autocorrélations du carré et de la valeur absolue des rendements sont significatives et décroissent lentement vers 0.
3. On remarque qu'en général, la suppression des valeurs extrêmes d'une série de rendements accroit la significativité de l'autocorrélation des rendements élévés au carré ou pris en valeur absolue.
4. L'autocorrélation la plus forte touche les valeurs absolues des rendements : celles relatives aux rendements élevés à une puissance quelconque sont moindres, mais significatives. Il s'agit de l'effet de Taylor, d'après Taylor (1986).
5. L'asymétrie de la volatilité : en général, la volatilité d'un actif s'accroit lorsque le rendements qui précèdent sont négatifs. On parle dans ce cas d'effet levier.
6. Rentabilité et volatilité pour une classe d'actif semble évoluer de pair.

7. La volatilité d'un actif manifeste en général des périodes de de volatilité importante et d'autres périodes de volatilité faibles (correspondant à ce qu'on appelle des *bear and bull* market.

Cette liste est bien évidement non exhaustive. Elle a cependant le mérite de mettre en lumière le fait que dès lors que l'on souhaite s'extraire des standards de la finance (la marche aléatoire des théories de l'efficience et de l'évaluation d'option), il est nécessaire de considérer l'ensemble de ces propriétés avec beaucoup d'attention.

Dans la suite de cette section, on considère un certain nombre de modèles visant à prendre en compte (d'un point de vue purement statistique) certains de ces faits stylisés de la finance.

5.3.2 Quelques mesures préliminaires de la variance

Par essence, la volatilité est un phénomène inobservable : un actif vanille (une action par exemple) n'est cotée qu'à l'aide de son prix. Il est alors nécessaire de construire un certain nombre de statistiques permettant de juger de la volatilité des rendements : celle-ci, pourtant essentielle à tout modèle financier, ne s'observe pas naturellement. Il est à noter que certaines classes d'actifs sont cotés en volatilité implicite (pour certaines classes d'options) et qu'il existe des index de volatilité pour certains marchés : Poon (2005) consacre ainsi son avant dernier chapitre à quelques précisions relatives au VIX, un indice de volatilité compilé par le Chicago Board of Option Exchange. Il s'agit d'un indice visant à capturer la volatilité du S&P 500, et à permettre un prévision de la volatilité future de ce même indice de marché. Le mode de calcul est simple : il s'agit d'une moyenne pondérée de la volatilité implicite des contrats d'options sur S&P500 (contre S&P100 pour son prédécesseur, le VXO) qui présentent la spécificité d'être *out of the money*. Ce nouvel index de volatilité commença a être publié en 2003 (l'ancien fut lancé en 1993).

Le fait que ce type d'index connaissent un succès évident ne fait que souligner le besoin qu'on les marchés financiers de disposer d'intruments de mesure de la volatilité. Ce index est l'une des tentatives d'établissement de ces mesures et semblent relativement robuste (cf. revue de littérature de Poon (2005)). Il ne s'agit pas de la seule façon de faire : on propose ici deux autres tentatives plus ou moins intéressantes : l'index *high-low* et le carré des rendements.

5.3.2.1 La mesure *high-low*

La mesure *high-low* de la volatilité constitue une méthode simple et relativement robuste de mesure la volatilité : high et low désignent naturellement le plus haut et le plus bas du cours d'un titre sur une journée de cotation. Là encore, on suppose que les rendements sont gaussiens : on mesure la volatilité journalière à l'aide de l'estimateur suivant :

$$\sigma_t^2 = \frac{(lnH_t - lnL_t)^2}{4ln2} \tag{5.182}$$

Cet estimateur est celui proposé par Bollen and Inder (2002) dans le cadre de rendement suivant un brownien géométrique. Il s'agit d'une application d'une mesure initiallement proposée par Parkinson (1980). Garman and Klass (1980) propose une amélioration de la mesure de Parkinson qui prend la forme suivante :

$$\sigma_t^2 = 0.5 \left(ln \frac{H_t}{L_t} \right)^2 - 0.39 \left(ln \frac{p_t}{p_{t-1}} \right)^2 \tag{5.183}$$

Ces estimateurs de la volatilité sont assez sensibles aux *valeurs extremes*, c'est à dire aux rendements anormalement importants (qu'il s'agissent de rendement négatifs ou positifs). L'importance de la probabilité d'occurence des événements extrèmes rend le processus effectivement suivi par les rendements incompatibles avec la loi normale. Dans cette mesure, les mesures H-L constitue une approximation intéressante de la variance des rendements.

Notons finalement que ces indices H-L modélisent d'une certaine façon la variance conditionnelle du processus des rendements : celle-ci change chaque jour. Il n'est par conséquent pas question d'utiliser ces résultats pour modéliser la variance de l'ensemble de la série des rendements d'un titre donnée, ce qui n'aurait de toute façon pas grand sens.

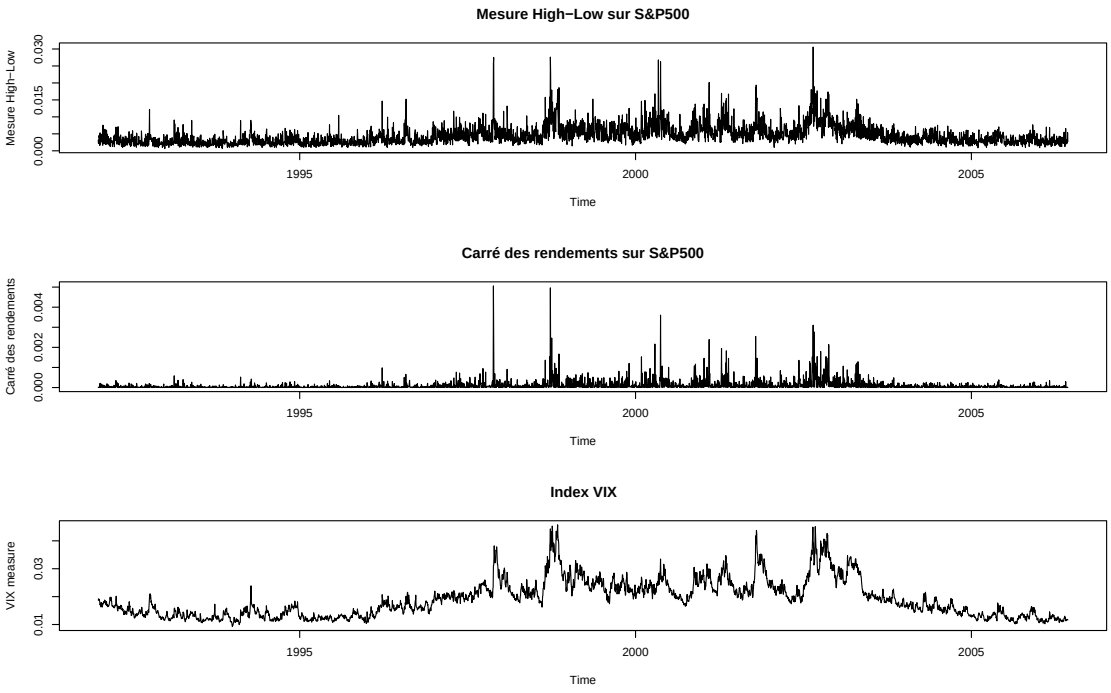

FIG. 5.21 – Différentes mesures de la volatilité appliquées au S&P 500

5.3.2.2 Le carré des rendements comme mesure de variance

Avant l'introduction de larges banques de données contenant des données intraday, de nombreux chercheurs se sont penchés sur l'utilisation d'un certain nombre de mesures

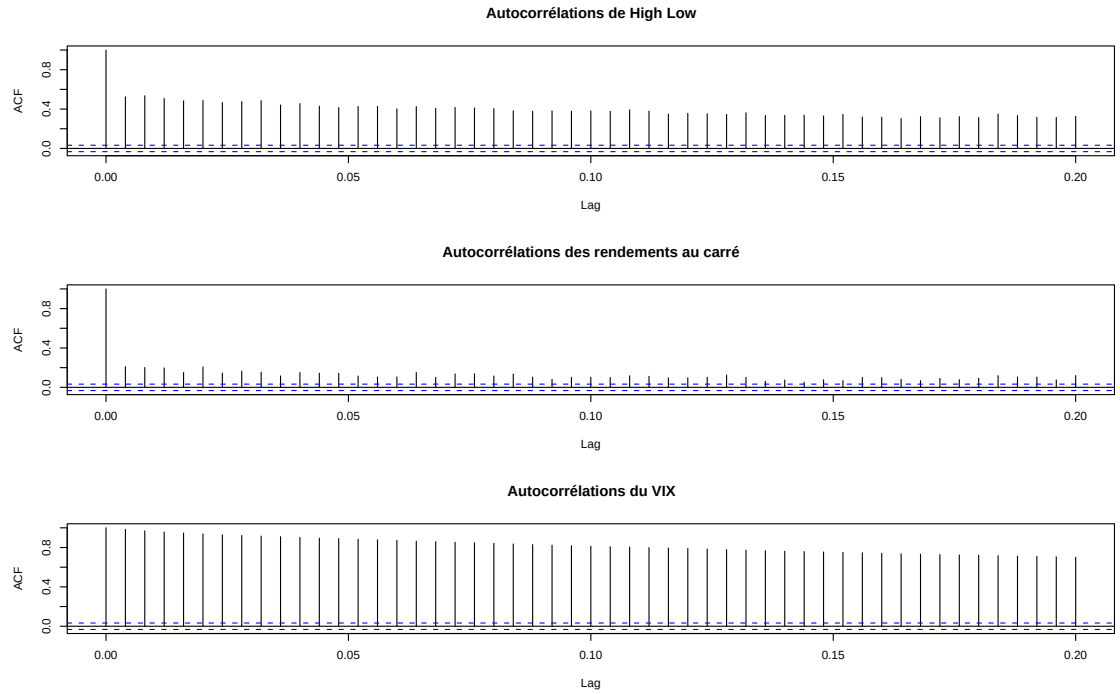

FIG. 5.22 – ACF des différentes mesures de la volatilité appliquées au S&P 500

de la volatilité des rendements sur la base de données journalières. Par exemple, Lopez (2001) propose le modèle suivant pour les rendements :

$$r_t = \mu + \sigma_t \epsilon_t \tag{5.184}$$

avec $\epsilon_t \sim N(0,1)$. On a alors :

$$\mathbb{E}[r_t^2 | \mathbb{F}_{t-1}] = \sigma_t^2 \tag{5.185}$$

Dans ce cas, le carré des rendements (en négligeant le drift) permet de mesurer la volatilité de ces mêmes rendements. Cependant, ϵ_t^2 suit naturellement une loi du χ^2 à un degré de liberté, dont la médiane est 0,455 : il s'en suit que ϵ_t^2 est inférieur à $\frac{\sigma_t^2}{2}$ dans plus de la moitié des cas. En effet, on a :

$$P\left(r_t^2 < \frac{\sigma_t^2}{2}\right) = P\left(\sigma_t^2 \epsilon_t^2 < \frac{\sigma_t^2}{2}\right) \tag{5.186}$$

$$= P\left(\epsilon_t^2 < \frac{1}{2}\right) \tag{5.187}$$

$$= 0.52 \tag{5.188}$$

Autrement dit, le carré des rendements (en l'absence de drift) est un estimateur sans biais de la variance des rendements. Cependant, dans plus de la moitié des cas, il sous-estime la variance des résidus. Ce résultat semble s'affaiblir lorsqu'on utilise le carré des rendements toutes les 5 minutes plutot que le carré des rendements journaliers.

5.3.3 Présentation des modèles ARCH-GARCH

Les processus ARCH visent également à rendre compte du fait que la variance conditionnelle n'est pas constante et proposent une façon de l'estimer basée sur le carré des rendements. D'après ce qui vient d'être dit, on traitera cette classe de modèle avec méfiance : il est possible que la volatilité soit non constante au cours du temps, mais qu'un modèle ARCH - ou leur généralisation GARCH - ne captent pas cet effet, voire concluent dans certains cas à l'absence de dépendance temporelle dans les rendements.

On présente dans ce qui suit les modèles ARCH et GARCH ainsi que leurs principales propriétés.

5.3.3.1 Pour commencer...

On a modélisé jusqu'à présent l'espérance conditionnelle dans le cadre de modèles linéaires simples (les modèles ARMA). Ces modèles ne sont guères applicables en finance car :

- La théorie financière repose sur la martingalité des prix : l'autocorrélation entre le rendement à la date t et à la date $t-1$ est une abhération financière, qui disparait peu à peu des rendements, à mesure que les marchés se liquéfient. Certains marchés comme les commodities (IPE BRENT par exemple) continuent d'exhiber de l'autocorrélation dans les rendements.
- Ces modèles reposent sur l'hypothèse de constance de la variance conditionnelle : dans les séries financières, la volatilité est généralement une fonction du temps. Il suffit d'étudier le processus de volatilité implicite d'une option pour se rendre compte de sa dépendance temporelle. C'est ce qu'on a observé à la fin de la section précédente.

Une façon simple d'illustrer les dangers liés à l'utilisation des rendements au carré pour mesurer la volatilité consiste à simuler un AR(1) et à représenter le carré du processus, en le confrontant à une série financière. L'illusion est parfaite : on aurait presque l'impression que les deux séries sont issues du même processus. C'est ce qu'on observe sur la figure 5.23.

Enfin, on est également à la recherche de modèles statistiques pour les rendements qui permettent de générer de la leptokurticité, i.e. des distributions à queues épaisses. On remarque en général que les séries financières présentent une kurtosis supérieure à 3, signifiant simplement que les queues de distribution de ces séries sont en général plus épaisses que celles de la loi normale.

Ce dernier fait a des implications importantes en terme de Risk Management, notamment au travers du calcul très simple de la Value at Risk. La VaR est simplement le quantile de la loi des rendements à $x\%$: il suffit de calculer ce quantile pour différents x et sur différentes lois pour se rendre compte de l'importance des queues de distribution. Le tableau 5.1 propose trois type de lois : la loi normale, qui n'est ni leptokutique, ni asymétrique ; la loi de Student qui est leptokutique, mais symétrique ; la loi de Laplace qui dans ce cas précis est à la fois leptokurtique et asymétrique (la queue basse est plus

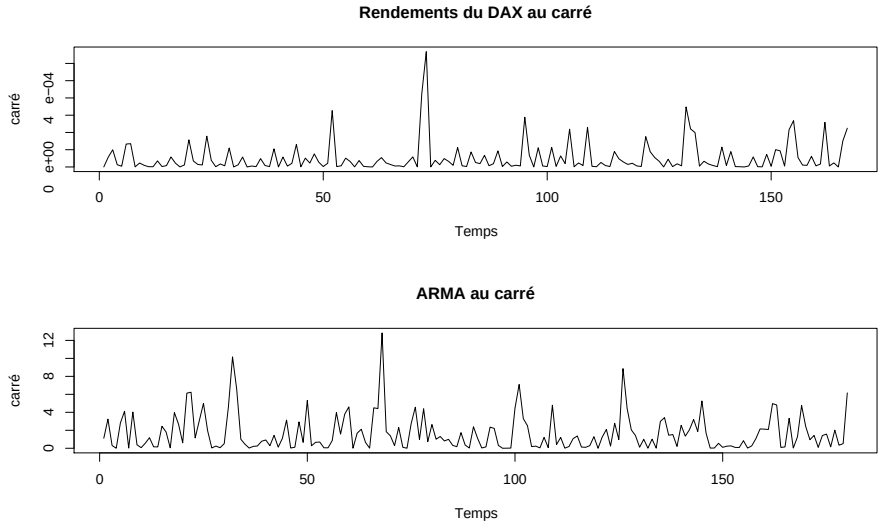

FIG. 5.23 – Rendements au carré et erreurs de mesure

	0%	5%	10%	50%	90%	95%	100%
Loi normale	-2,95	-1,62	-1,25	0,01	1,31	1,69	3,16
Loi de Student	-4,26	-1,76	-1,37	0,02	1,30	1,75	4,73
Loi Laplace	-13,89	-4,18	-3,15	0,08	1,79	2,42	5,27

TAB. 5.1 – Quantiles de différentes loi de moyenne nulle

épaisse que la queue haute). Les densités empiriques de ces trois lois sont présentées en figure 5.24. La VaR calculée pour différents seuils permet de se faire une idée plus précise des risques liés à calculer une VaR sur une hypothèse de gaussiannité des rendements, quand ceux-ci admettent des queues épaisses et/ou asymétriques.

La table 5.2 fournit une autre illustration de ces propriétés des séries financières : on présente espérance, écart type (volatilité), skewness et kurtosis des rendements de l'indice DAX ainsi que de divers call européens de strike différents, ayant tous le DAX pour sous-jacent. On remarque tout à fait l'existence de queues épaisses ainsi que d'une asymétrie.

	Espérance	Variance	Skewness	Kurtosis
DAX	0,001	0,007	-0,394	3,924
Option strike 4800	-0,001	0,173	-0,19	3,485
Option strike 4600	0,003	0,124	-0,131	3,599
Option strike 4400	0,004	0,09	-0,149	3,95
Option strike 4000	0,003	0,052	-0,268	4,241

TAB. 5.2 – Statistiques des rendements

Notons pour terminer cette section fourre-tout que ce qui vient d'être dit n'a pas que des implications pour le risk-management, mais aussi pour l'asset pricing. A peu de choses pret, comme le souligne très justement Cochrane (2002), le prix d'un actif est toujours et partout une espérance sous loi risque neutre actualisée. Dans le cas où les rendements sont gaussiens, la formule de Black Scholes semble tenir et est largement utilisée dans de nombreux domaines, dont les options sur action et les produits de taux. Cependant, l'existence des faits styliés qui viennent d'être mis en avant peut conduire à d'importantes erreurs de pricing : le modèle BS avec rendements GARCH introduit par Duan et par Heston et Nandi semble conduire à des prix d'option plus proche de la réalité que les prix BS standards. Ceci a d'importantes implications en terme d'asset management.

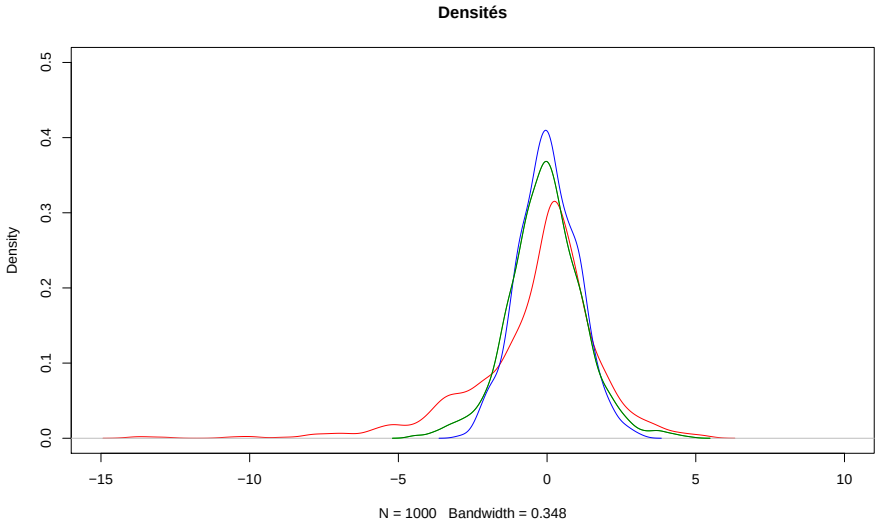

FIG. 5.24 – Densité de la loi de Laplace (rouge), normale (bleu) et Student (vert)

5.3.3.2 Introduction aux modèles ARCH-GARCH

5.3.3.2.1 La cas d'un ARCH(1)

Les modèles furent initiallement proposés par Engle (1982) et Bollerslev (1986), Tim Bollerslev étant le thésard de Robert Engle. Le premier modèle fut celui de Engle, et visait à obtenir une modélisation de la variance conditionnelle de l'inflation (en glissement mensuel) de la Grande Bretagne. Un modèle ARCH(1) est de la forme :

$$\begin{cases} x_t = \sqrt{h_t}\epsilon_t \\ h_t = \omega_0 + \omega_1 x_{t-1}^2 \end{cases} \tag{5.189}$$

avec $\epsilon_t \sim N(0,1)$. h_t représente la variance conditionnelle du processus x_t. Les moments conditionnels sont les suivants :

$$\mathbb{E}[x_t|h_t] = \mathbb{E}[\sqrt{h_t}\epsilon_t|h_t] \tag{5.190}$$
$$= \sqrt{h_t}\mathbb{E}[\epsilon_t|h_t] \tag{5.191}$$
$$= 0 \tag{5.192}$$

Il s'agit donc encore de processus applicables à des séries préalablement centrées, comme dans le cas des ARMA. Notons que les séries des rendements sont théoriquement naturellement centrées : il s'agit simplement d'une conséquence de la martingalité des prix.

La variance conditionnelle n'a plus rien à voir avec celle des ARMA :

$$\begin{aligned}
\mathbb{V}[x_t|h_t] &= \mathbb{V}[\sqrt{h_t}\epsilon_t|h_t] && (5.193)\\
&= h_t\mathbb{V}[\epsilon_t|h_t] && (5.194)\\
&= h_t \underbrace{\mathbb{E}[\epsilon_t^2|h_t]}_{=1 \text{ par hypothèse}} && (5.195)\\
&= h_t && (5.196)
\end{aligned}$$

Ainsi, contrairement aux modèles ARMA, la variance conditionnelle d'un processus ARCH n'est pas constante au cours du temps. C'est ce qui fait tout l'intérêt de ces processus, notamment pour les séries financières. Gardons cependant à l'esprit que ces modèles s'appuie sur une mesure de la variance proche de x_t^2. En effet, on a :

$$\mathbb{E}[x_t^2|h_t] = \mathbb{E}[h_t\epsilon_t^2|h_t] = h_t \tag{5.197}$$

Ceci tient simplement au fait que x_t soit naturellement un processus centré. Si ces modèles semblent d'un abord pratiques, il n'en reste pas moins qu'il produisent naturellement des erreurs de mesure sur la volatilité.

Le calcul des moments non conditionnels permet de déterminer quelques conditions à remplir afin de s'assurer de la stationnarité du processus. On détermine l'espérance à l'aide de la loi des espérances itérées :

$$\begin{aligned}
\mathbb{E}[x_t] &= \mathbb{E}[\mathbb{E}[x_t|h_t]] && (5.198)\\
&= \mathbb{E}[0] && (5.199)\\
&= 0 && (5.200)
\end{aligned}$$

Pour ce qui de la variance, il est possible de procéder par récurrence. Utilisons tout d'abord la loi de la décomposition de la variance :

$$\begin{aligned}
\mathbb{V}[x_t] &= \mathbb{E}[\mathbb{V}[x_t|h_t]] + \mathbb{V}[\mathbb{E}[x_t|h_t]] && (5.201)\\
&= \mathbb{E}[h_t] + 0 && (5.202)
\end{aligned}$$

On en déduit alors l'équation suivante :

$$\begin{aligned}
\mathbb{V}[x_t] &= \mathbb{E}[h_t] && (5.203)\\
&= \mathbb{E}[\omega_0 + \omega_1 x_{t-1}^2] && (5.204)\\
&= \omega_0 + \omega_1\mathbb{E}[x_{t-1}^2] && (5.205)\\
&= \omega_0 + \omega_1\mathbb{V}[x_{t-1}] && (5.206)
\end{aligned}$$

On obtient ainsi une formule de récurrence permettant de déterminer la variance non conditionnelle du processus. Il suffit, pour y parvenir, d'itérer la formule n fois, puis,

comme dans le cas des ARMA, de passer à la limite. On sait que :

$$\mathbb{V}[x_t] = \omega_0 + \omega_1 \mathbb{V}[x_{t-1}] \tag{5.207}$$
$$\mathbb{V}[x_{t-1}] = \omega_0 + \omega_1 \mathbb{V}[x_{t-2}] \tag{5.208}$$
$$\mathbb{V}[x_{t-2}] = \omega_0 + \omega_1 \mathbb{V}[x_{t-3}] \tag{5.209}$$

D'où :

$$\mathbb{V}[x_t] = \omega_0 + \omega_1(\omega_0 + \omega_1(\omega_0 + \omega_1 \mathbb{V}[x_{t-3}])) \tag{5.210}$$
$$= \omega_0(1 + \omega_1 + \omega_1^2) + \omega_1^3 \mathbb{V}[x_{t-3}] \tag{5.211}$$

D'où la formule générale :

$$\mathbb{V}[x_t] = \omega_0 \left(1 + \sum_{i=1}^{n} \omega_1^i\right) + \omega_1^{n+1} \mathbb{V}[x_{t-(n+1)}] \tag{5.212}$$
$$= \omega_0 \left(\sum_{i=0}^{n} \omega_1^i\right) + \omega_1^{n+1} \mathbb{V}[x_{t-(n+1)}] \tag{5.213}$$

D'où si $|\omega_1| < 1$, on a, lorsque $n \to \infty$:

$$\mathbb{V}[x_t] = \frac{\omega_0}{1 - \omega_1} \tag{5.214}$$

Cette dernière condition est nécessaire pour assurer l'existence de la variance, c'est à dire :

$$\mathbb{V}[x_t] < \infty \tag{5.215}$$

Cette condition est nécessaire pour obtenir un processus stationnaire (variance finie et indépendante du temps). Il est nécessaire d'imposer une seconde condition : la variance conditionnelle et non conditionnelle doivent être naturellement positives (la variance est le carré de l'écart type). La positivité de la variance conditionelle implique naturellement que :

$$\omega_0 > 0 \tag{5.216}$$
$$\omega_1 > 0 \tag{5.217}$$

Ces deux conditions impliquent naturellement que la variance non conditionnelle, dotée de la condition $|\omega_1| < 1$, soit positive. Remarquons finalement que, dans le cadre d'un processus ARCH, la variance conditionnelle ne coincide pas avec la variance non conditionnelle, ce qui est précisément ce que nous recherchions.

Ultime propriété d'un processus ARCH(1), il est possible de montrer que le carré du processus admet une représentation AR(1). On suit ici ce qui en est dit dans Poon (2005)[Chapitre 4]. Notons ν_t la différence entre x_t^2 et h_t. On a alors :

$$h_t = \omega_0 + \omega_1 x_{t-1}^2 \tag{5.218}$$
$$\Leftrightarrow x_t^2 - \nu_t = \omega_0 + \omega_1 x_{t-1}^2 \tag{5.219}$$
$$\Leftrightarrow x_t^2 = \omega_0 + \omega_1 x_{t-1}^2 + \nu_t \tag{5.220}$$

On retrouve ainsi un processus AR(1) sur les carrés des résidus. Ceci a plusieurs implications pratiques :

– D'une part, un processus ARCH(1) ne semble pas saisir de façon adéquate les processus de volatilité financière : on a vu lors des applications des processus ARMA que la volatilité de certains actifs semble présenter une structure plus proche des ARMA (retour à la moyenne en cas de choc importants) que des AR. Il sera donc nécessaire de complexifier légèrement la chose, afin d'accomoder cette caractéristique empirique.

– Seconde implications pratique, l'identification d'un ARCH(1) ne doit pas poser de problème, si l'on s'appuie sur ce qui a été dit plus haut au sujet des AR : il suffit d'étudier les fonctions d'autocorrélations simple et partielle pour se faire une idée de l'ordre du processus à retenir. On étudiera ceci au cours des applications empiriques proposées plus loin.

5.3.3.2.2 Les modèles ARCH(p)

Ce qui vient d'être dit au sujet des ARCH(1) peut se généraliser aisément au cas des processus ARCH(p). Un processus ARCH(p) est un processus x_t qui est de la forme :

$$x_t = \sqrt{h_t}\epsilon_t \tag{5.221}$$

$$h_t = \omega_0 + \sum_{i=1}^{p} \omega_{1,i} x_{t-i}^2 \tag{5.222}$$

avec $\epsilon \sim N(O, 1)$.

Comme précécement, on fournit les moments conditionnels :

$$\mathbb{E}[x_t|x_{t-1}] = 0 \tag{5.223}$$

$$\mathbb{V}[x_t|x_{t-1}] = h_t \tag{5.224}$$

Conditionnellement à l'information disponible à la date t, un processus GARCH est un processus de moyenne (conditionnelle) nulle et de variance égale à h_t. Qu'en est il des moments non-conditionnels ? L'espérance ne pose pas de problème, à la condition d'utiliser la loi des espérances itérées :

$$\mathbb{E}[x_t] = 0 \tag{5.225}$$

Pour ce qui est la variance, il est ici nécessaire de déterminer, comme précédement, une formule de récurrence pour parvenir finalement à exprimer la variance en passant à la limite. On ne refait pas ici les calculs : on se contente de fournir le résultat. Si $|\sum_{i=1}^{p} \omega_{1,i}| < 1$, alors la variance du processus existe et est de la forme :

$$\mathbb{V}[x_t] = \frac{\omega_0}{1 - \sum_{i=1}^{p} \omega_{1,i}} \tag{5.226}$$

5.3.3.2.3 Leptokurticité des processus ARCH(p)

Une propriété essentielle des processus ARCH est qu'ils génèrent des séries leptokurtiques. Il s'agit d'une propriété intéressante dans la mesure où conditionnellement, un processus ARCH est gaussien ! Ceci signifie qu'il n'est pas forcément nécessaire d'aller chercher des lois complexes ou méconnues pour rendre compte de la leptokurticité des

séries.

Il suffit donc de prouver que la kurtosis d'un processus ARCH est supérieure à 3. La preuve est simplissime et s'appuie sur le lemme de Jensen.

Rappel 3. Soit $f(x)$ une fonction convexe en x. On a alors :

$$\mathbb{E}\left[f(x)\right] > f\left(\mathbb{E}[x]\right) \tag{5.227}$$

La preuve d'établit comme suit, en se souvenant que $f(x) = x^2$ est bien convexe : soit x_t un processus ARCH(p) donné. Sa kurtosis est alors :

$$K_u(x) = \frac{\mathbb{E}[x_t^4]}{\mathbb{E}[x_t^2]^2} \tag{5.228}$$

car x_t est un processus centré. En appliquant le lemme des espérances itérées au numérateur, il vient :

$$\begin{aligned}
\mathbb{E}[x_t^4] &= \mathbb{E}\left[\mathbb{E}[x_t^4|x_{t-1}]\right] && (5.229)\\
&= \mathbb{E}\left[\mathbb{E}[h_t^2\epsilon_t^4|x_{t-1}]\right] && (5.230)\\
&= \mathbb{E}\left[h_t^2\mathbb{E}[\epsilon_t^4|x_{t-1}]\right] && (5.231)\\
&= \mathbb{E}\left[3h_t^2\right] && (5.232)\\
&= 3\mathbb{E}\left[h_t^2\right] && (5.233)
\end{aligned}$$

Cette dernière égalité tient au fait que ϵ_t suit une $N(0,1)$ donc la kurtosis est égale à 3. Dans la mesure où sa variance est égale à 1, constatons simplement que le dénominateur de la kurtosis est toujours égal à 1 pour une $N(0,1)$. D'où le résultat proposé. En utilisant Jensen, on a alors :

$$\mathbb{E}[x_t^4] = 3\mathbb{E}\left[h_t^2\right] > 3\mathbb{E}\left[h_t\right]^2 \tag{5.234}$$

La preuve s'achève en remarquant que :

$$\begin{aligned}
\mathbb{E}[x_t^2]^2 &= \mathbb{E}[h_t\epsilon_t^2]^2 && (5.235)\\
&= \mathbb{E}\left[\mathbb{E}[h_t\epsilon_t^2|x_{t-1}]\right]^2 && (5.236)\\
&= \mathbb{E}\left[h_t\mathbb{E}[\epsilon_t^2|x_{t-1}]\right]^2 && (5.237)\\
&= \mathbb{E}\left[h_t\right]^2 && (5.238)
\end{aligned}$$

En réinjectant ceci dans l'expression de la kurtosis, on prouve finalement :

$$K_u(x) = \frac{\mathbb{E}[x_t^4]}{\mathbb{E}[x_t^2]^2} > 3\frac{\mathbb{E}\left[h_t\right]^2}{\mathbb{E}\left[h_t\right]^2} = 3 \tag{5.239}$$

CQFD.

5.3.3.2.4 Quid de l'asymétrie ?

Les processus ARCH seraient parfaitement adaptés à la finance si ils étaient de plus capable de générer de l'asymétrie. On mesure d'ordinaire l'asymétrie à l'aide de la skweness :

$$S_k = \frac{\mathbb{E}[x_t^3]}{\mathbb{V}[x_t]^{3/2}} \tag{5.240}$$

Ceci est bien évidement vrai dans le cas où les séries sont centrées, comme c'est le cas pour les processus ARCH. Pour montrer que les processus GARCH ne sont pas asymétriques, il suffit de montrer que le numérateur est nul dans le cas d'un ARCH(p) :

$$\begin{aligned}
\mathbb{E}[x_t^3] &= \mathbb{E}[\mathbb{E}[x_t^3|x_{t-1}, x_{t-2}, ...]] && (5.241)\\
&= \mathbb{E}[\mathbb{E}[h_t^{3/2}\epsilon_t^3|x_{t-1}, x_{t-2}, ...]] && (5.242)\\
&= \mathbb{E}[h_t^{3/2}\mathbb{E}[\epsilon_t^3|x_{t-1}, x_{t-2}, ...]] && (5.243)\\
&= \mathbb{E}[h_t^{3/2} \times 0] && (5.244)\\
&= 0 && (5.245)
\end{aligned}$$

Non, les processus ARCH ne permettent pas de prendre en compte tous les faits stylisés de la finance : seule la leptokurticité est prise en compte.

5.3.3.3 Les modèles GARCH

Les modèles GARCH forment un légère complexification des modèles ARCH : on ajoute au processus de la variance les q valeurs passées de la variance, telle qu'elle est estimée par le modèle.

5.3.3.3.1 Le cas d'un GARCH(1,1)

Un modèle GARCH(1,1) s'écrit de la façon suivante :

$$\begin{aligned}
x_t &= \sqrt{h_t}\epsilon_t && (5.246)\\
h_t &= \omega_0 + \omega_1 x_{t-1}^2 + \omega_2 h_{t-1} && (5.247)
\end{aligned}$$

où $\epsilon_t \sim N(0,1)$. Comme précédement, on donne les moments conditionnels, à partir desquels on fournira finalement les moments non conditionnels par itération. L'espérance conditionnelle du processus est la suivante :

$$\mathbb{E}[x_t|\mathcal{F}_t] = 0 \tag{5.248}$$

où $\mathcal{F}_t$ est la filtration engendrée par les valeurs passées de x_t, de x_t^2 et de h_t. La variance conditionnelle est alors :

$$\begin{aligned}
\mathbb{V}[x_t|\mathcal{F}_t] &= \mathbb{V}[\sqrt{h_t}\epsilon_t|\mathcal{F}_t] && (5.249)\\
&= \sqrt{h_t}\mathbb{V}[\epsilon_t|\mathcal{F}_t] && (5.250)\\
&= h_t && (5.251)
\end{aligned}$$

h_t étant $\mathcal{F}_t$ mesurable. On détermine à présent les deux premiers moments conditionnels du processus. L'espérance s'obtient simplement à partir de la loi des espérances itérées :

$$\mathbb{E}[x_t] = \mathbb{E}[\mathbb{E}[x_t|\mathcal{F}_t]] \tag{5.252}$$
$$= 0 \tag{5.253}$$

Pour ce qui est de la variance non conditionnelle, on procède comme précédement, par itération. On sait que :

$$\mathbb{E}[x_t^2] = \mathbb{E}[h_t] \tag{5.254}$$
$$= \mathbb{E}[\omega_0 + \omega_1 x_{t-1}^2 + \omega_2 h_{t-1}] \tag{5.255}$$
$$= \omega_0 + \omega_1 \mathbb{E}[x_{t-1}^2] + \omega_2 \mathbb{E}[h_{t-1}] \tag{5.256}$$

Or, on sait que $\mathbb{E}[x_{t-1}^2] = \mathbb{E}[h_{t-1}]$. On peut donc réécrire la précédente égalité de la façon suivante :

$$\mathbb{E}[x_t^2] = \omega_0 + (\omega_1 + \omega_2)\mathbb{E}[x_{t-1}^2] \tag{5.257}$$

Comme précédement, on écrit cette formule de récurrence pour différents ordres :

$$\mathbb{E}[x_t^2] = \omega_0 + (\omega_1 + \omega_2)\mathbb{E}[x_{t-1}^2] \tag{5.258}$$
$$\mathbb{E}[x_{t-1}^2] = \omega_0 + (\omega_1 + \omega_2)\mathbb{E}[x_{t-2}^2] \tag{5.259}$$
$$\mathbb{E}[x_{t-2}^2] = \omega_0 + (\omega_1 + \omega_2)\mathbb{E}[x_{t-3}^2] \tag{5.260}$$

On trouve donc une formule de récurrence à l'ordre n suivante :

$$\mathbb{E}[x_t^2] = \omega_0 + (\omega_1 + \omega_2)\mathbb{E}[x_{t-1}^2] \tag{5.261}$$
$$= \omega_0 + (\omega_1 + \omega_2)(\omega_0 + (\omega_1 + \omega_2)(\omega_0 + (\omega_1 + \omega_2)\mathbb{E}[x_{t-3}^2])) \tag{5.262}$$
$$= \omega_0 + (\omega_1 + \omega_2)\omega_0 + (\omega_1 + \omega_2)^2\omega_0 + (\omega_1 + \omega_2)^3\mathbb{E}[x_{t-3}^2] \tag{5.263}$$

d'où pour n itération, on a la formule suivante :

$$\mathbb{E}[x_t^2] = \omega_0[1 + (\omega_1 + \omega_2) + (\omega_1 + \omega_2)^2 + ...] + (\omega_1 + \omega_2)^n\mathbb{E}[x_{t-n}^2] \tag{5.264}$$
$$= \omega_0 \sum_{i=0}^{n-1}(\omega_1 + \omega_2)^i + (\omega_1 + \omega_2)^n\mathbb{E}[x_{t-n}^2] \tag{5.265}$$

Comme précédement, à la condition que $|\omega_1 + \omega_2| < 1$, la série engendrée admet une limite (la variance existe) conduit à :

$$\mathbb{E}[x_t^2] = \frac{\omega_0}{1 - (\omega_1 + \omega_2)} \tag{5.266}$$

La positivité de la variance est assurée si ω_0 est du même signe que $1 - (\omega_1 + \omega_2)$.

On retrouve les mêmes propriétés de leptokurticité que pour les processus ARCH. On le montre de la même façon que précédement, en utilisant le lemme de Jensen. En

travaillant avec le quatrième moment du processus, il vient :

$$\mathbb{E}[x_t^4] = \mathbb{E}[\mathbb{E}[x_t^4|h_t]] \tag{5.267}$$
$$= \mathbb{E}[\mathbb{E}[h_t^2\epsilon_t^4|h_t]] \tag{5.268}$$
$$= \mathbb{E}[h_t^2\mathbb{E}[\epsilon_t^4|h_t]] \tag{5.269}$$
$$= \mathbb{E}[3h_t^2] \tag{5.270}$$
$$= 3\mathbb{E}[h_t^2] \geq 3\mathbb{E}[h_t]^2 \tag{5.271}$$

En divisant la dernière inégalité par $\mathbb{E}[h_t]^2$, on retrouve la kurtosis à gauche, et le résultat souhaité à droite. On a alors :

$$\frac{\mathbb{E}[x_t^4]}{\mathbb{E}[h_t]^2} \geq 3 \tag{5.272}$$

Si les GARCH sont à même de générer de la leptokurticité, ils sont en revanche incapable de générer de l'asymétrie. On le montre de la même façon que précédement, en utilisant la loi des espérances itérées :

$$\mathbb{E}[x_t^3] = \mathbb{E}[(\sqrt{h_t}\epsilon_t)^3] \tag{5.273}$$
$$= \mathbb{E}[\sqrt{h_t}^3\mathbb{E}[\epsilon_t^3|h_t]] \tag{5.274}$$
$$= \mathbb{E}[\sqrt{h_t}^3 \times 0] \tag{5.275}$$
$$= 0 \tag{5.276}$$

5.3.3.3.2 Les processus GARCH(p,q)

Un processus GARCH(p,q) se note de la façon suivante :

$$x_t = \sqrt{h_t}\epsilon_t \tag{5.277}$$
$$h_t = \omega_0 + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{i=1}^{q} \tag{5.278}$$

avec $\epsilon_t \sim N(0,1)$. Cette généralisation des GARCH(1,1) ne sera pas utilisée par la suite : le lecteur soucieux d'aller jusqu'au bout des calculs pourra appliquer ce qui a été fait précédement pour les ARCH(p).

5.3.4 Inférence des modèles ARCH-GARCH

5.3.4.1 Le cas d'un ARCH(1)

Comme précédement (pour les processus ARMA), on montre ici quelques rudiments nécessaires à l'estimation des ARCH(1). Il est essentiel de présenter les méthodes d'estimations, et de ne pas s'en remettre au fait que les logiciels de statistiques permettent une implémentation des ARCH sans connaissances particulières : l'interprétation des résultats ou la compréhension même des problèmes de convergence des estimateurs n'ont rien de trivial, comme de trop nombreux utilisateurs le pensent.

Sur le plan technique, l'estimation des ARCH n'est guère plus complexe que les ARMA : là encore, le processus x_t n'est pas i.i.d. et il est nécessaire de travailler conditionnellement au passé de x_t et de h_t, afin d'obtenir un processus i.i.d.. On connait les moments conditionnels d'un processus ARCH : celui-ci est conditionnellement gaussien, d'espérance nulle et de variance égale à h_t. Le problème ici, comme dans le cas des MA, est que la variance est "inobservable" : on a besoin des paramètres du processus ARCH pour obtenir la variance conditionnelle du processus. Comme précédement, la seule façon d'estimer ce type de processus est d'écrire la log-vraisemblance, puis de déterminer l'ensemble des dérivées dont on a besoin pour l'implémentation des méthodes numériques décrites au chapitre 5.

La loi conditionnelle de x_t conduit à la vraisemblance suivante :

$$L_i = \frac{1}{\sqrt{2\pi h_t}} exp\left\{-\frac{x_t^2}{2h_t}\right\} \tag{5.279}$$

La vraisemblance s'obtient en faisant le produit des différents L_i :

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi h_t}} exp\left\{-\frac{x_t^2}{2h_t}\right\} \tag{5.280}$$

et la log-vraisemblance *concentrée* est alors :

$$lnL = -\frac{1}{2}\sum_{i=1}^{n} log(h_t) - \frac{1}{2}\sum_{i=1}^{n} \frac{x_t^2}{h_t} \tag{5.281}$$

La méthode de Newton-Raphson s'appuie sur les dérivées premières qui s'obtiennent comme suit, pour un paramètre θ donné :

$$\frac{\partial lnL}{\partial \theta} = -\frac{1}{2}\sum_{i=1}^{n} \frac{\partial h_t/\partial \theta}{h_t} - \frac{\partial h_t}{\partial \theta}\frac{x_t^2}{h_t^2} \tag{5.282}$$

$$= -\frac{1}{2}\sum_{i=1}^{n} \frac{\partial h_t}{\partial \theta}\frac{1}{h_t}\left(1 - \frac{x_t^2}{h_t}\right) \tag{5.283}$$

En remarquant dans le précédent réarrangement que $\frac{x_t^2}{h_t} = \epsilon_t^2$, on trouve alors qu'à l'optimum, le score est nul. On rappelle que le score est l'espérance de la dérivée de la log-vraisemblance. En effet, on a, d'après ce qui vient d'être dit :

$$\mathbb{E}\left[\frac{x_t^2}{h_t}\right] = \mathbb{E}[\epsilon_t^2] = 1 \tag{5.284}$$

Afin d'obtenir les dérivées de la log-vraisemblance, il ne reste qu'à obtenir les dérivées de la variance conditionnelle par rapport à chacun des paramètres. Elles se calculent aisément :

$$\frac{\partial ht}{\partial \omega_0} = 1 \tag{5.285}$$

$$\frac{\partial ht}{\partial \omega_1} = x_{t-1}^2 \tag{5.286}$$

On trouve donc finalement les dérivées de la log-vraisemblance suivantes :

$$\frac{\partial lnL}{\partial \omega_0} = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{h_t}\left(1-\frac{x_t^2}{h_t}\right) \tag{5.287}$$

$$\frac{\partial lnL}{\partial \theta} = -\frac{1}{2}\sum_{i=1}^{n}\frac{x_{t-1}^2}{h_t}\left(1-\frac{x_t^2}{h_t}\right) \tag{5.288}$$

On est alors en mesure de mettre en oeuvre ce qui a été présenté dans le cadre du chapitre 5 : l'estimation se fait généralement par Newton-Raphson, en utilisant la matrice BHHH comme approximation de la variance de l'estimateur des paramètres. Il est cependant nécessaire d'en dire un peu plus : on a fait l'hypothèse que l'erreur du modèle (ϵ_t)suivait une loi normale. Qu'en est il si tel n'est pas le cas ? Greene (2002)[Chapitre 11] revient sur les implications de cette hypothèse : comme on l'a déjà précisé précédement, l'estimation par maximum de vraisemblance conditionel avec bruit gaussien conduit à l'obtention d'estimateurs consistent (i.e. qui converge presque surement vers les vraies valeurs des paramètres). Dans ce cas, on parle de *Pseudo Maximum de Vraisemblance*. Gourieroux et al. (1984) remarquent cependant que l'estimation de la matrice de variance/covariance des estimateurs par l'inverse de la matrice d'information de Fisher n'est pas bonne. Il est nécessaire d'uiliser l'approximation suivante :

$$\mathbb{V}[\theta] = H^{-1}FH^{-1} \tag{5.289}$$

avec :

$$H = -\mathbb{E}\left[\frac{\partial^2 lnL}{\partial\theta\partial\theta^\top}\right] \tag{5.290}$$

$$F = \mathbb{E}\left[\frac{\partial lnL}{\partial\theta}\frac{\partial lnL}{\partial\theta^\top}\right] \tag{5.291}$$

On remarque que la matrice 2×2 F est l'estimateur BHHH présenté au chapitre 5. La matrice F est en général très simple à calculer, puisque prise en espérance. Il suffit pour cela de déterminer les dérivées secondes de la log-vraisemblance. Dans le cas d'un ARCH(1), les calculs conduisent généralement aux résultats suivants :

$$\frac{\partial^2 lnL}{\partial \omega_0^2} = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{h_t^2}\left[2\frac{x_t^2}{h_t}-1\right] \tag{5.292}$$

$$\frac{\partial^2 lnL}{\partial \omega_0\partial \omega_1} = -\frac{1}{2}\sum_{i=1}^{n}\frac{x_{t-1}^2}{h_t^2}\left[2\frac{x_t^2}{h_t}-1\right] \tag{5.293}$$

$$\frac{\partial^2 lnL}{\partial \partial \omega_1^2} = -\frac{1}{2}\sum_{i=1}^{n}\frac{x_{t-1}^4}{h_t^2}\left[2\frac{x_t^2}{h_t}-1\right] \tag{5.294}$$

On ne calcule que l'une des dérivées croisées, dans la mesure où le lemme de Monge s'applique sans problème. En prenant l'espérance de ces dernières dérivées, et en remarquant encore une fois que $\mathbb{E}[\frac{x_t^2}{h_t}] = 1$, on obtient la matrice F :

$$F = \begin{pmatrix} -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{h_t^2} & -\frac{1}{2}\sum_{i=1}^{n}\frac{x_t^2}{h_t^2} \\ -\frac{1}{2}\sum_{i=1}^{n}\frac{x_t^2}{h_t^2} & -\frac{1}{2}\sum_{i=1}^{n}\frac{x_t^4}{h_t^2} \end{pmatrix} \tag{5.295}$$

On sait de plus calculer la matrice BHHH à partir des dérivées premières : on est alors en mesure d'estimer un modèle ARCH sans trop de difficulté. Dans la réalité (et dans R), l'estimation ne passe pas par Newton Raphson, mais par une méthode plus complexe garantissant de meilleurs résultats : la méthode BFGS. Celle-ci n'est pas développée ici.

5.3.4.2 Le cas d'un GARCH(1,1)

Tout ce qui vient d'être dit précédement au sujet de l'estimation des ARCH s'applique de la même façon aux modèles GARCH. La vraisemblance conditionnelle s'écrit encore une fois en constatant que la loi conditionnelle de x_t sachant h_t est connue :

$$x_t|_{h_t} \sim N(0, h_t) \tag{5.296}$$

Il s'agit donc du même point de départ qu'on modèle ARCH(1). L'estimation d'un modèle GARCH(1,1) nécessite cependant le calcul d'une dérivée première supplémentaire :

$$\frac{\partial h_t}{\partial \omega_2} = h_{t-1} \tag{5.297}$$

On obtient alors les dérivées de la vraisemblance suivantes :

$$\frac{\partial lnL}{\partial \omega_0} = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{h_t}\left(1-\frac{x_t^2}{h_t}\right) \tag{5.298}$$

$$\frac{\partial lnL}{\partial \omega_1} = -\frac{1}{2}\sum_{i=1}^{n}\frac{x_{t-1}^2}{h_t}\left(1-\frac{x_t^2}{h_t}\right) \tag{5.299}$$

$$\frac{\partial lnL}{\partial \omega_2} = -\frac{1}{2}\sum_{i=1}^{n}\frac{h_{t-1}}{h_t}\left(1-\frac{x_t^2}{h_t}\right) \tag{5.300}$$

On laisse à titre d'exercice le calcul des dérivées secondes. On renvoie à VonSachs and VanBellegem (2002) pour plus de précisions au sujet de l'estimation des GARCH. On propose le code R suivant, stricte application de ce qui vient d'être dit : il s'agit d'un code permettant d'estimer un GARCH(1,1) par Newton Raphson, avec matrice BHHH.

```
estim.garch<-function(theta,x){
G=matrix(1,3,1)
n=nrow(x)
check=theta
j=1;
while(sum(G^2)>0.0001){
# Computation of sigma2
sigma2=matrix(theta[1,1],n,1);
for (i in 2:n){sigma2[i,1]=theta[1,1]+theta[2,1]*sigma2[(i-1),1]+theta[3,1]*x[(i-1),1]^2}
comp=as.matrix((x^2-sigma2)/sigma2^2);
BHHH=cbind(comp[2:n,1],comp[2:n,1]*sigma2[1:(n-1),1],comp[2:n,1]*x[1:(n-1),1]^2);
H=(t(BHHH)%*%BHHH);
G[1,1]=sum(BHHH[,1]);
G[2,1]=sum(BHHH[,2]);
```

```
G[3,1]=sum(BHHH[,3]);
cat(j,"\n");
check=cbind(check,theta+solve(H)%*%G);
theta=theta+solve(H)%*%G;
j=j+1
}
var=sqrt(diag(solve(H)));
test=theta/var
return(list(theta=theta,check=check, test=test))
}
```

5.3.5 Premières Applications

Avant de se lancer dans des applications plus complexes utilisant quelques raffinements des modèles GARCH, on présente une application simple, visant à estimer un modèle GARCH sur les rentabilités mensuelles de l'indice DAX.

5.3.5.1 Etude de la volatilité sous-jacente de l'indice DAX

La figure 5.25 présente l'évolution de la rentabilité journlières du DAX depuis 2000, avec ses ACF et PACF. On remarque les quelques faits stylisés présentés plus haut : existence d'agrégats de volatilité et faiblesse de l'autocorrélation dans les rendements. Ce dernier point appelle à commentaire : l'existence d'une légère autocorrélation est en général le fait des séries longues. Celle-ci disparait aisément à mesure que l'on réduit la taille de l'échantillon et/ou que l'on utilise des données plus récentes. Quoiqu'il arrive, les prix des actifs sont en général martingale, et rien d'autre.

On étudie naturellement les ACF et PACF des rendements élevés au carré sur la figure 5.26. On remarque ce qui a été dit plus haut : l'autocorrélation dans les rendements au carré décroit lentement et la PACF admet quelques valeurs significativement différentes de 0. Ceci suggère naturellement l'estimation d'un modèle GARCH avec un ordre faible. On choisit ici délibérément un GARCH(1,1), en utilisant la fonction `R` fournie précédement.

La table 5.3.5.1 fournit les estimations d'un modèle GARCH(1,1) sur nos données. On remarque l'ensemble des paramètres est significatif à 95%. La figure 5.27 présente les résidus du modèle GARCH. On rappelle que les résidus d'un GARCH sont données par $\frac{x_t}{\sqrt{h_t}}$ et non par l'écart entre le modèle et les résidus. On constate que ces résidus ne présentent presque plus de phénomènes de cluster de volatilité. En revanche l'étude de la PACF et de l'ACF du carré des résidus (figure 5.28) conclue naturellement à l'insuffisance de l'ordre du GARCH choisit : il subsiste encore de la corrélation.

La figure 5.29 fournit l'évolution de la variance conditionnelle du DAX : encore une fois, la volatilité est une composante inobservable et les modèles GARCH constituent une approxiamtion permettant de la mettre à jour.

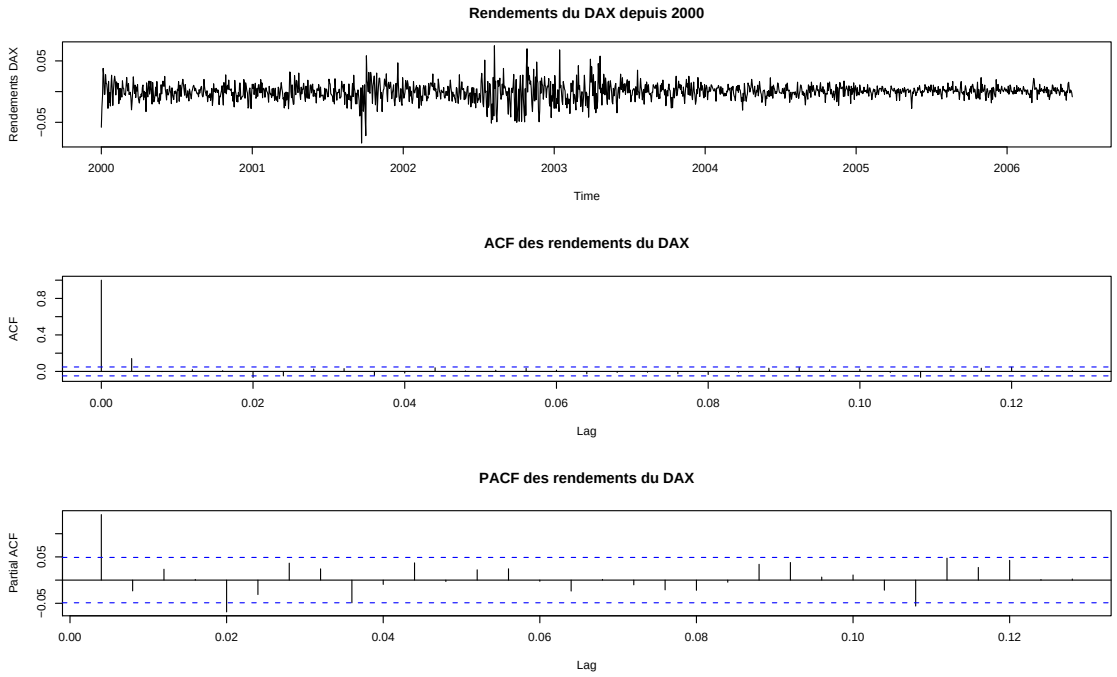

FIG. 5.25 – Rendements du DAX depuis 2000 : chronique, ACF et PACF

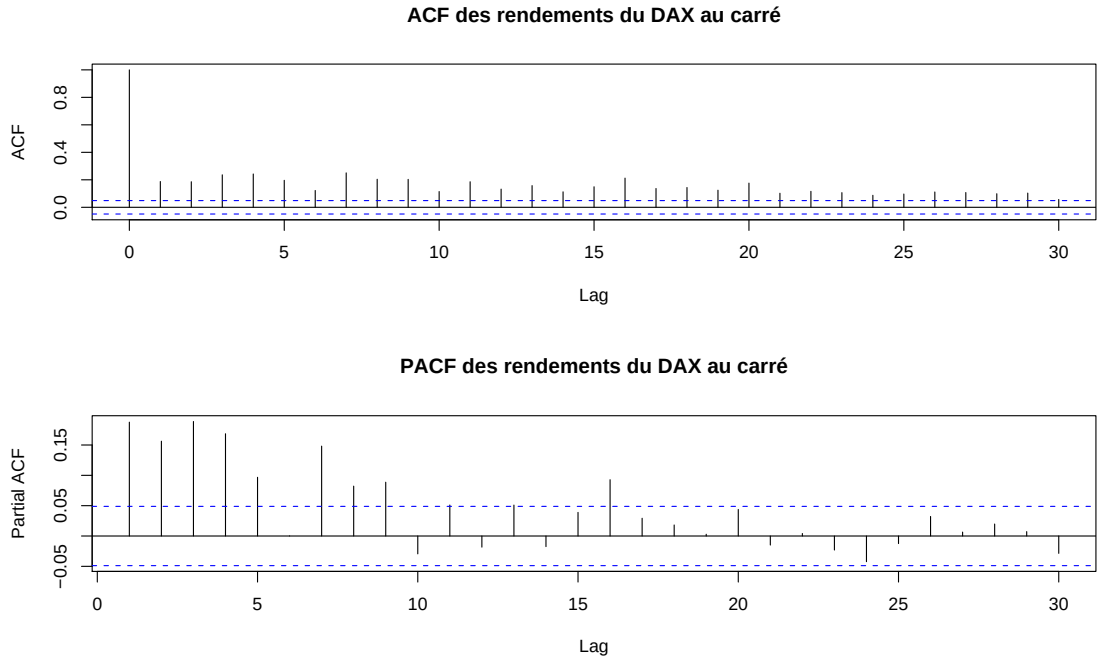

FIG. 5.26 – ACF et PACF des rendements du DAX élevés au carré

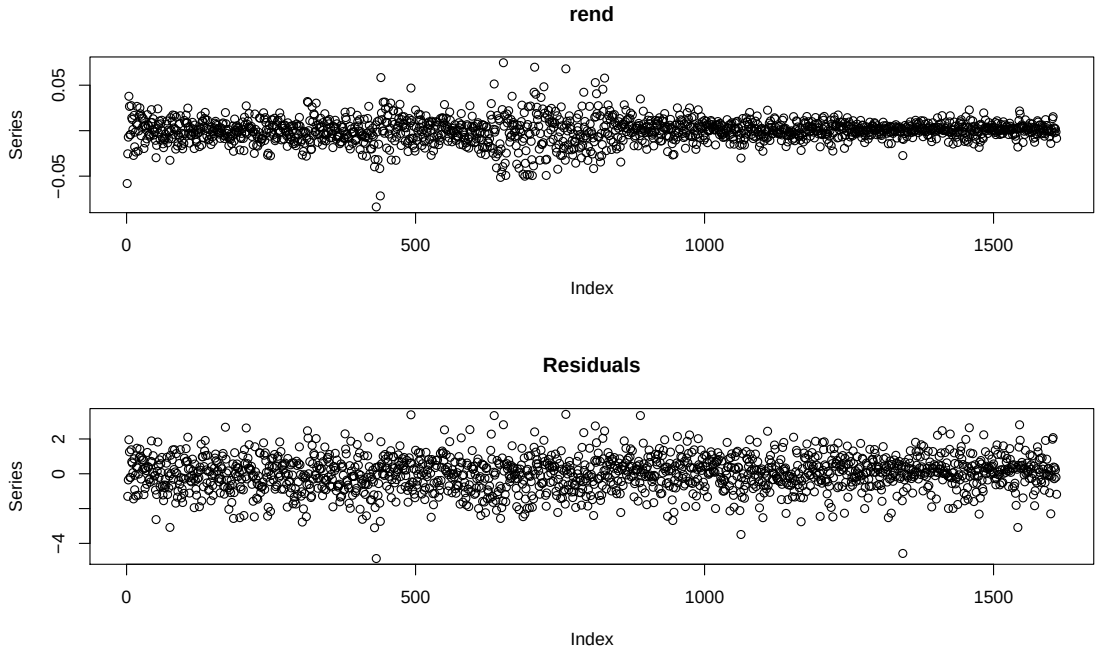

FIG. 5.27 – Rendements et résidus GARCH

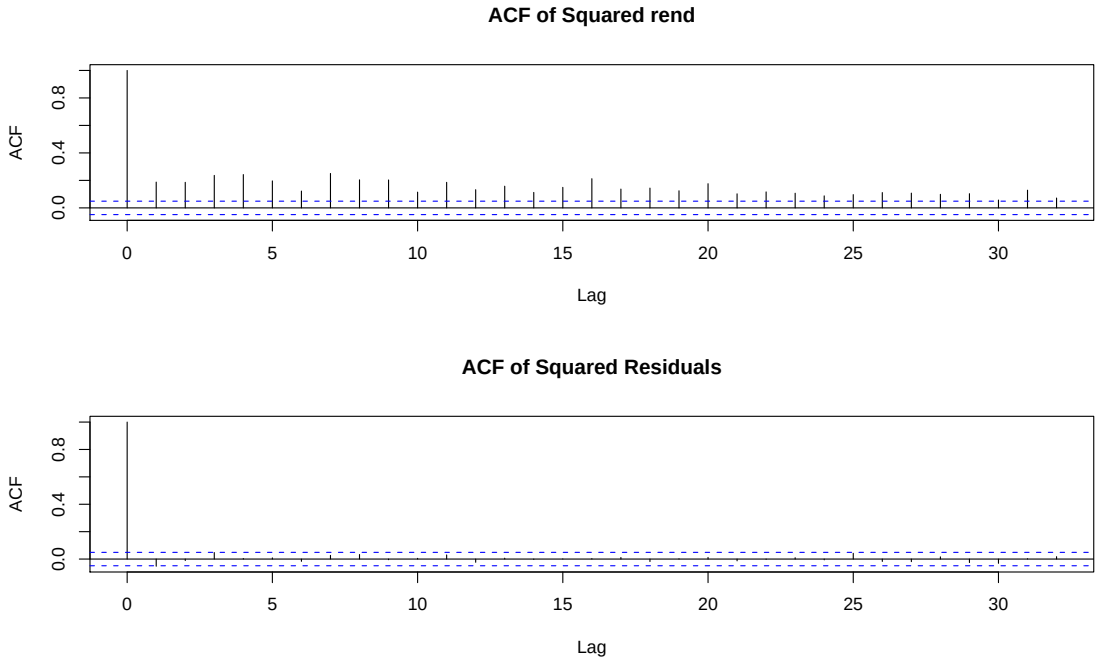

FIG. 5.28 – ACF et PACF des rendements du DAX et des résidus GARCH

	Estimate	**Std.Error**	**t-value**	**Pr**$(>|t|)$
ω_0	1.154e-06	3.047e-07	3.788	0.000152
ω_1	7.166e-02	8.113e-03	8.832	2,00E-16
ω_2	9.208e-01	8.827e-03	104.324	2,00E-16

TAB. 5.3 – Estimation d'un GARCH(1,1) sur données historiques du DAX

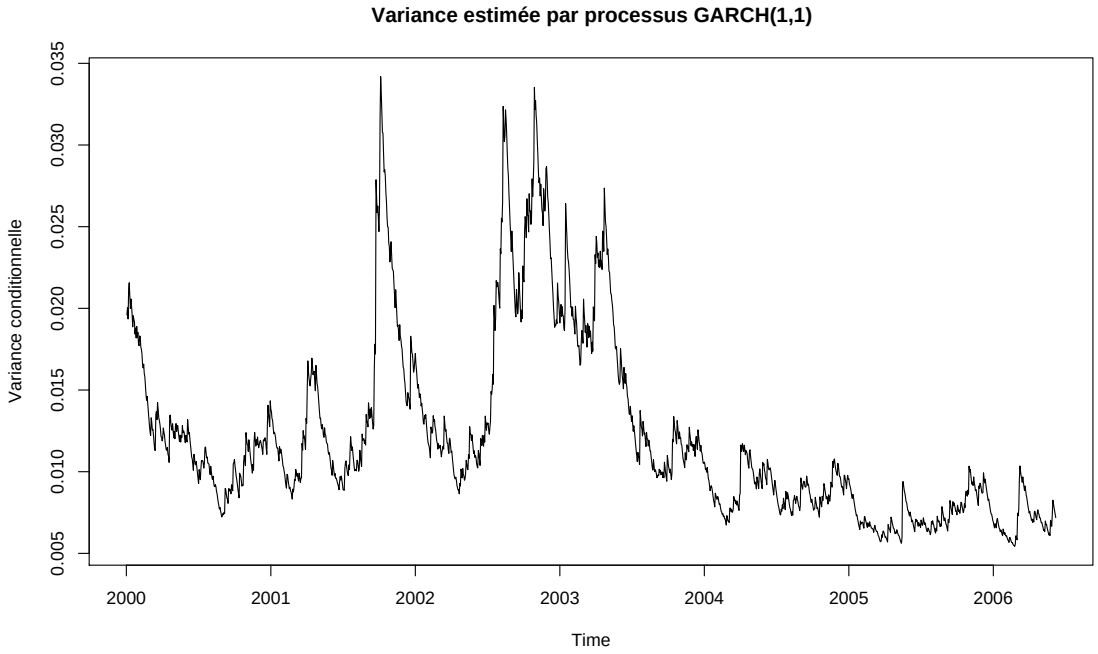

FIG. 5.29 – Volatilité estimée par processus GARCH sur DAX

5.3.5.2 Formule de Black Scholes avec processus GARCH : version ad-hoc

On propose parfois d'utiliser la volatilité conditionnelle dans le cadre de la formule de Black Schole : il s'agit d'une version ad-hoc de raffinements proposés par Heston et Nandi plus tard.

A l'aide d'un processus GARCH, on obtient $\sqrt{h_t}$ la volatilité conditionnelle pour chaque date d'existence de l'option étudiée. On s'interroge alors : l'erreur générée, i.e. l'écart entre prix de marché et prix Black Scholes, est elle moindre lorsque l'on utilise la l'écart type des rendements comme mesure de la volatilié ou la racine de la variance conditionnelle obtenue dans le cadre d'un modèle GARCH ? Pour répondre à cette question, on utilise généralement comme critère le *Root Mean Square Error*, c'est à dire la racine de l'erreur au carré moyenne. En notant P_t le prix réel de l'option et $\widetilde{P}_t$ le prix tel que la formule de Black Schole le propose, le RMSE est alors :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \widetilde{P}_i)^2} \tag{5.301}$$

Le tableau suivant fourni les erreurs des versions standards et GARCH ad-hoc générées par la formule de Black-Scholes.

Strike	Erreur GARCH	Erreur BS standard
4000	116,3457	115,7618
4300	25,5418	22,20917
4400	24,40097	20,82484
4500	21,94736	18,3525
4600	19,0568	15,68696
4700	16,8634	14,15297
4800	16,16095	14,37239
4900	15,71973	13,48666
5000	11,66557	9,937144

Le modèle BS standard semble surperformer nettement le modèle GARCH ad-hoc. Si cet avis est sans appel à la lecture de la table, l'observation du graphique 5.30 conduit à un tout autre jugement. On observe que le prix avec volatilité GARCH suit globalement bien mieux le vrai prix de l'option, sauf dans quelques rares cas pour lesquels il s'en écarte singulièrement, conduisant aux résultats donnés dans la table. L'explication à ce phénomène est simple : la formule de BS est très sensible à la volatilité. Pour les dates présentant ces écarts importants, la volatilité GARCH connait un saut, conduisant le prix BS à s'écarter du vrai prix de l'option.

5.3.5.3 Prévision de la volatilité et ses usages

Une fois un modèle GARCH calibré, il est possible de procéder à une analyse plus fine de la volatilité d'un sous-jacent. La volatilité est principalement utilisée dans le cadre de la Value at Risk et dans le cadre des modèles d'option. L'avantage d'un modèle GARCH est qu'il permet de construire une prévision naïve de la volatilité future du marché, sur une période courte. On présente ici quelques questions relatives à la prévision de la volatilité du rendement d'un actif.

Dans le cadre d'un modèle GARCH(1,1), de la forme suivante :

$$x_t = \sqrt{h_t}\epsilon_t \tag{5.302}$$
$$h_t = \omega_0 + \omega_1 x_{t-1}^2 + \omega_2 h_{t-1} \tag{5.303}$$

avec $\epsilon_t \sim N(0,1)$, il est aisé d'obtenir une prévision en date t de la variance conditionnelle en date $t+1$:

$$\mathbb{E}[h_{t+1}|h_t] = \omega_0 + \omega_1 \mathbb{E}[x_t^2|h_t] + \omega_2 h_t \tag{5.304}$$
$$= \omega_0 + (\omega_1 + \omega_2) h_t \tag{5.305}$$

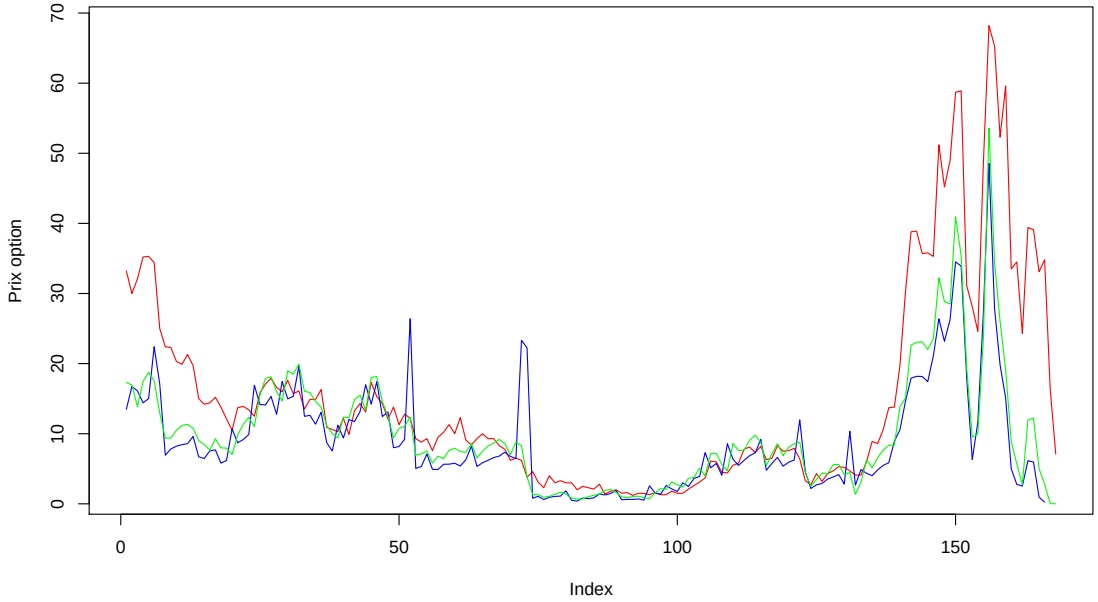

FIG. 5.30 – Prix du call DAX (rouge), prix BS standard (vert) et prix GARCH BS ad-hoc

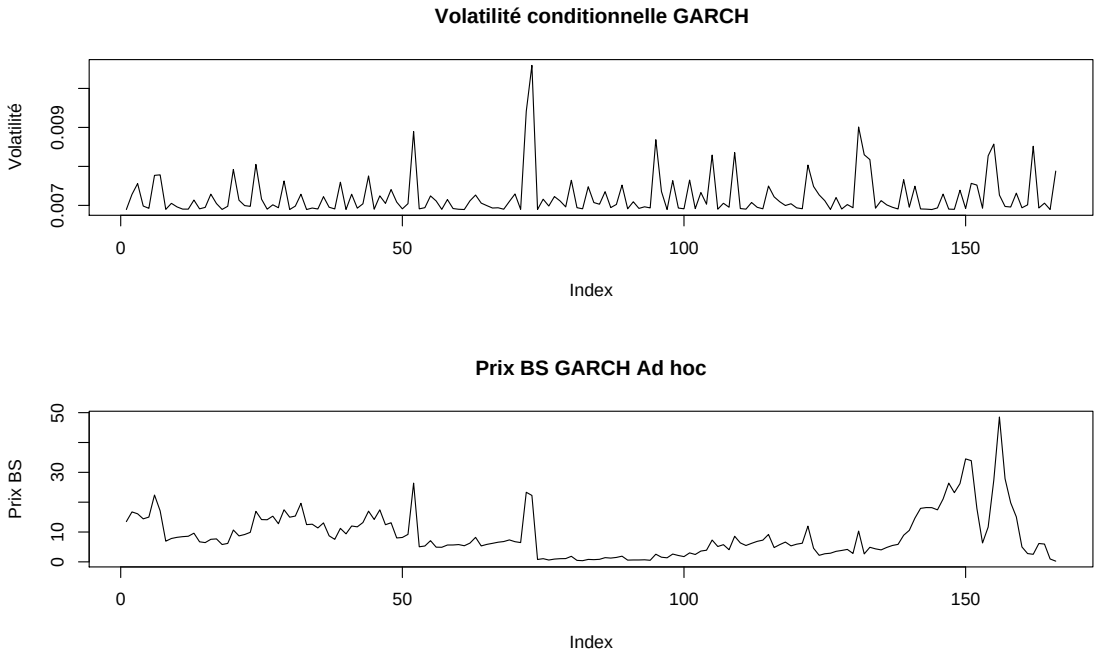

FIG. 5.31 – Prix théorique contre volatilité conditionnelle pour une option hors de la monnaie.

On déduit de façon récursive la suite des variances conditionnelles :

$$\begin{aligned}
\mathbb{E}[h_{t+2}|h_t] &= \omega_0 + \omega_1 \mathbb{E}[x_{t+1}^2|h_t] + \omega_2 \mathbb{E}[h_{t+1}|h_t] && (5.306)\\
&= \omega_0 + \omega_1 \mathbb{E}[h_{t+1}|h_t] + \omega_2 \mathbb{E}[h_{t+1}|h_t] && (5.307)\\
&= \omega_0 + (\omega_1 + \omega_2)\mathbb{E}[h_{t+1}|h_t] && (5.308)\\
&= \omega_0 + (\omega_1 + \omega_2)(\omega_0 + (\omega_1 + \omega_2)h_t) && (5.309)\\
&= \omega_0(1 + (\omega_1 + \omega_2)) + (\omega_1 + \omega_2)^2 h_t) && (5.310)\\
& && (5.311)
\end{aligned}$$

Comme on l'a montré lors du calcul da variance non conditionnelle, en procédant par récursion, on trouve la formule générale suivante :

$$\mathbb{E}[h_{t+h}|h_t] = \omega_0 \sum_{i=0}^{h-1} (\omega_1 + \omega_2)^i + (\omega_1 + \omega_2)^h h_t) \quad (5.312)$$

$$(5.313)$$

Comme dans le cas des ARMA, la prévision s'écrase rapidement contre la variance non conditionnelle. Ainsi l'horizon prédictif des modèles GARCH est limité : ces modèles sont bien mieux adaptés pour fournir une mesure de la volatilité, qui, rappelons le, est par essence inobservable.

Une des utilisations possibles de l'algorithme de prévision de la volatilité qui vient d'être développé conduit naturellement à une prévision de la Value at Risk. On rappelle brievement le sens et le mode de calcul de cette mesure de risque, sur la base de ce qui en est dit dans Tsay (2002).

5.3.5.3.1 La VaR

La VaR est le benchmark le plus utilisé quand il s'agit de juger de l'exposition maximale au risque de marché produit par une position donnée sur le marché. Le risque de marché est un terme générique permettant de décrire plusieurs situations possibles sur le marché, selon le montant des engagements produits. Une position conduisant à investir dans un seul et unique actif n'induit pas la même exposition au risque de marché qu'une position prise sur différents actifs *simultanément*. Dans le cas où l'on détient un portefeuille d'actifs, et en dépis des mécanismes bien connus de la diversification, la dépendance existant entre les différents actifs du portefeuille fait peser sur sa rentabilité une menace particulière, souvent présenté sous le titre de risque de corrélation.

La VaR peut être simplement définie comme la perte maximale liée à une position de marché, sur une fenetre de temps particulière et pour un niveau de probabilité donné : en supposant que l'on connaisse la loi du rendement du portefeuille, on est en mesure de donné le montant maximal de perte que l'on peut réaliser, dans le cadre d'une probabilité égale à 95%. Cela signifie simplement que dans 95% des cas, la perte ne devrait pas dépenser cette mesure de risque. Evidement, la perte ne se juge pas en terme de rentabilité ! Il est donc nécessaire de passer de la perte en terme de rentabilité à celle en terme de prix, en passant à l'exponentielle.

L'idée est donc : connaissant la loi de ${}_tr_{t+h}$, i.e. des rendements sur la période de temps allant de t à $t+h$ (la fenêtre de temps), on cherche la chutte maximale que peut connaitre le rendement du portefeuille sur ce même intervalle, avec une probabilité égale à 90% ou 95%. Plus formellement, ceci revient à chercher :

$$P\left({}_tr_{t+h} \geq VaR\right) = 95\% \tag{5.314}$$

Il est indifférent de rechercher la valeur VaR dans le cadre de la précédement formule ou dans le cas suivant :

$$P\left({}_tr_{t+h} \leq VaR\right) = 5\% \tag{5.315}$$

Ceci vient simplement du fait que :

$$P\left({}_tr_{t+h} \geq VaR\right) = 1 - P\left({}_tr_{t+h} \leq VaR\right) \tag{5.316}$$

En travaillant en terme de densités, les inégalités strictes et large ne produisent aucune différence de calcul. La VaR correspond simplement au quantile à 5% de la loi des rendements. A la différence des tests statistiques comme le test de Student qui sont des tests bilatéraux, il s'agit ici d'un quantile "unilatéral" : il n'est donc pas nécessaire de calculer un quantile à $x\%/2$, comme on le fait pour un test de Student.

Une fois ce quantilé obtenu (sous `R`, la function `quantile` permet d'obtenir une estimation non paramétrique de ce quantile), il est alors possible d'obtenir une estimation de la VaR en terme de prix, i.e. en terme de perte nette :

$$P\left({}_tr_{t+h} \leq VaR\right) = 5\% \tag{5.317}$$

$$\Leftrightarrow P\left(exp\{{}_tr_{t+h}\} \leq exp\{VaR\}\right) = 5\% \tag{5.318}$$

$$\Leftrightarrow P\left(P_t exp\{{}_tr_{t+h}\} \leq P_t exp\{VaR\}\right) = 5\% \tag{5.319}$$

$$\Leftrightarrow P\left(\alpha P_{t+h}\} \leq \alpha P_t exp\{VaR\}\right) = 5\% \tag{5.320}$$

On obtient ainsi la perte nette maximale obtenue pour une probabilité égale à 95%. Ceci amène plusieurs commentaires :

- α est ici un coefficient multiplicatif traduisant le montant de l'engagement dans le titre.

- Cette probabilité est en réalité une probabilité conditionnelle : sachant αP_t, le montant net de l'engagement à la date t dans l'actif étudié, quel serait la perte maximale rencontrée dans 95% des cas pour un horizon h.

- Il s'agit d'une VaR dans le cas d'une position courte : on a ici acheté le titre, et seule une baisse de son cours (un enchainement de rentabilité négatives) peut nous conduire à perdre de l'argent. On opère le raisonnement inverse dans le cas où l'on est placé dans le cadre d'une position longue (vendeur à découvert) dans le titre étudié. La perte dans ce cas viendra naturellement d'une montée du cours de l'actif. Il est alors nécessaire de calculer la hausse maximale que le titre peut enregistrer dans 95% des cas.

On rappelle simplement que la fonction quantile pour une variable aléatoire x est la fonction telle que :

$$x_p = inf\{x|F(x) \geq p\} \tag{5.321}$$

où p est une probabilité donnée.

Comme présenté dans Tsay (2002), le calcul de VaR est loin d'être aisé. Il nécessite de nombreux inputs :

- La probabilité utilisée : en général 5% ou 1%.
- L'horizon de calcul : une VaR à un mois n'a souvent rien à voir avec une VaR à 1 jours.
- La fréquence des données utilisées : travaille-t-on sur des rendements journaliers, hebdomadaires, mensuels ?
- La fonction de répartition des rendements : la fonction de répartition non conditionnelle peut être obtenue à l'aide d'une estimation non paramétrique alors qu'un fonction paramétrique peut être obtenue plus directement.
- Le montant et le sens de la position dans l'actif étudié.
- Dans le cas où il s'agit d'un portefeuille, il est nécessaire de déterminer la structure de dépendance liant les différents actifs en portefeuille.

Nombre de ces items sont en général fixés par le régulateur (accords Bâle II par exemple). Notons que le calcul de la VaR sert aux comptables pour déterminer le montant de provisions à passer pour couvrir le risque de marché. Il s'agit d'un montant placé dans les capitaux propres (Provisions pour risques et charges) afin de faire face à un décrochage violent mais temporaire du marché. Inutile de rappeler qu'un retournement puissant et durable du marché conduit à des effets systémiques d'une ampleur telle qu'il est impossible d'y faire face avec ce simple montant en poche. Ainsi, d'un point de vu pratique, on fixe en général de façon institutionnelle un montant de VaR qui sera passé en provision pour l'année à venir. Il sert ensuite à piloter le montant et le sens des engagements dans les actifs en portefeuille : on détermine la VaR de l'ensemble des titres détenus, et on vérifie que celle-ci ne dépasse pas le montant conventionnellement fixé. Dans le cas où la banque est engagée dans différents marchés, et possède donc différents *desks*, il est alors nécessaire d'allouer une part de la VaR conventionnelle à chacun de ces desks.

5.3.5.3.2 Calcul de la VaR à l'aide de modèles GARCH Comme le fait très justement remarqué Tsay (2002), la VaR est une prévision de la perte possible pour un horizon donné. Elle devrait toujours et partout calculée à l'aide d'une distribution prédictive (*predictive distribution*) du futur des rendements du portefeuille. Par exemple, une VaR pour un horizon d'un jour utilisant des rendements journaliers r_t devrait normalement être calculée en utilisant la distribution prédictive du rendements r_{t+1}, sachant l'information disponible à la date t. On imagine alors qu'il serait alors nécessaire de tenir compte de l'erreur possible d'estimation, comme on le fait à chaque fois que l'on travaille d'un point de vue économétrique. Dans la réalité, les méthodes utilisés ne se base pas sur ces distributions prédictives, et ne tiennent pas compte des

erreurs d'estimation. On propose néanmoins de montrer comment les modèles GARCH peuvent être utilisés afin de déterminer une VaR.

On reviendra plus loin sur la méthodologie proposées par RiskMetrics (méthode EWMA), dans la mesure où elle corresond à un modèle GARCH "raffiné" (modèle GARCH intégré).

5.3.5.3.2.1 VaR dans le cas univarié

Il est très aisé de déterminer une VaR pour un unique sous-jacent à l'aide d'un modèle GARCH. On prend encore ici le cas simple d'un GARCH(1,1), appliqué au CAC40 (les données sous tirées de la base EuStockMarkets, disponible dans `R`). On estime sur les rendements du CAC le modèle suivant :

$$r_t^{CAC} = \sqrt{h_t^{CAC}}\epsilon_t^{CAC} \tag{5.322}$$

$$h_t^{CAC} = \omega_0^{CAC} + \omega_1^{CAC} r_{t-1}^{CAC} + \omega_2^{CAC} h_{t-1}^{CAC} \tag{5.323}$$

Avec $\epsilon^{CAC} \sim N(0,1)$. Les estimations obtenues sont les suivantes :

	Estimate	**Std. Error**	**t-value**	**p-value**
ω_0	0,00001178	2,675E-06	4,406	0
ω_1	0,05916	0,01124	5,261	0
ω_2	0,8439	0,0315	26,791	0

A l'aide de ces résultats, on peut déterminer une VaR pour l'horizon souhaité. Pour une la VaR à un jour, on cherche :

$$P\left(r_{t+1}^{CAC} \leq VaR_1^{CAC}\right) = 5\% \tag{5.324}$$

$$\Leftrightarrow P\left(\sqrt{h_{t+1}^{CAC}}\epsilon_{t+1}^{CAC} \leq VaR_1^{CAC}\right) = 5\% \tag{5.325}$$

$$\Leftrightarrow P\left(\epsilon_{t+1}^{CAC} \leq \frac{VaR_1^{CAC}}{\sqrt{h_{t+1}^{CAC}}}\right) = 5\% \tag{5.326}$$

ϵ suivant une loi normale centrée réduite, on connait le quantile à 5% : il vaut -1,64. On en déduit donc :

$$\Leftrightarrow \frac{VaR_1^{CAC}}{\sqrt{h_{t+1}^{CAC}}} = -1,64 \tag{5.327}$$

$$\Leftrightarrow VaR_1^{CAC} = -1,64\sqrt{h_{t+1}^{CAC}} \tag{5.328}$$

On en déduit alors la VaR en terme de pertes :

$$P\left(r_{t+1}^{CAC} \leq -1,64\sqrt{h_{t+1}^{CAC}}\right) = P\left(exp\{r_{t+1}^{CAC}\} \leq exp\{-1,64\sqrt{h_{t+1}^{CAC}}\}\right) \quad (5.329)$$

$$= P\left(P_t^{CAC} exp\{r_{t+1}^{CAC}\} \leq P_t^{CAC} exp\{-1,64\sqrt{h_{t+1}^{CAC}}\}\right) \quad (5.330)$$

$$= P\left(P_{t+1}^{CAC} \leq P_t^{CAC} exp\{-1,64\sqrt{h_{t+1}^{CAC}}\}\right) \quad (5.331)$$

La fin des calculs s'appuie sur des données numériques : on a besoin du prix du CAC en date t ainsi que la prévision de la volatilité à la date $t+1$. On a procédé à ces quelques calculs sous `R`. On présente le résultat des opérations : on a calculé pour toutes les dates la VaR à un jour. Elle est présentée en figure 5.32.

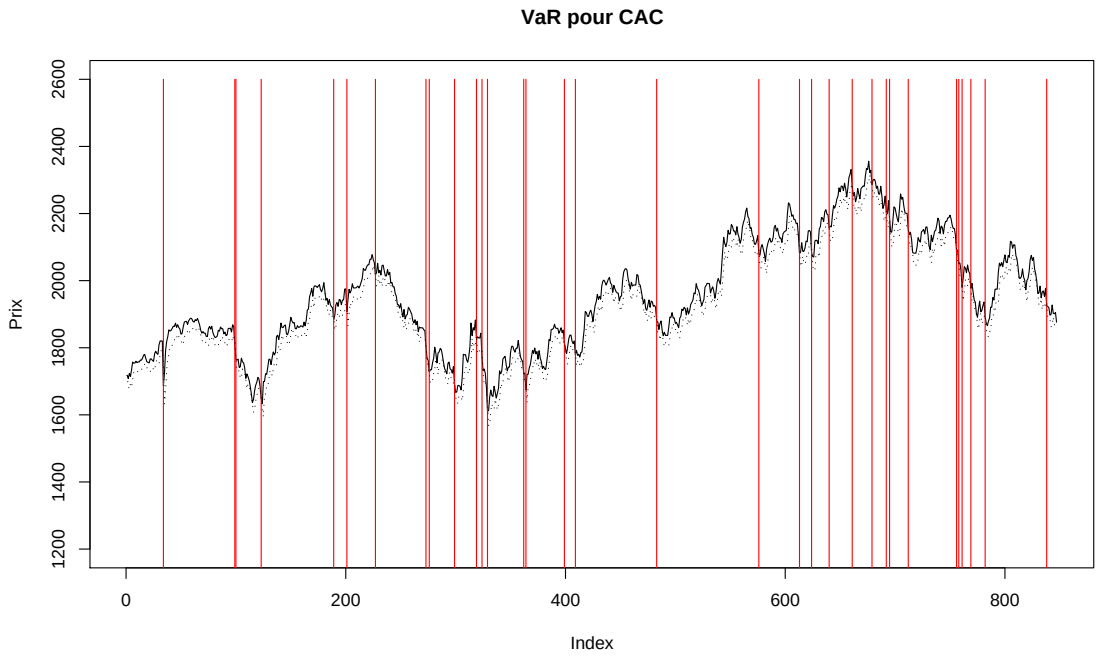

FIG. 5.32 – Value at Risk à un jour

Dans le cas où l'on souhaite obtenir une VaR pour un horizon $t+h$ quelconque, les choses se complexifient légèrement. On cherche à définir une borne basse pour P_{t+h}. On travaille en rendements, i.e. sur :

$$\tilde{r}_{t+h} = ln\frac{P_{t+h}}{P_t} \quad (5.332)$$

Il est alors nécessaire de travailler sur la distribution conditionnelle prédictive, i.e. sur

la loi de $\tilde{r}_{t+h}$. Pour cela, on remarque que :

$$ln\frac{P_{t+h}}{P_t} = ln\left(\frac{P_{t+1}}{P_t}\frac{P_{t+2}}{P_{t+1}}...\frac{P_{t+h}}{P_{t+h-1}}\right) \tag{5.333}$$

$$= \sum_{i=1}^{h} ln\left(\frac{P_{t+i}}{P_{t+i-1}}\right) \tag{5.334}$$

$$= \sum_{i=1}^{h} r_{t+i} \tag{5.335}$$

On peut déterminer la loi conditionnelle de cette somme de rendements :

$$\mathbb{E}\left[\sum_{i=1}^{h} r_{t+i}\right] = \sum_{i=1}^{h} \mathbb{E}\left[r_{t+i}\right] \tag{5.336}$$

$$\mathbb{V}\left[\sum_{i=1}^{h} r_{t+i}\right] = \sum_{i=1}^{h} \mathbb{V}\left[r_{t+i}\right] \tag{5.337}$$

$$= \sum_{i=1}^{h} \mathbb{E}\left[h_{t+i}\right] \tag{5.338}$$

L'ensemble de ces précédents calculs sont en réalité conditionnellement à l'information disponible à la date t. Pour alléger les notations, on a éliminé les notations conditionnelles. Il ne reste plus qu'à utiliser ce qui a été dit sur la prévision des variances au début de cette section pour être capable de déterminer la VaR sur les rendements futurs, conditionnellement à l'information disponible à la date t. On sait que :

$$\tilde{r}_{t+h} \sim N\left(0, \sum_{i=1}^{h} \mathbb{E}\left[h_{t+i}\right]\right) \tag{5.339}$$

On en déduit à la VaR en constatant que :

$$\frac{\tilde{r}_{t+h}}{\sqrt{\sum_{i=1}^{h} \mathbb{E}\left[h_{t+i}\right]}} \sim N(0,1) \tag{5.340}$$

D'où la VaR à 5% est égale à :

$$VaR = -1,64\sqrt{\sum_{i=1}^{h} \mathbb{E}\left[h_{t+i}\right]} \tag{5.341}$$

Comme précédement, on en déduit la VaR en terme de pertes nettes :

$$VaR_{P_{t+h}} = P_t exp\{-1,64\sqrt{\sum_{i=1}^{h} \mathbb{E}\left[h_{t+i}\right]}\} \tag{5.342}$$

Il suffit donc de calculer la somme des variances conditionnelles pour déterminer ensuite une VaR forward pour un horizon h. On représente la VaR et les moments où celle-ci

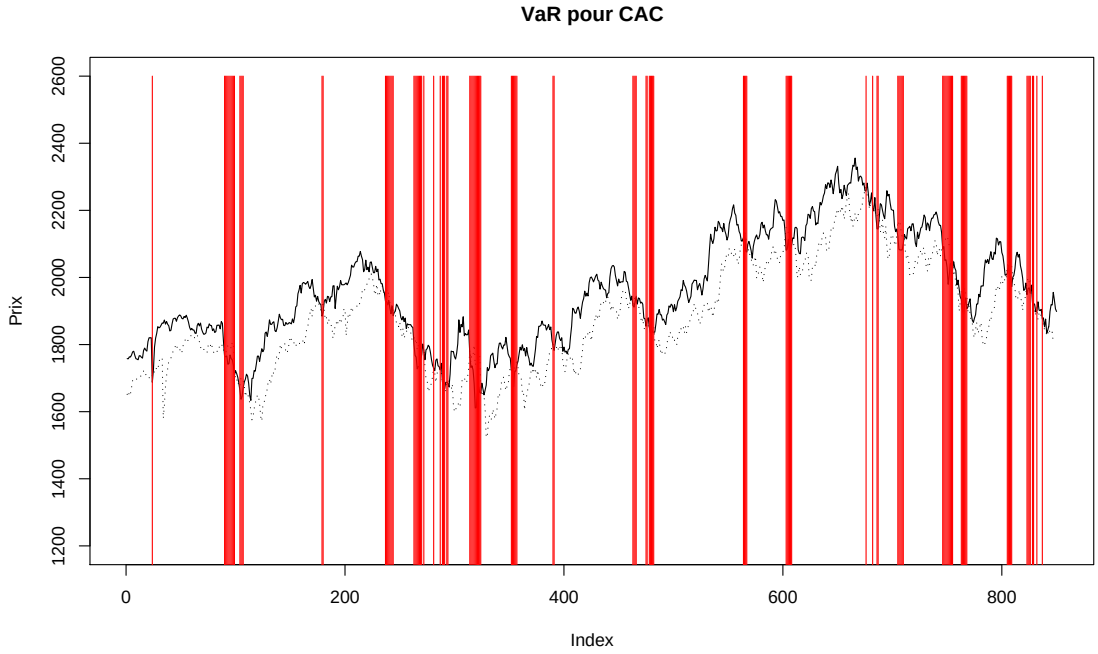

FIG. 5.33 – Value at Risk à dix jour

est violée sur la figure 5.33.

Les performances de l'approche à un jour et à dix jours appellent un simple commentaire. Dans le cas d'un VaR à un jour, le pourcentage de dépassement de la VaR est de 0,036 alors que dans le cas à 10 jours ce pourcentage est de 0,1168, ce qui est largement supérieur aux 5% autorisés. Ceci s'explique simplement par le fait que la prévision de la variance pour les modèles GARCH converge rapidement vers la la variance non conditionnelle, appauvrissant l'impact de la métode utilisée. Les modèles GARCH sont considérés comme des modèles à mémoire courte, i.e. qu'il retourne rapidement à la distribution non conditionnelle du processus. La méthode EWMA de Riskmetricks repose sur un modèle dit intégré, permettant de prendre en compte cet aspect mémoire longue de la volatilité.

5.3.5.3.2.2 VaR dans le cas bivarié : VaR par simulation

On se préocuppe à présent d'introduire un certain nombre d'idées concernant le calcul de la VaR d'un portefeuille de titre. Ces questions sont essentielles : ce sont elles qui sont en général utiles au quotidien des risks managers, dans la mesure où l'activité d'une banque ne se limite jamais à un seul et unique actif.

Avant toute chose, il est important de remarqué que la VaR d'un portefeuille, i.e. d'une somme d'actifs pondérés, n'est pas jamais la sommes pondérée des VaR des différents acifs. Ce qui rend le calcul de VaR d'un portefeuille complexe est la dépendance existante entre les différents actifs composant le portefeuille. Dans le cadre d'un modèle

conditionnellement gaussien comme c'est le cas pour un processus GARCH, cette dépendance est simplement mesurée par le coefficient de corrélation entre les deux termes d'erreurs. Soit deux actifs de rendement r_1 et r_2 suivant des processus GARCH(1,1) :

$$\begin{cases} r_{1,t} = \sqrt{h_{1,t}}\epsilon_{1,t} \\ h_{1,t} = \omega_{1,0} + \omega_{1,1} r_{1,t-1}^2 + \omega_{1,2} h_{1,t-1} \end{cases} \tag{5.343}$$

$$\begin{cases} r_{2,t} = \sqrt{h_{2,t}}\epsilon_{2,t} \\ h_{2,t} = \omega_{2,0} + \omega_{2,1} r_{2,t-1}^2 + \omega_{2,2} h_{2,t-1} \end{cases} \tag{5.344}$$

avec $(\epsilon_{i,t})_{i=1,2} \sim N(0,1)$ et $corr(\epsilon_{1,t}, \epsilon_{2,t}) = \rho$.

On se propose de déterminer la VaR de la somme de ces actifs, non par calcul direct, mais par simulation. Connaissant ρ, on est en mesure de simuler des ϵ corrélés, puis de simuler les processus r_1 et r_2 conditionnellement à l'information disponible en t. Enfin, on est en mesure de retrouver la VaR en terme de perte nette, comme on l'a fait précédement, en prenant l'exponentielle de chaque rendements.

On applique cette méthode à un portefeuille constitué d'une unité de DAX et d'une unité de CAC. On estime un GARCH(1,1) sur les rendements de ces actifs. Les résultats sont présentés dans la table suivante :

		Estimate	**Std. Error**	**t-value**	**p-value**
CAC	ω_0	0,00001178	0,000002675	4,406	0
	ω_1	0,05916	0,01124	5,261	0
	ω_2	0,8439	0,0315	26,791	0
DAX	ω_0	0,000004639	0,000000756	6,137	0
	ω_1	0,06833	0,01125	6,073	0
	ω_2	0,8891	0,01652	53,817	0

Une fois ceci fait, on estime la corrélation entre les résidus, en supposant que celle-ci n'est différente de 0 qu'instantanément (pas de corrélation entre les résidus pour un retard quelconque). On commence alors les simulations :

1. En partant d'un point donné dans le temps, on simule h ϵ_1 et h ϵ_2, pour l'instant décorrélés.
2. On transforme ces deux vecteurs de résidus, de façon à ce qu'ils soient corrélés. En notant M la matrice de corrélation entre ces deux résidus que l'on souhaite obtenir, on a :

$$M = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{5.345}$$

On utilise alors une décomposition de Choleski : il s'agit de déterminer une matrice F, telle que $F^\top F = M$. Pour obtenir deux séries d'innovations corrélés, il suffit

de multiplier la matrice composée de deux colonnes contenant les résidus simulés et pour l'instant décorrélés par F. En notant ϵ la matrice suivante :

$$\epsilon = \begin{pmatrix} \epsilon_{1,1} & \epsilon_{1,2} \\ \epsilon_{2,1} & \epsilon_{2,2} \\ \vdots & \vdots \\ \epsilon_{h,1} & \epsilon_{h,2} \end{pmatrix} \tag{5.346}$$

On calcule donc le produit ϵF.

3. On calcule ensuite de façon récursive les différentes valeurs de r_1 et r_2, en calculant au préalable la valeur de la variance conditionnelle.
4. Finalement, on calcule $\tilde{r_1}$ et $\tilde{r_2}$, la somme des rentabilités calculée pour le titre 1 et 2.

On répète ces simulations pour un nombre raisonnable de fois (un millier de fois conduit à des estimations précises). Enfin, on détermine la VaR de $\tilde{r_1}$ et $\tilde{r_2}$ à 5% de façon non paramétrique (commande `quantile` sous `R`). Une fois ces VaR univariées déterminées, on est en mesure de calculer la VaR en terme de perte nette du portefeuille détenu. Il suffit de calculer :

$$VaR_P = P_t^{CAC} \times exp\{VaR^{CAC}\} + P_t^{DAX} \times exp\{VaR^{DAX}\} \tag{5.347}$$

où VaR^i est la VaR calculée sur la somme des rendements du titre i. On répète l'ensemble des opérations pour toutes les dates d'intéret dans une approche backtesting.

On a procédé à une application sur le portefeuille CAC+DAX pour une VaR à 10 jours. On présente les résultats obtenus en figure 5.34. Les résultats semblent intéressants. Le principal problème reposant sur le fait que lorsque l'on observe une chutte brutale du prix, on est conduit à surestimer la VaR 10 jours après. Ce résultat est naturelle dans la mesure où la VaR à 10 jours est, à un coefficient multiplicatif pret, le prix d'aujourd'hui.

L'intérêt principal d'une démarche basée sur des simulations est qu'elle permet, en travaillant sur des séries univariées, d'incorporer la dépendance existant entre actifs. Mieux, comme on le détaillera dans la section sur les GARCH, on est implicitement amené à faire varier la covariance conditionnelle entre les deux actifs, en dépis du fait que l'on ai fixé la corrélation/covariance entre les résidus une fois pour toute. Le principal inconvénient est que, comme à chaque fois que l'on recourt à des simulations, le temps de calcul est plus long que dans le cas univarié simple, présenté plus haut.

Notons pour terminer qu'il est possible d'effectuer l'ensemble de ce qui vient d'être dit sur la base de calculs explicites. Dans la mesure où les deux actifs ont des rendements conditionnellement gaussiens, et en constatant que :

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2Cov(X,Y) \tag{5.348}$$

on est en mesure de décrire la loi conditionnelle de l'ensemble des deux rendements. En effet, on sait que :

$$\mathbb{E}[r_{1,t} + r_{2,t}|h_t] = 0 \tag{5.349}$$

$$\mathbb{V}[r_{1,t} + r_{2,t}|h_t] = h_{1,t} + h_{2,t} + 2\sqrt{h_{1,t}}\sqrt{h_{2,t}}\rho \tag{5.350}$$

$$\tag{5.351}$$

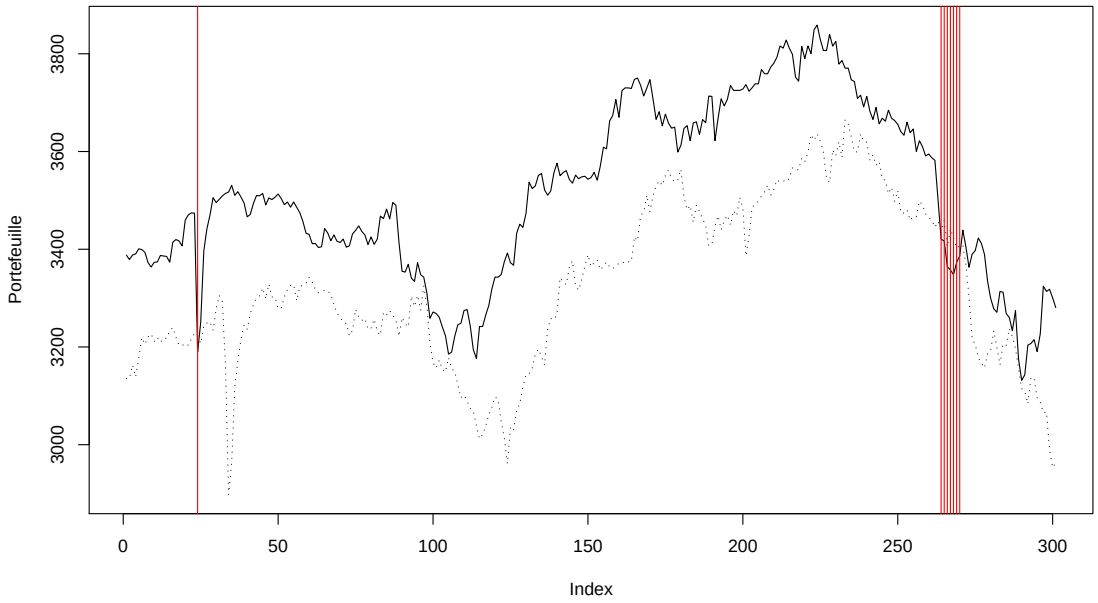

FIG. 5.34 – Value at Risk multivariée à dix jour

Il est alors possible d'écrire la VaR pour l'horizon souhaité, sans trop de problèmes (théoriques). Il évident que d'un point de vue pratique, l'utilisation de la corrélation comme mesure de dépendance est plus qu'un pari risqué ! On rappelle simplement que la corrélation n'est qu'une mesure de l'existance d'un lien linéaire entre actifs. Ces questions sont à mon sens - pour l'instant - bien trop avancées pour être intégrées dans ce cours. Les problèmes de mesure de dépendance sont abordées dans Embrechts et al. (1999) et Embrechts et al. (2001). L'une des réponses actuelles (quoique ceci commence à dater) à ces question de corrélations passe par les copules, outil statistique permettant de travailler de façon plus aisée avec des fonctions de répartition multidimensionelles. Le lecteur intéressé par les applications en finance de ce type d'outils lira Cherubini et al. (2005) avec intéret.

5.3.6 Bestiaire des GARCH

Il existe de nombreux processus GARCH, existant pour des raisons diverses : prise en compte de la lente décroissance de l'autocorrélation du processus de volatilité, prise en compte du lien risque/rentabilité, prise en compte de l'asymétrie observée dans les rendements ou liens avec des processus continus. On présente ici trois extensions possibles des processus GARCH : les modèles GARCH intégrés, les modèles GARCH-M, les modèles GARCH asymétriques ainsi que le modèle GARCH de Heston.

5.3.6.1 GARCH-M

Les modèles GARCH-M furent initialement introduits par Engle et al. (1987) : l'ambition de l'article est de présenter un modèle permettant de rétablir le lien entre rentabilité et rendement, cher à la finance de marché. L'idée est donc de faire dépendre la rentabilité conditionnelle du risque conditionnel lui-même, de façon linéaire le plus souvent. Le modèle GARCH-M(1,1) est le suivant :

$$x_t = \lambda h_t + \sqrt{h_t}\epsilon_t \tag{5.352}$$

$$h_t = \omega_0 + \omega_1 x_{t-1}^2 + \omega_2 h_{t-1} \tag{5.353}$$

avec $\epsilon_t \sim N(0,1)$. Seule la première équation est modifiée par rapport à un GARCH classique : on ajoute un terme multiplicatif de la variance, λh_t, que l'on appelle prime de risque. On donne les moments conditionnels :

$$\mathbb{E}[x_t|h_t] = \lambda h_t \tag{5.354}$$

$$\mathbb{V}[x_t|h_t] = \mathbb{V}[\lambda h_t + \sqrt{h_t}\epsilon_t|h_t] = h_t \tag{5.355}$$

L'introduction de la prime de risque n'a conduit qu'à la modification de l'espérance conditionnelle, laissant la variance conditionnelle inchangée. Cette modification ne modifiant que la valeur de l'espérance, ce processus reçu donc le nom de GARCH *in mean*, ou GARCH-M. Le calcul des moments non conditionnels est ici passé sous silence : ceux-ci sont naturellement plus complexes que ceux proposés dans le cadre des simples GARCH.

Les restrictions habituelles s'appliquent aux paramètres de la variance conditionnelle. L'inférence de ce type de processus ne pose pas de problème : on applique les mêmes méthodes que celles présentées plus haut, dans le cas des processus ARCH/GARCH. Là encore, on estime ces modèles par maximum de vraisemblance conditionnel, sachant que :

$$x_t|_{h_t} \sim N\left(\lambda h_t, h_t\right) \tag{5.356}$$

Il est alors aisé de déterminer la log-vraisemblance conditionnelle du processus, pour n observations :

$$lnL = -\frac{n}{2}ln(2\pi) - \frac{1}{2}\left(\sum_{i=1}^{n} ln(h_t) + \frac{(x_t - \lambda h_t)^2}{h_t}\right) \tag{5.357}$$

Ajoutons que les GARCH-M sont naturellement leptokurtiques, tout comme les GARCH. Qu'en est il cependant de l'asymétrie du processus ? Plutot que de se lancer dans des calculs longs et complexes, on a simplement procéder à des simulations : on simule des GARCH-M avec des paramètres égaux pour toutes les simulations et on calcule à chaque fois la skewness de la simulation. La figure 5.35 présente la densité estimée par noyau de l'estimateur de la skewness. Les simulations ont été effectuées à l'aide d'un λ négatif : on remarque que la série a de bonnes chances d'être asymétrique à gauche.

L'intéret de ces modèles est de permettre une mesure à chaque date de l'espérance et de la variance des rendements. Naturellement, ceci fait penser aux modèles de type

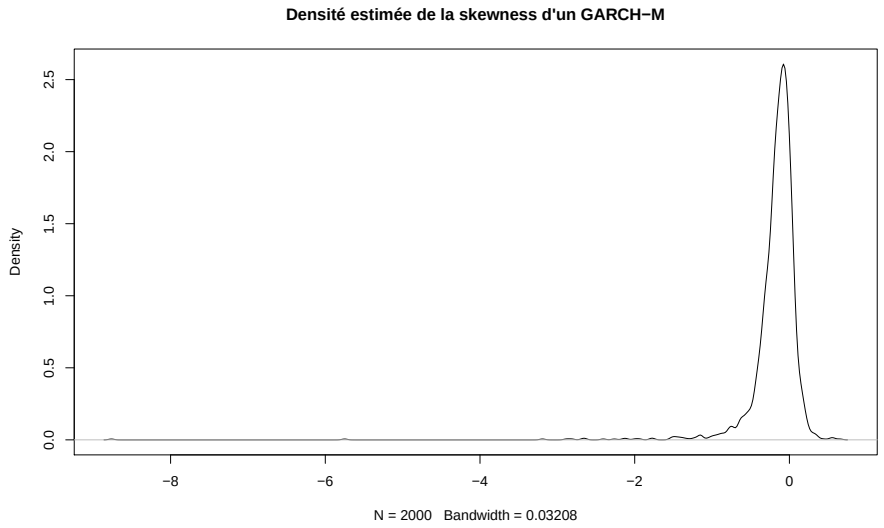

FIG. 5.35 – Densité non paramétrique de l'estimateur de la skewness

espérance/variance. On propose ici une application du modèle GARCH-M à la frontière de Markowitz. Soit r_1 et r_2, les rendements du titre 1 et du titre 2. Chacun de ses rendements est supposé suivre un modèle GARCH-M qui lui est propre. On a donc :

$$\begin{cases} r_{1,t} = \lambda_1 h_{1,t} + \sqrt{h_{1,t}}\epsilon_{1,t} \\ h_{1,t} = \omega_{1,0} + \omega_{1,1} r_{1,t-1}^2 + \omega_{1,2} h_{1,t-1} \end{cases} \tag{5.358}$$

$$\begin{cases} r_{2,t} = \lambda_2 h_{2,t} + \sqrt{h_{2,t}}\epsilon_{2,t} \\ h_{2,t} = \omega_{2,0} + \omega_{2,1} r_{2,t-1}^2 + \omega_{2,2} h_{2,t-1} \end{cases} \tag{5.359}$$

Toute la question est alors de proprement paramétrer le lien entre ϵ_1 et ϵ_2. On propose simplement de supposer qu'ils sont instantanément corrélés et que cette corrélation est constante. On a :

$$cor(\epsilon_1, \epsilon_2) = \rho \tag{5.360}$$

On connait déjà les deux premiers moments conditionnels univariés, pour chacun des deux actifs. Il ne reste plus qu'à déterminer la covariance conditionnelle entre r_1 et r_2. Il suffit de la calculer directement :

$$\begin{aligned} Cov(r_{1,t}, r_{2,t}|h_{1,t}, h_{2,t}) &= Cov(\lambda_1 h_{1,t} + \sqrt{h_{1,t}}\epsilon_{1,t}, \lambda_2 h_{2,t} + \sqrt{h_{2,t}}\epsilon_{2,t}|h_{1,t}, h_{2,t}) && (5.361) \\ &= Cov(\sqrt{h_{1,t}}\epsilon_{1,t}, \sqrt{h_{2,t}}\epsilon_{2,t}|h_{1,t}, h_{2,t}) && (5.362) \\ &= \sqrt{h_{1,t}}\sqrt{h_{2,t}} Cov(\epsilon_{1,t}, \epsilon_{2,t}|h_{1,t}, h_{2,t}) && (5.363) \\ &= \sqrt{h_{1,t}}\sqrt{h_{2,t}}\rho && (5.364) \end{aligned}$$

On obtient ainsi la corrélation conditionnelle de deux titres. On remarque que même si la corrélation entre les deux bruits est constante au cours du temps (autrement dit, les deux titres subissent les mêmes chocs et y répondent de façon similaire), la covariance conditionnelle, elle, varie au cours du temps. On a donc à disposition espérance,

variance et covariance conditionnelle pour chaque titre. La corrélation entre les titres, conditionnelle ou pas, reste la même :

$$cor(r_{1,t}, r_{2,t}|h_{1,t}, h_{2,t}) = \frac{Cov(r_{1,t}, r_{2,t}|h_{1,t}, h_{2,t})}{h_{1,t}h_{2,t}} = \rho \tag{5.365}$$

Les rendements étant de plus conditionnellement gaussiens, il est possible de mettre en oeuvre la méthodologie de Markowitz sans problème. On proposer ici de travailler sur des simulations. On simule des processus GARCH-M, liés entre eux par la corrélation entre les innovations. Puis on représente la frontière efficiente, basée sur les moments non conditionnels ainis que celles basées sur les moments conditionnels, évoluant au cours du temps. On observe le résultat de ces simulations en figure 5.36.

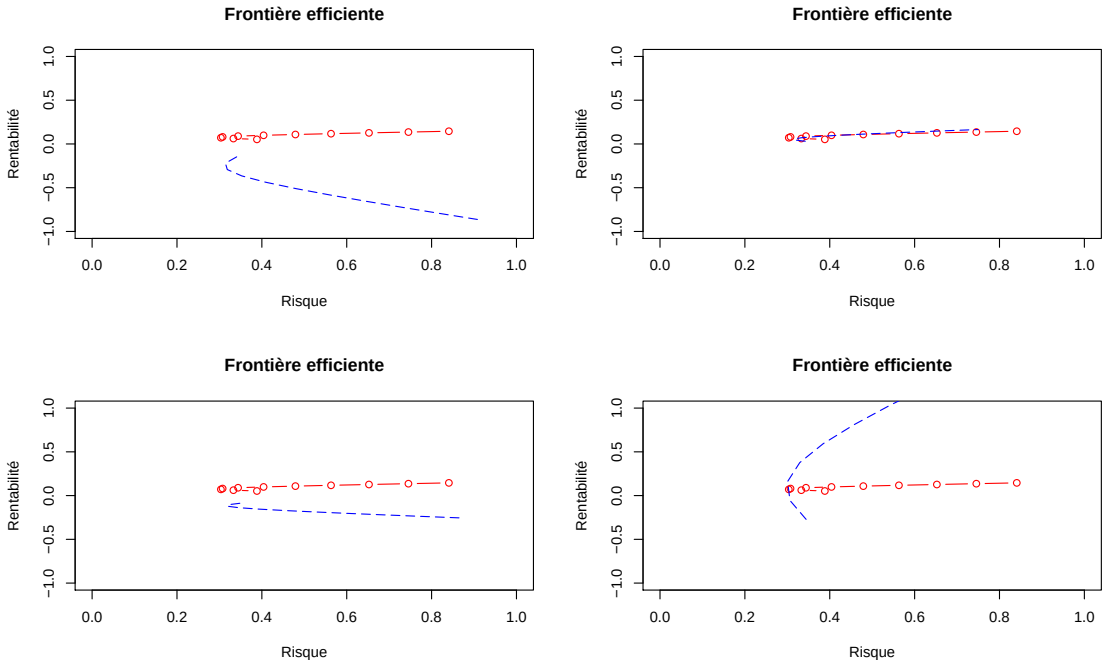

FIG. 5.36 – Frontière efficiente GARCH-M

Il est à noter ici que l'on a pas procédé à l'estimation de ces modèles sur des séries réelles, et ce pour plusieurs raisons. L'une des principales est que le lien rentabilité/risque est loin d'être stable selon les actifs étudiés. Il serait faux de croire que le lien espérance variance est toujours et partout validé. Il existe différents effets microstructurels qui peuvent expliquer ces phénomènes, au nombre desquels l'impact de la liquidité sur la formation du prix du risque. Ce type de considération dépasse cependant largement l'objet de ce cours.

Notons simplement que sur les données étudiées dans le cadre de la VaR, il est intéressant de remarquer que le lien risque/rentabilité existe et semble robuste. On calcule la corrélation entre les rendements et la variance conditionnelle telle qu'on l'obtient en estimant un modèle GARCH(1,1). En notant ρ cette corrélation, la statistique de test

permettant de tester que ρ est bel et bien significativement différent de 0 est la suivante :

$$\hat{t}_\rho = \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{n} \sim T_{n-1} \tag{5.366}$$

où T_{n-1} est une distribution de Student à $n-1$ degrés de liberté. En calculant le coefficient de corrélation ainsi que la statistique de test présentée, on obtient les résultats suivants :

	CAC	**DAX**
ρ	0,07229244	0,04553088
$\hat{t}$	3,125987	1,965682

Sur ces séries, il semble exister une corrélation significative et positive entre rendement et volatilité. En reprenant les estimations obtenues au cours de l'étude sur la VaR, on propose une estimation naive du paramètre λ pour les rendements du CAC et du DAX : on estime par MCO l'équation de la moyenne du GARCH-M, à équation de la variance connue. La table suivant fournit les résultats de l'estimation :

	Estimate	**Estimate**	**Std. Error**	**t-value**	**p-value**
CAC	ω_0	0,00001178	0,000002675	4,406	0,0000105
	ω_1	0,05916	0,01124	5,261	0,000000143
	ω_2	0,8439	0,0315	26,791	0
	λ	0.582470	0.186532	3.123	0.00182
	Constante	-0,005926	0,002056	-2,882	0,00399
DAX	ω_0	0,00001178	0,000002675	4,406	0,0000105
	ω_1	0,05916	0,01124	5,261	0,000000143
	ω_2	0,8439	0,0315	26,791	0
	λ	0.191232	0.097390	1.964	0.0497
	Constante	-0.001263	0.001007	-1.254	0.2099

5.3.6.2 GARCH intégrés

On aborde dans cette section l'un des principaux faits stylisés tirés des modèles GARCH : le fait que la volatilité soit un processus "intégré", plus particulièrement un processus I(1). Qu'est ce qu'un processus I(1) ? Il s'agit d'une notion tirée des séries temporelles que nous n'avons pas encore abordée, bien qu'elle occupe une place centrale de nos jours. Une littérature abondante s'est développée autour de cette notion de processus intégrés, dont l'ensemble des tests de racines unitaires, qui offrent davantage de portée académique que pratique.

Un processus intégré d'ordre 1 ou I(1) est un processus non-stationnaire (cf. fin de la section sur les ARMA), et dont la différence d'ordre 1 est, elle, stationnaire. Il s'agit donc d'un processus x_t, dont la moyenne, la variance ou l'autocovariance dépend du temps (n'est pas stable) mais dont la différence d'ordre 1, elle, ne dépend pas du temps. Cette différence est évidement $\Delta x_t = x_t - x_{t-1}$ (juste pour Pépino, pour le cas où). Un bon exemple de processus non stationnaire en finance est le prix d'un actif financier (hors électricité) : on sait que le prix du CAC n'est pas stationnaire, mais que le

	DAX	**SMI**	**CAC**	**FTSE**
ω_1	0,07	0,75	0,88	0,94
ω_2	0,89	0,11	0,05	0,05
Somme	0,96	0,86	0,93	0,99

TAB. 5.4 – Estimation de GARCH sur des indices européens

rendement (pas très loin de la différence d'ordre 1) est lui stationnaire (bruit blanc en général).

On a pu constater lors de l'estimation des modèles GARCH(1,1) que la condition de stationnarité n'était pas toujours remplie. On rappelle que pour un modèle GARCH(1,1) de la forme :

$$x_t = \sqrt{h_t}\epsilon_t \tag{5.367}$$

$$h_t = \omega_0 + \omega_1 x_{t-1}^2 \omega_2 h_{t-1} \tag{5.368}$$

ce processus n'est stationnaire (sa variance n'explose pas) que si $\omega_1 + \omega_2 < 1$. Dans de nombreuses estimations de GARCH, on observe en général que la somme de ω_1 et ω_2 très proche de 1, sans toutefois lui être supérieur. La table 5.4 présente l'estimation de modèles GARCH(1,1) sur les différents indices européens contenus dans la base de donnée `EuStockMarkets` disponible sous `R`. On remarque que la somme des coefficients est très souvent proche de 1 : on parle souvent de processus intégré pour la volatilité. Ceci fait au passage remarqué que la volatilité d'une actif financier n'est ni constante, ni une variable *lisse* : il s'agit d'un processus agité, sujet à des changements de régime le plus souvent. Il existe des modèles à changement de régime en séries temporelles, qui ne sont pas développés dans le cadre de ce maigre cours. Un lecteur intéressé trouvera une breve introduction dans Wang (2003)[Chapitre 5].

Ce qui suit s'inspire largement de Poon (2005)[Chapitre 4]. Un processus GARCH pour lequel on a $\omega_1 + \omega_2 = 1$ est un processus non stationnaire en variance, dans la mesure où sa variance non conditionnelle tend vers $+\infty$. Le processus r_t (les rendements) reste cependant stationnaire au sens stricte. On dit que la volatilité conditionnelle suit un processus IGARCH(1,1), et ni le moment d'ordre 2 ni le moments d'ordre 4 n'existent : ils divergent tous vers $+\infty$.

L'un des modèles de volatilité les plus utilisés en Risk Management est le modèle EWMA de RiskMetrics. EWMA signifie Exponentially Weighted Moving Average : il s'agit simplement d'un modèle spécifiant la volatilité comme une moyenne pondérée des volatilités passées, avec des pondérations décroissant exponentiellement dans le temps. D'une façon générale, le modèle s'écrit :

$$\sigma_{t+1}^2 = \frac{\sum_{i=0}^{\tau} \beta^i \sigma_{t-i-1}^2}{\sum_{i=0}^{\tau} \beta^i} \tag{5.369}$$

β est appelé paramètre de lissage. Il s'agit d'une approche visant à la prévision : le paramètre de lissage est estimé en minimisant l'erreur de prévision dans l'echantillon.

τ est l'horizon de mémoire du modèle. Ce modèle, outre le fait qu'il soit utilisé dans de nombreuses applications de RiskMetrics, a la particularité d'être proche d'un modèle GARCH intégré. Pour prouver cette propriété, il suffit de faire une petit coup de *pont d'Avignon* : on commence par retravailler l'expression d'un GARCH(1,1), puis on revient sur le modèle EWMA pour étudier la similarité.

Dans le cadre d'un GARCH(1,1), on a :

$$h_{t+1} = \omega_0 + \omega_1 x_{t-1}^2 + \omega_2 h_t \tag{5.370}$$
$$= \omega_0 + \omega_2\omega_0 + \omega_1 x_{t-1}^2 + \omega_2\omega_1 x_{t-1}^2 + \omega_2^2 h_{t-1} \tag{5.371}$$

En poursuivant les itérations, on trouve finalement :

$$h_{t+1} = \omega_0 \sum_{i=1}^{\tau} \omega_1^{i-1} + \omega_0 \sum_{i=1}^{\tau} \omega_1^{i-1} x_{t-i}^2 + \omega_2^{\tau} h_{t-\tau} \tag{5.372}$$

Alors, si $\omega_2 < 1$, lorsque $\tau \to \infty$, on a :

$$h_{t+1} = \frac{\omega_0}{1-\omega_2} + \omega_0 \sum_{i=1}^{\infty} \omega_1^{i-1} x_{t-i}^2 \tag{5.373}$$

Ceci est vrai, même dans le cas où $\omega_1 + \omega_2 = 1$. L'important est que $\omega_2 < 1$, ce qui semble vraisemblablement être le cas sur de nombreuses séries financières. On obtient alors une forme intéressante de volatilité. Comparons ceci au modèles EWMA :

$$\sigma_{t+1}^2 = \frac{\sum_{i=0}^{\tau} \beta^i \sigma_{t-i-1}^2}{\sum_{i=0}^{\tau} \beta^i} \tag{5.374}$$

Ici aussi, pour peu que $\beta < 1$, il est possible de passer à la limite pour trouver une forme analytique à la variance :

$$\sigma_{t+1}^2 = (1-\beta) \sum_{i=0}^{\infty} \beta^i \sigma_{t-i-1}^2 \tag{5.375}$$

A une constante pret, on retrouve notre dynamique de variance GARCH(1,1). Le modèle EWMA peut être vu à bien des égards comme un simple modèle GARCH(1,1) intégré. L'une des propriétés remarquable de ce type de processus est qu'il ne sont pas victimes de mean-reverting rapide comme c'est le cas pour un GARCH(1,1). Il ont au contraire tendance à conserver et à amplifier les chocs temporaires de volatilité : ils ont ainsi une mémoire plus longue que les simples modèles GARCH. On parle ainsi souvent de mémoire longue de la volatilité pour désigner ce fait stylisé. Il existe des processus de séries temporelles permettant de prendre en compte de façon statisfaisante cette mémoire longue (dit *processus à mémoire longue*). Il n'est cependant pas certain que leur application se justifie pleinement : de nombreux processus à mémoire courte (tel que les modèles à changement de régime) génèrent eux-aussi des faits stylisés de mémoire longue, sans être philosophiquement responsables. Il s'agit d'un terrain de recherche empirique depuis les années 2000.

Pour terminer, revenons sur le calcul de la VaR dans la perspective d'une volatilité intégrée. Tsay (2002)[chapitre 7] propose une ecriture alternative au modèle RiskMetrics. La présentation diffère légèrement de celle qui en est fournie dans Poon (2005), même reste globalement la même. Le modèle EWMA repose sur l'hypothèse que la distribution conditionelle des rendements soit normale, d'espérance nulle et de variance σ_t. La dynamique de la variance est décrite de la façon suivante :

$$\sigma_t^2 = \beta\sigma_{t-1}^2 + (1-\beta)r_{t-1}^2 \tag{5.376}$$

où $\beta < 1$. On retrouve donc bien un modèle intégré, avec une variance non conditionnelle qui n'est pas définie, dans la mesure où la somme des deux paramètres de la dynamique de la variance est égale à 1 par construction. Il s'agit par conséquent d'un modèle IGARCH(1,1), sans drift. Il est possible de fournir une prévision de la variance, comme on l'a fait dans le cas d'un processus GARCH classique. On construit ces prévisions de façon récursive, en rappelant que :

$$r_t = \sigma_t\epsilon_t \tag{5.377}$$

avec $\epsilon_t \sim N(0,1)$. On a alors :

$$\begin{aligned}
\sigma_{t+1}^2 &= \beta\sigma_t^2 + (1-\beta)(\sigma_t^2\epsilon_t^2) && (5.378)\\
&= \sigma_t^2 - \sigma_t^2 + \beta\sigma_t^2 + (1-\beta)(\sigma_t^2\epsilon_t^2) && (5.379)\\
&= \sigma_t^2 + (1-\beta)\sigma_t^2(\epsilon_t^2 - 1) && (5.380)
\end{aligned}$$

On sait que :

$$\mathbb{E}[\epsilon_t^2|\sigma_t] = 1 \tag{5.381}$$

On en déduit donc que :

$$\mathbb{E}[\sigma_{t+1}^2|\sigma_t] = \mathbb{E}[\sigma_t^2|\sigma_t] \tag{5.382}$$

Dans un modèle intégré, la meilleure prévision de la volatilité que l'on puisse formuler, *sachant que l'information dont on dispose en t est réduite à la seule la volatilité d'aujourd'hui* est σ_t. Une autre propriété remarquable découlant de ce qui vient d'être dit est la suivante :

$$\begin{aligned}
\mathbb{E}[\sigma_{t+h}^2|\sigma_t] &= \mathbb{E}[\sigma_{t+h-1}^2 + (1-\beta)\sigma_{t+h-1}^2(\epsilon_{t+h-1}^2 - 1)|\sigma_t] && (5.383)\\
&= \mathbb{E}[\sigma_{t+h-1}^2|\sigma_t] && (5.384)
\end{aligned}$$

En itérant la dernière équation, on trouve donc que la meilleure prévision que l'on puisse formuler de la volatilité future pour un horizon $t+h$ est la prévision que l'on pourrait faire à un jour.

Dans le cas où la filtration ne se réduit pas à σ_t, mais peut être étendue à $\{\sigma_t, r_t\}$, alors la prévision à un jour diffère de la volatilité d'aujourd'hui :

$$\mathbb{E}[\sigma_{t+1}^2|\sigma_t, r_t] = \beta\sigma_t^2 + (1-\beta)r_t^2 \tag{5.385}$$

On constate alors qu'un modèle intégré ne conduit pas à un retour de la prévision vers le niveau moyen de l'échantillon : la volatilité prévue à long terme n'est ni plus moins que celle d'aujourd'hui. Ceci est globalement en désaccord avec ce qu'on observe sur le marché. En général, une phase de forte volatilité (bull market) est suivie d'une retour au calme (bear market). Néanmoins, comme la table précédente le montrait, la volatilité, sans être complètement intégrée, n'est pas loin de l'être.

Fort de ces deux éléments ($\mathbb{E}[\sigma^2_{t+h}|\sigma_t, r_t] = \hat{\sigma}^2_{t+1}$ et $\hat{\sigma}^2_{t+1} = \beta\sigma^2_t + (1-\beta)r^2_t$), on est alors en mesure de déterminer une VaR pour un horizon h quelconque, en reprenant ce qui a été dit au cours des sections précédentes. On cherche la loi conditionnelle de $\sum_{i=1}^{h} r_{t+i}$, comme précédement. Sait faire de calcul, on sait qu'elle est gaussienne, d'espérance nulle et de variance égale à la somme des variances, du fait de l'hypothèse d'absence de corrélation sérielle dans les rendements. On a donc :

$$\mathbb{V}\left[\sum_{i=1}^{h} r_{t+i}|\sigma_t, r_t\right] = \sum_{i=1}^{h} \mathbb{V}\left[r_{t+i}|\sigma_t, r_t\right] \tag{5.386}$$

$$= \sum_{i=1}^{h} \hat{\sigma}^2_{t+1} \tag{5.387}$$

$$= h\hat{\sigma}^2_{t+1} \tag{5.388}$$

$$\tag{5.389}$$

Ainsi, la variance des rendements à h jours correspond simplement à h fois la variance à un jour. La VaR se calcule ensuite de façon évidente :

$$P\left(\sum_{i=1}^{h} r_{t+i} \leq VaR\right) = 5\% \tag{5.390}$$

$$\Leftrightarrow P\left(\frac{\sum_{i=1}^{h} r_{t+i}}{\sqrt{h\sigma^2_{t+1}}} \leq \frac{VaR}{\sqrt{h\sigma^2_{t+1}}}\right) = 5\% \tag{5.391}$$

$$\tag{5.392}$$

On trouve naturellement :

$$VaR_h = 1,64 \times \sqrt{h}\sigma_{t+1} \tag{5.393}$$

Il s'agit d'un règle bien connue par les risk managers et acceptée par Bâle II : la VaR à h jours est égale à $\sqrt{h}$ fois la VaR à un jour. Cette méthodologie repose bien évidement sur la gaussianité conditionnelle des rendements, et conduit naturellement à sous-estimer régulièrement la VaR "véritable". Mais quand les chiffres doivent tomber au quotidien, il n'en reste pas moins que la méthode est séduisante.

5.3.6.3 GARCH asymétriques

Dans cette section, on introduit brièvement les modèles GARCH asymétriques. Ceux-ci ont été introduits de façon à rendre compte du fait que les séries de rendements aient

en général un skewness inférieure à 0. On a vu qu'un modèle GARCH standard ne permettait pas d'obtenir une skewness différente de 0. Une abondante littérature s'est alors développée, proposant des modèles permettant de générer sous certaines conditions ce fait stylisé.

L'intérêt d'une distribution skewed vient de la présence de ce que la théorie financière a baptisé *effet levier*. Il s'agit d'un fait stylisé conduisant à l'observation (dans le cas d'effet levier au sens strict) d'un accroissement de la volatilité lorsque les rendements eux-même décroissent. Ceci s'interprète souvent en expliquant qu'une mauvaise nouvelle (rendements négatifs) a un impact positif important sur la volatilité future d'un titre. Gourieroux and Jasiak (2001)[chapitre 6] présente cette hypothèse en terme de corrélation, dans le cadre d'un modèle ARCH(1). Il s'agit de déterminer à quelle condition un ARCH(1) peut générer un effet levier. Le modèle général s'écrit :

$$r_t = (\omega_0 + \omega_1 r_{t-1}^2)^{1/2}\epsilon_t \tag{5.394}$$

avec ϵ_t suivant une loi inconnue, mais de variance égale à 1. On s'intéresse alors à la covariance suivante :

$$Cov(r_t - r_{t-1}, h_{t+1} - h_t | r_{t-1}) \tag{5.395}$$

L'effet de levier correspond à une covariance négative. On se demande à quelle condition on peut observer ce type de comportement dans le cadre d'un ARCH(1). Pour cela, il suffit de developper un tant soit peu les calculs :

$$\begin{aligned}
Cov(r_t - r_{t-1}, h_{t+1} - h_t | r_{t-1}) &= Cov(r_t, h_{t+1} | r_{t-1}) && (5.396)\\
&= Cov(r_t, \omega_0 + \omega_1 r_t^2 | r_{t-1}) && (5.397)\\
&= \omega_1 Cov(r_t, r_t^2 | r_{t-1}) && (5.398)\\
&= \omega_1 Cov(\sqrt{h_t}\epsilon_t, h_t\epsilon_t^2 | r_{t-1}) && (5.399)\\
&= \omega_1 h_t^{3/2} Cov(\epsilon_t, \epsilon_t^2 | r_{t-1}) && (5.400)\\
&= \omega_1 h_t^{3/2} \mathbb{E}[\epsilon_t^3] && (5.401)
\end{aligned}$$

Ainsi, un modèle ARCH(1) génère un effet levier à la condition que ses innovations soient asymétriques vers la gauche. Autrement dit, $Cov(r_t - r_{t-1}, h_{t+1} - h_t | r_{t-1}) < 0$ dans le cadre d'un ARCH(1) si $\mathbb{E}[\epsilon_t^3] < 0$, le reste des paramètres étants positifs.

On comprend alors mieux l'intéret de travailler sur des modèles permettant de générer de l'asymétrie. Ils permettent de rendre compte d'un fait stylisé important en finance, i.e. l'effet levier.

Il existe plusieurs modèles permettant de générer cet effet asymétrique. Ils reposent en général sur l'introduction d'une variable indicatrice (i.e. valant soit 0, soit 1) qui prend la valeur 1 dans le cas où les rendements sont négatifs. L'un des modèles les plus connus est le modèle GARCH de Glosten, Jagannathan et Runkle (Glosten et al. (1993)), dit GJR-GARCH. Il s'écrit de la façon suivante :

$$\begin{aligned}
r_t &= \sqrt{h_t}\epsilon_t && (5.402)\\
h_t &= \omega_0 + \omega_1 r_{t-1}^2 + \omega_2 h_{t-1} + \alpha \mathbb{1}_{r_{t-1}<0} r_{t-1}^2 && (5.403)
\end{aligned}$$

$\epsilon_t \sim N(0,1)$. $\mathbb{1}_{r_{t-1}<0}$ est une variable qui prend la valeur 1 lorsque les rendements en date $t-1$ sont négatifs. Dans ce cas là, on observe une volatilité h_t à la date suivante qui est plus importante que dans le cas où les rendements sont positifs, pourvu que $\alpha >$. On a encore un processus pour les rendements qui est conditionnellement centré. La variance conditionnelle quant à elle dépend des rendements à la date passé :

$$\mathbb{E}[h_t|r_{t-1} > 0] = \omega_0 + \omega_1 r_{t-1}^2 + \omega_2 h_{t-1} \tag{5.404}$$

$$\mathbb{E}[h_t|r_{t-1} < 0] = \omega_0 + (\omega_1 + \alpha) r_{t-1}^2 + \omega_2 h_{t-1} \tag{5.405}$$

Ce type de modèle se rapproche peu à peu de modèles à changement de régime, avec une specification naive du seuil sur les rendements à dépasser. Ces modèles ne sont encore une fois pas traités ici, bien qu'ils revetent un intéret certain en econométrie de la finance. Ils restent un sujet trop avancé pour être abordé ici.

5.3.6.4 Modèle GARCH de Heston

L'objet de cette section est de montrer les liens existants entre les processus GARCH(1,1) et les processus continus classiquement utilisés en finance. Cette section s'inspire grandement de Aboura (2005) ainsi que de Gourieroux (1992).

L'une des utilisations que l'on on aurait envie de faire de ces processus GARCH est bien évidement de les utiliser afin de valoriser des actifs financiers. Plus particulièrement, on sait que les options sont des actifs dont le prix est particulièrement sensible aux variations de la volatilité. Un certain nombres d'articles de recherche se sont ainsi tournés vers des méthodes permettant de valoriser des options vanilles (typiquement un call) en utilisant des dynamiques GARCH pour les rendements du sous-jacent. Dans le cadre de ce type de démarche, on butte sur deux types de difficultés :

- Il s'agit tout d'abord de déterminer un processus continu dont la discrétisation correspond bien à un processus GARCH. En général, on est souvent capable de passer d'un processus continu sa contrepartie discrète. On se souvient de la petit application du lemme d'Ito permettant de montrer que dans le cadre de Black-Scholes, la diffusion discrétisée correspond exactement à un bruit blanc. En revanche, lorsque l'on veut passer d'une dynamique discrète à sa version continue, rien n'est gagné d'avance. Il s'agit encore d'un sujet actuel de recherche : inutile d'aborder cette question en profondeur ici.
- Le second problème rencontré est de passer de la dynamique historique (celle que l'on observe) à la dynamique risque neutre. Un certain nombre de méthodes ont été proposés et nous aborderons brièvement celle proposée par Heston.

Commençons par nous intéressé à la la limite possible d'un processus GARCH. Bien évidement, il s'agit de la limite au sens financier du terme, i.e. pour un pas de temps infinitésimal. Il ne faut surtout pas confondre la limite en finance, avec une limite vers l'infini, souvent utilisée en statistique (par exemple dans l'énoncé de la loi des grands nombres). Engle et Ishida (2002) montrent comment trouver les diffusions correspondants à certains processus GARCH. On propose ici de développer le cas d'un GARCH(1,1).

On a le modèle suivant pour les rendements :

$$r_t = \sqrt{h_t}\epsilon_t \tag{5.406}$$
$$h_t = \omega_0 + \omega_1 r_{t-1}^2 + \omega_2 h_{t-1} \tag{5.407}$$

avec les hypothèses habituelles. On travaille ici sur la dynamique de la variance conditionnelle, h_t. Il s'agit d'une série d'astuces permettant de trouver par passage à la limite le processus effectivement suivi par la variance conditionnelle. On commence par réécrire la variance comme suit :

$$h_t = \omega_0 + (\omega_1 + \omega_2)h_{t-1} + \omega_1(r_{t-1} - h_{t-1}) \tag{5.408}$$

En posant alors $\gamma = \omega_1 + \omega_2$ et en remplaçant r_t^2 par son expression, il vient :

$$h_t = \omega_0 + \gamma h_{t-1} + \omega_1(h_{t-1}\epsilon_{t-1}^2 - h_{t-1}) \tag{5.409}$$
$$= \omega_0 + \gamma h_{t-1} + \omega_1 h_{t-1}(\epsilon_{t-1}^2 - 1) \tag{5.410}$$

On note alors $\eta_t = \epsilon_{t-1}^2 - 1$, un bruit de loi chi-deux, centré. On remplace :

$$h_t = \omega_0 + \gamma h_{t-1} + \omega_1 h_{t-1}\eta_t \tag{5.411}$$

Nouvelle astuce : on réécrie le processus en travaillant sur les différences de variance conditionnelle :

$$h_t - h_{t-1} = \omega_0 - h_{t-1} + \gamma h_{t-1} + \omega_1 h_{t-1}\eta_t \tag{5.412}$$
$$= \omega_0 - (1-\gamma)h_{t-1} + \omega_1 h_{t-1}\eta_t \tag{5.413}$$

On note à présent $k = (1 - \gamma)$. On remplace :

$$h_t - h_{t-1} = \omega_0 - k h_{t-1} + \omega_1 h_{t-1}\eta_t \tag{5.414}$$
$$= k\left(\frac{\omega_0}{k} - h_{t-1}\right) + \omega_1 h_{t-1}\eta_t \tag{5.415}$$

On note $\vartheta = \frac{\omega_0}{k} = \frac{\omega_0}{1-\omega_1}$. On n'aura pas manqué de remarquer que cette constante ressemble furieusement à la l'espérance non conditionnelle de la variance d'un processus GARCH(1,1). En réintroduisant, on obtient :

$$h_t - h_{t-1} = k\left(\vartheta - h_{t-1}\right) + \omega_1 h_{t-1}\eta_t \tag{5.416}$$

Cette dernière expression correspond à un cas particulier d'un processus général développé par Engle et Ishida, nommé CEV-GARCH, par analogie avec les modèles CEV (Constant Elasticity Variance) développés par la finance en temps continu. La version discrète d'un CEV-GARCH est :

$$h_t - h_{t-1} = k(\vartheta - h_{t-1}) + \omega_1 h_{t-1}^{\xi}\eta_t \tag{5.417}$$

et semble converger vers :

$$dh_t = k(\vartheta - h_t) + \omega_1 h_t^{\xi} dW_t \tag{5.418}$$

où W_t est un mouvement brownien standard. Dans notre cas, on trouve naturellement comme limite d'un GARCH(1,1) la diffusion suivante :

$$dh_t = k(\vartheta - h_t) + \omega_1 h_t dW_t \tag{5.419}$$

La question suivante, et qui n'est pas abordée ici, est de parvenir à déterminer la dynamique risque neutre des rendements. Heston[1993] propose le principe de Neutralisation Locale au Risque. Un lecteur soucieux d'en savoir plus lira avec intéret le dernier chapitre de Aboura (2005), ainsi que le papier de Heston. Ce point n'est pas développé dans la mesure où il s'appuie sur des modèles à volatilité stochastique qui n'ont pas encore été abordés.

5.3.7 Modèles exponentiels

Cette courte section a pour objectif d'introduire brièvement l'usage des modèles exponentiels et à volatilité stochastiques en séries temporelles. Les deux types de modèles peuvent être vus comme appartenant à la même famille, dans la mesure où ils reposent tous deux sur une paramétrisation de la volatilité sous la forme de l'exponentielle d'un processus, afin de garantir sa positivité.

5.3.7.1 Le modèle EGARCH

Les modèles EGARCH furent introduits par Nelson (1991) afin d'améliorer et d'assouplir les processus GARCH tel qu'on les a présenté jusqu'à présent. Il s'agit tout d'abord de trouver une façon d'éviter l'ensemble des contraintes pesant sur les paramètres des modèles GARCH : on a vu qu'il était nécessaire d'ajouter ces contraintes pour garantir la positivité et l'existence de la variance non conditionnelle. Nelson propose une méthode permettant d'éviter ces contraintes, et permettant de plus degénérer de la skewness négative.

On se borne ici à présenter le modèle, sans évoquer ni l'inférence, ni la prévision en utilisant ce modèle. Comme dans le cadre d'un GARCH classique, on a :

$$r_t = \sqrt{h_t}\epsilon_t \tag{5.420}$$

Ce qui change cette fois-ci, c'est la façon d'écrire l'équation de la variance. Celle-ci est écrite en terme de logarithme :

$$ln(h_t) = \omega_0 + \omega_1 ln(h_{t-1}) + \omega_2(|\epsilon_{t-1}| - \sqrt{2/\pi}) - \alpha\epsilon_{t-1} \tag{5.421}$$

avec $\epsilon_t \sim N(0,1)$. Dans ce cas $\mathbb{E}[|\epsilon_t|] = \sqrt{2/\pi}$, d'où le fait que l'on retire cette quantité dans le terme $\omega_2(|\epsilon_{t-1}| - \sqrt{2/\pi})$: on travaille sur une variable centrée. α est supposé être positif (mais précédé d'un signe négatif), afin de permettre les effets levier tels qu'on les a évoqué plus haut. On a ainsi :

$$\mathbb{E}[ln(h_t)|h_{t-1}, \epsilon_{t-1} > 0] = (\omega_0 - \omega_2\sqrt{2/\pi}) + \omega_1 ln(h_{t-1}) + (\omega_2 - \alpha)\epsilon_{t-1} \tag{5.422}$$

$$\mathbb{E}[ln(h_t)|h_{t-1}, \epsilon_{t-1} < 0] = (\omega_0 - \omega_2\sqrt{2/\pi}) + \omega_1 ln(h_{t-1}) - (\alpha + \omega_2)\epsilon_{t-1} \tag{5.423}$$

Ainsi, selon les valeurs des paramètres, la réponse de la volatilité à des chocs positifs ou négatifs peut différer, autorisant la modèlisation des fameux effets leviers. On se forge rapidement l'intuition que lorsque $\epsilon_{t-1} < 0$, on a alors une volatilité conditionnelle plus importante que dans le cas contraire. On n'en dira pas davantage au sujet de ces modèles, ce qui reste étant trop avancé.

5.3.7.2 Les modèles à volatilité stochastique

On présente encore plus brievement les modèles à volatilité stochastique, sur la base de ce qui en est dit dans Wang (2003)[chapitre 3 et 7]. Jusqu'à présent, on a supposé que la loi de la variance conditionnelle était intégralement dictée par la loi des r_t et qu'elle n'avait par conséquent pas de loi propre. On rappelle que d'après ce qui a été dit, la loi de la volatilité historique dans le cadre d'un GARCH ne devrait pas être loin d'un processus χ^2 décentré.

Les modèles de volatilté stochastique se présentent dans un cas particulier simple de la façon suivante :

$$r_t = \sigma_t \epsilon_t \tag{5.424}$$
$$\epsilon_t \sim N(0, \sigma_\epsilon) \tag{5.425}$$
$$h_t = ln\sigma_t^2 \tag{5.426}$$

A la différence des EGARCH, on propose une dynamique propre à la variance, avec une loi propre. Wang (2003) propose simplement de donner à $ln\sigma_t^2$ un pattern ARMA(p,q). On a alors dans le cas d'un ARMA(1,1), la volatilité suivante :

$$h_t = \alpha + \rho h_{t-1} + \nu_t + \theta \nu_{t-1} \tag{5.427}$$
$$\nu_t \sim N(0, \sigma_\nu) \tag{5.428}$$

Au final, on a donc deux bruits différents : un premier bruit lié à l'équation des rendements eux-même, et un second bruit venant de la volatilité elle-même. Dans ce cas précis, la loi de la volatilité est une loi log-normale, à support positif, tout comme le χ^2 que l'on obtient dans le cas d'un ARCH(1). On retrouve un processus proche d'un ARCH(1) dans le cas où $\theta = 0$.

L'inférence de ce type de processus repose sur des méthodes particulièrement avancées (filtres de Kalman) et il n'est pas question d'aborder ces questions ici. On remarque cependant que les modèles ARMA estimés sur les volatilité implicites présentées dans le cadre de la section consacrée aux ARMA peuvent être vu comme une façon de se rapprocher de ce type de modèles.

Terminons ce chapitre en remarquant qu'il est possible de fournir différentes paramétrisations pour la corrélation entre les deux bruits, et que cette paramétrisation permet de rendre compte de forme de smile assez variées. Cf. Aboura (2005)[Chapitre 2].

Chapitre 6

Boite à outils statistiques

Ce dernier chapitre présente quelques applications statistiques utiles à la finance : l'analyse en composantes principales, utile pour l'analyse des séries mutlivariées et des bases de données conséquentes ; les méthodes non-paramétriques, rarement utiles (elles consomment énormément de données), mais souvent utilisées dans les logiciels de statistique.

Il s'agit d'un chapitre relativement léger : on fournit à chaque fois un certain nombre de références pour tout lecteur soucieux d'approfondir ses connaissances (comme Pépino par exemple).

6.1 Méthodes non-paramétriques et application

Dans tout ce qui a été développé jusqu'à présent, on a spécifié une forme functionnelle intuitée à partir de différents éléments, tel que l'ACF du carré des rendements pour former l'intuition des modèles ARCH. Il n'est pas toujours possible de spécifier cette forme ex ante, et c'est précisément là qu'interviennent les méthodes non paramétriques.

6.1.1 Introduction aux méthodes non paramétriques

L'essence des modèles non paramétriques est le *lissage*, i.e. le fait d'utiliser des estimateurs permettant de lisser en quelques sortes les relations entre différentes variables. Supposons par exemple que l'on observe deux variables X et Y, liées par la relation suivante :

$$Y_t = m(X_t) + \epsilon_t \tag{6.1}$$

où ϵ_t est un bruit blanc, d'espérance nulle. $m(.)$ est une fonction arbitraire mais lisse de x_t. Plaçons nous en une date t, et supposons alors que $X = x$. Supposons également que pour cette date, l'on dispose de n observations pour Y_t. On peut alors écrire :

$$\frac{1}{n}\sum_{i=1}^{n} y_i = m(x) + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \tag{6.2}$$

On sait que par la loi des grands nombres, on a :

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \to 0 \tag{6.3}$$

On trouve donc que $\frac{1}{n}\sum_{i=1}^{n} y_i$ est un estimateur consistant de $m(x)$. Evidement, en finance, on dispose rarement (jamais) de ces n observations. En général, on a à disposition un processus joint $\{y_t, x_t\}$. Dans ces cas là, on propose d'utiliser un *compromis*, basé sur une moyenne pondérée des y_t à la place d'une simple moyenne. On propose alors une relation de la forme :

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n}\omega_i(x)y_i \tag{6.4}$$

où $\omega_t(x)$ est un poids qui est plus important pour les y_t pour lesquels on a des x_t proche de x. L'estimation est ainsi déterminée à la fois par la distance entre x_t et x et à la fois par le poids accordé à cette distance.

Tout ceci peut sembler plutot abscond, et est principalement utilisé lors de l'estimation des densités par noyau.

6.1.2 Estimateurs à noyau

Il est possible d'approcher l'estimation d'une densité, à l'aide d'un histogramme. Mais il est également possible, pour peu que l'on dispose de suffisament de données, de lisser cet histogramme à l'aide d'un noyau. Le noyau est une forme de fonction de lissage. On note $K(x)$ ce noyau. Comme il s'agit de poids, il est nécessaire d'avoir la propriété suivante :

$$\int K(z)dz = 1 \tag{6.5}$$

En général, on ajoute un paramètre d'échelle, appelé *bandwidth* et noté h, que l'on incorpore comme suit :

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right), \int K_h(z)dz = 1 \tag{6.6}$$

La fonction de poids est alors définie comme suit :

$$\omega_t(x) = \frac{K_h(x - x_t)}{\frac{1}{n}\sum_{t=1}^{n} K_h(x - x_t)} \tag{6.7}$$

Cette paramétrisation permet évidement d'obtenir des poids dont la somme soit égale à n. On obtient alors l'estimateur à noyau de Nadaraya-Watson, qui prend la forme suivante :

$$\hat{m}(x) = \frac{\sum_{t=1}^{n} K_h(x - x_t)y_t}{\sum_{t=1}^{n} K_h(x - x_t)} \tag{6.8}$$

Dans la pratique, il est nécessaire de spécifier une forme pour le noyau. On utilise dans le cas le plus courant un noyau gaussien qui est défini par :

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} exp\left(-\frac{x^2}{2h^2}\right) \tag{6.9}$$

Tout l'art de l'estimation par noyau consiste alors à déterminer un h optimal. Tsay (2002) présente quelques intuitions dans le cas d'un noyau d'Epanechnikov. On donne ici simplement la règle que l'on se forge rapidement en utilisant ce type d'estimation : un h trop faible conduit à ne pas assez lisser l'histogramme. Au contraire, un h trop important lisse plus que de raison l'histogramme empirique.

Le choix de h ne donc si être trop petit, ni trop grand. La fonction `density` de `R` détermine ce h de façon optimale, à l'aide de méthodes qui sont trop avancées pour être présentées ici. Notons pour conclure que ces méthodes non paramétriques ne sont applicables qu'à partir du moment où l'on dispose d'un nombre de données suffisantes (i.e. au moins 1000 observations). On présente un exemple en figure 6.1.

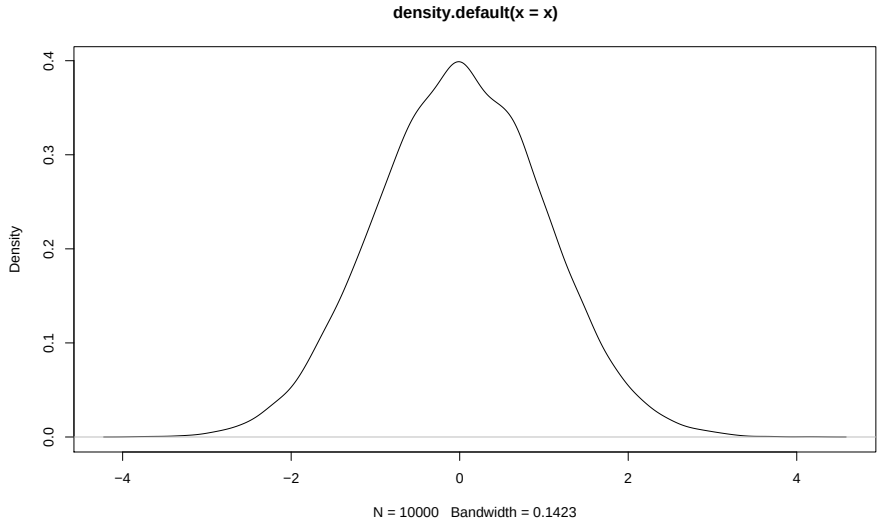

FIG. 6.1 – Densité estimée par méthode des noyaux

Voici un exemple de fonction `R` permettant d'appliquer la méthode des noyaux. Elle est relativement simple et nécessite d'optimiser le h à taton.

```
kernel<-function(X,n,h,min,max){
k=length(X);
support<-seq(min,max,length=n);
fonct.rep<-numeric(n);
for (i in 1:n){fonct.rep[i]=1/(k*h)*sum(1/sqrt(2*pi)*exp(-((X-support[i])/h)^2))};
return(list(fonct.rep=fonct.rep, t=support))
}
```

La figure 6.2 présente l'estimation de la densité d'une loi normale centrée réduite dans le cas où h est mal choisi. On a utilisé le code précédent pour construire ces graphiques.

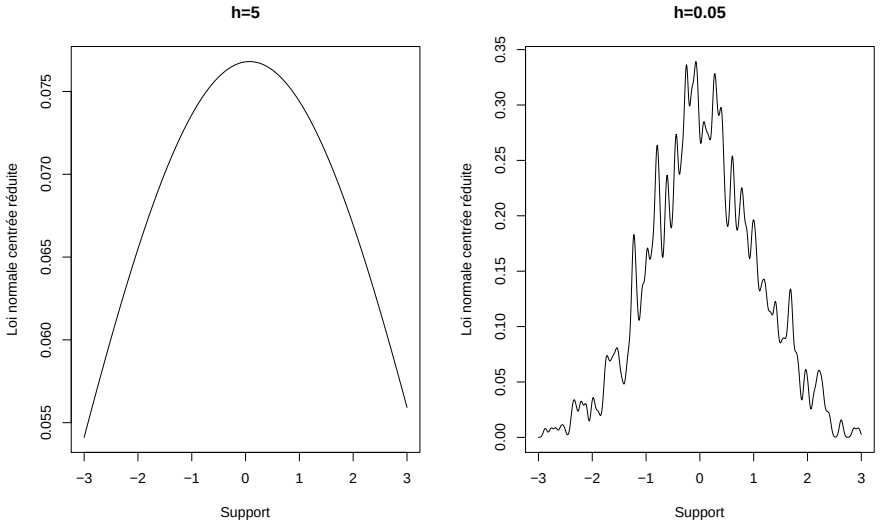

FIG. 6.2 – Densité estimée par méthode des noyaux

6.2 Analyse des données

L'analyse des données fournit un spectre de méthodes robustes permettant de faire du *data-mining*, i.e. de l'exploration de données. Il arrive assez souvent que l'on se retrouve face à un nombre important de séries de données, dont on se sait pas grand chose, sinon qu'elles son liées. L'analyse en composantes principales permet de passer d'une base de données contenant un grand nombre de variables, à une base de données *résumée* par un nombre de facteurs limité. Il est alors nettement plus facile d'analyser ce nombre réduit de facteurs que de travailler sur l'ensemble des séries de données. On présente dans ce qui suit la méthode de l'ACP, ainsi qu'une application bien connue à la courbe des taux.

6.2.1 Analyse en composante principales

Nous avons jusqu'ici accordé peu d'attention aux séries mutlivariées, i.e. à l'étude de p séries de données simultanément. L'ACP permet d'explorer ces p séries de données, en formant des facteurs orthogonaux à partir de la matrice de corrélation des p séries.

On travaille à présent sur une matrice X de taille $\mathcal{M}(n,p)$, contenant les n observations des p séries que l'on souhaite étudier. Une ACP propose de déterminer un nombre réduit de facteurs qui sont des combinaisons linéaires des éléments de la matrice X, de façon à expliquer les liens existants entre les différentes séries. Ce faisant, on parvient en quelques sortes à expliquer la matrice de corrélations de ces p séries. L'important dans cette méthode est de permettre la construction d'un nombre de facteurs qui soit

inférieur à p. L'idée est donc de réduire la dimensionalité du problème.

Il existe un certain nombre de bonnes références, plus ou moins avancées, permettant de comprendre l'ACP ou d'approfondir ce qui va être présenté. On pourra se reporter par exemple à Tsay (2002)[chapitre 8], Pagès (2005) et Saporta (1988). Ce sont les ouvrages utilisés lors de l'élaboration de ces notes de cours.

On travaille à présent sur notre matrice X qui est telle que :

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & \vdots & \vdots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \tag{6.10}$$

On commence par un petit peu de calcul matriciel. On appelle *centre de gravité* le vecteur composé des moyennes empiriques, variable par variable. On le note :

$$g^\mathsf{T} = \begin{pmatrix} \bar{x}_{.,1} & \bar{x}_{.,2} & \dots & \bar{x}_{.,p} \end{pmatrix} = \frac{1}{n} X^\mathsf{T} \mathbb{1}_n \tag{6.11}$$

Avec $\mathbb{1}_n$ un vecteur colonne composé de 1. On note que ceci revient à accorder autant d'importance à toutes les observations. Il est possible d'utiliser des poids différents de $\frac{1}{n}$, permettant d'accorder plus ou moins d'importance aux observations, selon, par exemple, leur écart à la moyenne (métrique inverse variance).

On en déduit Y, la matrice des p observations centrées :

$$Y = X - \mathbb{1}_n g^\mathsf{T} \tag{6.12}$$

Il est alors aisé d'obtenir la matrice de variance/covariance des séries. Il s'agit simplement de :

$$V = \frac{1}{n} X^\mathsf{T} X - g g^\mathsf{T} \tag{6.13}$$

$$= \frac{1}{n} Y^\mathsf{T} Y \tag{6.14}$$

La matrice des observations centrées réduites s'écrit donc naturellement :

$$Z = Y D_{1/s}, \text{ avec } D_{1/s} = \begin{pmatrix} \frac{1}{s_1} & 0 & \dots & 0 \\ 0 & \frac{1}{s_1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \frac{1}{s_p} \end{pmatrix} \tag{6.15}$$

où $D_{1/s}$ est la racine de l'inverse de la première diagonale de la matrice V, i.e. une matrice composée de l'inverse des écart-types de chaque variable. La matrice de corrélation des variables s'obtient alors de façon aisée :

$$R = \frac{1}{n} Z^\mathsf{T} Z \tag{6.16}$$

Il est possible de procéder à une ACP sur la matrice V ou sur la matrice R. Cependant, l'utilisation de V présente le désavantage de ne pas fournir des facteurs stables à toute combinaison linéraire de chaque variable. En clair : si on modifie de façon linéaire une variable particulière, l'ACP sur V ne fournira pas ex-post les mêmes facteurs. On préférera alors utiliser R pour l'ACP : R est insensible à toute transformation linéaire variable par variable. Ceci revient en réalité à réaliser une ACP sur V, en utilisant une *métrique* particulière, i.e. une mesure de distance entre observations, i.e. la métrique définie par $D_{1/s}$. On en dira pas plus sur ce point plus théorique que pratique.

On appelle intertie totale de X la moyenne pondérée des carrés des distances des observations au centre de gravité. On a donc :

$$I_g = \frac{1}{n}\sum_{i=1}^{n}\left(X_{i,.} - g^{\mathsf{T}}\right) D_{1/s}\left(X_{i,.} - g^{\mathsf{T}}\right)^{\mathsf{T}} \tag{6.17}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Z_{i,.} Z_{i,.}^{\mathsf{T}} \tag{6.18}$$

L'inertie peut être vue comme une sorte de variance totale de X : il s'agit de l'écart entre chaque observation et le centre de gravité. Cette notion présente un certain nombre de propriétés intéressantes. On commence par réécrire l'inertie de la façon suivante :

$$I_g = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} Z_{i,j}^2 \tag{6.19}$$

$$= \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} Z_{i,j}^2 \tag{6.20}$$

$$= \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}\frac{(X_{i,j} - g_j)^2}{s_j^2} \tag{6.21}$$

$$= \sum_{j=1}^{p}\frac{1}{s_j^2}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - g_j)^2}_{s_j^2} \tag{6.22}$$

$$= \sum_{j=1}^{p}\frac{s_j^2}{s_j^2} \tag{6.23}$$

$$= Tr(R) \tag{6.24}$$

$$= p \tag{6.25}$$

Avec la métrique utilisée, l'inertie totale est toujours la même : elle est égale au nombre de variables présentes dans X. L'idée sous-jacente à l'ACP est de proposer une méthode permettant de résumer un grand nombre de variables en un nombre réduit de facteurs, ces facteurs ayant été composés suivant une règle basée sur l'inertie. Le critère pour former les facteurs est simple : on cherche k facteurs orthogonaux, de façon à former une nouvelle base orthogonalisée permettant de représenter nos p variables. On cherche ces facteurs de façon à ce que l'inertie de X dans cette nouvelle base soit la plus importante possible.

En général, on présente ceci sous la forme de projections : on cherche un projecteur P, i.e. une matrice permettant de transformer des variables dans X pour en faire des variables dans la nouvelle base. Un projecteur présente un certain nombre de propriétés, telles que :

$$P^2 = P \tag{6.26}$$
$$P^\mathsf{T} D_{1/s^2} = D_{1/s^2} P \tag{6.27}$$

On détermine alors l'inertie de $X^\mathsf{T} P$, i.e. de X dans la nouvelle base. En travaillant sur les variables centrées (i.e. Y), on a :

$$V' = \left(YP^\mathsf{T}\right)^\mathsf{T} D_{1/s} \left(YP^\mathsf{T}\right) \tag{6.28}$$
$$= PVP^\mathsf{T} \tag{6.29}$$

Comme précédement, l'inertie du nuage vaut :

$$Tr(PVPD_{1/s^2}) = Tr(PVD_{1/s^2}P) \tag{6.30}$$
$$= Tr(VD_{1/s^2}P^2) \tag{6.31}$$
$$= Tr(VD_{1/s^2}P) \tag{6.32}$$
$$\tag{6.33}$$

Le problème de l'ACP est donc de trouver P de façon à ce que l'inertie de YP^T soit maximale. On ne détaille pas la solution à ce problème, on se contente du résultat suivant : les k facteurs réalisant cette condition ont pour coordonnées les k premiers vecteurs propres de VD_{1/s^2}, c'est à dire de R. Si le $i^{\text{ème}}$ vecteur propre est u_i, alors le $i^{\text{ème}}$ facteur est égal à :

$$f_i = Zu_i \tag{6.34}$$

On appelle également ces facteurs *composantes principales*. Elles présentent la particularité suivante :

$$V(f_i) = \frac{1}{n}(Zu_i)^\mathsf{T}(Zu_i) \tag{6.35}$$
$$= \frac{1}{n} u_i^\mathsf{T} Z^\mathsf{T} Z u_i \tag{6.36}$$
$$= u_i^\mathsf{T} \frac{1}{n} Z^\mathsf{T} Z u_i \tag{6.37}$$
$$= u_i^\mathsf{T} R u_i \tag{6.38}$$
$$= \lambda_i u_i^\mathsf{T} u_i \tag{6.39}$$
$$= \lambda_i \tag{6.40}$$

où λ_i est la valeur propre associée au vecteur propre u_i. On obtient ces derniers résultats car par définition, une valeur propre est telle que :

$$Ru_i = \lambda_i u_i \tag{6.41}$$

On a de plus par construction :

$$u_i^{\mathsf{T}} u_i = 1 \tag{6.42}$$

Terminons par quelques remarques : on a vu tout d'abord que la variance d'un facteur est égale à la valeur propre associée au vecteur propre permettant d'obtenir ce facteur. Ces facteur étant orthogonaux, la variance de la somme des facteurs est la somme de la variance des facteurs, autrement dit la somme des valeurs propres. On a donc :

$$V\left(\sum_{i=1}^{k} f_i\right) = \sum_{i=1}^{k} V\left(f_i\right) \tag{6.43}$$

$$= \sum_{i=1}^{k} \lambda_i \tag{6.44}$$

Un autre résultat classique de l'algèbre linéaire, la trace d'une matrice carré est égale à la somme de ses valeurs propres. Dans notre cas, on a donc :

$$Tr(R) = \sum_{i=1}^{k} \lambda_i \tag{6.45}$$

On trouve donc que l'inertie associée à X (du moins à sa version centrée réduite) est exactement égale à l'inertie dans la nouvelle base. La différence tient au fait que dans la nouvelle base, les facteurs sont ordonnés par part d'inertie décroissante : du fait de l'alogorithme de maximisation permettant de trouver la solution à notre problème (alogrithme qui n'a pas été présenté ici), les vecteurs propres que l'on détermine arrive par valeurs propres décroissantes. Autrement dit, le facteur 1 est associé au vecteur propre pour lequel on a la plus grande valeur propre.

Pour résumer, l'ACP revient à remplacer les variables composant X qui sont corrélées, par de nouvelles variables, les composantes principales qui sont combinaisons linéaires des variables composant X, non corrélées entre elles, de variance maximale et les plus liées en un certain sens aux variables composant X. L'ACP est ce qu'on appelle une méthode factorielle linéaire.

6.2.2 Applications : les facteurs de la courbe des taux

On présente une courte application aux taux d'intéret de la méthode de l'ACP. On dispose dans une matrice X de n observations pour des taux swap de maturité constante, allant de 1 an à 30 ans. On obtient la matrice de corrélation suivante :

	1 an	**2 an**	**3 an**	**4 an**	**5 an**	**6 an**	**7 an**	**8 an**	**9 an**	**10 an**	**15 ans**	**20 ans**	**30 ans**
1 an	1,00	0,21	0,24	0,27	0,27	0,23	0,25	0,25	0,26	0,23	0,24	0,22	0,24
2 an	0,21	1,00	0,92	0,89	0,86	0,83	0,81	0,79	0,77	0,80	0,77	0,74	0,70
3 an	0,24	0,92	1,00	0,93	0,90	0,88	0,86	0,84	0,83	0,85	0,81	0,79	0,75
4 an	0,27	0,89	0,93	1,00	0,95	0,91	0,91	0,89	0,88	0,88	0,85	0,83	0,79
5 an	0,27	0,86	0,90	0,95	1,00	0,94	0,94	0,92	0,91	0,90	0,87	0,85	0,81
6 an	0,23	0,83	0,88	0,91	0,94	1,00	0,96	0,97	0,97	0,90	0,88	0,86	0,83
7 an	0,25	0,81	0,86	0,91	0,94	0,96	1,00	0,95	0,95	0,91	0,86	0,84	0,82
8 an	0,25	0,79	0,84	0,89	0,92	0,97	0,95	1,00	0,98	0,90	0,89	0,87	0,85
9 an	0,26	0,77	0,83	0,88	0,91	0,97	0,95	0,98	1,00	0,91	0,90	0,88	0,86
10 an	0,23	0,80	0,85	0,88	0,90	0,90	0,91	0,90	0,91	1,00	0,90	0,88	0,86
15 ans	0,24	0,77	0,81	0,85	0,87	0,88	0,86	0,89	0,90	0,90	1,00	0,98	0,94
20 ans	0,22	0,74	0,79	0,83	0,85	0,86	0,84	0,87	0,88	0,88	0,98	1,00	0,94
30 ans	0,24	0,70	0,75	0,79	0,81	0,83	0,82	0,85	0,86	0,86	0,94	0,94	1,00

Les deux graphiques suivantes présentent les valeurs propres et les vecteurs propres extraits de la base de données. Idéalement, on ne souhaite retenir que quelques uns de ces facteurs pour décrire le comportement de la courbe des taux. La méthode usuelles est la *méthode du coude* : on ne retient que les valeurs propres précédant la formation d'un coude sur le graphique. Ici, on ne retiendrait que 3 ou 4 facteurs. On considère en général que la courbe des taux est guidée par trois facteurs, dont le premier étant la source principale de ses mouvements. Ce facteur est en général identifié comme la politique monétaire.

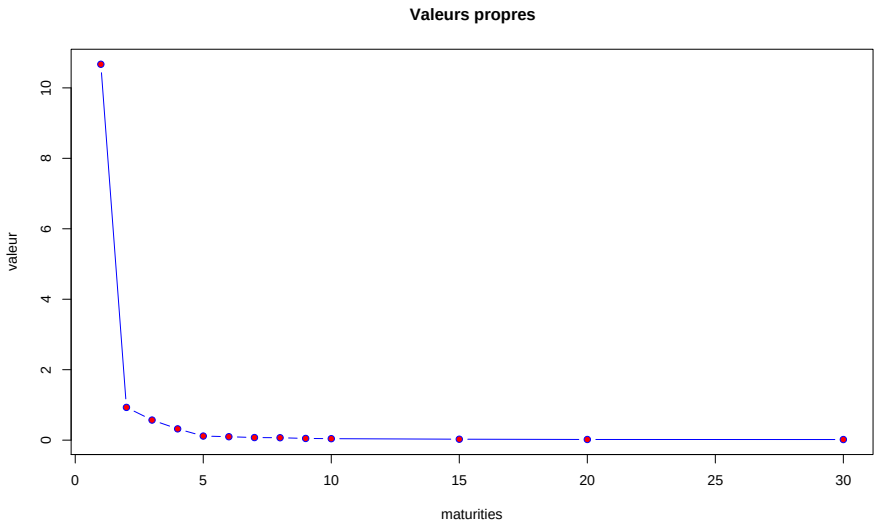

FIG. 6.3 – Valeurs propres

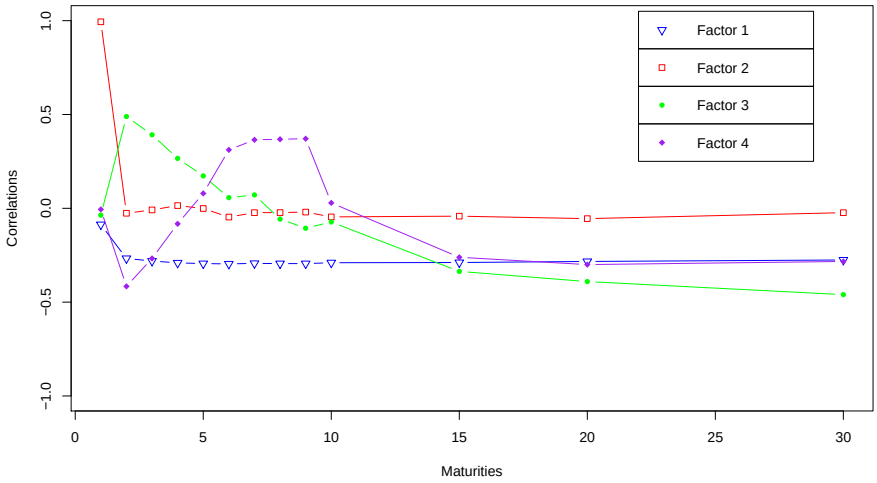

FIG. 6.4 – Vecteurs propres

Bibliographie

Aboura, S. (2005). Les modèles de volatilité et d'options. *Publibook*.

Berndt, E. K., Hall, R. E., Hall, B., and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, 3 :653–665.

Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, (81) :637–654.

Bollen, B. and Inder, B. (2002). Estimating Daily Volatility in Financial Market Utilizing Intra Day Data. *Journal of Empirical Finance*, 9 :551–562.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroscedasticity. *Journal of Econometrics*, 31 :307–328.

Cherubini, U., Luciano, E., and Vecchiato, W. (2005). *Copula Methods in Finance*. Wiley.

Cochrane, J. (2002). Asset Pricing. *Princeton University Press*.

Cochrane, J. (2005). Time series for macroeconomics and finance - Manuscrit. http ://gsbwww.uchicago.edu/fac/john.cochrane/research/Papers/time_series_book.pdf.

Crépon, B. (2005). Econométrie linéaire. *Cours ENSAE deuxième année*, http ://www.crest.fr/pageperso/crepon/poly052005.pdf.

Davidson, R. and MacKinnon, J. (1993). Estimation et inférence en Économétrie. *Oxford University Press*. http ://russell.vcharite.univ-mrs.fr/EIE/.

Deschamps, P. (2004). Cours d'économétrie. *Université de Neuchâtel*. http ://mypage.bluewin.ch/Philippe_Deschamps/Notes0405.pdf.

Duflo, M. (1996). *Algorithmes stochastiques*. Springer, Berlin, Heidelberg, New York.

Embrechts, P., Lindskog, F., and McNeil, A. (2001). Modelling dependance with copulas and applications to risk management. Preprint ETZH.

Embrechts, P., McNeil, A., and Straumann, D. (1999). Correlation and dependency in risk management : properties and pitfalls. Departement of Mathematik , ETHZ, Zürich, Working Paper.

Engle, R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50 :987–1007.

Engle, R., Lilien, D., and Robbins, R. (1987). Estimating Time Varying Risk Premia in the Term Structure : the ARCH-M Model. *Econometrica*, 55 :391–407.

Faraway, J. (2002). Practical regression and anova using R. http ://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Garman, M. and Klass, M. (1980). On the Estimation of Security Price Volatilities from Historical Data. *Journal of Business*, 53 :67–78.

Glosten, L., Jagannathan, R., and Runkle, D. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance*, 48 :1779–1801.

Gourieroux, C. (1992). Modèles ARCH et applications financières. *Economica*.

Gourieroux, C. and Jasiak, J. (2001). Econometrics of Finance. *Princeton University*.

Gourieroux, C., Montfort, A., and Trognon, A. (1984). Pseudo Maximum Likelihood Methods : Applications to Poisson Models. *Econometrica*, 52 :701–720.

Greene, W. H. (2002). Econometric analysis. *Prentice Hall*.

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

Harvey, A. C. (1990). The econometric analysis of time series. *LSE Handbooks in economics*.

Lopez, J. (2001). Evaluating the Predictive Accuracy of Volatility Models. *Journal of Forecasting*, 20 :87–109.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, M., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21 :1087–1092.

Münk, C. (2004). Asset Pricing Theory. *Lectures Notes for PhD Students*.

Nelson, D. (1991). Conditional Heteroscedasticity in Asset Returns : a New Approach. *Econometrica*, 59 :347–370.

Pagès, J. (2005). Statistiques Générales pour Utilisateurs - Tome 1 & 2. *Presses Universitaires de Rennes*.

Paradis, E. (2005). R pour les débutants. http ://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf.

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *Journal of Business*, 53 :61–65.

Poon, S.-H. (2005). A Practical Guide to Forecasting Financial Market Volatility. *Wiley Finance*.

Quinn, K. (2001). The newton raphson algorithm for function optimization. http ://www.stat.washington.edu/quinn/classes/536/notes/Newton.pdf.

Robert, C. (1996). *Méthodes de Monte Carlo par Chaînes de Markov*. Economica.

Saporta, G. (1988). Probabilités, Analyse des Données et Statistiques. *Technip*.

Taylor, S. (1986). Modelling Financial Time Series. *Wiley*.

Tsay, R. S. (2002). Analysis of Financial Time Series. *Wiley*.

VonSachs, R. and VanBellegem, S. (2002). Méthodes stochastiques appliquées à la prévision, séries chronologiques. *Université Catholique de Louvain*, http ://www.stat.ucl.ac.be/cours/stat2414/syllabus.pdf.

Wang, P. (2003). Financial Econometrics. *Routledge*.