Universitat de Barcelona
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Modeling Volatility using ARCH and GARCH Processes

Autor: Àlex Martorell i Locascio

Director: Dr. Josep Vives i Santa-Eulàlia
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, June 20, 2021

Abstract

The main goal of this thesis is to present ARCH and GARCH models, key concepts in Time Series Analysis. These models are extremely useful when describing the behavior of Financial time series (i.e. Stock returns, Exchange rates, Economic indexs...), because they deal with volatility (variation of price in a delimited time period). The discipline of Time Series Analysis as a whole is introduced, as well as its main concepts, which are needed to further explore the useful properties behind ARCH and GARCH. We put special emphasis in developing a consistent theory of Estimation and Forecasting (i.e determinate the parameters of a model and predict future values) for Time Series Analysis, which serves as a justification for estimating and predicting ARCH and GARCH models. Practical examples are given using the `R` statistics package.

2020 Mathematics Subject Classification. 62M10, 62P05

"S'ha d'escriure amb llibertat, amb gust, amb plaer,
però amb la màxima observació possible"
Josep Pla, *Notes del Capvesprol* (1979)

"The aims of life are the best defense against death"
Primo Levi, *The Drowned and the Saved* (1986)

Acknowledgements

I want to thank my advisor Dr. Josep Vives Santa-Eulalia, who through many meetings has helped me overcome many of the issues and problems that came up whilst writing this thesis. His advice and multiple suggestions have made this Final Degree Project something that I am proud of. Also, I want to thank my family who has given me unconditional support during all these years and especially during these last months.

Finally, I want to mention my brothers in arms: Maria, Alex, Marc, to whom I can only express gratitude for all continued help and support they have always given me. Thank you for sticking with me through thick and thin, highs and lows.

Contents

1 Introduction

Observation and intuition tells us that there are many phenomena that evolve through time. Whether it is the population of a country, the global temperature, company earnings, seismic waves or financial returns, we are in front of a broad range of measurements whose values change as time passes. Plotting the data according to its time t clearly showed us if there was an existing tendency or not. So the next question was: Could a model be inferred? Could we predict future values? The field of Time Series Analysis was born.

But a reasonable question might come to mind. Considering that Time Series Analysis is a fairly modern discipline in Mathematics and Statistics, what are its precedents and why do they fail? The answer is not trivial, but it has to do with the fact that previous models assumed that the observations or adjacent points in time are independent and identically distributed. It was not hard to see in a given plot that there was an underlying correlation between observed values, that in some way x_n depends on x_1. The concept *Auto-covariance* tackles this idea. The auto-covariance of a set of observations $\{x_1, \ldots, x_n\}$ if is inferred from the statistical notion of covariance ($\overline{x}$ is the sample mean of n observations)

$$\widehat{\gamma(h)} = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \overline{x})(x_t - \overline{x}).$$

We refer to h as the *lag*. The lag indicates the time difference between two values. Autocovariance (or autocorrelation, as we will see) is key, but another issue specific to time series rose: do they exhibit any sort of regularity over time? One can argue that financial returns do not feature this at a first glance. We will delve into this idea of regularity, which in the Time Series context we define as *stationarity*.

Stochastic Processes is a major field of interest in modern mathematics. We could call Time Series Analysis a subset (a large one, noticeably!) of this global set. Simplifying, a stochastic process is a sequence $\{X_t\}_{t\geq 0}$ of random variables in a probability set, which are indexed by time t, which is roughly the same as a time series. Hence, Time Series Analysis, inherits all the probability and statistical theory behind Stochastic Processes, but it focuses on developing mathematical results that serve the needs of analytical research: An estimation of a model that describes a time series, and how to predict future values at time $t+1, \ldots, t+h$.

Classical regression, which is a valid model for many situations, usually falls short for some Time Series. In the early 20th century, *Autoregressive models* became a revolution in the sense that they were not only considering the value at time t, but past values $t-1, \ldots, t-p$. Therefore, AR(p) (Autoregressive models of order p) expressed X_t as a function of past values:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \omega_t$$

where $\phi_1, \ldots, \phi_p$ are constants and ω_t is a white noise (mean 0 and finite variance).

Today, we can find an extensive and rigorous theory on autoregressive models. Many books and papers have made them the purpose of their investigation. However, there are not our object of study. 40 years go, mathematicians and economists noticed that AR(p) models were failing to model most cases of financial-type returns. Whether these were stock returns, exchange rates, or monetary measures (GDP, GNP), volatility was not part of the equation. This is because AR(p) type models are *homoskedastic*, meaning that the conditional variance (which is time-dependent) was constant. This does not work well for finance, since instability usually is clustered in very specific moments of time. As one notices, this does not happen in a population time series or in a study of average temperatures since brusque changes hardly ever occur. A need for new modeling options was growing.

Robert F. Engle (1942-), who is currently a professor at NYU Stern, is responsible for having developed ARCH in the 1980s, new models that captured volatility. Professor Engle was awarded the Nobel Prize in Economics in 2003 for this major discovery. Tim Bollerslev (1958-), instructor at Duke University, is credited for having furthered the investigation of ARCH models, and in his thesis presented in 1986 under the supervision of prof. Engle himself, he presented GARCH models, which expressed variance not only as a function of the past returns, but of the past variances as well.

This thesis is divided into two main parts. We start by presenting extensive theoretical background for the study of financial returns, which means we start with a thorough presentation of ARCH and GARCH models. This gives way to chapter 3, in which we give a strong mathematical theory concerning Estimation and Prediction of Time Series. As we mentioned earlier, these are two crucial aspects when studying object of this kind. The second part (end of chapter 3 and chapter 4) is devoted to a deep study of a few real situations of financial returns. Analysis is performed with the `R` statistical package, an excellent tool for Time Series and for any other statistical needs. We end our journey asking ourselves if deficiencies have been found in ARCH and GARCH models, and as we move into the 21st century, what is the current state of investigation in stochastic volatility models.

2 Time Series. ARCH and GARCH models

2.1 Basic Concepts in Time Series Analysis

In this chapter we will go over the basic definitions that are needed to begin our study of ARCH and GARCH models and their properties, which is the first goal of our research. Proofs of important results will be given, and for the rest we will refer to relevant literature.

Definition 2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability set. A *stochastic process* X is a map

$$X : (\omega, t) \in \Omega \times \mathbb{T} \longrightarrow X_t(\omega) \in \mathbb{R}$$

that is mesurable (i.e. $X^{-1}(B) \in \mathcal{F} \times \mathcal{B}$ for all $B \in \mathcal{B}(\mathbb{R})$) Note that we have a function dependent not only on time but also on an element that belongs to a probability set. Intuitively, we can say a stochastic process is a collection of random variables $X(t)$ indexed by $t \in \mathbb{T}$, where $\mathbb{T} = \mathbb{N}, \mathbb{Z}$. Since our focus are time series, from now on we will refer to the collection $\{X_t\}_{t\in\mathbb{N}}$ as a *time series*. We now introduce a few fundamental notions concerning the study of time series.

Definition 2.2. Let $\{X_t\}$ be a time series. For all $s, t \in \mathbb{T}$, the *autocovariance function* is:

$$\gamma(s,t) = E\big[\big(X_s - E(X_s)\big)\big(X_t - E(X_t)\big)\big].$$

Observation 2.3. Note the following:

(1) $\gamma(t,t) = E\big[\big(X_t - E(X_t)\big)^2\big]$ is the second-order moment.

(2) The term autocovariance can be used interchangeably with the term covariance. Autocovariance may be prefered since we are in a time series context.

The following lemma is a useful recall from Statistics.

Lemma 2.4. $\gamma(s,t) = E(X_sX_t) - E(X_s)E(X_t)$.

Proof.

$$\begin{aligned}
\gamma(s,t) &= E\big[\big(X_s - E(X_s)\big)\big(X_t - E(X_t)\big)\big] \\
&= E\big[X_sX_t - X_sE(X_t) - E(X_s)X_t + E(X_s)E(X_t)\big] \\
&= E(X_sX_t) - E(X_s)E(X_t) - E(X_s)E(X_t) + E(X_s)E(X_t) \\
&= E(X_sX_t) - E(X_s)E(X_t).
\end{aligned}$$

□

For simplicity, sometimes we want to express the set of all the autocovariance functions $\forall s, t \leq n$ We can express the time series as a vector of random variables
$X = (X_1, \dots, X_n)^T$

Definition 2.5. The *variance-covariance matrix* or *autocovariance matrix* is defined as

$$\Gamma(X) = E[(X - E(X))(X - E(X))^T].$$

An alternative expression which is easily deduced is:

$$\Gamma(X) = \begin{pmatrix} \text{Var}(X_1) & \gamma(1,2) & \dots & \gamma(1,n) \\ \gamma(2,1) & \text{Var}(X_2) & \dots & \gamma(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n,1) & \gamma(n,2) & \dots & \text{Var}(X_n) \end{pmatrix}.$$

Now let us consider combinations of random variables, which will be necessary in section 3. Given $X = (X_1, \dots, X_n)^T$ a vector of random variables, we write the linear combination of random variables as $L(X_1, \dots, X_n) = a_1X_1 + \dots + a_nX_n$ or in abbreviated notation as $L(X_1, \dots, X_n) = a^TX$ where $a = (a_1, \dots, a_n)^T$ is a vector of $a_1 \dots a_n \in \mathbb{R}$ constants.

Lemma 2.6. $Var(a^TX) = 0$ for $a \neq 0$ if and only if Γ is singular.

Proof. By definition of variance we have:

$$\begin{aligned} \text{Var}(a^TX) &= E\big(a^TX(a^TX)^T\big) - E(a^TX)(E(a^TX))^T \\ &= a^TE(XX^T)a - a^TE(X)E(X)^Ta \\ &= a^T\Big(E(XX^T) - E(X)E(X)^T\Big)a \\ &= a^T\Gamma(X)a \end{aligned}$$

the last equal sign is a consequence of a generalized version of Lemma 2.4. Also, since $\text{Var}(a^TX) = 0$ we have $a^T\Gamma(X)a = 0$ which means that $\Gamma(X)$ has a 0 eigenvalue and hence it is singular. We recall from a Linear Algebra result that a matrix A is invertible if and only if every eigenvalue is nonzero. We now have both implications, which concludes the proof. □

Definition 2.7. Let X be a random variable with $E(X) = \mu$ and variance $\text{Var}(X) = \sigma^2$. The *skewness* and *kurtosis* are respectively defined as:

- $S(X) = E\left(\dfrac{X-\mu}{\sigma}\right)^3$
- $K(X) = E\left(\dfrac{X-\mu}{\sigma}\right)^4$

Both skewness and kurtosis refer to the shape of our distribution. Skewness measures the asymmetry of a given distribution whereas kurtosis depicts the tail thickness of a given distribution. If $S(X) > 0$ the asymmetry is to the right (the right tail is longer), and the opposite holds for $S(X) < 0$. Kurtosis larger than 3 means that our distribution is"heavy-tailed" and we call it *leptokurtic.* On the contrary, if the kurtosis is smaller than 3 our distribution is "light-tailed" and it is called *platykurtic.*

Definition 2.8. A time series $\{Y_t \ t \in \mathbb{Z}\}$ is *strictly stationary* if the law of the set $\{Y_{t_1}, Y_{t_2}, \dots Y_{t_k}\}$ is identical to the law of the shifted set $\{Y_{t_1+h}, Y_{t_2+h}, \dots Y_{t_k+h}\}$ (where $h \geq 0$).

Definition 2.9. A time series $\{Y_t \ t \in \mathbb{Z}\}$ is *stationary* if the following conditions are fulfilled:

1. $E(Y_t) = \mu$ i.e. the expected value is constant and does not depend on time.
2. $\gamma(s,t)$ depends only on the difference $|s-t|$ noted by l, which we call the *lag*.

Observation 2.10. In the case of a stationary time series, we shall write $\gamma(l) = \gamma(s,t)$ where $l = s - t$.

Definition 2.11. The *autocorrelation function* of a stationary time series is:

$$\rho(l) = \frac{\gamma(l)}{\gamma(0)}, \quad l \geq 0$$

Proposition 2.12. Let $\{Y_t\ t \in \mathbb{Z}\}$ be a stationary time series. The following holds:

(a) $\gamma(l) = \gamma(-l)$ (symmetry)

(b) $|\gamma(l)| \leq \gamma(0)$

(c) $-1 \leq \rho(l) \leq 1$

Proof.

(a) $\gamma(l) = \gamma(t+l-t) = E\big[\big(Y_{t+l} - E(Y_{t+l})\big)\big(Y_t - E(Y_t)\big)\big]$
$= E\big[\big(Y_t - E(Y_t)\big)\big(Y_{t+l} - E(Y_{t+l})\big)\big] = \gamma(t-(t+l)) = \gamma(-l)$.

(b) $|\gamma(l)| = \big|E\big(Y_t - E(Y_t)\big)\big(Y_{t+l} - E(Y_{t+l})\big)\big| \leq E\big(\big|Y_t - E(Y_t)\big|\big|Y_{t+l} - E(Y_{t+l})\big|\big)$
$\leq E\big(\big|Y_t - E(Y_t)\big|^2\big)^{\frac{1}{2}}\big(\big|Y_{t+l} - E(Y_{t+l})\big|^2\big)^{\frac{1}{2}} = \sqrt{\gamma(0)}\sqrt{\gamma(0)} = \gamma(0)$, where in the first inequality we have used Jensen's inequality and in the second inequality we apply Cauchy-Schwarz's Inequality.

(c) Using (b) we have $\left|\frac{\gamma(l)}{\gamma(0)}\right| \leq \frac{\gamma(0)}{\gamma(0)} = 1$.

□

Examples 2.13. Relevant Time Series examples

1. **White Noise (WN)**. Uncorrelated sequence of random variables $\{\varepsilon_t\}$ with $E(\varepsilon_t) = 0$ and finite and constant variance $\mathrm{Var}(\varepsilon_t) = \sigma^2$ for all t. If the random variables are independently and identically distributed we shall write ε_t ∼**iid noise**. If the random variables have a normal law, we call it a **Gaussian White Noise (GWN)**.

2. **AR(p)**. An *autoregressive model of order* p is defined as the time series $\{Y_t\ t \in \mathbb{Z}\}$ such that

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \phi_p Y_{t-p} + \omega_t$$

where ω_t is a GWN. We can rewrite this as $\phi(B)Y_t = \omega_t$ where $\phi(B)$ is called the *autoregressive operator* and is defined as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

where B is the *backshift operator*: $BY_t = Y_{t-1}$.

3. **MA(q)**. A *moving average model of order q* is defined as the time series $\{Y_t\ t \in \mathbb{Z}\}$ such that

$$Y_t = \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_q \omega_{t-q}$$

where ω_t is GWN. As in the previous example, we consider writing the model as $Y_t = \theta(B)\omega_t$ where

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$$

is the moving average operator.

4. **ARMA(p,q)** An *autoregressive moving average model* is a stationary time series $\{Y_t\ t \in \mathbb{Z}\}$ such that

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \phi_p Y_{t-p} + \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_q \omega_{t-q}$$

where ω_t is GWN. It is evident that AR(p) and MA(q) models are restrictions of the ARMA(p,q) when $q = 0$ and $p = 0$, respectively. We can abbreviate the expression above by using the operators introduced, we write: $\phi(B)Y_t = \theta(B)\omega_t$

ARMA models are crucial for a theory of Time Series, though they are not the main focus of this thesis. However, there are two main concepts of ARMA models which are worth mentioning. These are causality and invertibility. We introduce the ARMA invertible model now, and we will talk about the causal ARMA model in section 3.5.

Definition 2.14. Let $\{Y_t\}$ be an ARMA model with its corresponding polynomial representation $\phi(B)Y_t = \theta(B)\omega_t$. We say that $\{Y_t\}$ is *invertible* if there exists a sequence of constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$\omega_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}, \quad t = 0, 1 \ldots, n.$$

The following result assesses the invertibility of ARMA processes. We include it because the proposition that states the condition for stationary GARCH's will use it.

Theorem 2.15. Let $\{Y_t\}$ be an ARMA(p,q) process. Given the polynomial representation of this process we suppose that the polynomials $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeroes. Then $\{Y_t\}$ is invertible if and only if $\theta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficients $\{\pi_j\}$ (Definition 2.14) are determined by the relation

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)} \quad |z| \leq 1. \tag{2.1}$$

Proof. By hypothesis, $\theta(z) \neq 0$ if $|z| \leq 1$. Then, $1/\theta(z)$ can be expanded as a power series

$$\frac{1}{\theta(z)} = \sum_{j=0}^{\infty} \eta_j z^j = \eta(z), \quad |z| < 1 + \varepsilon$$

for $\varepsilon > 0$. If we apply $\eta(B)$ to both sides of the equation $\phi(B)Y_t = \theta(B)\omega_t$ we get

$$\eta(B)\phi(B)Y_t = \eta(B)\theta(B)\omega_t = \omega_t.$$

This would give us

$$\omega_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}$$

where the sequence $\{\pi_j\}$ is the one determined in (2.1). We got to the definition of an invertible process as we wanted to. Evidently, these last steps are not as immediate as they seem. One has to argue that the operators $\eta(z) = \sum_{j=0}^{\infty} \eta_j z^j$ for $\sum_{j=0}^{\infty} |\eta_j| < \infty$ inherit the algebraic properties of power series (in this case, the product). This, is a consequence of a technical probability result which shows that series of the type $\psi(B)Y_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j}$ converges absolutely with probability one and in mean square to the same limit (Y_t has to be stationary). In addition, if we write $Z_t = \psi(B)Y_t$ then the process is stationary with autocovariance function $\gamma_Z(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \gamma(h-j+k)$. In other words, stationarity and the autocovariance function are preserved by the operator $\psi(B)$. This clearly exceeds our scope, but it must be mentioned.
Conversely, we now suppose that $\{Y_t\}$ is invertible, which means $\omega_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}$ for a sequence of coefficients $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$. Then:

$$\phi(B)\omega_t = \pi(B)\phi(B)Y_t = \pi(B)\theta(B)\omega_t$$

If we define $\xi(z) = \pi(z)\theta(z) = \sum_{j=0}^{\infty} \xi_j z^j$ with $|z| \leq 1$, we can write this as

$$\sum_{j=0}^{p} \theta_j \omega_{t-j} = \sum_{j=0}^{\infty} \xi_j \omega_{t-j}$$

and if we multiply on both sides by ω_{t-k} and take the expected value we obtain $\xi_k = \theta_k$, for $k = 0, \ldots, p$ and $\xi_k = 0$ for $k > p$. Therefore:

$$\phi(z) = \xi(z) = \pi(z)\theta(z), \quad |z| \leq 1.$$

By hypothesis we have that $\phi(z)$ and $\theta(z)$ have no common zeroes, and since $|\pi(z)| \leq \infty$ for $|z| \leq 1$, we conclude that $\theta(z)$ cannot be zero for $|z| \leq 1$. □

2.2 Financial theory basics

Definition 2.16. The *simple return* of an asset at time t is defined by the following quotient:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

For instance, if we intend to work with daily returns, we subtract the closing price with the opening price and divided it by the opening price. If we want to calculate the average return over k periods, this is defined as follows:

$$R_t(k) = \left(\prod_{j=0}^{k-1} (1 + R_{t-j}) \right)^{\frac{1}{k}} - 1.$$

Definition 2.17. The *log return* at time t is defined as the following quotient:

$$r_t = \log \frac{P_t}{P_{t-1}} = \log(1 + R_t).$$

If we want to average the return for k periods we get

$$r_t(k) = \log(1 + R_t(k)) = \frac{1}{k} \log \prod_{j=0}^{k-1} (1 + R_{t-j}) = \frac{1}{k} \sum_{j=0}^{k-1} \log(1 + R_{t-j}) = \frac{1}{k} \sum_{j=0}^{k-1} r_{t-j}.$$

We note that the final expression is the arithmetic average. One might ask themselves how simple and log returns compare. Taylor approximation of $\log(1 + x)$ shows us that the difference is negligible:

$$\log(1 + x) = x + O(x^2).$$

This translates as $\log(1 + R_t) \approx R_t$ for R_t close to zero. A rule of thumb is that if one is studying time series with a high frequency, i.e. daily values, if R_t are under 10% we can use simple returns and log returns indistinctly.

In finance, asset returns may not follow a linear trend, and the variation in price may be substantial in short periods of time. This is what volatility tackles. Heuristically, we can say that volatility is a measurement that shows us the degree of variation in prices. Mathematically this reminds us of standard deviation, but we will refine this definition in the next section.

2.2.1 Stylized facts

The term "stylize" means to conform to a particular style, to conventionalize. So a *stylized fact* tends to be a broad generalization that is fundamentally true, but may contain some imprecisions in the detail. These facts are employed in many social sciences, also in the economic market. Prof. Rama Cont, in [3] developed a thorough list, which we summarize below.

- "Heavy tails": The unconditional distribution of returns has heavier / fatter tails than a normal distribution. This is because crashes and booms happen often. Cont also notes that even after correcting the returns, which means that we take into account volatility clustering (GARCH models) the conditional distribution still shows heavy tails.
- "Asymmetry": The unconditional distribution of returns is negatively skewed. This means that negative returns (i.e. losses) are more common than positive ones.
- "Aggregated Gaussianity": If we should increase the time interval in which we calculate the returns (weekly, monthly etc.) we notice that the unconditional returns get closer to a normal distribution.
- "Absence of correlation / Strong correlation in absolute returns": For the most part, returns will not show autocorrelation except for very small time intervals. However, if we consider the absolute returns time series, we will observe strong autocorrelation. Cont describes that it decays over time similar to an exponential distribution with $\beta \in [0.2, 0.4]$.
- "Volatility clustering": This means that periods of volatility are clustered, which causes a positive autocorrelation. Large returns will be followed by large returns, positive or negative (recall previous fact). This has to do with *heteroskedasticity.*

We will see this in more detail, but if $\{\mathcal{F}_{t-1}\}$ is the "information available" at time $t-1$, we have

$$\operatorname{Var}(r_t \mid \mathcal{F}_{t-1}) \neq \operatorname{Var}(r_{t-1})$$

meaning that the variability depends on the recent changes.

- "Correlation varies through time": It may be high during periods with increased volatility.

2.3 Review and Relevant Results in Probability Theory

What follows is a brief probability review that helps us to define the notion of volatility in financial markets, but most importantly, it allows us to deepen our study of ARCH and GARCH models. We also include some important results which we will need in order to prove some theorems for ARCH and GARCH.

Definition 2.18. Let $(\Omega, \mathcal{F}, P)$ be a probability set. If $\mathcal{G} \subseteq \mathcal{F}$ is a σ-algebra, we define the random variable

$$E(X \mid \mathcal{G})$$

as the *conditional expectation.* Using the Radon-Nykodim theorem we can prove the existence and uniqueness of the variable, as well as its measurability, but this exceeds our purpose.

Under the same conditions, we define the *conditional variance* as

$$\operatorname{Var}(X \mid \mathcal{G}) = E[(X - E(X|\mathcal{G}))^2)|\mathcal{G}].$$

Lemma 2.19. Under the previous assumptions, let X, Y be integrable random variables. The following holds.

(1) $E(E(X \mid \mathcal{G})) = E(X)$ (law of total expectation)

(2) If Y is $\mathcal{G}$-mesurable and $X \cdot Y$ is integrable, then $E(XY \mid \mathcal{G}) = YE(X \mid \mathcal{G})$

(3) $\operatorname{Var}(X) = E(\operatorname{Var}(X \mid \mathcal{G})) + \operatorname{Var}(E(X \mid \mathcal{G}))$ (law of total variance) □

Definition 2.20. Let $(\Omega, \mathcal{F}, P)$ be a probability set. A *filtration* is a sequence of σ-algebras $F := \{\mathcal{F}_n, n \in \mathbb{T}\}$ such that

- $\mathcal{F}_n \subseteq \mathcal{F}$, $\forall n \in \mathbb{T}$
- $\mathcal{F}_{n-1} \subseteq \mathcal{F}_n$, $\forall n \in \mathbb{T}$.

Theorem 2.21. (Kolmogorov's Strong Law of Large Numbers) Let $\{X_n\ n \geq 1\}$ be a sequence of random i.i.d variables. Let us suppose $E(X_1) < \infty$. Then:

$$\lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = E(X_1) \quad \text{a.s.}$$

Proof. cf [10], pp.175-185. □

Theorem 2.22. (Chung-Fuchs) Let $\{X_n \; n \geq 1\}$ be a sequence of random i.i.d variables. If $E(X_1) = 0$ and $E(|X_1|) > 0$, then $\limsup \sum_{j=1}^{n} X_i = +\infty$.

Proof. cf [13]. □

We gave an heuristic introduction to volatility in the previous section, but the question is if there is a way to model it in mathematical terms. As we will show in detail, ARCH and GARCH models appeared in the 80s because previous models were grossly failing when modeling financial returns. A coherent but not widely accepted definition of volatility is the subsequent definition.

Definition 2.23. The volatility of a time series $\{X_t\}$ is defined as the conditional variance $\mathrm{Vol} = \mathrm{Var}(X_t|\mathcal{F}_{t-1})$.

2.4 ARCH and GARCH models

The question we must assess is the growing need for ARCH and GARCH models. A naive answer to this question could start by stating that ARMA type models assume constant conditional variance. This assumption becomes a an inconvenient when we try to model the volatility of the finance market, since the definition of volatility and the stylized facts go against it. This was one of the reasons led to the appearance of ARCH and GARCH models, which we owe to Engle in 1982, and later developments by Bollerslev in 1986. In 2003, Engle was awarded the Nobel Prize in Economics. ARCH models were cited as one of his most important contributions.

Definition 2.24. A time series is *homoskedastic* if $Var(Y_t|\mathcal{F}_{t-1}) = \sigma^2$ for all t. If this does not hold, the time series is *heteroskedastic.*

Example 2.25. It is easily shown that the AR(1) model is homoskedastic. Note also that in an AR(1) model $E(Y_t \mid \mathcal{F}_{t-1}) = \phi Y_{t-1}$ and $\mathrm{Var}(Y_t \mid \mathcal{F}_{t-1}) = \mathrm{Var}(\varepsilon_t) = \sigma^2$.

Definition 2.26. A time series $\{Y_t \; t \in \mathbb{Z}\}$ is an ARCH(p) model (Auto Regressive Conditional Heteroskedastic) if its second order moment is finite and causal (not future dependent) such that:

$$Y_t = \sigma_t \varepsilon_t \quad t \in \mathbb{Z}$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i}^2 \quad p \geq 1$$

where ε_t is a GWN$(0,1)$ and $\alpha_0, \ldots, \alpha_p$ are the parameters of the model, which are positive.

Observation 2.27. The definition above of an ARCH model is the most widely used, and we follow many texts in Time Series such as [1] that use it. However, other sources give other less restrictive definitions of an ARCH model. In the case of [5], the authors distinguish between strong ARCH, semi-strong ARCH or weak ARCH. For instance, in the case of the semi-strong ARCH, we do not require for ε_t to be i.i.d. Other authors consider that ε_t need not be GWN(0,1), and relax the condition to i.i.d. with mean 0 and variance 1. These considerations may alter some theoretical results.

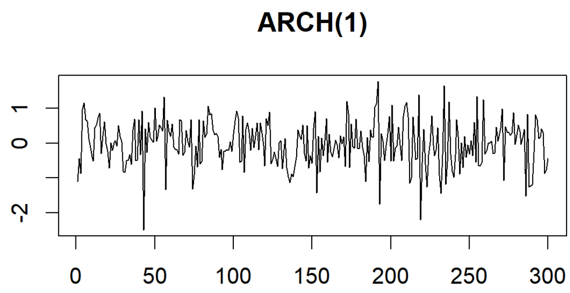

Figure 2.1: Generated ARCH(1) model, with $\alpha_0 = 0.25$ and $\alpha_1 = 0.5$

Proposition 2.28. The conditional variance of the process Y_t is σ_t^2, and it is autoregressive. Moreover, ARCH(p) is heteroskedastic.

Proof. Let $\mathcal{F}_t = \sigma\{Y_k \ k \leq t\}$ be the natural filtration of the process. We want to compute the conditional variance of Y_t, as it follows:

$$\begin{aligned}\operatorname{Var}(Y_t \mid \mathcal{F}_{t-1}) &= E(Y_t^2 \mid \mathcal{F}_{t-1}) - E\big(Y_t \mid \mathcal{F}_{t-1}\big)^2 = E(\sigma_t^2\varepsilon_t^2 \mid \mathcal{F}_{t-1}) - E\big(\sigma_t\varepsilon_t \mid \mathcal{F}_{t-1}\big)^2 \\ &= E(\sigma_t^2\varepsilon_t^2 \mid \mathcal{F}_{t-1}) - \sigma_t E\big(\varepsilon_t \mid \mathcal{F}_{t-1}\big)^2 = \sigma_t^2 E(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) - \sigma_t E\big(\varepsilon_t\big)^2 \\ &= \sigma_t^2 E(\varepsilon_t^2) = \sigma_t^2\end{aligned}$$

where have used that since ε_t is GWN. Also, by definition of the variance of ε_j, $E(\varepsilon_t^2) = 1$ which gives us our conclusion. Note that the conditional variance is autorregressive by its definition. Since it is not a constant, this model is heteroskedastic. □

Proposition 2.29. The returns Y_t are a *martingale difference*: they have mean zero and are uncorrelated.

Proof. Let $\mathcal{F}_t = \sigma\{Y_k \ k \leq t\}$ be the natural filtration of the process. Then, using Lemma 2.19 (1):

$$E(Y_t) = E\big(E(\sigma_t\varepsilon_t \mid \mathcal{F}_{t-1})\big).$$

Let us calculate the conditional mean:

$$E(Y_t \mid \mathcal{F}_{t-1}) = E(\sigma_t\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t E(\varepsilon_t \mid \mathcal{F}_{t-1}) = E(\varepsilon_t) = 0$$

where we have used property 2.19 (2) and the fact the σ_t^2 is $\mathcal{F}_{t-1}$-measurable. We also note that ε_t is independent of $\mathcal{F}_{t-1}$. Going back to our initial calculation we now get $E(Y_t) = E(\sigma_t E(0)) = 0$. We now compute the autocovariance function

$$\begin{aligned}\gamma(l) &= \operatorname{cov}(Y_t, Y_{t+l}) = E(Y_t - E(Y_t))(Y_{t+l} - E(Y_{t+l})) \\ &= E(Y_t Y_{t+l}) = E(E(Y_t Y_{t+l} \mid \mathcal{F}_{t+l-1})) \overset{(1)}{=} E(Y_t E(Y_{t+l} \mid \mathcal{F}_{t+l-1})) = 0.\end{aligned}$$

(1) follows from Lemma 2.19 (2), since Y_t is $\mathcal{F}_{t+l-1}$-measurable (or in other words, it belongs to the generated filtration). The final equality is inferred by what we just saw before. □

Proposition 2.30. The squared returns Y_t^2 of an ARCH(1) follow an AR(1) model.

Proof. We start by squaring Y_t and writing the ARCH(1) equations as follows

$$Y_t^2 = \sigma_t^2 \epsilon_t^2$$
$$\alpha_0 + \alpha_1 Y_{t-1}^2 = \sigma_t^2.$$

Subtracting them we get $Y_t^2 - \alpha_0 - \alpha_1 Y_{t-1}^2 = \sigma_t^2(\epsilon_t^2 - 1)$ Writing $v_t = \sigma_t^2(\epsilon_t^2 - 1)$ gives us $Y_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + v_t$, which is an AR(1) model as long as we prove v_t is a white noise. We must show then that the expected value is 0, the variance is finite and constant and last, that given a lag l, the two different instances are uncorrelated. As said, the expected value and the variance must be constant.

$$E(v_t) = E(\sigma_t^2(\epsilon_t^2 - 1)) = E(E(\sigma_t^2(\epsilon_t^2 - 1) \mid \mathcal{F}_{t-1})) = E(\sigma_t^2 E(\epsilon_t^2 - 1 \mid \mathcal{F}_{t-1})) = \\ E(\sigma_t^2)E(\epsilon_t^2 - 1) = E(\sigma_t^2)\big(E(\epsilon_t^2) - E(1)\big) = 0$$

where we have applied that σ_t^2 is $\mathcal{F}_{t-1}$-measurable and ε_t is independent from $\mathcal{F}_{t-1}$ because it is GWN.

$$\mathrm{Var}(v_t) = E(v_t^2) - E(v_t)^2 = E(E(v_t^2 \mid \mathcal{F}_{t-1})) = E(E(\sigma_t^4(\varepsilon_t^2 - 1)^2 \mid \mathcal{F}_{t-1})) = \\ E(\sigma_t^4 E((\varepsilon_t^2 - 1)^2 \mid \mathcal{F}_{t-1})) = E(\sigma_t^4)E((\varepsilon_t^2 - 1)^2) = E(\sigma_t^4)\mathrm{Var}(\varepsilon_t^2) = 2E(\sigma_t^4)$$

where we have used $E((\varepsilon_t^2 - 1)^2) = E(\varepsilon_t^2 - E(\varepsilon_t^2)) = \mathrm{Var}(\varepsilon_t^2) = Var(\chi_1^2) = 2$. The variance of v_t is constant, since $E(Y_t^4) = E(\sigma_t^4 \varepsilon_t^4) = E(\sigma_t^4) \cdot 3$. This is inferred from $\mathrm{Var}(\varepsilon_t^2) = E(\varepsilon_t^4) - E(\varepsilon_t)^4$ Finally, we check the autocovariance function (i.e. v_t is a centered white noise):

$$\gamma(l) = \mathrm{cov}(v_t, v_{t+l}) = E(\sigma_t^2(\varepsilon_t^2 - 1)\sigma_{t+l}^2(\varepsilon_{t+l}^2 - 1)) = E(E[\sigma_t^2(\varepsilon_t^2 - 1)\sigma_{t+l}^2(\varepsilon_{t+l}^2 - 1) \mid \mathcal{F}_{t+l-1}]) = \\ E[\sigma_t^2(\varepsilon_t^2 - 1)E(\sigma_{t+l}^2(\varepsilon_{t+l}^2 - 1) \mid \mathcal{F}_{t+l-1})] = 0$$

because ε_t^2 is independent of $\mathcal{F}_{t+l-1}$ and σ_t^2 is $\mathcal{F}_{t+l-1}$-measurable. □

Theorem 2.31. If $\alpha \geq 0$ and $\alpha \in [0, 1)$, the process:

$$Y_t = \sqrt{\sum_{l=0}^{\infty} \alpha_0 \alpha_1 \varepsilon_t^2 \varepsilon_{t-l}^2}$$

is a strictly stationary process that solves ARCH(1) equations.

Proof. Found in [2]. □

Corollary 2.32. The expected value and the variance of the returns are constant provided $\alpha_1 \in [0, 1)$.

Proof. This is immediate: Since Y_t is strictly stationary, the expected value and the variance are constant. Let us calculate $E(Y_t^2)$ (which is the actual variance, since we have shown $E(Y_t) = 0$). If we define $\mathcal{F}_t = \sigma\{Y_k \ k \leq t\}$,

$$E(Y_t^2) = E(E(Y_t^2|\mathcal{F}_{t-1})) = E(E(\sigma_t^2\varepsilon_t^2|\mathcal{F}_{t-1})) = E(\sigma_t^2 E(\varepsilon_t^2|\mathcal{F}_{t-1})) = E(\sigma_t^2 E(\varepsilon_t^2)) = E(\sigma_t^2).$$

Therefore:

$$E(Y_t^2) = E(\sigma_t^2) = E(\alpha_0 + \alpha_1 Y_{t-1}^2).$$

If we apply what we just saw, which is $E(Y_t^2) = E(Y_{t-1}^2)$

$$E(Y_t^2) = \frac{\alpha_0}{1-\alpha_1}.$$

□

Proposition 2.33. The returns Y_t follow a leptokurtic distribution, and we have

$$K(Y_t) = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2}.$$

Proof. We start by observing:

$$E(\sigma_t^4) = E((\alpha_0 + \alpha_1 Y_{t-1})^2) = \alpha_0^2 + \alpha_1^2 E(Y_{t-1}^4) + 2\alpha_0\alpha_1 E(Y_{t-1}^2).$$

By the previous corollary we have:

$$E(Y_t^2) = \frac{\alpha_0}{1-\alpha_1}.$$

Since Y_t is strictly stationary, we know that $E(Y_{t-1}^4)$ is constant for any t. Recall also from 2.30 that $E(Y_t^4) = 3E(\sigma_t^4)$. If we write $k = E(Y_{t-1}^4)$

$$\frac{k}{3} = \alpha_0^2 + \alpha_1^2 k + 2\alpha_0\alpha_1\frac{\alpha_0}{1-\alpha_1}$$
$$(1-3\alpha_1^2)k = 3\alpha_0^2 + 6\alpha_0^2\alpha_1\frac{1}{1-\alpha_1}.$$

Finally, we get:

$$E(Y_t^4) = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1)^2}.$$

Observe this quantity is well defined as long as $\alpha_1 < \frac{1}{\sqrt{3}} \simeq 0.5774$. By definition of kurtosis:

$$K(Y_t) = \frac{E(Y_t^4)}{(\mathrm{Var}(Y_t^2))^2} = \frac{3(1-\alpha_1)(1+\alpha_1)}{(1-3\alpha_1^2)} = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2} \geq 3.$$

□

The kurtosis is greater or equal than 3, which is the definition of a leptokurtic distribution. This satisfies the stylized fact which stated that returns have ”heavy-tails”.

Example 2.34. We consider ARCH(1) model generated with parameters $\alpha_0 = 0.25$ and $\alpha_1 = 0.5$ shown in Figure 2.1. Now we can deepen our study thanks to results obtained above. If we calculate the mean of the returns, we get:

```
> mean(y)
[1] -0.0146146
```

This is close to zero. As it is obvious, calculating the mean at a given time t will never yield an exact zero value. If we plot the ACF we see that the returns are uncorrelated. We build a plot for the volatility so we can see that the conditional variance changes through time which means the process is heteroskedastic.

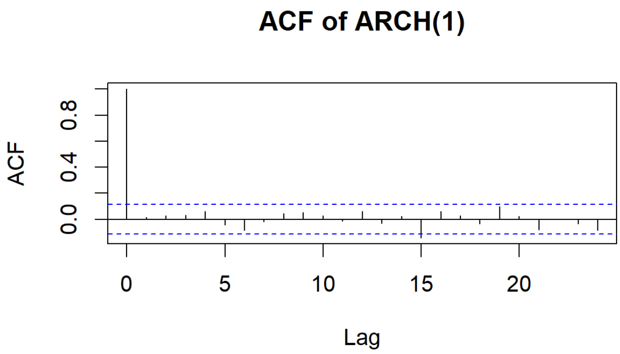

Figure 2.2: ACF function for the generated ARCH(1) model , with $\alpha_0 = 0.25$ and $\alpha_1 = 0.5$

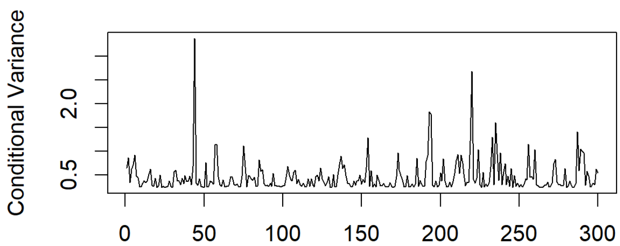

Figure 2.3: Volatility plot of the ARCH(1) model generated in figure 2.1

Bollerslev wrote in 1986 an article [12], where he developed an extension of the ARCH model called the GARCH model (G stands for Generalized). The main difference is that GARCH is a function of two parameters, but it maintains the underlying structure of the ARCH process.

Definition 2.35. GARCH (1,1) is defined as:

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \end{aligned} \tag{2.2}$$

where ε_t is GWN(0,1) and the parameters $\alpha_0, \alpha_1, \beta_1$ are the parameters of the model and greater than 0.

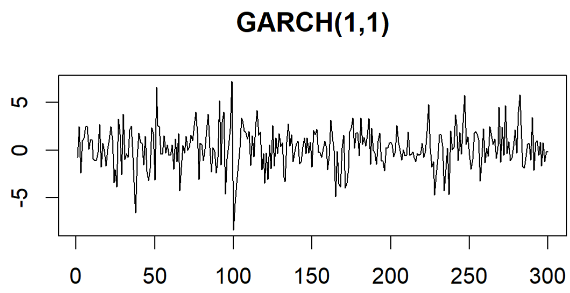

Figure 2.4: Generated GARCH(1,1) model, with $\alpha_0 = 0.5$, $\alpha_1 = 0.2$ and $\beta_1 = 0.7$

Lemma 2.36. The squared returns of a GARCH(1,1) model can be rewritten as an ARMA(1,1) model

Proof. Substracting the term σ_t^2 on both sides of the equation in (2.2) gives us

$$Y_t^2 - \sigma_t^2 = \sigma_t^2(\varepsilon_t^2 - 1). \tag{2.3}$$

For time $t-1$ we consider the same substraction multiplied by β_1:

$$\beta_1(Y_{t-1}^2 - \sigma_{t-1}^2) = \beta_1\sigma_{t-1}^2(\varepsilon_{t-1}^2 - 1). \tag{2.4}$$

Substracting (2.3) - (2.4) we get:

$$\beta_1(Y_{t-1}^2 - \sigma_{t-1}^2) - Y_t^2 + \sigma_t^2 = \beta_1\sigma_{t-1}^2(\varepsilon_{t-1}^2 - 1) - \sigma_t^2(\varepsilon_t^2 - 1).$$

We define $v_t = \sigma_t^2(\varepsilon_t^2 - 1)$:

$$\beta_1(Y_{t-1}^2 - \sigma_{t-1}^2) - Y_t^2 + \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1\sigma_{t-1}^2 = \beta_1 v_{t-1} - v_t$$

Finally, rearranging the terms gives us: $Y_t^2 = \alpha_0 + (\alpha_1 + \beta_1)Y_{t-1}^2 + v_t - \beta_1 v_{t-1}$. ☐

Proposition 2.37. Let $\{Y_t\}$ be a GARCH(1,1) model. The following holds:

(1) If $\mathrm{Var}(Y_t) = \sigma^2$ then $\mathrm{Var}(Y_t) = \dfrac{\alpha_0}{1 - \alpha_1 - \beta_1}$.

(2) $K(Y_t) = \dfrac{3(1 + \alpha_1 + \beta_1)(1 - \alpha_1 - \beta_1)}{1 - \beta_1^2 - 2\alpha_1\beta_1 - 3\alpha_1^2}$ and it is a leptokurtic distribution.

Proof.

(1) Using similar arguments as in Proposition 2.33 we have

$$\begin{aligned}\sigma^2 = \mathrm{Var}(Y_t) = E(Y_t^2) &= E(\alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1\sigma_{t-1}^2)\\ &= \alpha_0 + \alpha_1\sigma^2 + \beta_1 E(\sigma_{t-1}^2) = \alpha_0 + \sigma^2(\alpha_1 + \beta_1)\end{aligned}$$

It easy to prove that $E(\sigma_{t-1}^2) = \sigma^2$ as long as we use that $\sigma_{t-1}^2 = Var(Y_{t-1}|\mathcal{F}_{t-2}$. Rearranging the terms gives us

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}$$

Note that we need $\alpha_1 + \beta_1 < 1$, which is the condition for second-order stationary, as we will see later on.

$\square$

After a few basic properties for ARCH(1) and GARCH(1,1) have been discussed, we can go on a never-ending path of complex mathematical results that have been . One of the main and most remarkable ones is from 1990 and was proven by Nelson in [7] and can also be found [6]. We include it below.

Lemma 2.38. (Root test) Given $\sum a_n$, infinite series, if the quantity $\limsup_{n\to\infty} \sqrt[n]{|a_n|}$ is less than 1 the series converges absolutely. It diverges if the quantity is strictly bigger than 1.

Theorem 2.39. Suppose $\alpha_0 > 0$ and $\alpha_1, \beta_1 \geq 0$. The GARCH(1,1) equations have a strictly stationary solution if and only if $E[\log(\alpha_1\varepsilon_{t-1}^2 + \beta_1)] < 0$, which is unique and determined by:

$$\sigma_t^2 = \alpha_0\Bigg(1 + \sum_{j=1}^{\infty}\prod_{i=1}^{j}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\Bigg). \tag{2.5}$$

Proof. We write the conditional variance as $\sigma_t^2 = \alpha_0 + (\alpha_1\varepsilon_{t-1}^2 + \beta_1)\sigma_{t-1}^2)$. If we reiterate the expression in the following manner:

$$\sigma_t^2 = \alpha_0 + (\alpha_1 + \varepsilon_{t-1}^2 + \beta_1)\Bigg(\alpha_0 + (\alpha_1\varepsilon_{t-2}^2 + \beta_1)\Bigg(\alpha_0 + (\alpha_1\varepsilon_{t-3}^2 + \beta_1)\sigma_{t-3}^2\Bigg)\Bigg)$$
$$\ldots$$

we get for the k-th step:

$$\begin{aligned}\sigma_t^2 &= \alpha_0\Bigg(1 + \sum_{j=1}^{k}\prod_{i=1}^{j}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\Bigg) + \Bigg(\prod_{i=1}^{k+1}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\Bigg)\sigma_{t-k-1}^2 \\ &= f_t(k) + \Bigg(\prod_{i=1}^{k+1}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\Bigg)\sigma_{t-k-1}^2\end{aligned}$$

where $f_t(k)$ is just to simplify the notation. Note that $f_t = \lim_{k\to\infty} f_t(k)$ belongs to $\overline{\mathbb{R}^+}$. Similiar to what we saw before, we can define:

$$f_t(k) = \alpha_0 + (\alpha_1\varepsilon_{t-1}^2 + \beta_1)f_t(k-1)$$

and as $k \to \infty$

$$f_t = \alpha_0 + (\alpha_1\varepsilon_{t-1}^2 + \beta_1)f_t.$$

Let us suppose $E[\log(\alpha_1\varepsilon_{t-1}^2 + \beta_1)] < 0$. We apply by Strong Law of Large Numbers (Theorem 2.21) to the sequence $\{\log(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\}$, which is i.i.d because ε_{t-i} are i.i.d

$$\left(\prod_{i=1}^{j}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\right)^{\frac{1}{j}} = \exp\left(\frac{1}{j}\sum_{i=1}^{j}\log(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\right) \longrightarrow e^{E[\log(\alpha_1\varepsilon_t^2+\beta_1)]} \quad \text{a.s}$$

with $j \to \infty$. If we go to expression (2.5) and apply the Root Test to the infinite series in f_t. we know the series converges a.s. (Note that $e^{E[\log(\alpha_1\varepsilon_t^2+\beta_1)]} < 1$)) so the limit process f_t is finite. This means that if we define $Y_t = \sqrt{f_t}\varepsilon_t$ we have a strictly stationary

GARCH(1,1) model.
Now we have to prove the uniqueness. Let $\widetilde{Y}_t = \widetilde{\sigma}_t^2 \varepsilon_t$ be another strictly stationary solution.

$$\widetilde{\sigma}_t^2 = f_t(k) + \left(\prod_{i=1}^{k+1} (\alpha_1 \varepsilon_{t-i}^2 + \beta_1) \right) \widetilde{\sigma}_{t-k-1}^2.$$

Subtracting the limit f_t on both sides

$$\widetilde{\sigma}_t^2 - f_t = f_t(k) - f_t + \left(\prod_{i=1}^{k+1} (\alpha_1 \varepsilon_{t-i}^2 + \beta_1) \right) \widetilde{\sigma}_{t-k-1}^2.$$

Obviously $f_t(k) - f_t \to 0$ as $k \to \infty$. Since the series that defines f_t converges a.s., we have that

$$\prod_{i=1}^{k} (\alpha_1 \varepsilon_{t-i}^2 + \beta_1) \widetilde{\sigma}_{t-k-1}^2 \longrightarrow 0 \quad \text{in probability 1}$$

as k tends to infinity, where have also used that $\widetilde{\sigma}_{t-k-1}^2$ is independent of k by stationarity of the series. In conclusion, this gives us $P(\widetilde{\sigma}_t^2 - f_t > \epsilon) = 0$, as $k \to \infty$, this means $\widetilde{\sigma}_t^2 = f_t$ a.s. Let us discard the remaining cases. If $E[\log(\alpha_1 \varepsilon_t^2 + \beta_1)] > 0$, if we repeat the reasoning with the Strong Law of Large Numbers we get that $e^{E[\log(\alpha_1 \varepsilon_t^2 + \beta_1)]} > 1$ a.s., leading to the divergence of the series a.s when $k \to \infty$. Therefore if $\alpha_0 > 0$ we have $f_t = +\infty$ a.s. For $E[\log(\alpha_1 \varepsilon_t^2 + \beta_1)] = 0$, we argue by contradiction. We suppose that there exists a strictly stationary solution. Note that

$$\sigma_t^2 \geq \alpha_0 \left(1 + \sum_{j=1}^{k} \prod_{i=1}^{j} (\alpha_1 \varepsilon_{t-i}^2 + \beta_1) \right).$$

Applying the same idea as before which used the Strong Law of Large Numbers we have that $\alpha_0 \prod_{i=1}^{j} (\alpha_1 \varepsilon_{t-i}^2 + \beta_1) \to 0$ a.s. Alternatively

$$\sum_{i=1}^{j} \log(\alpha_1 \varepsilon_{t-i}^2 + \beta_1) + \log \alpha_0 \to -\infty$$

as $j \to \infty$. But the Chung-Fuchs theorem tells us that $\limsup \sum_{i=1}^{j} \log(\alpha_1 \varepsilon_{t-i}^2 + \beta_1) = +\infty$ with probability 1, which is a contradiction. □

Corollary 2.40. A GARCH(1,1) is second order stationary if and only if $\alpha_1 + \beta_1 < 1$.

Proof. Let us suppose that $\alpha_1 + \beta_1 < 1$. By Jensen's inequality for concave functions we have $E[\log(\alpha_1 \varepsilon_t^2 + \beta_1)] \leq \log(E[\alpha_1 \varepsilon_t^2 + \beta_1]) = \log(\alpha_1 + \beta_1) < 0$. By theorem 2.39, we have that σ_t^2 is a stationary solution for the GARCH(1,1) equations.

Conversely, we compute $E(\sigma_t^2)$. Since GARCH(1,1) is a stationary process:

$$\begin{aligned} E(\sigma_t^2) &= E\Bigg[\alpha_0\Bigg(1 + \sum_{j=1}^{\infty}\prod_{i=1}^{j}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\Bigg)\Bigg] \\ &= \alpha_0\Bigg(1 + \sum_{j=1}^{\infty} E[\prod_{i=1}^{j}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)]\Bigg) \\ &= \alpha_0(1 + \sum_{j=1}^{\infty}(\alpha_1 + \beta_1)^j) \\ &= \frac{\alpha_0}{1 - (\alpha_1 + \beta_1)} \end{aligned}$$

which is finite if $\alpha_1 + \beta_1 < 1$. □

Corollary 2.41. For a GARCH(1,1) model, with initial conditions at $t = 0$, we have:

$$E[\log(\alpha_1\varepsilon_t^2 + \beta_1)] > 0 \quad \text{then} \quad \sigma_t^2 \overset{t\to\infty}{\longrightarrow} +\infty \quad \text{a.s.}$$

In addition, if $E[(\log(\varepsilon_t^2)] < \infty$, then:

$$E[\log(\alpha_1\varepsilon_t^2 + \beta_1)] > 0 \quad \text{then} \quad Y_t^2 \overset{t\to\infty}{\longrightarrow} +\infty \quad \text{a.s.}$$

These are called the *conditions for explosion.*

Proof. Recall from Theorem 2.39 that we have the following inequalities:

$$\sigma_t^2 \geq \alpha_0\Bigg(1 + \sum_{j=1}^{t-1}\prod_{i=1}^{j}(\alpha_1\varepsilon_{t-i}^2 + \beta_1)\Bigg) \geq \alpha_0\prod_{i=1}^{t-1}(\alpha_1\varepsilon_{t-i}^2 + \beta_1).$$

If we take the $\liminf$ on both sides:

$$\liminf_{t\to\infty}\frac{1}{t}\log(\sigma_t^2) \geq \liminf_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t-1}\log(\alpha_1\varepsilon_{t-i}^2 + \beta_1) = \gamma.$$

Which gives us $\log\sigma_t^2 \to \infty$ and $\sigma_t^2 \to \infty$. Similarly:

$$\liminf_{t\to\infty}\frac{1}{t}\log(Y_t^2) = \liminf_{t\to\infty}\frac{1}{t}(\log\sigma_t^2 + \log\varepsilon_t^2) \geq \gamma + \liminf_{t\to\infty}\frac{1}{t}\log(\varepsilon_t^2) = \gamma.$$

□

3 Estimation and Prediction

This chapter is devoted to Estimation and Prediction of ARCH and GARCH models. This is the fundamental objective when studying time series, once we try to fit a specific model to observed values $\{x_t\}_{\geq 0}$, can we estimate the parameters of our model ? If we have $x_1, \ldots, x_n$ values, can we predict what will the value x_{n+1} be?. This is the main focus of this chapter.

We will develop our theory in the context of Hilbert spaces, which allows us to rigorously present the Classical Regression problem. In Chapter 4 we will see why this method fails for most financial time series, despite having already seen some answers to this question in the previous chapter. Also, functional analysis works well in developing a consistent theory about Prediction and Estimation with time series. Simple examples in `R` will given in order to illustrate the theoretical results.

3.1 Sampled parameters and Estimation

In the previous chapter, we went over theoretical properties of time series. In Chapter 4, we will be working with real data, hence we will be considering a finite number of observations $\{y_1, \ldots, y_n\}$. A few statistical notions must be introduced, which differ slightly from the theoretical definition.

Definition 3.1. The sample mean and sample variance are defined as:

$$\overline{y} = \frac{1}{n}\sum_{t=1}^{n} y_t \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_t - \overline{y})^2.$$

Recall that the sample mean is an unbiased estimator of the expected value whereas the sample variance is not.

Definitions 3.2. The *sample autocovariance function* is defined as

$$\widehat{\gamma(l)} = \frac{1}{n}\sum_{t=1}^{n-l}(y_{t+l} - \overline{y})(y_t - \overline{y}).$$

And the *sample autocorrelation function* (ACF) is defined as

$$\widehat{\rho(l)} = \frac{\widehat{\gamma(l)}}{\widehat{\gamma(0)}}.$$

Theorem 3.3. Let $\{Y_t\}$ be iid and fourth-moment finite, with n sufficiently large. The sample ACF $\widehat{\rho(l)}$ where $l = 1, \ldots, h'$ (h' arbitrarily fixed) is approximately normally distributed with zero mean and standard deviation $\sigma_{\widehat{\rho(l)}} = \dfrac{1}{\sqrt{n}}$.

Proof. Can be found in [1] pp. 482-490. □

3.2 The Classical Regression Model

Recall basic functional analysis definitions: A *pre-Hilbert space* is a vector space with an inner product. A *Banach space* is a normed space such that the distance induced by the norm is complete (i.e. every Cauchy sequence converges). A *Hilbert space* is a pre-Hilbert space if it is a Banach space with the norm induced by the inner product.

Theorem 3.4. (Projection theorem) Let H be a Hilbert space and $Y \subset H$ a closed subset. Then $H = Y \oplus Y^{\perp}$, i.e. given $x \in H$ there exists a unique $y_x \in Y$ and $z_x \in Y^{\perp}$ such that $x = y_x + z_x$. In addition:

(a) y_x is the only vector of Y that satisfies $x - y_x \perp Y$

(b) $d(x, Y) = \|x - y_x\|$ and $d(x, Y^{\perp}) = \|x - z_x\| = \|y_x\|$

Also, $d(x, Y) = \|x - y_x\| = \inf_{y \in Y} \|x - y\|$ □

Corollary 3.5. Under the hypothesis of Theorem 3.4 we have the *Prediction Equations*:

$$\langle x - y_x, y \rangle = 0 \quad \forall y. \tag{3.1}$$

Also, $P_Y x$ denotes the projection of x on a closed subspace Y of a Hilbert space H, and it fulfills the following properties:

Proposition 3.6. Let H be a Hilbert space, $Y \subset H$ closed and P_Y the projection in Y (called projection operator / projection mapping) The following properties hold.

(1) P_Y is linear, i.e. $P_Y(\alpha x + \beta x') = \alpha P_Y x + \beta P_Y x'$.

(2) $P_Y x = x$ for all $x \in Y$. Hence $P_Y^2 x = P_Y x$, meaning P_Y is a projection of H on Y.

(3) We have $Q_Y x = (I - P_Y)x \in Y^{\perp}$ for all $x \in H$ and $Q_Y x = x$ for all $x \in Y^{\perp}$.

(4) For every $x \in H$ we have $x = P_Y x + Q_Y x$ and $\|x\|^2 = \|P_Y x\|^2 + \|Q_Y x\|^2$ for all $x \in H$

(5) If $Y_1 \subseteq Y_2$ if and only if $P_{Y_1} P_{Y_2} x = P_{Y_1} x$ for all $x \in H$. □

Theorem 3.7. Let $H = \mathbb{R}^n$, $x_i \in \mathbb{R}^n$ for $i = 1 \dots m$ and $Y = \mathrm{span}[x_1, \dots, x_m]$. Then:

$$P_Y x = X\beta$$

where $x \in \mathbb{R}^n$, X is a $n \times m$ matrix whose j-th column is x_j and

$$X^T X \beta = X^T x.$$

The previous equation has at least one solution for $\beta \in \mathbb{R}^n$ but $X\beta$ is the same for all solutions. There is exactly one solution if and only if $X^T X$ is invertible and in this case

$$P_Y x = X(X^T X)^{-1} X^T x.$$

Proof. We start by noting that $P_Y x \in Y$, so we write it as $P_Y x = \sum_{i=1}^{n} \beta_i x_i = X\beta$ for $\beta = (\beta_1, \dots, \beta_m)^T \in \mathbb{R}^n$. Using Corollary 3.5 we write $\langle X\beta, x_j \rangle = \langle x, x_j \rangle$, for $j = 1, \dots, m$. Equivalently, in matrix form we write $X^T X \beta = X^T x$. The existence of β follows from the existence of $P_Y x$. $X\beta$ is the same for all solutions because $P_Y x$ is unique by Theorem 3.4. □

In the context of the multiple linear regression model, the response variable Y_t is a function of parameters called *regressors*: $Y_t = \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_p X_{t,p} + \varepsilon_t$, being $\epsilon_t \sim$ iid noise$(0, \sigma^2)$. (these restrictions imposed on ϵ_t are sometimes referred to Gauss-Markov conditions) We want to estimate the vector β. This is done by least squares estimation, which minimizes the following quantity:

$$SSR = \sum_{t=1}^{n} \varepsilon_t^2 = \sum_{t=1}^{n} \left(Y_t - \sum_{i=1}^{p} \beta_i X_{t,i} \right)^2 .$$

Note that SSR stands for Sum of Squared residuals. To simplify notation, it is not uncommon to use matrices and vectors:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} X_{1,1} & \dots & X_{1,p} \\ X_{2,1} & \dots & X_{2,p} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \dots & X_{n,p} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Now the least squares estimation problem looks like:

$$SSR = \sum_{t=1}^{n} \varepsilon_t^2 = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta) = \|Y - X\beta\|^2.$$

Thanks to the Projection theorem we notice that minimizing the sum is done by finding the projection y on to the linear space $M = \text{span}\{X_{\cdot,1}, \dots X_{\cdot,p}\}$ where $X_{\cdot,j}$ signifies the j-th column of the matrix. For each instance, the vector found minimizes the SSR. By theorem 3.7 we have:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Definition 3.8. The *residuals* of the regression are defined as:

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta} = MY = M(X\beta + \varepsilon) = (MX)\beta + M\varepsilon = M\varepsilon$$

defined as $M = I_n - P$. By Proposition 3.6 we have that M defines a projection onto $Y^{\perp}$.

Proposition 3.9. Under the Gauss-Markov conditions (i.e. $\varepsilon_t \sim$ i.i.d noise with mean 0 and variance σ^2)), we have:

(a) The estimator $\hat{\beta}$ is unbiased.

(b) The estimator for the variance of $\hat{\beta}$ is $\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.

Proof. Let us prove (a),
$E(\hat{\beta}) = E((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X\beta = \beta.$ □

3.3 Prediction theory

The idea behind prediction theory is simple: Given $y_1, \dots, y_t$ observations of $Y_1, \dots, Y_t$ random variables, can we predict $Y_{t+1}, \dots, Y_{t+p}$?

Lemma 3.10. $L^2(\Omega, \mathcal{F}, \mathcal{P})$ is a Hilbert space with the inner product

$$\langle X, Y \rangle = E(XY) = \int_{\Omega} XY dP$$

for all $X, Y \in L^2$. (i.e. X, Y are square-integrable) □

Definition 3.11. Given $\{X_n,\, n \geq 1\}$ a sequence of random variables of $L^2(\Omega, \mathcal{F}, \mathcal{P})$. The variables converge in *mean square convergence* to a random variable X if

$$\lim_{n\to\infty} E(|X_n - X|^2) = 0.$$

Naively, if $M \subset L^2$, and $Y \in L^2$ we can assume that finding the best mean-square predictor Y consists in finding $\hat{Y}$ such that:

$$\|Y - \hat{Y}\|^2 = \inf_{Z\in M} E(|Y - Z|^2).$$

We observe that the conditions are quasi-identical to the projection theorem. So the best mean square predictor is given by the projection of Y into subset M. We now introduce an alternative definition for conditional expectation, which we must interpret in the following manner: If $\mathcal{G} \subset \mathcal{F}$, the best predictor of Y is based in the information of $\mathcal{G}$. So the following definition is consistent.

Definition 3.12. Let M be a closed subset of L^2. If $X \in L^2$ we define the *conditional expectation* of X given M as

$$E_M X = P_M X.$$

Note that this definition of Conditional Expectation is more restricted than the one given in Definition 2.19. Despite having the general definition, since all this section works under the L^2 space structure, Definition 3.11 suffices.
Now, if $Z_1, \ldots, Z_n \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $X \in L^2$ we can extend the definition 3.11 to $E_{M(Z_1,\ldots,Z_n)}X$ where $M(Z_1, \ldots, Z_n)$ is defined as the closed subspace of L^2 generated by the variables in L^2 of the form $\phi(Z_1, \ldots, Z_n)$ for some Borel function $\phi : \mathbb{R}^n \to \mathbb{R}$. Now, by the projection theorem the Conditional Expectation $E_{M(Z_1,\ldots,Z_n)}$ is the best mean square predictor for X. By the prediction equations (3.1):

$$\begin{aligned}
&\langle X - P_M X, W\rangle = 0 \quad \forall W \in M \\
&\langle X, W\rangle = \langle P_M X, W\rangle \\
&E(XW) = E(E_M XW).
\end{aligned}$$

We avoid going into the details of this definition because of the technicalities it presents. Extensive information is found in [4].

3.3.1 The best linear predictor

The prediction equations above give us a final expression that can be very hard to calculate. We consider an alternative: If $X_1, \ldots, X_n \in L^2$ and $\text{span}(1, X_1, \ldots, X_n) \subseteq M(X_1, \ldots, X_n)$ then we write

$$P_{\overline{\text{span}}\{1,X_1,\ldots,X_n\}} Y = \sum_{i=0}^{n} \alpha_i X_i, \quad X_0 = 1. \tag{3.2}$$

By the prediction equations in (3.1), (3.2) satisfies:

$$\begin{aligned}
&\left\langle \sum_{i=0}^{n} \alpha_i X_i, X_j \right\rangle = \langle Y, X_j\rangle \quad j = 0, \ldots n \\
&\sum_{i=0}^{n} \alpha_i E(X_i X_j) = E(Y X_j) \quad j = 0, \ldots n
\end{aligned}$$

These equation have a unique solution thanks to the Projection Theorem.

Notation 3.13. $P(Y \mid X_1, \dots X_n) = \widehat{\alpha_0} + \widehat{\alpha_1} X_1 + \dots + \widehat{\alpha_n} X_n$ denotes the *optimal linear predictor*.
We have now seen a general statement for our problem. Given that $E(Y \mid X_1, \dots, X_n)$ can be a tedious calculation, we defined $f(X_1, \dots, X_n) = E(Y \mid X_1, \dots, X_n) = \alpha_0 + \alpha_1 X_1 + \dots \alpha_n X_n$ which has a solution guaranteed by the projection theorem. Recall that our objective is to minimize prediction errors, which are defined as follows.

Definition 3.14. The expressions $E[(Y - E(Y|X_1, \dots, X_n))^2]$ or $E[(Y - P(Y|X_1, \dots, X_n))^2]$ are called *prediction errors*.

Let us start solving the optimal linear predictor problem. We want to find $P(Y \mid X_1, \dots X_n) = \hat{\alpha_0} + \hat{\alpha_1} X_1 + \dots + \hat{\alpha_n} X_n$. The Prediction Equations (3.1) give:

$$\begin{aligned} &\langle Y - P(Y|X_1, \dots, X_n, 1\rangle = 0 \\ &\langle Y - P(Y|X_1, \dots, X_n), X_t\rangle = 0 \quad t = 1, \dots n. \end{aligned}$$

The previous equations are equivalent to

$$\begin{aligned} &\alpha_0 + \alpha_1 E(X_1) + \dots \alpha_n E(X_n) = E(Y) \\ &\alpha_0 E(X_1) + \alpha_1 E(X_1^2) + \dots + \alpha_n E(X_1 X_n) = E(X_1 Y) \\ &\vdots \\ &\alpha_0 E(X_n) + \alpha_1 E(X_1) + \dots + \alpha_n E(X_n^2) = E(X_n Y) \end{aligned} \tag{3.3}$$

To simplify notation, we define $\alpha^T = (\alpha_0, \dots, \alpha_n)$, $\mu_X^T = (E(X_0), \dots, E(X_n))$. This gives us:

$$\alpha_0 = E(Y) - \alpha^T \mu_X$$

If we multiply the equation by $E(X_t)$ for $t = 1, \dots, n$.

$$\alpha_0 E(X_t) = E(X_t)E(Y) - E(X_t)\alpha^T \mu_X \quad \forall t = 1, \dots n.$$

Substituting this expression in the the equations $1, \dots, n$ from the system (3.3) we get:

$$\Big(E(X_t)E(Y) - E(X_t)\alpha^T \mu_X\Big) + E(X_t)E(X_1)\alpha_1 + \dots + E(X_t)E(X_n) = E(X_n)E(Y) \quad t = 1, \dots n.$$

Recall that Lemma 2.5 gives us $\gamma(s,t) = E(X_s X_t) - E(X_t)E(X_s)$ Reorganizing the terms gives us the family of equations:

$$\alpha_1 \gamma(t, 1) + \dots + \alpha_n \gamma(t, n) = \operatorname{cov}(Y, X_t) \quad t = 1, \dots, n. \tag{3.4}$$

Note that if we simplify the expressions

$$\Gamma \alpha = \operatorname{cov}(Y, X)$$

where $\Gamma = \{\gamma(i,j)\}_{i,j=1,\dots n}$ and $\operatorname{cov}(Y, X)$ is the covariance vector $\operatorname{cov}(Y, X_i)$ $i = 1 \dots, n$. Assuming that Γ is invertible we get:

$$\widehat{\alpha} = \Gamma^{-1} \operatorname{cov}(Y, X). \tag{3.5}$$

The best linear predictor is finally:

$$P(Y|X_1, \dots X_n) = \widehat{\alpha_0} + \widehat{\alpha_1} X_1 + \dots + \widehat{\alpha_n} X_n = E(Y) - \widehat{\alpha}^T \mu_X + \alpha^T X = E(Y) + \widehat{\alpha}^T (X - \mu_X).$$

And the prediction error (PE_Y) is:

$$\begin{aligned}
PE_Y &= E\big[(Y - P(Y|X_1,\dots,X_n))^2\big] = E\big[(Y - E(Y) - \widehat{\alpha}^T(X-\mu_X))^2\big] \\
&= E\big[(Y-E(Y))^2 - 2(Y-E(Y))\widehat{\alpha}^T(X-\mu_X) + (\widehat{\alpha}^T(X-\mu_X))^2\big] \\
&= \mathrm{Var}(Y) + \widehat{\alpha}^T\Gamma\widehat{\alpha} - 2\widehat{\alpha}\mathrm{cov}(Y,X)^T \\
&= \mathrm{Var}(Y) - \widehat{\alpha}^T\mathrm{cov}(Y,X)
\end{aligned}$$

where have applied Lemma 2.6 and the definition of $\Gamma\widehat{\alpha} = \mathrm{cov}(Y,X)$. Recall that $\mathrm{cov}(Y,X) = E\big[(Y-E(Y))(X-\mu_X)^T\big]$. We have also used the following property for the vectors $\widehat{\alpha}^T(X-\mu_X) = (X-\mu_X)^T\widehat{\alpha}$.

Proposition 3.15. The following properties for $P(Y|X_1,\dots,X_n)$ hold:

(1) $P(X_j|X_1,\dots,X_n) = X_j$

(2) $P(\alpha + \beta Y_1 + \gamma Y_2|X_1,\dots,X_n) = \alpha + \beta P(Y_1|X_1,\dots,X_n) + \gamma P(Y_2|X_1,\dots,X_n)$

(3) If $\mathrm{cov}(Y,X_t) = 0$ for $t = 1,\dots n$ then $P(Y|X_1,\dots,X_n) = E(Y)$.

(4) $P(P(Y|X_1,X_2|X_1) = P(Y|X_1)$.

Proof. Since $P(Y|X_1,\dots,X_n)$ is the orthogonal projection onto $\overline{\mathrm{span}}\{1,X_1,\dots,X_n\}$ the listed properties are immediate from Proposition 3.7 and the definition of expected value. □

Now let us consider the Time Series case. First, we start by assuming that our time series is second order stationary (i.e. first and second order moments are constant through time). The expression for

$$P(X_{n+1}|X_1,\dots,X_n) = \mu + \widehat{\alpha}^T(X-\mu)$$

since $\{X_t\}$ is stationary. The prediction error calculated before now is:

$$\begin{aligned}
PE_{n+1} &= E\big[(X_{n+1} - P(X_{n+1}|X_1,\dots,X_n))^2\big] \\
&= E\big[(X_{n+1} - \mu - \widehat{\alpha}^T(X-\mu))^2\big] \\
&= \mathrm{Var}(X_{n+1}) - \mathrm{cov}(X_{n+1},X)^T\widehat{\alpha} = \gamma(0) - \gamma_n^T\widehat{\alpha}
\end{aligned}$$

where $\gamma_n = (\gamma(1),\dots,\gamma(n))^T$. We will use this notation in other results.

Proposition 3.16. If $\{X_t\}$ is stationary time series with mean zero, then:
$P_{\overline{\mathrm{span}}\{1,X_1,\dots,X_n\}}X_{n+1} = P_{\overline{\mathrm{span}}\{X_1,\dots,X_n\}}X_{n+1}$.

Proof. Let us write $X_0 = 1$. We know $P_{\overline{\mathrm{span}}\{X_0,X_1,\dots,X_n\}}X_{n+1} = \alpha_0X_0 + \sum_{i=1}^n\alpha_iX_i$. By the prediction equations

$$\langle \alpha_0X_0 + \sum_{i=1}^n \alpha_iX_i, X_j\rangle = \langle X_{n+1}, X_j\rangle \quad j = 0,1\dots,n$$

$$\alpha_0E(X_0X_j) + \sum_{i=1}^n \alpha_iE(X_1X_j) = E(X_{n+1}X_j).$$

But $\{X_t\}$ is a stationary time series with mean zero, which means $E(X_0 X_j) = E(X_j) = 0$, hence:

$$\sum_{i=1}^{n} \alpha_i E(X_1 X_j) = E(X_{n+1} X_j).$$

And if we reverse the process we get $P_{\overline{\text{span}}\{X_1,\dots,X_n\}} X_{n+1} = \sum_{i=1}^{n} \alpha_i X_i$. $\square$

This result allows to work with $\overline{\text{span}}\{X_1,\dots,X_n\}X_{n+1}$ as long as our time series is stationary.

Definition 3.17. Given a stationary time series, the *partial autocorrelation function* (PACF) $\alpha(\cdot)$ is defined by

$$\begin{aligned}
&\alpha(1) = \text{corr}(X_2, X_1) = \rho(1) \\
&\alpha(k) = \text{corr}(X_{k+1} - P_{\overline{\text{span}}\{1,X_2,\dots,X_k\}} X_{k+1}, X_1 - P_{\overline{\text{span}}\{1,X_2,\dots,X_k\}} X_1), \quad k \geq 2.
\end{aligned}$$

This definition gives us more understanding of dependence structure of a stationary process. We can think of it as the correlation between X_{k+1} and X_1 that has been adjusted because of the presence of $X_2,\dots,X_k$ variables. In other words, we are "removing" the effect of $X_2,\dots,X_k$, which is why the linear projection makes sense.
Note also that this is not a very operational definition, which is why we adequately consider an alternative definition for the PACF:
Let $\{X_t\}$ be a stationary time series with $E(X_t) = 0$ and ACF such that $\gamma(h) \to 0$ for $h \to \infty$ (not a restrictive hypothesis). Let us suppose that $P_{\overline{\text{span}}\{X_1,\dots,X_k\}} X_{k+1} = \sum_{t=1}^{k} b_t X_{k+1-t}$
By the prediction equations we know that:

$$\langle X_{k+1} - P_{\overline{\text{span}}\{X_1,\dots,X_k\}} X_{k+1}, X_t \rangle = 0, \quad t = k,\dots,1.$$

Dividing by $\gamma(0)$ on both sides (we assume $\gamma(0) > 0$ wlog) we get:

$$\begin{pmatrix} \rho(0) & \rho(1) & \rho(2) & \dots & \rho(k-1) \\ \rho(1) & \rho(0) & \rho(1) & \dots & \rho(k-2) \\ \vdots & & & & \vdots \\ \rho(k-1) & \rho(k-2) & \rho(k-3) & \dots & \rho(0) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k) \end{pmatrix}$$

Definition 3.18. (alternative) The PACF $\alpha(k)$ of $\{X_t\}$ at lag k is

$$\alpha(k) = b_k \quad k \geq 1$$

where b_k is uniquely determined by the system of equations above. Note that we can defined the sample PACF analogously.

Definition 3.19. Let us consider observations $\{x_1,\dots x_n\}$. The sample PACF $\widehat{\alpha(k)}$ of $\{X_t\}$ at lag k is $\widehat{\alpha(k)} = \widehat{b}_k$. Note that $\widehat{b}_k$ is determined through the sample ACF $\widehat{\gamma}(t)$ for $t = 1,\dots k$. Also, we need at least $x_i \neq x_j$ for some i and j.

Note that we made an important assumption in the last steps of our calculations. The following proposition asserts that this is not a major issue.

Lemma 3.20. Given A rank n real symmetric matrix, it decomposes into $A = QDQ^T$ with Q orthogonal (i.e. $Q^T = Q^{-1}$ meaning $QQ^T = Id$) and $D = diag(\lambda_1, \ldots, \lambda_n)$ with strictly positive eigenvalues.

Proposition 3.21. Let $\{X_t\}$ be a stationary time series with $E(X_t) = 0$. If $\gamma(0) > 0$ and $\gamma(h) \to 0$ for $h \to \infty$, the variance-covariance matrix Γ_n is non-singular for every n.

Proof. Let us suppose that Γ_n is singular for some n. By hypothesis $E(X_t) = 0$, so there exists non-zero $r \geq 1$ and constants $a_1, \ldots, a_r$ such that Γ_r is non-singular and:

$$X_{r+1} = \sum_{t=1}^{r} a_t X_t. \tag{3.6}$$

This is due to Lemma 2.6: Given $X = (X_1, \ldots, X_{r+1})$, we have $\mathrm{Var}(a^T X) = 0$ (The covariance matrix Γ_{r+1} is singular). Now, we apply the fact that $\{X_t\}$ is a stationary time series. This means the variance of the linear combination remains constant, meaning that $\mathrm{Var}(a^T X) = Var(a^T X_l) = 0$ where $l > 0$ is the lag. This gives us:

$$X_{r+h} = \sum_{t=1}^{r} a_t X_{t+h-1} \tag{3.7}$$

for a given lag $h \geq 1$.
In consequence, for $n \geq r+1$ there exist $a_1^{(n)}, \ldots, a_r^{(n)} \in \mathbb{R}$ constants such that

$$X_n = a^{(n)^T} X_r \tag{3.8}$$

where $X_r = (X_1, \ldots, X_r)$. Lemma 2.6 gives us $Var(a^{(n)^T} Xr) = a^{(n)^T} \Gamma_r a^{(n)}$. If we apply Lemma 3.20

$$\gamma(0) = a^{(n)^T} \Gamma_r a^{(n)} = a^{(n)^T} QDQ^T a^{(n)}$$

where D has $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_r$ strictly positive eigenvalues. We can write:

$$\gamma(0) \geq \lambda_1 a^{(n)^T} QQ^T a^{(n)} = \lambda_1 \sum_{t=1}^{r} (a^{(n)})^2$$

which gives an upper bound for $a_t^{(n)}$ for each fixed t. By definition of $\gamma(0) = \mathrm{cov}(X_n, \sum_{t=1}^{r} a_t^{(n)} X_t)$, we have the bound:

$$\gamma(0) \leq \sum_{t=1}^{r} |a_t^{(n)}||\gamma(n-t)|.$$

A contradiction arises since it is not possible that $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$. $\square$

Corollary 3.22. Under the hypothesis of Proposition 3.21 the best linear predictor $\hat{X}_{n+1}$ of X_{n+1} given $X_1, \ldots, X_n$ is

$$\hat{X}_{n+1} = \sum_{t=1}^{n} \alpha_t X_{n+1-t}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^T = \Gamma_n^{-1} \mathrm{cov}(X_{n+1}, X)$ where Γ_n is the autocovariance matrix and $X = (X_1, \ldots, X_n)^T$.

Proof. Proposition 3.21 guarantees the existence of Γ_n^{-1}. The rest has already been shown. □

From numerical analysis we know that calculating the inverse of a matrix is costly. The Durbin-Levinson Algorithm solves this problem. It is one of the most used methods in Prediction, which is why we include it in our study. There are many more recipes ,such as the Innovations Algorithm, but we will not consider them. A strong reason to include the Durbin-Levinson algorithm is because it will allow us to prove the equivalency between definitions 3.17 and 3.18.

Lemma 3.23. Let $\{X_t\}$ be a stationary time series with zero mean.
Let $H = \overline{\text{span}}\{X_1, \ldots, X_n\}$, we define $H_1 = \overline{\text{span}}\{X_2, \ldots, X_n\}$ and $H_2 = \overline{\text{span}}\{X_1 - P_{H_1}X_1\}$. Then H_1 and H_2 are orthogonal. Moreover, if $Y \in L^2(\Omega, \mathcal{F}, \mathcal{P})$ then:

$$P_{H_n}\hat{X}_{n+1} = P_{H_1}X_{n+1} + P_{H_2}X_{n+1} = P_{H_1}X_{n+1} + a(X_1 - P_{H_1}X_1)$$

where

$$a = \frac{\langle X_{n+1}, X_1 - P_{H_1}X_1\rangle}{\|X_1 - P_{H_1}X_1\|^2}.$$

Proof. By the Prediction Equations in (3.1), we get $\langle X_i, X_1 - P_{H_1}X_1\rangle = 0\ \forall i = 2, \ldots n$. This shows that H_1 and H_2 are orthogonal. Note that $H = H_1 + H_2$. Since the sum of closed orthogonal subspaces is a closed subspace, we conclude that $H \subset L^2$ is a closed subspace. A Linear Algebra results shows us that if $H = H_1 + H_2$, where H_1 and H_2 are closed subspaces, $P_{H_1} + P_{H_2} = P_H$ is a projection if and only if $H_1 \perp H_2$. Now, given $\hat{X}_{n+1} = P_H\hat{X}_{n+1}$ we have $\hat{X}_{n+1} = P_{H_1}X_{n+1} + P_{H_2}X_{n+1} = P_{H_1}X_{n+1} + a(X_1 - P_{H_1}X_1)$ Note that if we want the expression derived in the last equal sign, we must determine a. We know that:

$$0 = \langle \hat{X}_{n+1} - X_{n+1}, X_i\rangle \quad \forall i > 1$$
$$0 = \langle \hat{X}_{n+1} - X_{n+1}, X_1 - P_{H_1}X_1\rangle = \langle P_{H_1}X_{n+1} - X_{n+1}, X_1 - P_{H_1}X_1\rangle + a\|X_1 - P_{H_1}X_1\|^2$$

If we observe that $\langle P_{H_1}X_{n+1}, X_1 - P_{H_1}X_1\rangle = 0$ (which is a variation of what we saw in the beginning of the proof), we isolate a and get:

$$a = \frac{\langle X_{n+1}, X_1 - P_{H_1}X_1\rangle}{\|X_1 - P_{H_1}X_1\|^2}.$$

□

Proposition 3.24. Let $\{X_t\}$ be a stationary time series with zero mean. Then

$$\hat{X}_{n+1} = P_H X_{n+1} = b_{n1}X_n + \ldots + b_{nn}X_1 \; n \geq 1$$

We denote the prediction error $PE_n = E(X_{n+1} - \hat{X}_{n+1})^2$ for $n \geq 1$ The ACF satisfies $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$. Also $b_{11} = \dfrac{\gamma(1)}{\gamma(0)}$ and $PE_0 = \gamma(0)$, we define recursively:

$$b_{nn} = \left[\gamma(n) - \sum_{t=1}^{n-1} b_{n-1,t}\gamma(n-t)\right] PE_{n-1}^{-1} \quad PE_n = PE_{n-1}[1 - b_{nn}^2]$$
$$b_{nt} = b_{n-1,t} - b_{nn}b_{n-1,n-t} \quad t = 1 \ldots n-1.$$

Proof. Since the time series is stationary, the covariance matrix for $X_1, \ldots, X_n$ is the same as $X_n, \ldots, X_1$ and $X_2, \ldots, X_{n+1}$. This gives us:

$$\begin{aligned} P_{H_1} X_1 &= \sum_{t=1}^{n-1} b_{n-1,t} X_{t+1} \\ P_{H_1} X_{n+1} &= \sum_{t=1}^{n-1} b_{n-1,t} X_{n+1-t} \end{aligned} \tag{3.9}$$

and if we calculate the prediction error

$$\|X_1 - P_{H_1} X_1\|^2 = \|X_{n+1} - P_{H_1} X_{n+1}\|^2 = \|X_n - \hat{X}_n\|^2 = PE_{n-1}.$$

Joining (3.6) equations and Lemma 3.23 we have:

$$\hat{X}_{n+1} = aX_1 + \sum_{t=1}^{n-1} [b_{n-1,t} - ab_{n-1,n-t}] X_{n+1-t}.$$

Using the value a determined in Lemma 3.23 and (3.6) again we get:

$$a = \frac{\langle X_{n+1}, X_1 \rangle - \sum_{t=1}^{n-1} b_{n-1,t} \langle X_{n+1}, X_{t+1} \rangle}{\|X_1 - P_{H_1} X_1\|^2} = \left[\gamma(n) - \sum_{t=1}^{n-1} b_{n-1,t} \gamma(n-t) \right] PE_{n-1}^{-1}$$

The results obtained in Proposition 3.21 gives us the certainty that the following representation is unique:

$$\hat{X}_{n+1} = \sum_{t=1}^{n} b_{n,t} X_{n+1-t}.$$

The two expressions we have for $\hat{X}_{n+1}$ give us $b_{nn} = a$ and $b_{nt} = b_{n-1,t} - ab_{n-1,n-t}, t = 1 \ldots n-1$. We are left with the last claim to prove, that is $PE_n = PE_{n-1}[1 - b_{nn}^2]$. By definition of PE_n:

$$\begin{aligned} PE_n &= \|X_{n+1} - \hat{X}_{n+1}\|^2 = \|X_{n+1} - P_{H_1} X_{n+1} - P_{H_2} X_{n+1}\|^2 \\ &= \|X_{n+1} - P_{H_1} X_{n+1}\|^2 + \|P_{H_2} X_{n+1}\|^2 - 2\langle X_{n+1} - P_{H_1} X_{n+1}, P_{H_2} X_{n+1} \rangle. \end{aligned}$$

From Lemma 3.23, we have $P_{H_2} X_{n+1} = a(X_1 - P_{H_1} X_1)$. Plus the definition of a gives us:

$$PE_n = PE_{n-1} + a^2 PE_{n-1} - 2a\langle X_{n+1}, X_1 - P_{H_1} X_1 \rangle = PE_{n-1} + a^2 PE_{n-1} - 2a^2 PE_{n-1}.$$

We obtain $PE_n = PE_{n-1}(1 - a^2)$ which completes the proof. $\square$

Corollary 3.25. Under the same conditions of Proposition 3.24 we have

$$b_{nn} = \operatorname{corr}(X_{n+1} - P_{\overline{\text{span}}\{1, X_2, \ldots, X_n\}} X_{n+1}, X_1 - P_{\overline{\text{span}}\{1, X_2, \ldots, X_n\}} X_1). \tag{3.10}$$

Proof. By definition $P_{H_1} X_{n+1} \perp (X_1 - P_{H_1} X_1)$. We know by many results obtained from Proposition 3.24:

$$\begin{aligned} b_{nn} = a = \frac{\langle X_{n+1}, X_1 - P_{H_1} X_1 \rangle}{\|X_1 - P_{H_1} X_1\|^2} &= \frac{\langle X_{n+1} - P_{H_1} X_{n+1}, X_1 - P_{H_1} X_1 \rangle}{\|X_1 - P_{H_1} X_1\|^2} \\ = \frac{\langle X_{n+1} - P_{H_1} X_{n+1}, X_1 - P_{H_1} X_1 \rangle}{\|X_1 - P_{H_1} X_1\| \|X_{n+1} - P_{H_1} X_{n+1}\|} &= \operatorname{corr}(X_{n+1} - P_{H_1} X_{n+1}, X_1 - P_{H_1} X_1). \end{aligned}$$

$\square$

Observation 3.26. Recall that from definition 3.18 $\alpha(k) = b_n$. If we look at how we defined the prediction there, we can quickly infer $b_n = b_{nn}$. This proves the equivalence between definitions 3.17 and 3.18.

3.3.2 Prediction in ARCH and GARCH models

Let $Y_1, \ldots, Y_n$ be realizations of an ARCH model defined in 2.26. We want to study how to predict Y_{n+l}^2 for $l > 0$. Recall that Proposition 2.30 shows us that squared returns follow an AR(p) model, so we could apply the Durbin-Levinson Algorithm. Although we will explore a different approach, which is by recursive prediction, inspired by results in [8]. In section 3.2, we saw that the optimal prediction is defined as the conditional expectation on the subspace $M(Y_1, \ldots, Y_{n+p-1})$, in other words:

$$\widehat{Y_{n+l}^2} = P_{M(Y_1,\ldots,Y_{n+l-1})} Y_{n+l}^2 = E_{M(Y_1,\ldots,Y_{n+l-1})} Y_{n+l}^2 \quad l > 0.$$

Knowing that $\widehat{Y_{n+l}^2} = \hat{\sigma}_{t+1}^2 \hat{\varepsilon}_{t+1}^2$. We write $\widehat{Y_{n+l}}$ as a function of previous observations:

$$\begin{aligned}
\widehat{Y_{n+1}} &= \hat{\varepsilon}_{n+1}\sqrt{\alpha_0 + \alpha_1 Y_n^2 + \ldots + \alpha_p Y_{n-p+1}^2} = f_1(Y_n, \ldots, Y_{n-p+1}, \hat{\varepsilon}_{n+1}) \\
\widehat{Y_{n+2}} &= \hat{\varepsilon}_{n+2}\sqrt{\alpha_0 + \alpha_1 \widehat{Y_{n+1}^2} + \ldots + \alpha_p Y_{n-p+1}^2} \\
&= \hat{\varepsilon}_{n+2}\sqrt{\alpha_0 + \alpha_1\left(\hat{\varepsilon}_{n+1}^2(\alpha_0 + \alpha_1 Y_n^2 + \ldots + \alpha_p Y_{n-p+1}^2)\right) + \ldots + \alpha_p Y_{n-p+2}^2} \\
&= f_2(Y_n, \ldots, Y_{n-p+1}, \hat{\varepsilon}_{n+1}, \hat{\varepsilon}_{n+2}) \\
\widehat{Y_{n+l}} &= \ldots
\end{aligned}$$

Which gives us $\widehat{Y_{n+l}} = f_l(Y_1, \ldots, Y_n, \varepsilon_1, \ldots, \varepsilon_n)$. Note that since $Y_1, \ldots, Y_n$ are known values, we have $\widehat{Y_{n+l}} = f_l(\varepsilon_1, \ldots, \varepsilon_n)$. By definition of the optimal prediction (see section 3.2):

$$\widehat{Y_{n+1}^2} = E_{M(Y_1,\ldots,Y_{n+p-1})} \varepsilon_{n+1}^2 \sigma_{n+1}^2 = \alpha_0 + \alpha_1 Y_n^2 + \ldots + \alpha_p Y_{n-p+1}^2.$$

Where we have used that $E_{M(Y_1,\ldots,Y_{n+p-1})} \varepsilon_{n+1}^2 = E(\varepsilon_{n+1}^2) = 1$. By reiteration:

$$\begin{aligned}
\widehat{Y_{n+2}^2} &= \alpha_0 + \alpha_1 \sigma_{n+1}^2 + \ldots + \alpha_p Y_{n-p+2}^2 \\
&\cdots \\
\widehat{Y_{n+l}^2} &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \sigma_{n+l-i}^2
\end{aligned}$$

with $\sigma_{n+l-i}^2 = Y_{n+l-i}^2$ if $l - i \leq 0$.

For the GARCH(p,q) model, we proceed in a similar manner. We will focus on finding the prediction for a GARCH(1,1) model for two reasons: for a major simplicity in the equations and because we restrict our study to these models. If we rewrite the definition of a GARCH(1,1) model given in 2.35 we get:

$$\begin{aligned}
\sigma_n^2 &= \alpha_0 + \alpha_1 Y_{n-1}^2 + \beta_1 \sigma_{n-1}^2 \\
&= \alpha_0 + \alpha_1 \sigma_{n-1}^2 \varepsilon_{n-1}^2 + \alpha_1 \sigma_{n-1}^2 - \alpha_1 \sigma_{n-1}^2 \\
&= \alpha_0 + (\alpha_1 + \beta_1)\sigma_{n-1}^2 + \alpha_1 \sigma_{n-1}^2 (\varepsilon_{n-1}^2 - 1).
\end{aligned}$$

As we did with the ARCH model,

$$\widehat{Y_{n+1}^2} = E_{M(Y_1,\ldots,Y_{n+p-1})} \varepsilon_{n+1}^2 \sigma_{n+1}^2 = \alpha_0 + (\alpha_1 + \beta_1)\sigma_{n-1}^2.$$

Where have applied $E_{M(Y_1,\ldots,Y_{n+p-1})}(\varepsilon_{n+1}^2 - 1) = E(\varepsilon_{n+1}^2 - 1) = 0$. By reiteration:

$$\begin{aligned}\widehat{Y_{n+2}^2} &= \alpha_0 + (\alpha_1+\beta_1)\sigma_{n+1}^2 = \alpha_0 + (\alpha_1+\beta_1)\big[\alpha_0 + (\alpha_1+\beta_1)\sigma_n^2\big]\\ &= \alpha_0 + \alpha_0(\alpha_1+\beta_1) + (\alpha_1+\beta_1)^2\sigma_n^2\\ \widehat{Y_{n+l}^2} &= \alpha_0(1 + (\alpha+\beta)^1 + \ldots + (\alpha_1+\beta_1)^{l-1}) + (\alpha_1+\beta_1)^l\sigma_n^2.\end{aligned}$$

By the sum of a geometric progression of $l-1$ terms we get

$$\begin{aligned}\widehat{Y_{n+l}^2} &= \alpha_0\frac{1-(\alpha_1+\beta_1)^l}{1-(\alpha_1+\beta_1)} + (\alpha_1+\beta_1)^l\sigma_n^2\\ &= \frac{\alpha_0}{1-\alpha_1-\beta_1} + (\alpha_1+\beta_1)^l\left(\sigma_n^2 - \frac{\alpha_0}{1-\alpha_1-\beta_1}\right).\end{aligned}$$

Using the definition of σ^2 obtained in Proposition 2.37

$$\widehat{Y_{n+l}^2} = \sigma^2 + (\alpha_1+\beta_1)^l(\sigma_n^2 - \sigma^2).$$

This expression shows us that if $l \to \infty$, we get $\widehat{Y_{n+l}^2} \to \sigma^2$, provided that $\alpha_1 + \beta_1 < 1$.

In section 4, we will delve into this, as well as deal extensively deal with the case $\alpha_1 + \beta_1 \approx 1$. This will bring us to some newer developments in the GARCH family, which were discovered in the 2000s.

3.4 Noteworthy Tests for Time Series Analysis

Recall basic statistics: *Hypothesis tests* allow us to confirm or deny an assumption or theory, given a set of observed values. We call H_0 the null hypothesis, which is the assumption we want to confirm. This is usually a value or a parameter, i.e. Is the coin fair? (50% chance of getting heads or tails). H_1 refers to the alternative hypothesis i.e. The coin is biased. A test with its correspondent distribution will give us a certain value, which we must interpret.

The *significance level* or α level is the probability of rejecting the null hypothesis when it is true. This error is called Type I error. The *p-value* is the probability of obtaining test results as extreme as the results observed, assuming that the null hypothesis is correct. Recall that once we have calculated the *p-value*, if it is less than or equal to the significance level, we discard the null hypothesis.

3.4.1 Testing for ARCH effects

Engle in [11] proposed the usage of the LM-Test (or Breusch-Godfrey test) to check for ARCH effects. LM stands for Lagrange Multiplier. Recall from Proposition 2.30 the AR(p) representation for an ARCH(p) model, meaning that the squared returns Y_t^2 follow and AR(p) model. We apply the OLS method to the following autoregression

$$Y_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + \ldots + \alpha_p Y_{t-p}^2 + v_t$$

where $v_t = \sigma_t^2(\varepsilon_t^2 - 1)$ is a white noise. In section 3.2, we went into detail on how to determine $\alpha_0, \ldots, \alpha_p$. Now, for the LM test, we have define the null and alternative hypothesis as:

$$H_0 : \alpha_1 = 0, \alpha_2 = 0, \ldots, \alpha_p = 0 \qquad H_1 : \alpha_1 \geq 0, \alpha_2 \geq 0, \ldots, \alpha_p \geq 0.$$

with at least one strict inequality. The LM Test Statistic is $(T-p)R^2$ where T is the number of observations, p is the parameter of the model and R^2 is the "R-squared" coefficient (R^2 statistic). We have the following result:

$$(T-p)R^2 \sim \chi^2_p$$

if the null hypothesis holds. If $(T-P)R^2 \geq \chi^2_{p,1-\alpha}$ we reject the null hypothesis, which means that there are ARCH effects.

3.4.2 More Tests

In order to reach a profound understanding once we model stock returns, we need the tools given by hypothesis tests. We want to study normality, correlation, volatility. The following survey displays the main hypothesis tests we will be working with and points out the main aspects concerning their usage.

- **Shapiro-Wilk Test:** The null hypothesis H_0 is that the observed values $Y_1, \ldots Y_n$ are normally distributed.
- **Q-Q Plot:** This test is graphical, in the sense that we plot the quantiles of two distributions we wish to compare. In general, we compare sorted data to a normal distribution. Usually the y-axis coordinates refer to the observed values, and the x-axis is the value correspondent to a normal distribution. If all points (x, y) are close to the straight line $y = x$ we can conclude that the our observed values are normally distributed.
- **Jarque Bera Test:** (1981) We calculate the quantity

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

  where S and K are the sample skewness and kurtosis (refer to the definitions in section 2.1). Note that $JB \sim \chi^2_2$ Our goal is also to check if our sampled data comes from a normal distribution. The null hypothesis here is that the sample skewness and excess kurtosis $(K-3)$ are zero. It easy to see that the quantity JB will increase the further away S and K are from zero. One of the limitations of this test is that works poorly if the samples are not large. Another drawback is that sample skewness and kurtosis must be calculated empirically, which may result in an inexact measurement. Newer alternatives are available, but they exceed our purpose.
- **Portmanteau Tests:** These tests is characterized by a clear null hypothesis, whereas the alternative stays flexible. The main tests in this family are the **Box-Pierce test** and **Ljung-Box test**. The latter one is an evolution of the first one and is most widely used. Here the null hypothesis H_0 asserts that the returns are uncorrelated. The alternative hypothesis H_1 would be that there is some sort of serial correlation between returns. We calculate the Q-statistic:

$$Q = n(n+2)\sum_{l=1}^{H} \frac{\hat{\rho}^2(l)}{n-l}$$

where n is the sample size and H is a value chosen arbitrarily smaller than n. Under H_0, we have $Q \sim \chi^2_{H-p-q}$. We reject the null hypothesis at significance level α if $Q > \chi^2_{H-p-q}$ in the $(1-\alpha)$ quantile. p, q are included in case that we are under an ARIMA model.

Akaike then introduced a magnitude which masures goodness of fit for a given model.

Definition 3.27. The *Akaike Information Criterion* (AIC) is:

$$AIC = 2k - 2\log(\hat{L})$$

where k is the number of estimated parameters of the model and $\hat{L}$ is the maximum value of the log-likelihood function. The smaller the obtained value, the beter the model. We can see the term $2k$ as a “penalty” for over-fitting a model, since more parameters usually means a better fit. There is not complete agreement on the penalty term, and there exists specific literature devoted to this question.

Definition 3.28. The *Bayesian Information Criterion*(BIC) is:

$$BIC = k\log(n) - 2\log(\hat{L})$$

Here, n is the number of observations.

3.5 Estimation

The results in this section are valid considering we are in $L^2(\Omega, \mathcal{F}, \mathcal{P})$.

3.5.1 The causal ARMA(p,q) model

In section 2.1, we briefly introduced AR(p) models. Let us consider the simple example of an AR(1) model, which is $Y_t = \phi Y_{t-1} + \omega_t$ where $\omega_t \sim \mathrm{WN}(0, \sigma^2)$. If we iterate k times we get:

$$\begin{aligned} Y_t &= \phi(\phi(Y_{t-2} + \omega_{t-1})) + \omega_t = \phi^2 Y_{t-2} + \phi\omega_{t-1} + \omega_t \\ &= \dots \\ &= \phi^k Y_{t-k} + \sum_{j=0}^{k-1} \phi^j \omega_{t-j} \end{aligned}$$

Since $\{Y_t\}$ is stationary, we have that $\|Y_t\|^2 = E(Y_t^2)$ (given $Y_t \in L^2$) is constant. If we subtract X_t and the sum of white noises obtained above we get:

$$\left\| Y_t - \sum_{j=0}^{k-1} \phi\omega_{t-j} \right\|^2 = \phi^{2k}\|X_t\|^2 \to 0 \quad \text{as } k \to \infty$$

which is true iff $|\phi| < 1$. We then know that

$$Y_t = \sum_{j=0}^{\infty} \phi\omega_{t-j} \tag{3.11}$$

thanks to the Lemma below. Note that we have substituted k for ∞.

Lemma 3.29. Given $\{U_k \ k \geq 1\}$ a sequence of centered, pairwise uncorrelated and square-integrable random variables we have

$$\sum_{k\geq 1} U_k \quad \text{is convergent in } L^2 \text{ if and only if} \quad \sum_{k\geq 1} \mathrm{Var}(U_k) < \infty$$

(Recall the notion of convergence in L^2 given in definition 3.10)

Proof. Found in [2] $\square$

Note that representation (3.11) is useful because it represents en AR(p) process as a linear model. We quickly deduce the stationarity of the process since

$$E(X_t) = \sum_{j=0}^{\infty} \phi E(\omega_{t-j}) = 0$$

In section 2.1 we also went over ARMA(p,q) models. The compact form is simple:

$$\phi(B)Y_t = \theta(B)\omega_t \quad t = 0, \ldots, n$$

Definition 3.30. An ARMA(p,q) process is *causal* if we can write the time series $\{Y_t\}$ as a one sided linear process. Hence, there exists a sequence of constants $\{\psi_j\}$ such that

$$Y_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \quad t = 0, 1 \ldots \tag{3.12}$$

where $\sum_{j=0}^{\infty} |\psi_j| < \infty$.
The idea behind causality is the problem that future-dependent models present. In some cases, which we do not intend to cover, some models have a future-dependent expression. A very simple example is the AR(1) model $Y_t = -\sum_{j=1}^{\infty} \phi^{-1} Y_{t+j}$. Note that forecasting is useless here, since Y_t depends on future values. This is why the causal restriction on a model is important. The following theorem characterizes the coefficients ψ_j, as well as given necessary and sufficient conditions for the causality of an ARMA model.

Theorem 3.31. Given $\{Y_t\}$ and ARMA(p,q) process, we suppose the polynomials $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeros. Then $\{Y_t\}$ is causal if and only if $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficients $\{\psi_j\}$ in (3.11) are determined as follows:

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j \frac{\theta(z)}{\phi(z)} \quad |z| \leq 1$$

Proof. Can be found in [4], pp.83-86. $\square$

We begin tackling Estimation. Given a $\{Y_t\}$ causal AR(p) process with $E(X_t) = 0$, using the representation given in 2.13

$$X_t = \phi_1 X_{t-1} + \ldots \phi_p X_{t-p} + \omega_t \quad \omega_t \sim \mathrm{WN}(0, \sigma^2) \tag{3.13}$$

we wish to estimate $\phi_1, \ldots, \phi_p$. Using similar notation as in section 3.3, we define $\phi = (\phi_1, \ldots, \phi_p)^T$. Our goal is to find ϕ and the variance σ^2. for the white noise. Since AR(p) is causal:

$$X_t = \sum_{j=0}^{\infty} \psi_j \omega_{t-j}$$

Note that since we are under an AR(p) model, by Theorem 3.31 we get $\psi(z) = \sum_{j=0}^{\infty} \psi_j \frac{1}{\phi(z)}$ for $|z| \leq 1$. This is because $\theta(z) \equiv 1$. Now, if we multiply expression (3.12) by X_{t-h} for $h = 0, \ldots, p$ we get

$$X_t X_{t-h} + \phi_1 X_{t-1} X_{t-h} + \ldots + \phi_p X_{t-p} X_{t-h} = \omega_t X_{t-h} \quad h = 0, 1 \ldots p$$

Taking expectations on both sides we obtain

$$\gamma(h) + \phi_1 \gamma(h-1) + \ldots + \phi_p \gamma(h-p) = 0 \quad h = 0, 1 \ldots p \tag{3.14}$$

where we have noted that

$$E(\omega_t X_{t-h}) = E\bigg(Z_t \sum_{j=0}^{\infty} \psi_j \omega_{t-h-j}\bigg) = 0.$$

We can write (3.14) in an abbreviated form which is.

$$\Gamma_p \phi = \gamma_p$$

where Γ_p is the variance-covariance matrix, and $\gamma_p = (\gamma(1), \ldots, \gamma(p))^T$.

Definition 3.32. Equations defined in (3.14) are called the *Yule-Walker* equations.

Now, since we are trying to estimate ϕ, we must take into consideration that in our model we can only calculate sample covariances $\widehat{\gamma(h)}$. So we call $\hat{\phi}$ and $\hat{\sigma}^2$ the Yule-Walker estimators, which we find be solving:

$$\hat{\Gamma}_p \hat{\phi} = \hat{\gamma}_p$$

Theorem 3.33. Under the conditions of the AR(p) process defined in (3.13), if we try to estimate ϕ by the Yule-Walker equations, the estimator $\hat{\phi}$ then satisfies

$$\sqrt{n}(\hat{\phi} - \phi) \rightarrow N(0, \sigma^2 \Gamma_p^{-1}).$$

3.6 Estimation of ARCH(1) and GARCH(1,1) models

Consider the AR(1) model from Definition 2.26

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 Y_{t-1}^2 \end{aligned}$$

Proposition 2.30 showed us that Y_t^2 follow an AR(1) model. We estimate the parameters α_0 and α_1 using MLE. The next steps are to be followed:

1. If $Y = (Y_1, \ldots, Y_n)$. we calculate the Likelihood function $L(Y, \alpha_0, \alpha_1) = f_{\alpha_0, \alpha_1}(Y)$. (Note the ambiguity between the random variable and the observed value, which does not affect our reasoning). Trivially, we know that $Y_1, \ldots Y_n$ are not independent, so we must consider the conditional likelihood functions, since the variables are conditionally independent (c.f. Proposition 2.29 (the returns are uncorrelated)). This yields:

$$L(Y_1, \alpha_0, \alpha_1) = \prod_{t=2}^{n} f_{\alpha_0, \alpha_1}(Y_t \mid Y_{t-1})$$

We saw before that $Y_t \mid Y_{t-1} \sim \mathrm{N}(0, \alpha_0 + \alpha_1 Y_{t-1}^2)$. Which means that

$$f_{\alpha_0,\alpha_1}(Y_t \mid Y_{t-1}) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{Y_t^2}{2\sigma_t^2}\right)$$

Consequently

$$L(Y_1, \alpha_0, \alpha_1) = \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{Y_t^2}{2\sigma_t^2}\right)$$

Taking logarithms and arranging terms :

$$-\frac{n}{2}\log(2\pi) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\left(\sum_{t=2}^{n} \log(\sigma_t^2) + \frac{Y_t^2}{\sigma_t^2}\right)$$

It is easy to note that multiplication by a constant in the Likelihood function does not affect the calculation of the extrema. Therefore, we can discard the first two terms in the previous equation, yielding:

$$\log L(Y_1, \alpha_0, \alpha_1) = -\frac{1}{2}\sum_{t=2}^{n}\left(\log(\sigma_t^2) + \frac{Y_t^2}{\sigma_t^2}\right) = -\frac{1}{2}\sum_{t=2}^{n}\left(\log(\alpha_0 + \alpha_1 Y_{t-1}^2) + \frac{Y_t^2}{\alpha_0 + \alpha_1 Y_{t-1}^2}\right)$$

2. We derivate with respect to parameters α_0 and α_1 and get:

$$\begin{aligned}
\frac{\partial \log L(Y, \alpha_0, \alpha_1)}{\partial \alpha_0} &= -\frac{1}{2}\sum_{t=2}^{n}\left[\frac{1}{\sigma_t^2}\frac{\partial}{\partial \alpha_0}(\sigma_t) - 2Y_t^2 \frac{\partial}{\partial \alpha_0}\left(\frac{1}{\sigma_t^2}\right)\right] = -\frac{1}{2}\sum_{t=2}^{n}\left[\frac{1}{\sigma_t^2} - \frac{Y_t^2}{\sigma_t^4}\right] \\
\frac{\partial \log L(Y, \alpha_0, \alpha_1)}{\partial \alpha_1} &= -\frac{1}{2}\sum_{t=2}^{n} Y_{t-1}^2\left[\frac{1}{\sigma_t^2} - \frac{Y_t^2}{\sigma_t^4}\right]
\end{aligned}$$

In theory, we should equate the derivatives to zero, but we see this is a complex equation to solve.

3. We also calculate the second derivatives:

$$\begin{aligned}
\frac{\partial^2 \log L(Y, \alpha_0, \alpha_1)}{\partial \alpha_0^2} &= -\frac{1}{2}\sum_{t=2}^{n}\left[-\frac{2}{\sigma_t^3}\frac{1}{2\sigma_t} + -2Y_t^2\frac{\partial}{\partial \alpha_0}\left(\frac{1}{\sigma_t^4}\right)\right] \\
&= -\frac{1}{2}\sum_{t=2}^{n}\left[-\frac{1}{\sigma_t^4} + \frac{2Y_t^2}{\sigma_t^6}\right]
\end{aligned}$$

$$\frac{\partial^2 \log L(Y, \alpha_0, \alpha_1)}{\partial \alpha_1^2} = -\frac{1}{2}Y_{t-1}^4\sum_{t=2}^{n}\left[-\frac{1}{\sigma_t^4} + \frac{2Y_t^2}{\sigma_t^6}\right]$$

$$\frac{\partial^2 \log L(Y, \alpha_0, \alpha_1)}{\partial \alpha_0 \alpha_1} = -\frac{1}{2}\sum_{t=2}^{n} Y_{t-1}^2\left[-\frac{1}{\sigma_t^4} + \frac{2Y_t^2}{\sigma_t^6}\right]$$

Where we have used

$$\begin{aligned}
\frac{\partial \sigma_t}{\partial \alpha_0} &= \frac{1}{2\sigma_t} \\
\frac{\partial \sigma_t}{\partial \alpha_1} &= \frac{1}{2\sigma_t}Y_{t-1}^2
\end{aligned}$$

Now that we have calculated the Jacobian and the Hessian Matrix, we go over a few methods avaiable for finding the vector of parameters $\theta = (\alpha_0, \alpha_1)^T$. All methods we consider are a variation of the Newthon-Raphson algorithm for finding zeros, studied in any standard Numerical Analysis course.

Definition 3.34. Under regularity conditions $I = -E\left[\left(\frac{\partial^2}{\partial\theta_i\theta_j}\log L(x,\theta)\right)_{i,j=1\div n}\right]$ is the *Fisher information matrix*, where θ is the vector of parameters.

Theorem 3.35. Under some technical assumptions, an MLE estimator $\hat{\theta}$ is asymptotically normal:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, I^{-1})$$

As long as the process is strictly stationary. We avoid going into the technical assumptions because it has little to do with our focus. Summarizing, we could say that we need a regular statistical model as well as some other restrictions.

Proof. cf. Proposition 8.2.6 in [9]. □

The Fisher-Scoring Algorithm. Given an inital value θ_0, which can be the Yule-Walker Estimator, one can approximate using a Taylor expansion:

$$J_f(\theta) \approx J_f(\theta_0) + H_f(\theta_0)(\theta - \theta_0).$$

Note that J_f and H_f denote that Jacobian and Hessian matrix for $f = \log L(Y, \alpha_0, \alpha_1)$ respectively. Solving for θ we get

$$\theta = \theta_0 - H_f(\theta_0)^{-1}J_f(\theta_0).$$

Now we construct the following iterative method:

$$\begin{aligned} &\theta_0 = \hat{\theta}_0 \\ &\theta_{j+1} = \theta_j - H_f(\theta_j)^{-1}J_f(\theta_j) \quad j = 0, 1, 2, \ldots \end{aligned} \tag{3.15}$$

until convergence, i.e. $\|\theta_{j+1} - \theta_j\| \leq \epsilon$. We note that (3.14) is the classical Newton-Raphson iteration. Most of the time, we substitute the Hessian matrix for the Fisher information matrix defined previously. This gives us:

$$\begin{aligned} &\theta_0 = \hat{\theta}_0 \\ &\theta_{j+1} = \theta_j - I(\theta_j) - 1_f^J(\theta_j) \quad j = 0, 1, 2, \ldots \end{aligned} \tag{3.16}$$

This means we evaluate the Fisher Information Matrix (FIM) at θ_j. It is important not to be confused by the notation of θ_j. Recall that in the definition for the FIM it denotes the parameter index, whereas in the algorithm it refers to the iteration number.

Engle in article [11] where he introduced ARCH models for the first time, used the BHHH (Brendt-Hall-Hall-Hausman) algorithm to estimate parameters. Since we have already covered Fisher Scoring extensively and also for restricted writing space, we opt for leaving out BHHH, as well as many other possibilities.

Now, we consider the case of a GARCH(1,1) model. The log likelihood function looks like:

$$\begin{aligned}
\log L(Y, \alpha_0, \alpha_1, \beta_1) &= -\frac{1}{2}\sum_{t=2}^{n}\left(\log(\alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1\sigma_{t-1}^2) + \frac{Y_t^2}{\alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1\sigma_{t-1}^2}\right) \\
&= -\frac{1}{2}\sum_{t=2}^{n}\left(\log(\sigma_t^2) + \frac{Y_t^2}{\sigma_t^2}\right).
\end{aligned}$$

The derivatives are identical to the ARCH(1) although now we incorporate the derivative with respect to β_1:

$$\begin{aligned}
\frac{\partial \log L(Y, \alpha_0, \alpha_1, \beta_1)}{\partial \beta_1} &= -\frac{1}{2}\sum_{t=2}^{n}\sigma_{t-1}^2\left[\frac{1}{\sigma_t^2} - \frac{Y_t^2}{\sigma_t^4}\right] \\
\frac{\partial^2 \log L(Y, \alpha_0, \alpha_1, \beta_1)}{\partial \beta_1^2} &= -\frac{1}{2}\sum_{t=2}^{n}\sigma_{t-1}^4\left[-\frac{1}{\sigma_t^4} + \frac{2Y_t^2}{\sigma_t^6}\right]
\end{aligned}$$

where we have noted that

$$\frac{\partial \sigma_t}{\partial \beta_1} = \frac{1}{2\sigma_t^2}\sigma_{t-1}^2.$$

However, a general derivative with respect to $\theta = (\alpha_0, \alpha_1, \beta_1)^T$ is adequate for the GARCH(1,1) case. Note that we can write all the first derivatives with respect to $\alpha_0, \alpha_1, \beta_1$ in the summarized form:

$$\frac{\partial \log L(\theta)}{\partial \theta} = -\frac{1}{2}\sum_{t=2}^{n}\frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta}\left(\frac{Y_t^2}{\sigma_t^2} - 1\right).$$

We note that

$$\frac{\partial \sigma_t^2}{\partial \theta} = (1, Y_{t-1}^2, \sigma_{t-1}^2)^T + \frac{\partial \sigma_{t-1}^2}{\partial \theta}.$$

The Hessian Matrix is harder to compute and we omit its calculation, but once we have it we can adapt Theorem 3.35 to GARCH(1,1) which gives us asymptotic normality for estimator $\hat{\theta}$:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, I^{-1})$$

where I is the fisher information matrix. This theorem is valid as long as the GARCH(1,1) process is strictly stationary.

4 Statistical Analysis of Stock Returns

4.1 Plan Outline

As it was stated in the introduction, we wish to put the mathematical theory developed in Chapters 2 and 3 into practice. Volatility is seen regularly in finance, and especially in stock returns. The main focus of this section will be to give an in depth statistical analysis of the returns observed in the S&P500 index. This is a stock market index that lists the 500 largest companies in the United States. More information can be found at `https://en.wikipedia.org/wiki/S%26P_500` A robust analysis is possible thanks to the usage of the we use the R statistical package. What follows is a detailed list of steps that we will go over in our analysis in order to reach consistent conclusions. However, ARCH and GARCH models do not adjust well to every sample. Our goal is to study if they work better for shorter or longer samples of data. We will compare a sample of data for 1 year and 11 years.

1. *Calculating statistical parameters.* Once we have defined the time frame of our data, we proceed to calculate main statistical parameters. This is done to verify the stylized facts. We will graph the auto-correlation function for the returns and squared returns and use results obtained in Section 2 that help us identify volatility.

2. *Fitting a model.* We will start by showing why classical regression fails when modeling time series of this nature, which might seem as an obvious claim at this point, but nonetheless it i useful to explore the reason why alternative models were needed. We will test for ARCH or GARCH effects, and adequately fit them. (estimation of parameters).

3. *Prediction.* Using the results in section 3, we will try to predict the future values of the conditional variance. We will discuss the certainty or not of the results.

4. *Alternatives and shortfalls.* Evidently, any prediction is far away from perfect. We will analyze where and when ARCH and GARCH models fail and provide information on some recent developments in the area of volatility models and the new contributions they have brought upon.

4.2 Data Summary

Data is retrieved from Yahoo Finance at `https://stooq.com/q/d/?s=%5Espx`. For the short sample, we consider the daily returns registered from May 26th 2020 to May 24th 2021 (1 year approximately). This consists of a total of 252 returns. For our long sample, we use daily return data starting May 26th 2010 up to May 24th 2021, a total of 2770 returns.

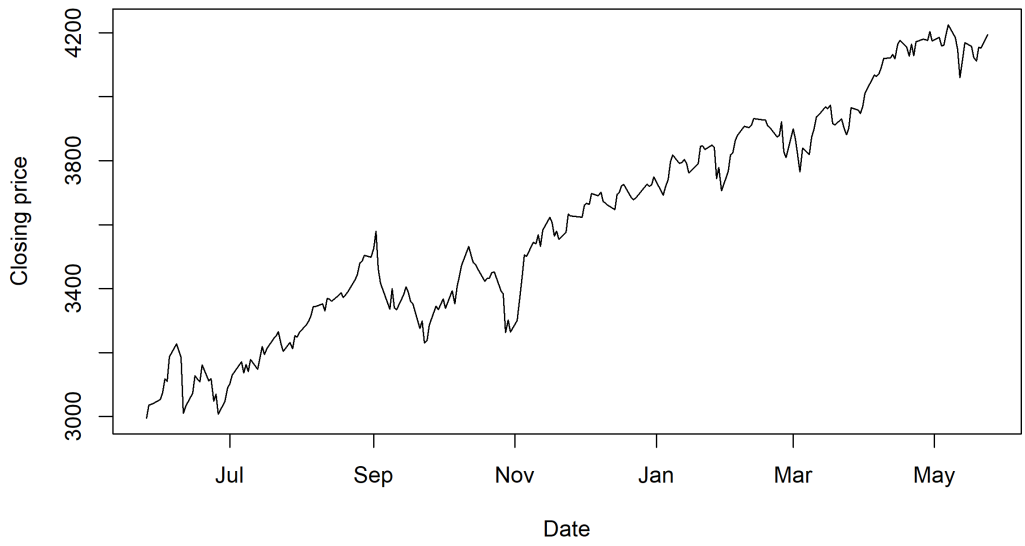

Figure 4.1: Stock value from May 26th 2020 up to May 24th 2021

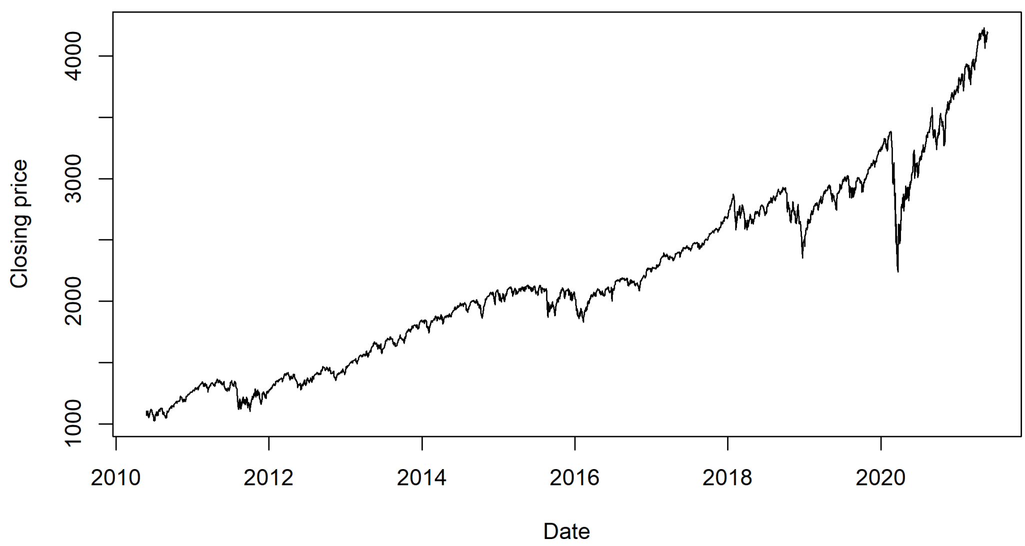

Figure 4.2: Stock value from May 26th 2010 up to May 24th 2021

If we plot the returns (see definition 2.16) we get the following:

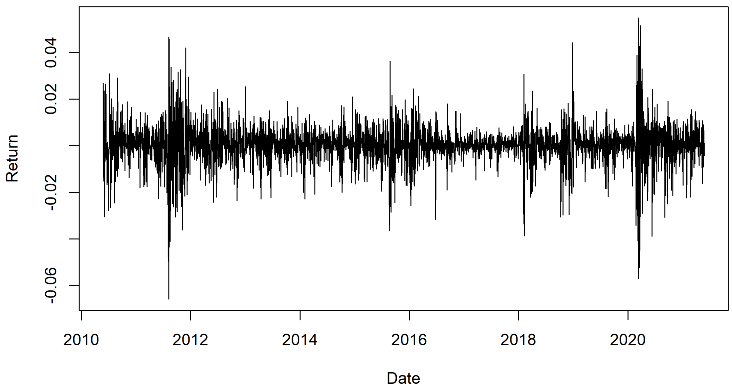

Figure 4.3: Returns from May 26th 2010 up to May 24th 2021

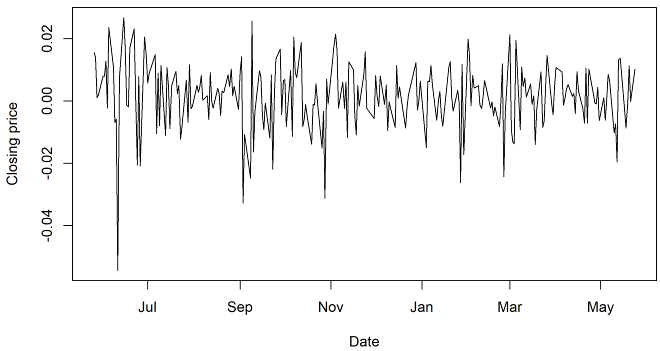

Figure 4.4: Returns from May 26th 2020 up to May 24th 2021

The following tables summarizes what can be seen in the plot, as well as outlays significant statistical parameters

Long Sample	Min	Max	Mean	Median	Std. Dev.	Kurtosis	Skewness
returns	-0.06593	0.05487	0.00032	0.00054	0.00908	6.17622	-0.39274
squared returns	0	0.00435	8.261e-05	1.579e-05	0.000235	90.0769	7.97224
log-returns	-0.12765	0.08968	0.00049	0.00066	0.01098	16.7477	-0.87358

Table 1: Magnitudes for S&P500 returns (large sample) , logarithmic returns and squared returns

A histogram of the returns gives us an accurate distribution of the returns for both samples

Short Sample	Min	Max	Mean	Median	Std. Dev.	Kurtosis	Skewness
returns	-0.0545	0.0268	0.00146	0.00165	0.01061	2.95522	-0.8549
squared returns	0	0.00297	0.00011	4.427e-05	0.00024	77.8456	7.40841
log-returns	-0.05674	0.02422	0.00135	0.00185	0.01079	3.28209	-1.08676

Table 2: Magnitudes for S&P500 returns (short sample), logarithmic returns and squared returns

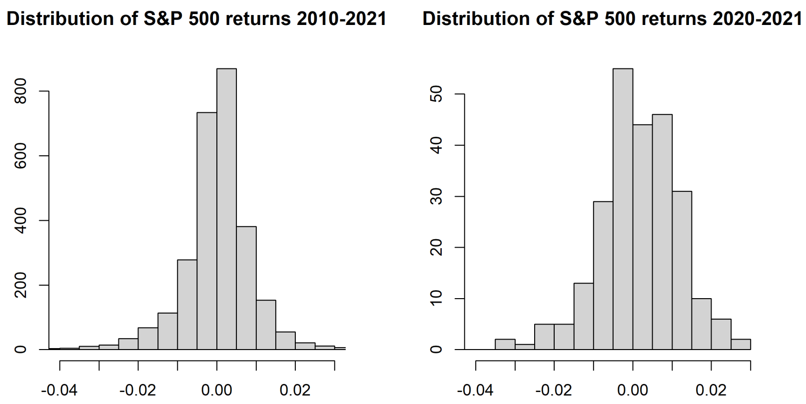

Figure 4.5: Histogram of the returns

The ACF for the returns and absolute returns are:

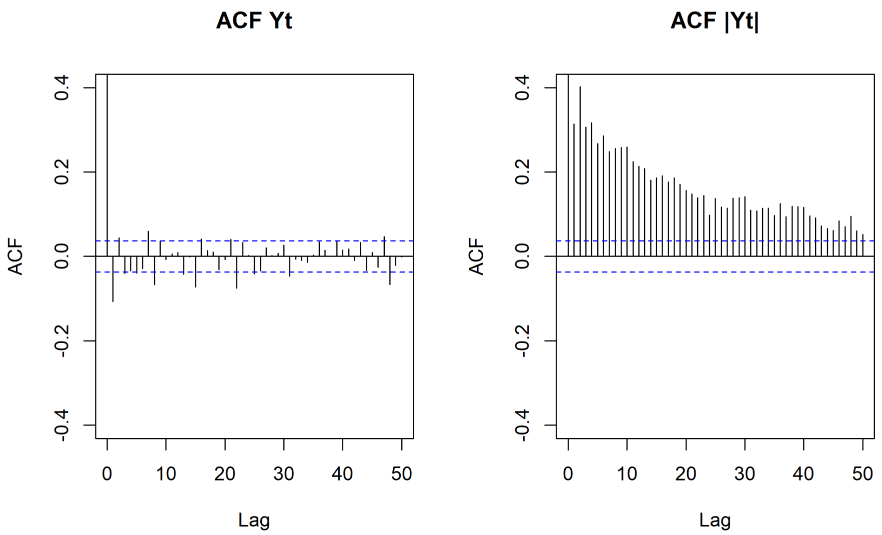

Figure 4.6: ACF functions up to $l = 50$ for returns and absolute returns (long sample)

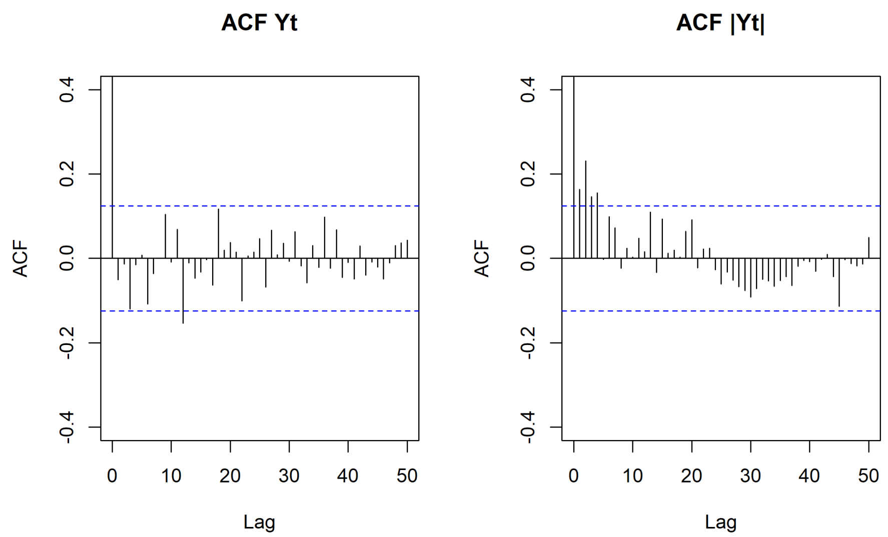

Figure 4.7: ACF functions up to $l = 50$ for returns and absolute returns (short sample)

If we plot the ACF for the squared returns, we get the following results. Note that for the long case, a modeling with ARCH and GARCH Processes is suggested in virtue of Proposition 2.30 and 2.36. In following table, which can be found in [1] summarizes the behavior of the ACF and PACF for ARMA models, which will help us verify the results.

	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Cuts after lag q	Tails off
PACF	Cuts after lag p	Tails off	Tails off

Table 3: ACF and PACF functions for ARMA(p,q) models

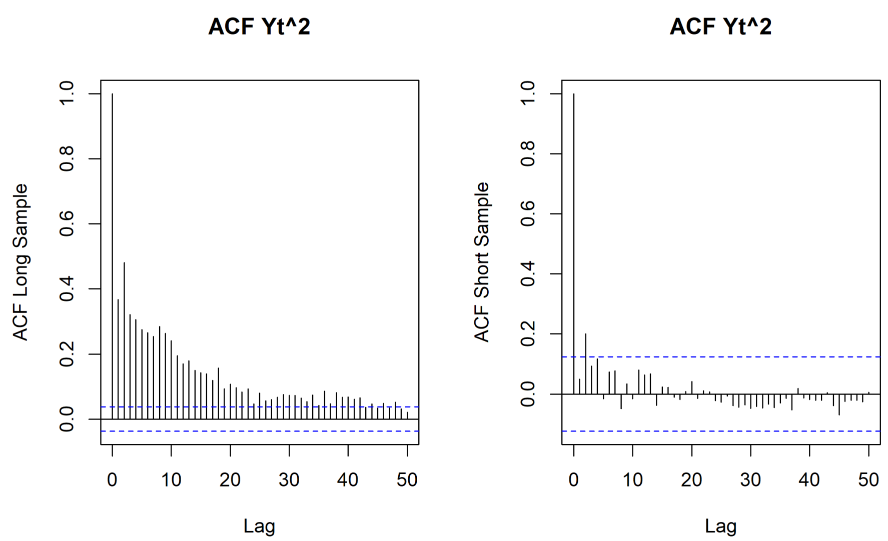

Figure 4.8: Squared ACF up to $l = 50$ for short and long sample

Let us go over the stylized facts described in section 2.2.1. Through a qualitative analysis based on the plots above, we will try to verify them.

- The functions `skewness()` and `kurtosis()` are included in the moments R package. We start by noticing that the distribution is negatively skewed, which corresponds to the "Asymmetry" stylized fact. It may not be evident by just graphing the histograms, but if we take the following fact: In a negatively skewed distribution, the mean is smaller than the median, which is at the same time is smaller than the mode. The results both in Table 1 and 2 satisfy this claim (mean < median) We note that it is almost a negligible difference.

- The blue dotted line in Figure 4.4 indicates where the ACF is significantly different from 0. Note that it is for very few lags that the dotted line is exceeded. There is no significant autocorrelation for either the short sample or the long one; however, for the latter one correlation is significantly smaller. This is due to the bigger picture that a longer sample gives. In conclusion, our returns behave similarly to a white noise. However, in the absolute returns we see autocorrelation decreasing in a steady manner as the lag l increases, but only for the large sample. This is another stylized fact, and it is noteworthy to observe that the short sample does not in any way get close to following it (except for the first 4 lags). This is will interfere once why try to model.

- We perform R analysis on the long sample:

  ```
  > which.max(ret); which.min(ret)
  [1] 2467 [1] 304
  > head(order(ret)); tail(order(ret))
  [1] 304 2466 2472 2468 302 306
  [1] 2474 2162 307 305 2476 2467
  ```

  This gives us a good idea of volatility clustering. `head(order(ret))` returns the position of lowest values, the first one being the smallest. `tail(order(ret))` returns the position of the highest values, the last one being the biggest. Observe that except value "2162", all the other ones belong to two periods. For instance, note how values 302, 304 and 306 belong to the lower returns end, and 307 and 305 belong the higher returns end. This clearly verifies the volatility clustering claim.

4.3 Fitting and Prediction Results

The problems with Linear Regression. The `lm()` function allows us to fit a line $y = \beta_0 + \beta_1 x + \varepsilon$ to our data. Recall that ε is assumed to be iid noise with mean 0 and variance σ^2. It is easy to notice that this kind of analysis does not capture absolutely any specific characteristics of our time series.

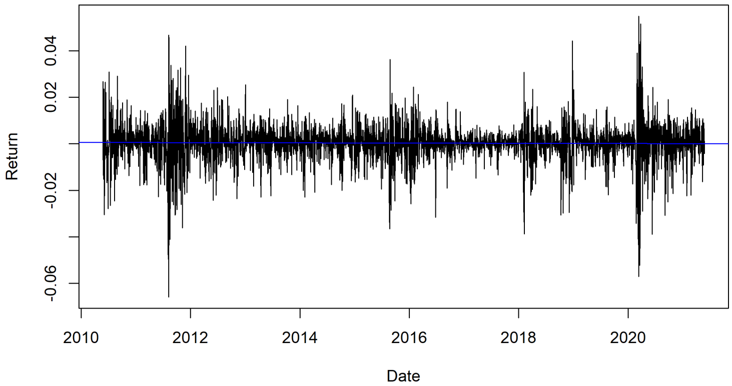

Figure 4.9: Linear regression applied to S&P500 returns

The results are very poor as the Ajusted R-squared statistic is -0.0001299, showing no correlation between variables. We know that is true, since the ACF function showed us that for any lag $l > 1$ significant correlation did not exist. So why linear regression will not work ? Because it does not offer any significant information on conditional variance. Once again, the model we are using for our data is:

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t \quad Var(Y_t|\mathcal{F}_{t-1}) = \sigma^2.$$

So we are assuming homoskedasticity. Another major assumption in the model that does not work well for us is independence of the residuals. Evidently, returns are not independent, as return Y_t depends somehow from the previous one. In other words, the dependent variable is influenced by the past independent variables. As we know, independence implies uncorrelated, but not way around.

Before exploring ARCH and GARCH, we recall that ARMA(p,q) tend to be poor fits for financial return, since they assume correlation between variables and homoskedasticity.

Testing for ARCH Effects. In section 3.4.1, we briefly went over on how to test for ARCH effects via the LM- Test. A regression $Y_t = \beta_0 + \varepsilon_t$ will allow us to obtain a vector for the residuals. This technique is called regression on a constant. After this, we calculate the squared residuals and perform a linear regression, which has as many parameters as the ARCH(p) model we want to fit. Recall the null hypothesis is $\alpha_0 = \alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$. For the S&P 500 time series, we consider fitting an ARCH of order 1, 2 and 3. The `ArchTest()` function from the `finTS()` package performs this test. The results for the long and short sample are respectively: The table clearly states

	ARCH(1)	ARCH(2)	ARCH(3)
p-value	<2.2e-16	<2.2e-16	<2.2e-16

Table 4: Long sample p-value test results for ArchTest() function

the difference we started to see with the short sample and the long sample. We reject the null hypothesis in all cases of the long sample since the p-value is 0. However, with the short sample and considering we ahave a significance level of $\alpha = 0.05$, we reject the null hypothesis for order 2 and 3. This means that ARCH(2) and ARCH(3) are still good options to consider.

	ARCH(1)	ARCH(2)	ARCH(3)
p-value	0.5726	0.005208	0.01002

Table 5: Short sample p-value test results for ArchTest() function

Fitting an ARCH(2). The `fGarch` package includes the `garchFit()` function which gives us the possibility to fit ARCH(p) and GARCH(p,q) models. The R console prints the estimate parameters and the results from statistics tests:

	Estimate	Std. Error	t value	p-value
mu	5.766e-04	1.255e-04	4.594	4.35e-06
omega	3.204e-05	1.611e-06	19.885	<2e-16
alpha1	2.601e-01	3.378e-02	7.699	1.38e-14
alpga2	3.863e-01	3.640e-02	10.612	<2e-16

Table 6: Estimated parameter results for ARCH(2)

	Estimate	Std. Error	t value	p-value
mu	1.734e-03	5.970e-04	2.905	0.00367
omega	6.776e-05	1.230e-05	5.507	3.66e-08
alpha1	1.609e-01	1.270e-01	1.267	0.20505
alpha2	2.784e-01	1.234e-01	2.257	0.02402

Table 7: Estimated parameter results for ARCH(2) (short sample)

We start by noting that `mu` is the mean of the series. It has always been common in our discussion to assume that the mean is zero. However, if this is not the case, we use the transformation `x - mean(x)`. Also, `omega` is what we defined as α_0. The function itself performs the t-test for hypothesis, which in general, consists in computing the following quantity:

$$t = \frac{\hat{\beta}}{\text{s.e}(\hat{\beta})}$$

where $\hat{\beta}$ is the estimator for β. s.e$(\hat{\beta})$ is the standard error of the estimator. Statistical packages such as `R` suppose $H_0 : \beta = 0$, although this can be altered. In our case, the standard error is the standard deviation of the estimate. It is clear from the definition that the t-statistic measure the distance from 0 of the parameter. The larger the quantity is, the further away from 0. A rule of thumb for the t-statistic is that if it is larger than 2 we conclude that the parameter is not 0. The p-value works in a similar matter. If the level of significance is $\alpha = 0.05$, we check if $P(> |t|) < 0.05$ which is the definition of the p-value. If this holds, we conclude that the parameters are not zero. Note that the `garchFit()` function gives us the p-value for each parameters, as it is shown on the table. Note that in the long sample case, all the relevant p-values, except are way below 0.05. The fitting for the short sample, however, is not as good. There is not enough returns for the fitting to be good enough.

Fitting a GARCH(1,1) model. Using the same function as before, we get similar results:

	Estimate	Std. Error	t value	p-value
mu	5.094e-04	1.156e-04	4.407	1.05e-05
omega	2.707e-06	4.349e-07	6.224	4.85e-10
alpha1	1.849e-01	1.984e-02	9.321	<2e-16
beta1	7.881e-01	1.964e-02	40.124	<2e-16

Table 8: Estimated parameter results for GARCH(1,1) when modeling the long sample

Note that our results are $\alpha_1 + \beta_1 = 0.1849 + 0.7881 \approx 0.973 < 1$. This means that we have fitted a stationary model to our data. However, it is close to 1. In the next section we will study the case when the sum is equal to 1. An important consideration is how the `fGarch` function works and models. It assumes the returns Y_t follow:

$$\begin{aligned} Y_t &= \mu + Y_t' \\ Y_t' &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \end{aligned}$$

where μ is a constant and Y_t' is a GARCH process exactly how we defined it in Chapter 2. This explains why in the summary of our fitting a `mu` value is given. This is because although we can prove that the returns have mean zero (as we did for ARCH models), this is not the case for sampled data. We call Y_t' the residuals. If we add the residuals to μ we get the original return vector. The `R` output also gives us the Standardized Residuals Tests Table and the Information Criterion.

			Statistic	p-value
Jarque-Bera Test	R	Chi^2	960.4235	0
Shapiro-Wilk Test	R	W	0.9675186	0
Ljung-Box Test	R	Q(10)	12.27823	0.266867
Ljung-Box Test	R	Q(15)	19.78009	0.1805318
Ljung-Box Test	R	Q(20)	30.72414	0.0589316
Ljung-Box Test	R^2	Q(10)	10.16003	0.4265676
Ljung-Box Test	R^2	Q(15)	13.37055	0.5737004
Ljung-Box Test	R^2	Q(20)	15.47306	0.74873
LM Arch Test	R	TR^2	12.16496	0.4325245

Table 9: Standardized Residuals Tests

AIC	BIC
-6.973774	-6.965215

Table 10: Information Criterion Statistics

Let us analyze the results given. We apply the explanation from section 3.4 of the tests. These tests are applied to the standardized residuals, which are Y_t'/σ_t. The Shapiro-Wilk test tests the residuals and gives us a p-value of zero (with a significance level of $\alpha = 0.05$), meaning that the residuals do not come from a normal distribution. The Jarque-Bera test for residuals gives us a high number, and a p-value of zero. The null hypothesis in this test is also that the kurtosis and skewness come from a normal distribution. However, we already stressed how the distribution is quite different than a normal distribution (excess kurtosis of 0, in our case 6), we also saw through that it is negatively skewed.

We reject the null hypothesis. The Ljung-Box Test assesses correlation of returns. The null hypothesis is that the returns are independently distributed. Some of the p-values are high, which tells us that we can affirm that the residuals are independent from each other.

Predicting for GARCH(1,1) We restrict our prediction to GARCH(1,1). We will estimate volatility, using the results from Section 3.3.

$$\sigma_{t+l}^2 = \sigma^2 + (\alpha_1 + \beta_1)^l(\sigma_t^2 - \sigma^2)$$

Using `R` we get:

We plot the both estimated and real results on the same graph

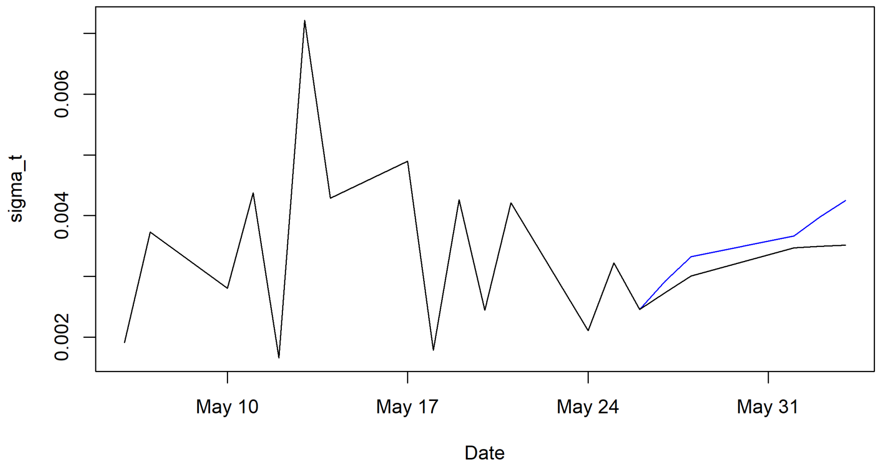

Figure 4.10: Volatility prediction for the next 5 steps

Obviously, prediction is not perfect. However, by applying the formula we get higher values than the real values end up being, but for the most part the trend is respected.

4.4 The shortfalls of ARCH and GARCH models. The IGARCH effect, and EGARCH models

A question that is of major importance and we have yet to assess thoroughly is why did Bollerslev go ahead in perfect the ARCH's, giving way to GARCH. A practical reason is what we have just observed: For some examples of financial returns, ARCH only works with higher p, and as one might think, it becomes more costly to estimate the parameters. GARCH, which innovates by including a dependence on past conditional variances, it allows for a better and more stable fitting, while keeping the parameters low.

As we saw in Corollary 2.40, the condition for stationarity for a GARCH(1,1) model is $\alpha_1 + \beta_1 < 1$. When $\alpha_1 + \beta_1 \approx 1$, this is called the *Integrated GARCH* (IGARCH) effect. This seems to happen with longer samples. With a small amount of observations, the sum of α_1 and β_1 is significantly lower than 1.

Contradicting stationarity means that some of the conditions inherent to the definition are not being fulfilled. This means that either the expected value or the second moment are not constant.

Definition 4.1. An IGARCH(1,1) process is a GARCH(1,1) process with the added condition $\alpha_1 + \beta_1 = 1$.

Proposition 4.2. Strictly stationary IGARCH(1,1) processes exist if $P(\varepsilon_t^2 = 1) < 1$.

Proof. We have $\gamma = E \log(\alpha_1 \varepsilon_t^2 + \beta_1) \leq \log(E(\alpha_1 \varepsilon_t^2 + \beta_1)) = 0$. If the term $\alpha_1 \varepsilon_t^2 + \beta_1$ is a constant almost surely, the inequality is strict. But we know by definition $E(\alpha_1 \varepsilon_t^2 + \beta_1) = 1$ if and only if $\alpha_1 \varepsilon_t^2 + \beta_1 = 1$, which means $\varepsilon_t^2 = 1$ a.s. If we supposed the inequality is strict, hence contradicting Theorem 2.40, we conclude the proof. $\square$

What about the shortcomings of GARCH? Even before 1980s it had been observed that the assumption that positive and negative returns had the same effect on volatility was often not true. In other words, volatility increased more when there were negative returns. This was named the *Leverage effect.* Nelson & Cao developed an extension of the GARCH model called the EGARCH model (*Exponential* GARCH), whose goal was to include this asymmetry of the volatility.

Definition 4.3. We define the EGARCH(1,1) model as:

$$\begin{aligned} &Y_t = \sigma_t \varepsilon_t \\ &\log \sigma_t^2 = \alpha_0 + \alpha_1 g(\varepsilon_{t-1}) + \beta_1 \log \sigma_{t-1}^2 \end{aligned}$$

where $g(\varepsilon_{t-1}) = \theta \varepsilon_{t-1} + \varsigma(|\varepsilon_{t-1}| - E(|\varepsilon_{t-1}|))$ and $\alpha_0, \alpha_1, \beta_1, \theta, \varsigma$ are real constants.
First of all, we take notice of the multiplicative form of the volatility (in contrast to the additive aspect of the volatility in a traditional GARCH model). We have:

$$\sigma_t^2 = e_0^{\alpha} + e^{\alpha_1 g(\varepsilon_{t-1})} (\sigma_{t-1}^2)^{\beta_1}.$$

Let us examine the g function. We suppose w.l.o.g that $\alpha = 1$. If $\varepsilon_{t-1} < 0$ (meaning $Y_{t-1} < 0$). Note that if ε_{t-1} is decreasing, we have a change of rate $\varsigma - \theta$ for $\log \sigma_t^2$ ($\varsigma + \theta$ if ε_{t-1} increases). So σ_t^2 increases iff $\varsigma - \theta > 0$ and $\varsigma + \theta > 0$). We get $-\varsigma < \theta < \varsigma$.

Having said this, where is the asymmetry effect shown ? Note that a negative shock $\varsigma - \theta$ has a bigger effect than a positive shock $\varsigma + \theta$ if and only if $\theta < 0$. In conclusion the coefficient θ reflects asymmetry, the term $\theta \varepsilon_t$ determines sign effect and the latter term the size of the effect. Finally, note that it is immediate to see that $E(g(\varepsilon_t)) = 0$.

5 Conclusions

To put an end to this thesis after a long journey is not a trivial matter, we have seen that our field of study acts as a sort of Pandora's Box, in a good way! Since its early developments in the 1980's, its consolidation with the appearance of GARCH in 1986, a never ending and increasing investigation has brought mathematical developments of financial time series analysis to a new level. It was the main goal of this dissertation to offer a vision of what were its beginnings.

Having said this, we must offer some concluding thoughts. Engle revolutionized Time Series with the discovery of ARCH models. In Chapter 2, we proved most of the basic properties associated with ARCH and studied its relationship with AR(p) models via squared returns. Moving to Bollerslev's discovery of the GARCH models, we proved some of its attributions and studied in detail what are the conditions for strict stationarity in GARCH, one of the central results of this thesis. As we have stressed in many occasions, stationarity is key to most properties in Time Series and allows for more fruitful results, since constant mean and variance are a huge gain for our analysis.

It was one of our main interests to provide a strong background for Predicting or Forecasting Time Series. The extension of basic probability theory into the solid structure provided by Functional Analysis and Hilbert spaces allowed for a thorough description of prediction and how it is inextricably intertwined with Time Series. It is worth mentioning that such a presentation was possible thanks to the teachings of the subject *Anàlisi real i funcional*. The second part of this chapter was focused on Estimation of parameters, where we presented the Maximum Likelihood Estimation for ARCH and GARCH models.

Finally, in Chapter 4, we aimed for a hands-on approach, meaning that the theoretical results obtained in Time Series can be interesting and elegant, but there is a call for a practical presentation of these results. The analysis of an example of stock returns allowed us to see many of the lessons of the previous chapter put into practice. It is clear that the study could have been a bit more thorough, however, limitations imposed by time and scarce knowledge of advanced R have made it too much of a challenge. The intention of the last section was to introduce some of the new challenges posed by the deficiencies of GARCH models, a new realm of possibilities which has become the object of attention in current investigations.

6 Appendix

R code

1. Code for ARCH(1,1) and GARCH(1,1)

```
#Simulated ARCH(1) model with n=300
alpha0 <-0.25
alpha1 <-0.5
y <-numeric(300)
y[1]<-rnorm(1)
for(i in 2:300)
{
  epsilon <-rnorm(1)
  sigmasq <-alpha0 + alpha1*(y[i-1]*y[i-1])
  y[i] <-epsilon*sqrt(sigmasq)

}

#Simulated GARCH(1,1) model with n=300
alpha0 <-0.5
alpha1 <-0.2
beta1 <- 0.7
y <-numeric(300)
sigma <-numeric(300)
y[1]<-rnorm(1)
sigma[1]<-sqrt( alpha0 / (1-alpha1-beta1)) # we initialize it to the uncond. v
control <-numeric(300)
for(i in 2:300)
{
  epsilon <-rnorm(1)
  sigma[i] <-alpha0 + alpha1*(y[i-1]*y[i-1]) + beta1*(sigma[i-1])
  y[i] <-epsilon*sqrt(sigma[i])

}
```

2. Code for Example?? We used libraries `ggplot2`, `ggfortify` for plotting and `astsa` for the ARIMA modelling function

```
GNP <-read.csv("GNP.csv")
GNPts <-ts(GNP[,2], frequency=4,
        start=c(1947,1), end=c(2002,3)) #transform data frame -> time series
GNPret <-diff(log(GNPts)) #diff calculates difference Yt - Yt-1
plot(GNPret)
GNPretdt <-data.frame(GNPret) #for function ggplot we need a data frame
GNPretdt$GNPret
autoplot(GNPret)
ggplot(data=GNPretdt, aes(x=time(GNPret), y=GNPret))+ geom_point()
        + geom_line()
        + geom_hline(yintercept=mean(GNPret), color="red", size=1)
```

```
    + labs(x="Time", y="Log Return")
par(mfrow=c(1,2))
acf(GNPret, lag.max=50, main="")
pacf(GNPret, lag.max=50, main="", ylim=c(-0.4, 0.4))
sarima(GNPret, 0,0,2) #Fits an MA(2) model and prints diagnostics
```

3. Main code for Chapter 4

```
#SP500 returns analysis
SP500short<-read.csv("SP500.csv") #data
SP500<-read.csv("SP500long.csv") #data from Yahoo Finance
na.omit(as.numeric(SP500[,2])) #we eliminate blank returns
complete.cases(SP500) #shows if all lines are complete
#return = closing - opening / opening
ret <-(SP500[,5]-SP500[,2]) / (SP500[,2])
retshort <-(SP500short[,5]-SP500short[,2]) / (SP500short[,2])
which.max(ret) ; which.min(ret)
max(ret); min(ret); mean(ret); median(ret)
library(moments)
sd(ret); kurtosis(ret); skewness(ret)
which.max(retshort) ; which.min(retshort)
max(retshort); min(retshort); mean(retshort); median(retshort)
library(moments)
sd(retshort); kurtosis(retshort); skewness(retshort)
kurtosis(retshort)
#time series date format
dateret <- as.Date(as.character(SP500[,1]),format="%Y-%m-%d")
dateretshort <- as.Date(as.character(SP500short[,1]),format="%Y-%m-%d")

Sys.setlocale(category = "LC_ALL", locale = "english")
#plots SP500index evolution
png(filename="SP500_0.png", res=300, width = 2400, height = 1500)
plot(dateret, SP500[,5], xlab="Date", ylab="Closing price"
    , type="l")
while (!is.null(dev.list())) dev.off()

png(filename="SP500short_0.png", res=300, width = 2400, height = 1500)
plot(dateretshort, SP500short[, 5], xlab="Date", ylab="Closing price",
    type="l")
while (!is.null(dev.list())) dev.off()

#plots return
png(filename="SP500_1.png", res=300, width = 2400, height = 1500)
plot(dateret,ret, xlab = "Date", ylab="Return", type="l")
dev.off()

png(filename="SP500short_1.png", res=300, width = 2400, height = 1500)
plot(dateretshort, retshort, xlab="Date", ylab="Closing price", type="l")
while (!is.null(dev.list())) dev.off()
```

```
#return histogram
png(filename="SP500_2.png",  res=300, width = 2600, height = 1500)
par(mfrow=c(1,2))
hist(ret, breaks=20, xlim = c(-0.04, 0.03), xlab = "", ylab="",
     main = "Distribution of S&P 500 returns 2010-2021")
hist(retshort, breaks=20, xlim = c(-0.04, 0.03), xlab = "", ylab="",
     main =  "Distribution of S&P 500 returns 2020-2021")
dev.off()

#ACF and PACF for returns and absolute returns (long sample)
retabs <- abs(ret)
png(filename="SP500_3.png",  res=300, width = 2400, height = 1500)
par(mfrow=c(1,2))
acf(ret, lag.max=50, main="ACF Yt", ylim=c(-0.4, 0.4))
acf(retabs, lag.max=50, main="ACF |Yt|", ylim=c(-0.4, 0.4))
dev.off()

#ACF and PACF for returns and absolute returns (short sample)
retshortabs <- abs(retshort)
png(filename="SP500_3short.png",  res=300, width = 2400, height = 1500)
par(mfrow=c(1,2))
acf(retshort, lag.max=50, main="ACF Yt", ylim=c(-0.4, 0.4))
acf(retshortabs, lag.max=50, main="ACF |Yt|", ylim=c(-0.4, 0.4))
dev.off()

#ACF^2
retsquared<-ret*ret
acf(retsquared, lag.max=50, main="ACF Yt^2")
which.max(retsquared) ; which.min(retsquared)
max(retsquared); min(retsquared); mean(retsquared); median(retsquared)
sd(retsquared); kurtosis(retsquared); skewness(retsquared)

logret<-diff(log(SP500[,5]))
which.max(logret) ; which.min(logret)
max(logret); min(logret); mean(logret); median(logret)
sd(logret); kurtosis(logret); skewness(logret)

#ACF^2 plotting
retsquared<-ret*ret
retsquaredshort<-retshort*retshort

png(filename="SP500_6.png",  res=300, width = 2400, height = 1500)
par(mfrow=c(1,2))
acf(retsquared, lag.max=50, main="ACF Yt^2", ylab="ACF Long Sample")
acf(retsquaredshort, lag.max=50, main="ACF Yt^2", ylab="ACF Short Sample")
dev.off()


#Trying to fit linear regression y = b0 + b1*t + error
```

```
fitreg<-lm(ret~dateret)
fitSP500<-lm(SP500[, 5]~dateret)
summary(fitSP500)
plot(fitSP500)
summary(fitreg)

png(filename="SP500_4.png", res=300, width = 2400, height = 1500)
plot(dateret,ret, xlab = "Date", ylab="Return", type="l")
abline(lm(ret ~ dateret), col="blue")
dev.off()

plot(dateret, SP500[,5], xlab = "Date", ylab="Return", type="l")
abline(lm(SP500[, 5] ~ dateret))

#Testing for ARCH effects
library(FinTS)
SP500.archTest1 <- ArchTest(ret, lags=1, demean=TRUE)
SP500.archTest1
SP500.archTest2 <- ArchTest(ret, lags=2, demean=TRUE)
SP500.archTest2
SP500.archTest3 <- ArchTest(ret, lags = 3, demean = TRUE)
SP500.archTest3

library(FinTS)
SP500.archTest1 <- ArchTest(retshort, lags=1, demean=TRUE)
SP500.archTest1
SP500.archTest2 <- ArchTest(retshort, lags=2, demean=TRUE)
SP500.archTest2
SP500.archTest3 <- ArchTest(retshort, lags = 3, demean = TRUE)
SP500.archTest3

#Fitting an ARCH(2)
library(fGarch)
fita2= garchFit(~garch(2, 0), data=ret)
summary(fita2)
fita2short = garchFit(~garch(2, 0), data=retshort)
summary(fita2short)

#Fitting a GARCH(1,1)
library(fGarch)
fitg = garchFit(~garch(1, 1), data=ret)
summary(fitg)

quest <-coef(fitg)[1] + residuals(fitg) #gives us the original ret


#Predictions volatility GARCH(1,1)
# We create two cond var vectors
condvarGARCH1<-numeric(length(ret) + 5)
```

```
condvarGARCH1true<-numeric(length(ret) + 5)
sigma<-coef(fitg)[2] /(1-coef(fitg)[3]-coef(fitg)[4]) #uncond var.
condvarGARCH1[1] <-sigma
condvarGARCH1true[1] <-sigma

for(i in 2:length(ret))
{
  condvarGARCH1[i] <-coef(fitg)[2] + coef(fitg)[3]*retsquared[i-1]
   + coef(fitg)[4]*condvarGARCH1[i]
  condvarGARCH1true[i]<-coef(fitg)[2] + coef(fitg)[3]*retsquared[i-1]
  + coef(fitg)[4]*condvarGARCH1true[i]
}

#Prediction for next 5 values
diffsigma <- condvarGARCH1[length(ret)] - sigma
for(i in 1:5){
  sum1 <- (coef(fitg)[3] + coef(fitg)[4])
  l<-1
  while(l<i){
    print(l)
    sum1<- sum1*(coef(fitg)[3] + coef(fitg)[4])
    l<- (l+1)
  }
  condvarGARCH1[length(ret)+i] <- sigma + sum1*diffsigma
}

SP500ahead<-read.csv("SP500ahead2.csv") #data from Yahoo Finance
retahead <- (SP500ahead[, 5]- SP500ahead[, 2]) / SP500ahead[, 2]
dateretahead <- as.Date(as.character(SP500ahead[,1]),format="%Y-%m-%d")
retaheadsq <-retahead*retahead
lengthret<-length(ret)
for(i in (lengthret+1):(lengthret+5)){
  print(i)
  condvarGARCH1true[i] <- coef(fitg)[2]
        + coef(fitg)[3]*retaheadsq[-lengthret+i]
        + coef(fitg)[4]*condvarGARCH1true[i-1]
}

volatilitydates<-c(dateret[2756:2770], dateretahead[1:5])

Sys.setlocale(category = "LC_ALL", locale = "english")
png(filename="SP500_5.png",  res=300, width = 2400, height = 1500)
plot(volatilitydates, tail(sqrt(condvarGARCH1), 20), type="l" ,
        ylab="sigma_t", xlab="Date", col="blue")
lines(volatilitydates,tail(sqrt(condvarGARCH1true), 20),
        type="l", col="black")
dev.off()
```

References

[1] Shumway, Robert H. & Stoffer, David S. *Time Series and Its Applications with R examples.* New York: Springer, 2017 . Springer Texts in Statistics, ISBN: 978-3-319-52452-8

[2] Vives Santa-Eulalia, Josep. *Time Series Analysis. Unpublished notes*

[3] Cont, Rama. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance.* United Kingdom, IOP Publishing, October 2020, Volume 1, pp. 223-236.

[4] Brockwell, Peter J. & Davis, Richard A. *Time Series: Theory and Methods* 2nd edition. New York: Springer, 1991. Springer Text in Statistics, ISBN: 3-540-97429-6.

[5] Franke, Jürgen, et al. *Statistics of Financial Markets. An introduction* 3rd edition. Berlin: Springer-Verlag, 2011. Universitext, ISBN: 978-3-642-16520-7

[6] Francq, Christian & Zakoian, Jean-Michel. *GARCH models : structure, statistical inference and financial applications.* Second Edition. USA: John Wiley & Sons, 2019. ISBN: 978-1-119-31356-4

[7] Nelson, Daniel B. Stationarity and Persistence in the GARCH(1,1) Model. *Econometric Theory* [online] United Kingdom: Cambridge University Press, September 1990, Volume 6, Issue 3, pp. 318-334. `https://www.jstor.org/stable/3532198?seq=1`

[8] Chen, J., Politris, D.N. Optimal Multi-Step-Ahead Prediction of ARCH/-GARCH Models and NoVaS Transformation. *Econometrics* MDPI AG, August 2019, Volume 7. `https://www.researchgate.net/publication/335077533_Optimal_Multi-Step-Ahead_Prediction_of_ARCHGARCH_Models_and_NoVaS_Transformation`

[9] Marquez-Carreras, David & Julià de Ferran, Olga. *Un primer curs d'Estadística.* Barcelona: Publicacions i Edicions de la Universitat de Barcelona, 2011. Col·lecció: Universitat, ISBN: 974-8-447-53483-8.

[10] Sanz i Solé, Marta. *Probabilitats.* Barcelona: Edicions de la Universita de Barcelona, 1999. Col·lecció: UB, ISBN: 84-8338-091-9

[11] Engle, Robert F. Autoregressive Conditional heteroskedasticity with Estimates of the Variance of United Kindgom Inflation. *Econometrica* [online] United States: The Econometric Society, July 1982, Volume 50, Issue 4, p. 987- 1008 `http://www.econ.uiuc.edu/~econ536/Papers/engle82.pdf`

[12] Bollerslev, Tim. Generalized Autoregressive Connditional heteroskedasticity. *Journal of Econometrics.* Netherlands: Elsevier, February 1986, Volume 31, p.307-327. `https://public.econ.duke.edu/~boller/Published_Papers/joe_86.pdf`

[13] Chow, Yuan Shih & Teicher, Henry. *Probability theory. Independence, Interchangeability, Martingales.* New York: Springer, 1997. ISBN: 978-1-4612-1950-7