

# PROJECT

## Financial Econometrics

Louis Monier

Master 2 : Data Science

Manuel Griseri

Master 2 : Quantitative Finance

Academic year 2025/2026

# Contents

<b>1 Test the effect of pandemic on the volatility in different countries</b>	<b>2</b>
1.1 Exploratory data analysis . . . . .	2
1.2 Correcting for autocorrelation in volatility tests . . . . .	4
<b>2 Volatility for daily time series</b>	<b>7</b>
2.1 Exploratory data analysis . . . . .	7
2.2 Variation of GARCH models and the leverage effect . . . . .	8
<b>3 High-frequency time series modelling</b>	<b>11</b>
3.1 Exploratory data analysis . . . . .	11
3.2 Functional data analysis for intraday cumulative log returns . . . . .	12
<b>A Methodological appendix and additional results</b>	<b>16</b>
A.1 CLT and Slutsky's theorem . . . . .	16
A.2 Lag truncation parameter selection . . . . .	17
A.3 Autocorrelation and Ljung-Box test . . . . .	18
A.4 Stationarity and Augmented Dickey-Fuller test . . . . .	19
A.5 Log-likelihood and Exponential-GARCH . . . . .	20
A.6 Brownian motion and Jarque-Bera test . . . . .	21
<b>References</b>	<b>23</b>

# Chapter 1

## Test the effect of pandemic on the volatility in different countries

In this first chapter, we conduct an analysis to quantitatively assess the difference in volatility between the pre-COVID-19 period (2015–2019) and the post-COVID-19 period (2020–2024) for France and the United States, as well as the difference between the two countries themselves. As datasets, we use the main stock market indices of each country, namely the CAC 40 for France and the S&P 500 for the United States.

Section 1.1 presents an exploratory data analysis to provide a stronger intuition about the results we may expect from the statistical tests. Subsequently, section 1.2 demonstrates how to correctly perform a t-test with autocorrelation correction, as it is necessary in this case.

### 1.1 Exploratory data analysis

Intuitively, we expect the volatility in the post-COVID period to be higher than in the pre-COVID period, for several reasons. International political events and macroeconomic developments have made the world considerably riskier after 2020 : not only due to the COVID-19 pandemic, but also because of the war in Ukraine since February 24, 2022, the conflict in the Middle East since October 7, 2023, and, on a macroeconomic level, the surge in inflation, the consequent monetary policies implemented by central banks, and the growing concerns over public debt sustainability in both the United States and France. Let us now see whether this intuition is reflected in the data.

In Figure 1.1, the empirical distributions of the S&P 500 returns for the pre- and post-COVID periods are plotted together (similar shape for CAC 40). We can clearly see that the empirical distribution of the post-COVID period exhibits heavier tails and higher frequencies of extreme events compared to the pre-COVID period. By plotting in Figure 1.3 the two time series overlaid, we gain a more precise idea of the relationship between the volatilities of the two countries. The two series are clearly correlated, reflecting the strong economic ties between the United States and France. However, it appears that during the pre-COVID period, the CAC 40 was more volatile than the S&P 500. This may be due to factors such as the greater robustness of an index composed of 500 stocks compared to one of only 40, as well as to geopolitical events, for instance, it is evident that Brexit (June 2016) had a much stronger impact on France. Everything that is observed graphically can also be seen numerically when computing the annualized volatilities, which are reported below.

This preliminary analysis helps us better understand the behavior of the data, but it does not yet provide any quantitative tool to formally assess whether the differences in volatility are statistically significant. We therefore move on to the inferential analysis.

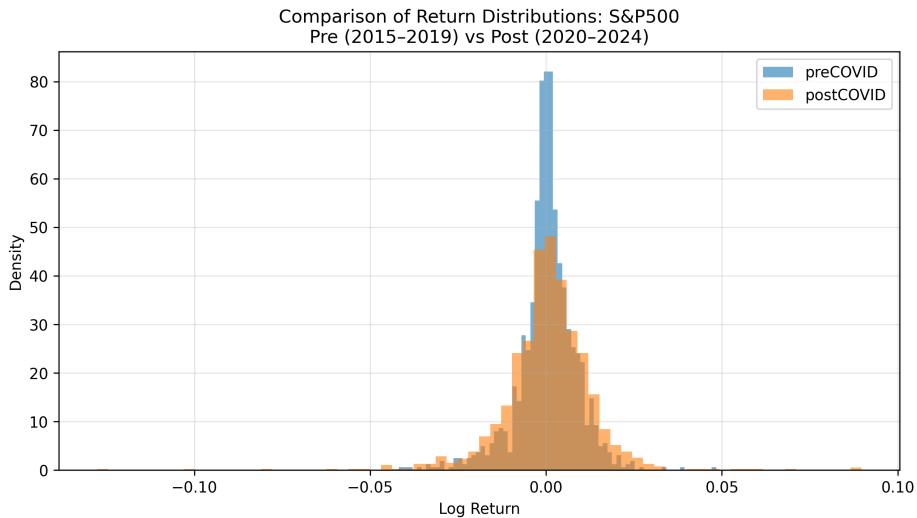


Figure 1.1: Comparison pre- vs post-COVID empirical distributions (S&P 500)

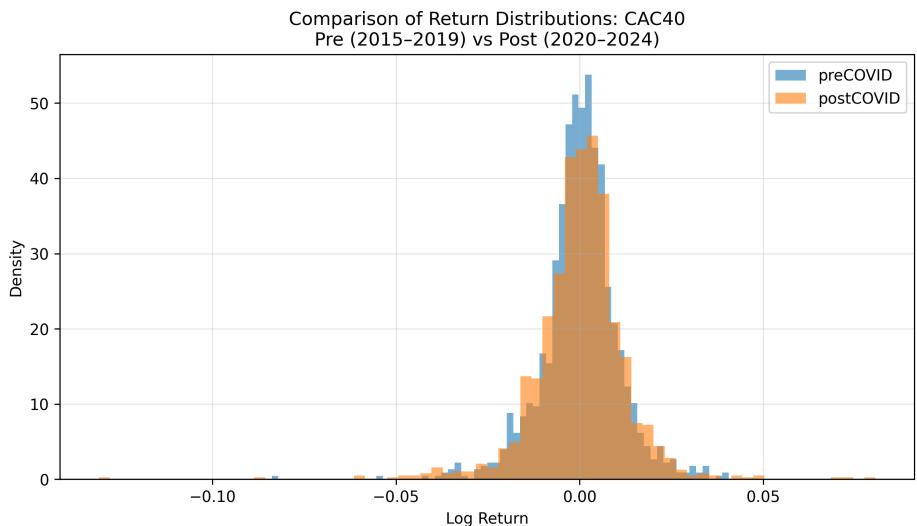


Figure 1.2: Comparison pre- vs post-COVID empirical distributions (CAC 40)

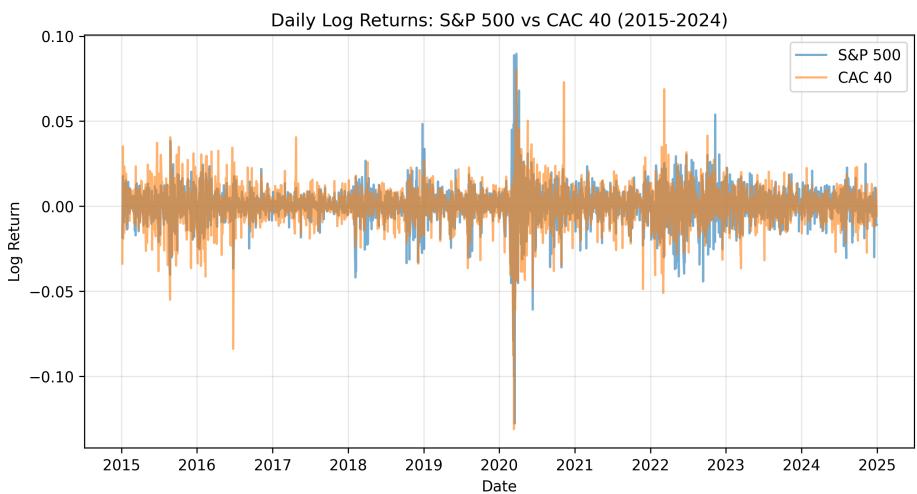


Figure 1.3: Daily log returns

```

===== S&P500: Annualized Volatility =====
preCOVID: 0.1346
postCOVID: 0.2143

===== CAC40: Annualized Volatility =====
preCOVID: 0.1685
postCOVID: 0.2057

```

Figure 1.4: Annualized volatility

## 1.2 Correcting for autocorrelation in volatility tests

First of all, let us formally state the null and alternative hypotheses for the first test we want to perform, namely, the test of difference in volatility between the pre- and post-COVID periods, first for the United States and then for France. Let  $r_t$  denote the log-return at date  $t$ , Pre the set of dates from 2015 to 2019, and Post the set of dates from 2020 to 2024. Since log-returns have zero mean, the expected value of their square corresponds to their variance. Hence, the hypotheses can be written as follows :

$$H_0 : \mathbb{E}[r_t^2 | t \in \text{Post}] = \mathbb{E}[r_t^2 | t \in \text{Pre}] \quad \text{vs} \quad H_1 : \mathbb{E}[r_t^2 | t \in \text{Post}] \neq \mathbb{E}[r_t^2 | t \in \text{Pre}] \quad (1.1)$$

We assume these conditional expectations don't depend on  $t$ . Clearly, testing whether the two expected values are equal is equivalent to testing whether their difference is equal to zero. As an estimator of the expected value, we use the sample mean, defined for instance, for  $\mathbb{E}[r_t^2 | t \in \text{Post}]$ , as :

$$\overline{r^2}_{\text{Post}} = \frac{1}{|\text{Post}|} \sum_{t \in \text{Post}} r_t^2 \quad (1.2)$$

To obtain a rejection criterion for  $H_0$ , we need to know the sampling distribution of this estimator. One might consider applying the Central Limit Theorem (CLT) and performing an asymptotic t-test, but this would be a mistake, since the assumption of independence among sample observations is simply false for time series data, there is almost always autocorrelation (see Appendix A.1 for the wrong results). Therefore, we turn to a HAC test, i.e. a t-test corrected for autocorrelation. We will rely on an asymptotic theorem that extends the CLT to a sequence of correlated random variables.

If  $(X_t)_{t \in \mathbb{Z}}$  is weakly stationary with mean  $\mu$ , autocorrelation function  $\rho(\cdot)$  and autocovariance function  $\gamma(\cdot)$ ,  $\sum_h |\rho(h)| < \infty$  and some mixing condition, then as  $n \rightarrow \infty$  :

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nu) \quad (1.3)$$

where  $\nu$  is the long-run variance :

$$\nu = \sum_{j=-\infty}^{\infty} \gamma(j) = \gamma(0) + 2 \sum_{j=1}^{\infty} \gamma(j) \quad (1.4)$$

An estimator of it is the weighted sum of sample covariances

$$\widehat{\nu} = \sum_{|j| \leq q} w\left(\frac{j}{q}\right) \widehat{\gamma}_n(j) \quad (1.5)$$

where  $w$  is a symmetric weighting (kernel) function with  $w(0) = 1$ , and  $q$  is the bandwidth satisfying  $q \rightarrow \infty$  and  $q/n \rightarrow 0$  as  $n \rightarrow \infty$  (further details in Appendix A.2). An example of kernel function is the Bartlett kernel :  $w(x) = (1 - |x|)\mathbf{1}_{|x|<1}$ .

Thus, the standard error (SE) for the difference in means (also called Newey-West SE) is

$$\text{SE} = \sqrt{\frac{\hat{\nu}_{\text{Post}}}{|\text{Post}|} + \frac{\hat{\nu}_{\text{Pre}}}{|\text{Pre}|}} \quad (1.6)$$

where  $\hat{\nu}_{\text{Post}}$  and  $\hat{\nu}_{\text{Pre}}$  denote the estimators which converge in probability to the long-run variances of the series ( $r_t^2$ ) over the post-COVID and pre-COVID samples, respectively. Consequently, by Slutsky's theorem, under the null hypothesis  $H_0$ , the asymptotic distribution of the test statistic  $T_1$  is standard normal :

$$T_1 = \frac{\bar{r}_{\text{Post}}^2 - \bar{r}_{\text{Pre}}^2}{\text{SE}} \xrightarrow[n \rightarrow \infty]{\mathcal{L} \mid H_0} \mathcal{N}(0, 1) \quad (1.7)$$

Let  $t_1$  denote the empirical realization of the statistic computed from the sample. The corresponding  $p$ -value is then given by

$$p\text{-value} = \mathbb{P}(|T_1| > t_1 \mid H_0 \text{ true}) = 2[1 - \Phi(t_1)] \quad (1.8)$$

where  $\Phi(\cdot)$  denotes the cumulative distribution function (CDF) of a standard normal variable.

Regarding the test of equality in volatility between the two countries (first for the pre-COVID period and then for the post-COVID period), the hypotheses are :

$$H_0 : \mathbb{E}^{[\text{SP}]} r_t^2 = \mathbb{E}^{[\text{CAC}]} r_t^2 \quad \text{vs} \quad H_1 : \mathbb{E}^{[\text{SP}]} r_t^2 \neq \mathbb{E}^{[\text{CAC}]} r_t^2 \quad (1.9)$$

The reasoning is analogous to the previous test, but simpler. By linearity of the expectation operator, we consider the stochastic sequence  $Y_t = \text{SP}r_t^2 - \text{CAC}r_t^2$  and denote by  $\hat{\nu}$  its long-run variance estimator and by  $n$  the sample size. Then, the corresponding test statistic is given by

$$T_2 = \frac{\bar{Y}\sqrt{n}}{\sqrt{\hat{\nu}}} \xrightarrow[n \rightarrow \infty]{\mathcal{L} \mid H_0} \mathcal{N}(0, 1) \quad (1.10)$$

We now turn to our sample. Each series contains more than 1000 observations, so we can safely regard ourselves as operating in the asymptotic regime. We verify the presence of autocorrelation by plotting the ACFs of the S&P 500 and the CAC 40 in Figures 1.5 and 1.6, and by reporting the Ljung-Box test  $p$ -values in Figures 1.7 and 1.8 (details in Appendix A.3).

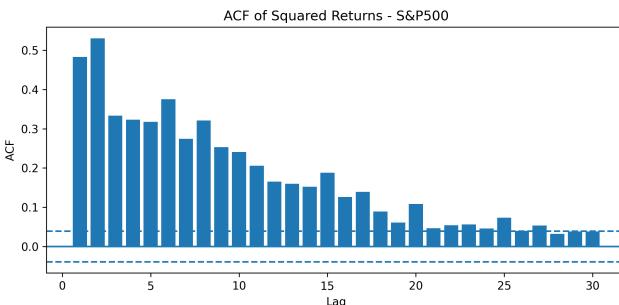


Figure 1.5: ACF (S&P 500)

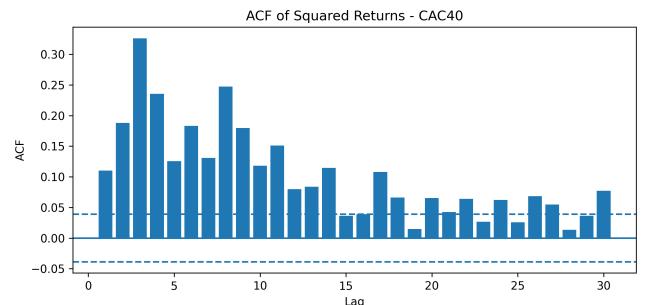


Figure 1.6: ACF (CAC 40)

```
==== S&P500: Ljung-Box on squared returns ====
      lb_stat      lb_pvalue
1    586.640906  1.347605e-129
5   2090.381214  0.000000e+00
10   3197.020264  0.000000e+00
20   3727.642852  0.000000e+00
```

Figure 1.7: Results LB test (S&P 500)

```
==== CAC40: Ljung-Box on squared returns ====
      lb_stat      lb_pvalue
1    31.034445  2.534897e-08
5   575.402064  4.169342e-122
10   978.655434  7.403176e-204
20  1163.734646  4.252488e-234
```

Figure 1.8: Results LB test (CAC 40)

Next, in order to apply the modified CLT, we require stationarity. We assess this with an Augmented Dickey-Fuller (ADF) test; the (favorable) result (rejecting the unit-root null) is shown in Figure 1.9 (details in Appendix A.4).

```
S&P500 preCOVID: ADF stat = -9.224, p-value = 1.7516832851600235e-15
S&P500 postCOVID: ADF stat = -5.944, p-value = 2.2304080287528021e-07
CAC40 preCOVID: ADF stat = -8.320, p-value = 3.586859809517114e-13
CAC40 postCOVID: ADF stat = -6.501, p-value = 1.161550962361604e-08
```

Figure 1.9: Results ADF test

Having satisfied these prerequisites, we can apply our asymptotic results and compute the *p*-values for the four tests of interest. The outcomes are summarized in Figure 1.10.

```
==== S&P500 - Pre vs Post COVID ====
Mean diff: -0.000110, SE: 0.000041
t (NW): -2.669, p-value: 0.0076 (n1=1257, n2=1257)

==== CAC40 - Pre vs Post COVID ====
Mean diff: -0.000055, SE: 0.000031
t (NW): -1.790, p-value: 0.0734 (n1=1276, n2=1282)

==== preCOVID - CAC40 vs S&P500 ====
Mean diff: 0.000044, SE: 0.000009
t (NW): 4.967, p-value: 0.0000 (n=1244)

==== postCOVID - CAC40 vs S&P500 ====
Mean diff: -0.000012, SE: 0.000018
t (NW): -0.685, p-value: 0.4934 (n=1247)
```

Figure 1.10: Results t-test

As anticipated from the exploratory analysis, for the S&P 500 the pre- vs post-COVID test is highly significant; for the CAC 40 it is somewhat less so, but in both cases there is sufficient empirical evidence to reject the null hypothesis of equal volatilities. Regarding the cross-country comparison, again as expected, there is a statistically significant difference in volatility in the pre-COVID period. However, we cannot reject the null of equal volatilities in the post-COVID period because the corresponding *p*-value is too large.

# Chapter 2

## Volatility for daily time series

This second chapter will focus on GARCH models, which are particularly useful for analyzing the volatility of financial time series. In particular, we will concentrate on a variation of the model called the Threshold GARCH (T-GARCH) model, which is more appropriate when there exists a difference in the behavior of volatility depending on whether the series exhibits losses or gains.

In section 1.1, we present a brief exploratory analysis of the new dataset, while in section 1.2 we perform a statistical test to determine whether there is an asymmetric effect in the data, that is, whether a standard GARCH model or a Threshold GARCH model provides a more suitable representation of the underlying volatility process.

### 2.1 Exploratory data analysis

In this analysis, we use as our dataset the log-return series of Tesla stock from 2015 to 2024. The time evolution of the series and its volatility can be seen in Figure 2.1.

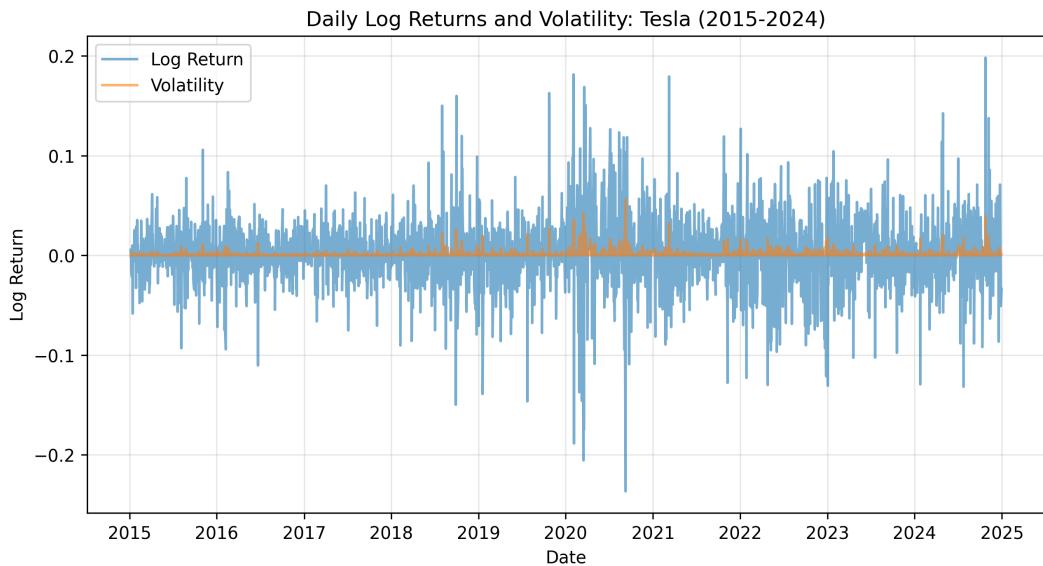


Figure 2.1: Daily log returns of Tesla stock

In general, although not always true, financial time series tend to exhibit higher volatility during periods of losses than during gains. This behavior reflects the attitude of investors, who typically react more strongly to negative shocks. To verify whether this intuition holds in our data, we computed the average of squared returns (a common proxy for volatility) conditional on

the sign of the previous log-return. Specifically, we compared the mean of squared returns when the previous return was negative with the mean when it was positive. The results, displayed in Figure 2.2, confirm our intuition : volatility following losses is indeed slightly higher than after gains, although the difference appears modest. In the next part, we will test whether this difference is statistically significant.

```
==== Descriptive Asymmetry Analysis: Tesla (2015–2024) ====
- The average squared log-return after a GAIN is 0.001259.
- The average squared log-return after a LOSS is 0.001325.
```

Figure 2.2: Asymmetric effect

## 2.2 Variation of GARCH models and the leverage effect

The Threshold-GARCH(1,1) model is defined as

$$\epsilon_t = \sigma_t Z_t, \quad (Z_t)_t \sim \text{i.i.d. } \mathcal{N}(0, 1) \quad (2.1)$$

$$\sigma_t^2 = \omega + \alpha_- \epsilon_{t-1}^2 \mathbf{1}_{\{\epsilon_{t-1} < 0\}} + \alpha_+ \epsilon_{t-1}^2 \mathbf{1}_{\{\epsilon_{t-1} > 0\}} + \beta \sigma_{t-1}^2 \quad (2.2)$$

We want to test the null hypothesis

$$H_0 : \alpha_- = \alpha_+ \quad \text{vs} \quad H_1 : \alpha_- \neq \alpha_+ \quad (2.3)$$

The test can be rewritten in vector form as follows. Let  $R = (0, 1, -1, 0)'$  and  $\theta_T = (\omega, \alpha_-, \alpha_+, \beta)'$ ,

$$H_0 : R' \theta_T = 0 \quad \text{vs} \quad H_1 : R' \theta_T \neq 0 \quad (2.4)$$

Now, let  $\Theta_T = \mathbb{R}_+^4$ ,  $\Theta_G = \{(a, b, c, d)' \in \mathbb{R}_+^4 : b = c\}$ ,  $l(\theta)$  denote the log-likelihood function, and  $\hat{\theta}_T$  be the MLE estimator of  $\theta_T$  under  $H_1$ ,  $\hat{\theta}_G$  be the MLE estimator of  $\theta_T$  under  $H_0$ , that is,

$$\hat{\theta}_T = \arg \max_{\theta \in \Theta_T} l(\theta), \quad \hat{\theta}_G = \arg \max_{\theta \in \Theta_G} l(\theta) \quad (2.5)$$

Using a Taylor approximation of  $l(\theta)$  around  $\theta_T$ , we have

$$l(\theta) \approx l(\theta_T) + \nabla l(\theta_T)'(\theta - \theta_T) + \frac{1}{2}(\theta - \theta_T)' H l(\theta_T)(\theta - \theta_T) \quad (2.6)$$

where  $H l(\theta_T)$  is the Hessian matrix of the log-likelihood. Since  $\nabla l(\theta_T) = 0$  and

$$H l(\theta_T) = \frac{\partial^2 l}{\partial \theta \partial \theta'}(\theta_T) = -n I(\theta_T) \quad (2.7)$$

where  $n$  is the sample size and  $I(\theta_T)$  is the Fisher information matrix, it follows that

$$l(\theta) \approx l(\theta_T) - \frac{1}{2}(\theta - \theta_T)' n I(\theta_T)(\theta - \theta_T) \quad (2.8)$$

This result allows us to simplify the maximization problem, since  $\max_{\theta \in \Theta_G} l(\theta)$  is equivalent to

$$\min_{\theta \in \Theta_G} (\theta - \hat{\theta}_T)' I(\theta - \hat{\theta}_T) \quad (2.9)$$

which is a constrained optimization problem. Indeed, it can be rewritten as

$$\min_{\theta \in \Theta_T} (\theta - \hat{\theta}_T)' I(\theta - \hat{\theta}_T) \quad \text{sub} \quad R'\theta = 0 \quad (2.10)$$

To solve this problem, we write the Lagrangian function as

$$\mathcal{L}(\theta, \lambda) = (\theta - \hat{\theta}_T)' I(\theta - \hat{\theta}_T) + 2\lambda R'\theta \quad (2.11)$$

The first-order conditions are both necessary and sufficient, since we are minimizing a convex function. Hence,

$$\frac{\partial \mathcal{L}}{\partial \theta} = 2I(\theta - \hat{\theta}_T) + 2R\lambda = 0$$

and

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 2R'\theta = 0$$

From the first equation, we obtain

$$\theta = \hat{\theta}_T - I^{-1}R\lambda$$

Substituting this expression into the second equation, we get

$$\lambda = \frac{R'\hat{\theta}_T}{R'I^{-1}R}$$

where both the numerator and denominator are scalars. Therefore,

$$\hat{\theta}_G = \hat{\theta}_T - I^{-1}R \frac{R'\hat{\theta}_T}{R'I^{-1}R} \quad (2.12)$$

Let  $l_T = l(\hat{\theta}_T)$ . Then, from equation (2.8) and using equation (2.12), we have

$$2(l_T - l_G) = n(\hat{\theta}_G - \hat{\theta}_T)' I(\hat{\theta}_G - \hat{\theta}_T) = n \frac{(R'\hat{\theta}_T)^2}{R'I^{-1}R} \quad (2.13)$$

From the asymptotic normality of the MLE estimator, it follows that

$$\sqrt{n} R'(\hat{\theta}_T - \theta_T) \xrightarrow{\mathcal{L}} \mathcal{N}(0, R'I^{-1}R), \quad \text{as } n \rightarrow \infty \quad (2.14)$$

Under the null hypothesis  $R'\theta_T = 0$ , this implies that

$$\frac{\sqrt{n} R' \hat{\theta}_T}{\sqrt{R'I^{-1}R}} \xrightarrow{\mathcal{L}|H_0} \mathcal{N}(0, 1) \quad (2.15)$$

In equation (2.13), we have shown that  $2(l_T - l_G)$  is the square of this quantity. Therefore,

$$2(l_T - l_G) \xrightarrow{\mathcal{L}|H_0} \chi^2(1) \quad (2.16)$$

This will be our test statistic LR (for Likelihood Ratio). From an interpretative point of view, the larger the difference between the two log-likelihoods, the more it indicates that the parameters of the T-GARCH model provide a better explanation of the volatility dynamics in the data compared to the standard GARCH model (so stronger evidence to reject  $H_0$ ). Denoting by  $LR_{\text{obs}}$  the observed value of the test statistic and by  $\chi_1^2(\cdot)$  the CDF of a  $\chi^2(1)$ , we have

$$p\text{-value} = \mathbb{P}(LR > LR_{\text{obs}} \mid H_0 \text{ true}) = 1 - \chi_1^2(LR_{\text{obs}}) \quad (2.17)$$

Returning to the Tesla stock data, we can observe the fitting results of a GARCH(1,1) model, a TGARCH(1,1) model, and the outcome of the LR test shown in Figure 2.3. The test is statistically significant, which allows us to confirm the presence of a leverage effect in the data, although in the opposite direction to what was initially suggested by the exploratory analysis (the higher parameter being  $\alpha^+$ ). In the Appendix A.5, an EGARCH(1,1) model, another variant of the GARCH framework that also accounts for asymmetric effects, is also fitted and tested. Therefore, modeling the volatility of the series with a TGARCH(1,1) model proves to be more accurate than with a simple GARCH(1,1), as significant asymmetric effects are present in the data. Figure 2.4 displays the series of percentage log-returns along with the volatility bands obtained from the fitted TGARCH(1,1) model.

```

==== GARCH(1,1) (constraint alpha_- = alpha_+) ====
{'omega': 2e-05, 'alpha': 0.051831, 'beta': 0.932962}
logLik: 4961.817, success=True, nit=14

==== T-GARCH(1,1) (two alphas) ====
{'omega': 1.1e-05, 'alpha_plus': 0.034162, 'alpha_minus': 0.033351, 'beta': 0.957136}
logLik: 4964.611, success=True, nit=71

==== LR test H0: alpha_- = alpha_+ ====
LR statistic: 5.586, df=1, p-value: 0.018103

```

Figure 2.3: Results LR test

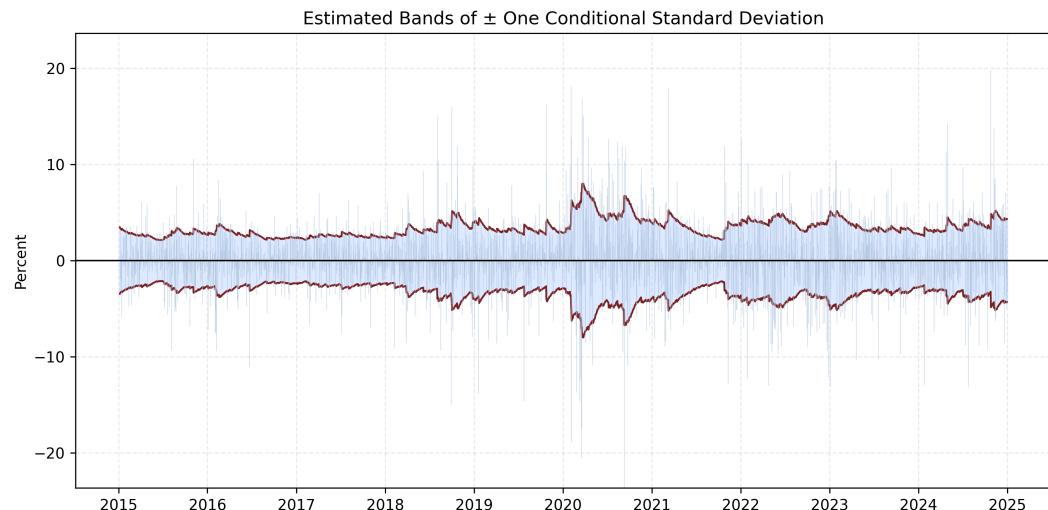


Figure 2.4: Bands of Tesla's stock

# Chapter 3

## High-frequency time series modelling

After having addressed how to handle problems related to volatility (tests in the first chapter and fitting in the second) on daily data, in this final chapter we will analyse the difficulties associated with high-frequency intraday data, with a focus on their prediction through functional data analysis.

The first section will, as usual, be devoted to presenting the data and evaluating whether we can obtain non-rigorous insights; the second will describe the various theoretical steps (together with the corresponding results) required to model and forecast high-frequency data.

### 3.1 Exploratory data analysis

Let us begin by explaining the idea of functional modelling and then take a graphical look at the data. We denote by  $R_n(t_{nj})$  the log returns of the asset on day  $n$  at the  $j$ -th tick time  $t_{nj}$ . The index  $n$  ranges from 1 to  $N$ , and  $j$  from 1 to  $J_n$  (the number of ticks on day  $n$ ). The object of our modelling, however, will be the cumulative log return, defined as

$$Y_n(t_{nj}) = R_n(t_{nj}) - R_n(t_{n1}) \quad (3.1)$$

Several trajectories, one for each day, can be observed on the left-hand side of Figure 3.1. On the right-hand side, we display the realization of the empirical mean  $\hat{\mu}(t_j)$ , obtained by averaging across all days the intraday observations at each tick  $j$ . We can see that it is very close to zero; however, we would like to know if there is additional information conveyed by volatility, which could be highly valuable for prediction.

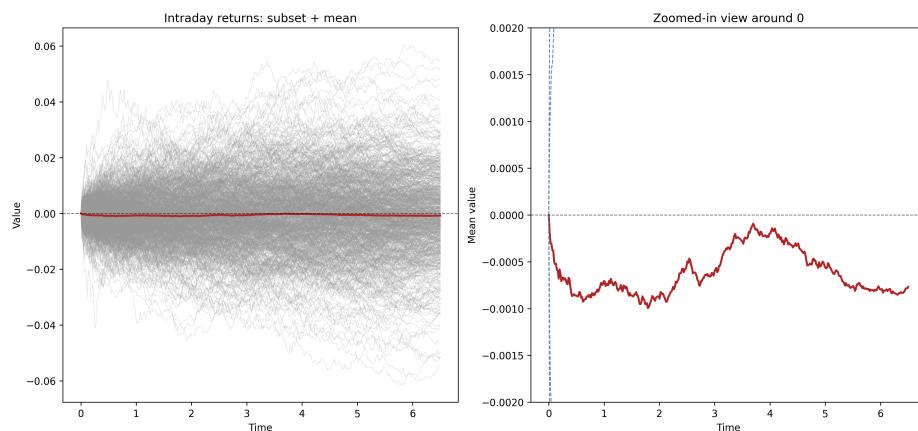


Figure 3.1: Intraday cumulative log returns trajectories and focus on the mean

### 3.2 Functional data analysis for intraday cumulative log returns

So, under the functional data framework, the data can be expressed as

$$Y_n(t_{nj}) = X_n(t_{nj}) + \varepsilon_{nj} \quad (3.2)$$

where  $\varepsilon_{nj} \sim (0, \sigma^2)$  and

$$X_n(t) = \mu(t) + \sum_{k=1}^p \xi_{nk} \psi_k(t) \quad (3.3)$$

In this setting,  $(X_n)$  is a discrete-time process (day 1, 2, ...), but with the particular feature that each random variable of the sequence takes values in a space of functional realizations. These realizations correspond to the time-dependent trajectories that the cumulative log-returns may follow on a given day (the mean  $\mu(t)$  is itself a function). Therefore, when making predictions, we will in fact forecast entire functions rather than individual scalar observations. The expression for  $X_n(t)$  derives from the Karhunen-Loëve expansion, truncated at order  $p$ , which will be selected through Functional Principal Component Analysis (FPCA) as the number of eigenfunctions  $\psi_k$  explaining 95% of the total variance. The scores are defined as  $\xi_{nk} = \langle X_n - \mu, \psi_k \rangle \sim (0, \lambda_k)$ , where  $\lambda_k$  are the eigenvalues. They do not depend on the time variable  $t$  but only on the day  $n$ , and thus determine how the trajectory changes from day to day. The eigenfunctions, in contrast, are the same for all days but vary with  $t$ ; they provide the dynamic component of the representation relative to the mean, capturing additional information when the variability represents not only noise but also signal.

The algorithm used to perform Functional Principal Component Analysis (FPCA) is based on the empirical approximation of the covariance operator in the space of discretized trajectories. As always,  $Y_n(t)$  denote the cumulative log-return for day  $n$  on a regular grid  $t_1, \dots, t_m \in [0, 1]$ . The observations available for FPCA are therefore the centered curves

$$M_n(t_j) = Y_n(t_j) - \hat{\mu}(t_j), \quad n = 1, \dots, N, \quad j = 1, \dots, m \quad (3.4)$$

Stacking these curves into a matrix  $M \in \mathbb{R}^{N \times m}$ , the empirical covariance operator in discretized form is approximated by the sample covariance matrix

$$\hat{C} = \frac{1}{N-1} M^\top M \quad (3.5)$$

which corresponds to a Riemann-sum approximation of the theoretical covariance operator on  $L^2[0, 1]$ . Since the discretization uses an equally spaced grid, the spacing is  $\Delta t = 1/(m-1)$  and is used for  $L^2$  normalisation. The spectral decomposition of  $\hat{C}$  provides eigenvalues and eigenvectors,

$$\hat{C} \psi_k = \lambda_k \psi_k, \quad k = 1, \dots, m \quad (3.6)$$

where  $\psi_k$  is the discretized estimate of the  $k$ -th eigenfunction of the covariance operator, and  $\lambda_k$  is the corresponding eigenvalue. After the decomposition, each eigenvector is normalised so that it satisfies the discrete analogue of the  $L^2$  constraint :

$$\sum_{j=1}^m \psi_k(t_j)^2 \Delta t = 1 \quad (3.7)$$

Ordering the eigenvalues in decreasing order  $\lambda_1 \geq \lambda_2 \geq \dots$ , the proportion of variance explained by the first  $p$  components is

$$\text{EV}(p) = \frac{\sum_{k=1}^p \lambda_k}{\sum_{k=1}^m \lambda_k} \quad (3.8)$$

The value of  $p$  is selected so that  $\text{EV}(p)$  exceeds 95%, following the standard FPCA criterion. The figures below illustrate these results. On the left, we see that the 95% variance threshold is exceeded with  $p = 3$ . On the right, the first three eigenfunctions are plotted. Their interpretation is rather clear : the first represents the general trend, the second captures the contrast between the first and the second part of the day, and the third reflects the similarity in behaviour between the morning and the evening and the difference with the midday period.

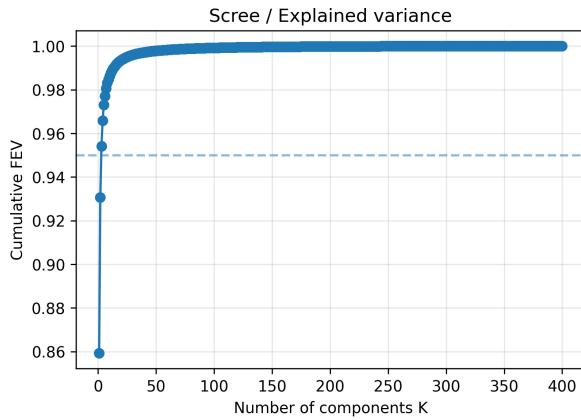


Figure 3.2: Cumulative explained variance

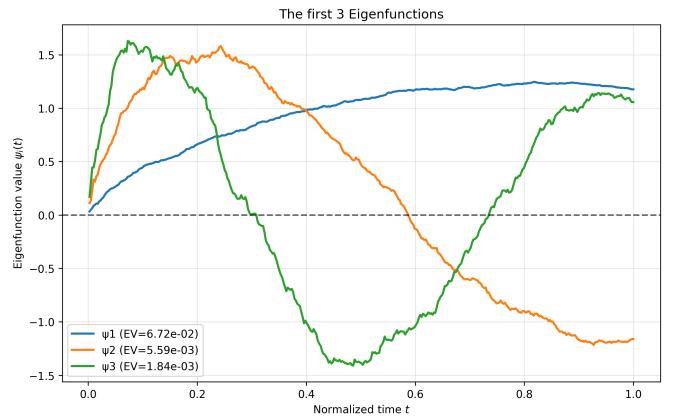


Figure 3.3: Eigenfunctions

In order to assess the quality of the fit, we use the Root Mean Squared Error (RMSE) as metric. Its definition is

$$\text{RMSE} = \left( \frac{1}{Nm} \sum_{n=1}^N \sum_{j=1}^m (Y_n(t_j) - X_n(t_j))^2 \right)^{1/2} \quad (3.9)$$

Figure 3.4 displays how the error behaves as we increase the number of eigenfunctions included in the model. The elbow method confirms that the choice  $p \approx 3$  is essentially optimal.

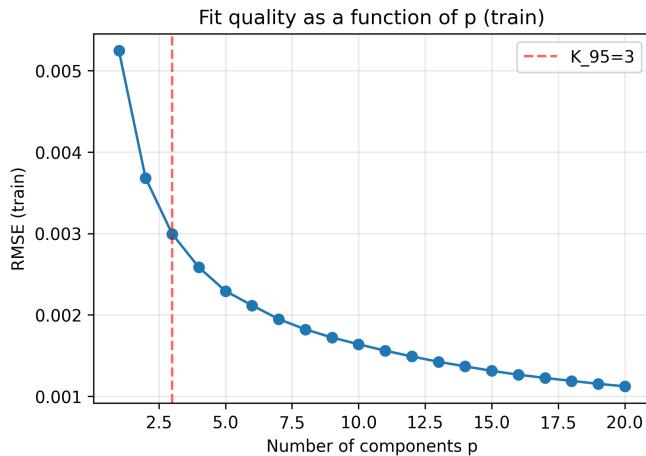


Figure 3.4: RMSE as a function of p

The results were obtained using the first year of data as the training set, while the second year will serve as the test set on which we will attempt to make predictions. However, in order to carry out prediction in a rigorous way, a preliminary analysis of the training results is required. First, we must verify that prediction is not hopeless : high-frequency data are often heavily affected by noise, which may prevent the extraction of any meaningful information. A standard way to check this is to compare the total variance of the  $X_n$ 's with the variance of the noise  $\sigma^2$ . The total variance of the signal is given by the trace of the empirical covariance matrix, i.e., by the sum of the eigenvalues, which in our case amounts to approximately 0.0782. The noise level, instead, can be estimated using the result

$$\frac{\widehat{RV}_j}{j} \xrightarrow{P} 2\sigma^2 \quad (3.10)$$

where  $\widehat{RV}_j$  is the realized volatility estimator

$$\widehat{RV}_j = \frac{1}{N} \sum_{n=1}^N \sum_{i \leq j} (Y_n(t_i) - Y_n(t_{i-1}))^2 \quad (3.11)$$

From this expression we obtain an estimate equal to 0.0012, which is smaller than the sum of the eigenvalues. This suggests that contamination by noise is negligible, making the extraction of informative structure from the data plausible.

Secondarily, we analyse the empirical distributions of the three scores. As a first step, it is useful to determine whether they behave as white noise series. From the Ljung-Box test results in Figure 3.5, we can conclude that the first two scores are indeed white noise, while the third one is not. However, as the scores may exhibit informative cross-correlations, a VAR(1) model was fitted. In addition, Figure 3.6 reports the results of a Jarque-Bera (JB) test, from which we may conclude that only the third score can be regarded as normally distributed. Its histogram and QQ-plot are shown in Figure 3.7. There is thus some residual evidence of normality, and given also the similarity between the estimated eigenfunctions and those of a Brownian motion, one may argue that modelling this score as a Wiener process could be appropriate (see Appendix A.6 for further details).

Jarque-Bera normality test on FPCA scores (per component)			
k	JB stat	p-value	Decision ( $\alpha=0.05$ )
1	6.0328	0.04898	Reject $H_0$ (non-normal)
2	25.3019	3.205e-06	Reject $H_0$ (non-normal)
3	1.0400	0.5945	Fail to reject $H_0$ (normal)

Figure 3.5: Results Ljung-Box test

Figure 3.6: Results Jarque-Bera test

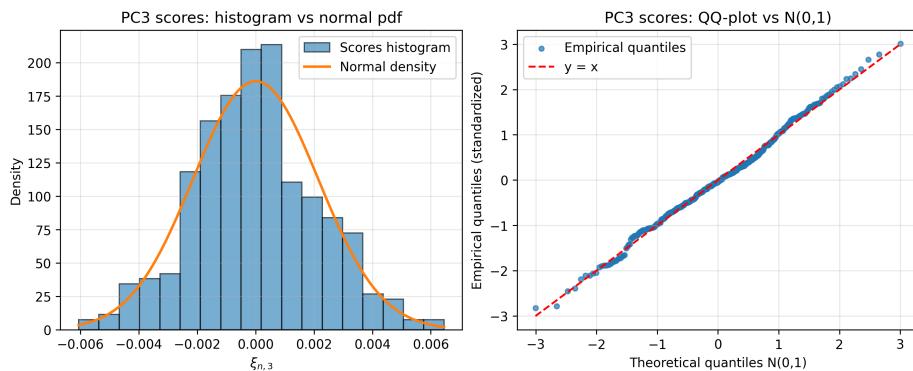


Figure 3.7: Histogram vs density + QQ-plot

We now have all the necessary components to proceed with prediction on the second year of data (the last day of the first year is denoted by  $N$ ). Our forecasting method is therefore

$$\widetilde{X}_{N+h|N}(t) = \mu(t) + \sum_{k=1}^3 \tilde{\xi}_{N+h,k} \psi_k(t) \quad (3.12)$$

$$\begin{pmatrix} \tilde{\xi}_{N+h,1} \\ \tilde{\xi}_{N+h,2} \\ \tilde{\xi}_{N+h,3} \end{pmatrix} = \begin{pmatrix} -5.64 \times 10^{-5} \\ 1.30 \times 10^{-5} \\ 6.17 \times 10^{-6} \end{pmatrix} + \begin{pmatrix} -0.0140 & 0.0427 & 0.2461 \\ 0.0079 & 0.0756 & 0.0164 \\ -0.0082 & -0.0136 & -0.0886 \end{pmatrix} \begin{pmatrix} \xi_{N+h-1,1} \\ \xi_{N+h-1,2} \\ \xi_{N+h-1,3} \end{pmatrix} \quad (3.13)$$

From Figure 3.8 we observe that the predicted trajectories are very close to the mean function, since all score values are extremely small. Indeed, Figure 3.10 shows that there is a small performance difference between forecasting with the FPCA-based model and simply using the empirical mean of the training set. Nevertheless, the slightly better behaviour of our method, together with the previous analysis, suggests, although not conclusively, that some latent signal in the data has been successfully captured.

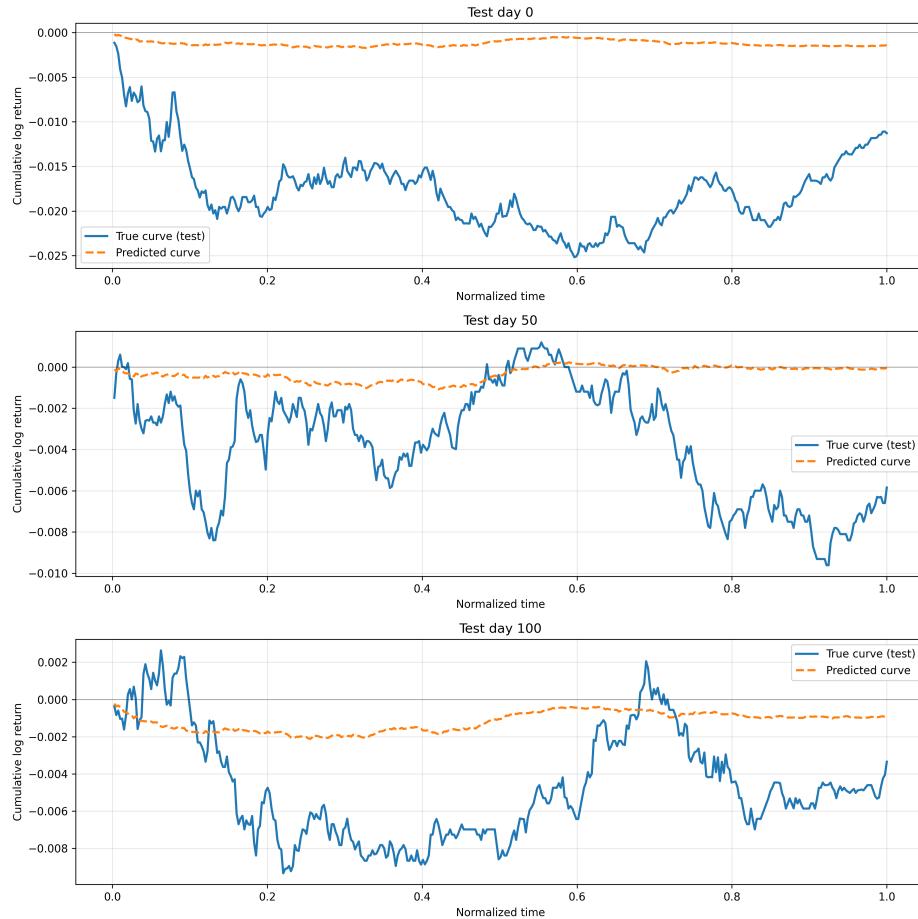


Figure 3.8: Predictions of cumulative log returns for some test days

Global prediction RMSE (full functions): 0.0109142  
Average RMSE per day: 0.0090179

RMSE baseline (zero curve): 0.0109227  
RMSE of the model (FPCA + VAR): 0.0109142  
Model / baseline ratio: 0.9992247

Figure 3.9: Performance on the test

Figure 3.10: Our model vs trivial baseline

# Appendix A

## Methodological appendix and additional results

### A.1 CLT and Slutsky's theorem

**Theorem A.1.1** (Central Limit Theorem). *Let  $X_1, X_2, \dots, X_n$  be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean  $\mu = \mathbb{E}[X_i]$  and finite variance  $\sigma^2 = \text{Var}(X_i) < \infty$ . Define the sample mean as*

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

*Then, as  $n \rightarrow \infty$ ,*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

*where  $\xrightarrow{\mathcal{L}}$  denotes convergence in law and  $\mathcal{N}(0, 1)$  is the standard normal distribution.*

The theorem states that the standardized sample mean converges in distribution to a standard normal variable, regardless of the underlying distribution of  $X_i$ , provided it has finite variance.

**Theorem A.1.2** (Slutsky's Theorem). *Let  $\{X_n\}$  and  $\{Y_n\}$  be sequences of random variables such that :*

$$X_n \xrightarrow{\mathcal{L}} X \quad \text{and} \quad Y_n \xrightarrow{P} c$$

*where  $\xrightarrow{P}$  denotes convergence in probability, and  $c$  is a constant. Then, as  $n \rightarrow \infty$  :*

1.  $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$ ,
2.  $X_n Y_n \xrightarrow{\mathcal{L}} cX$ ,
3. if  $c \neq 0$ , then  $\frac{X_n}{Y_n} \xrightarrow{\mathcal{L}} \frac{X}{c}$ .

Slutsky's theorem allows substitution of consistent estimators for unknown parameters in statistics that converge in distribution. For example, combining the Central Limit Theorem and Slutsky's theorem yields :

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \tag{A.1}$$

where  $\hat{\sigma}^2$  is a consistent estimator of  $\sigma^2$ .

If we mistakenly assumed that the log-returns in Chapter 1 were independent and naively applied the classical Central Limit Theorem for the tests (that is, using the standard variance instead of the long-run variance in the standard error), we would obtain the results shown in the Figure A.1.

```
==== S&P500 - Pre vs Post COVID ====
Mean diff: -0.000110, SE: 0.000021
t: -5.221, p-value: 0.0000 (n1=1257, n2=1257)

==== CAC40 - Pre vs Post COVID ====
Mean diff: -0.000055, SE: 0.000020
t: -2.725, p-value: 0.0064 (n1=1276, n2=1282)

==== preCOVID - CAC40 vs S&P500 ====
Mean diff: 0.000044, SE: 0.000008
t: 5.755, p-value: 0.0000 (n=1244)

==== postCOVID - CAC40 vs S&P500 ====
Mean diff: -0.000012, SE: 0.000017
t: -0.724, p-value: 0.4688 (n=1247)
```

Figure A.1: Results classic t-test

We can observe that, compared to the tests corrected for autocorrelation, there is an underestimation of the standard errors, which leads to too liberal tests with smaller  $p$ -values, although the final decisions of the four tests remain unchanged.

## A.2 Lag truncation parameter selection

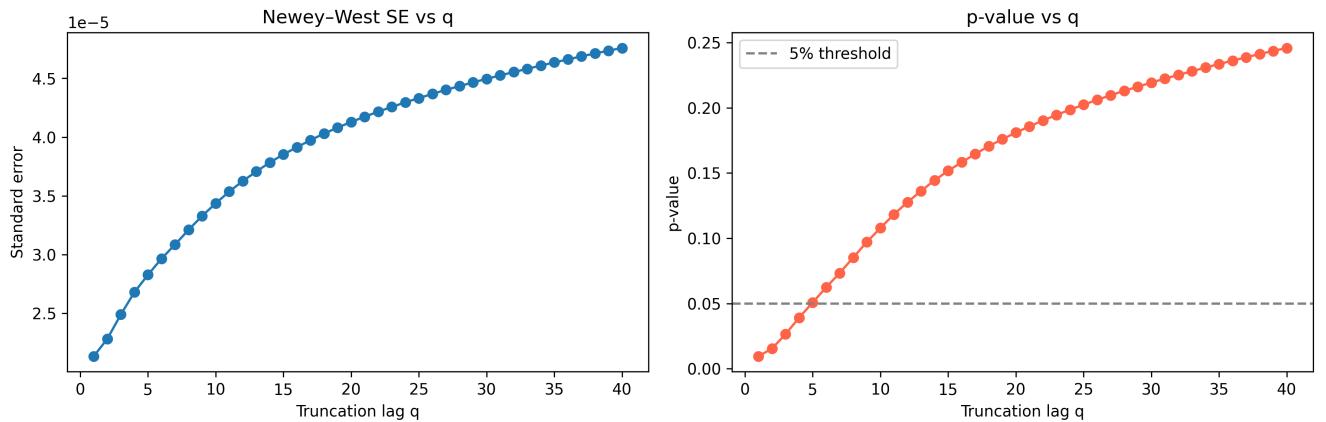


Figure A.2: NW SE and  $p$ -value vs  $q$

In the Figure A.2, we plot the evolution of the Newey-West standard error and the  $p$ -value (for the pre-post CAC 40 test in this specific case, although a similar pattern is observed for the other tests) as functions of the maximum lag truncation parameter  $q$ . We can observe that the relationship is increasing, and that there is a trade-off to be balanced when selecting the optimal value of  $q$ , since :

- If  $q$  is too small, there is a risk of underestimating the standard error, which makes the test too liberal, i.e. it leads to an inflated type I error probability (rejecting the null hypothesis even when it is true).
- If  $q$  is too large, there is a risk of overestimating the standard error, making the test too conservative, i.e. leading to an inflated type II error probability (failing to reject the null even when it is false).

In practice, simulations are often performed to tune optimal values for the truncation parameter. For example, when using Bartlett weights, a commonly used rule-of-thumb formula is (see page 11 of [3]) :

$$q = \lfloor 4(n/100)^{2/9} \rfloor \quad (\text{A.2})$$

### A.3 Autocorrelation and Ljung-Box test

Let  $\{X_t\}_{t=1}^n$  be a weakly stationary time series with mean  $\mu = \mathbb{E}[X_t]$ . The autocovariance function at lag  $k$  is defined as

$$\gamma(k) = \text{Cov}(X_t, X_{t-k}) = \mathbb{E}[(X_t - \mu)(X_{t-k} - \mu)] \quad (\text{A.3})$$

The corresponding autocorrelation function (ACF) is

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} \quad (\text{A.4})$$

where  $\gamma(0) = \text{Var}(X_t)$  is the variance of the process. In practice, the sample autocorrelation at lag  $k$  is estimated as

$$\hat{\rho}(k) = \frac{\sum_{t=k+1}^n (X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2} \quad (\text{A.5})$$

The Ljung-Box test is used to assess whether a time series exhibits significant autocorrelation up to a given lag  $K$ .

$$H_0 : \rho(1) = \rho(2) = \dots = \rho(K) = 0 \quad \text{vs} \quad H_1 : \rho(j) \neq 0 \text{ for some } 1 \leq j \leq K \quad (\text{A.6})$$

Under  $H_0$ , as  $n \rightarrow \infty$ ,

$$Q = n \sum_{k=1}^K \hat{\rho}^2(k) \xrightarrow{\mathcal{L}} \chi^2(K) \quad (\text{A.7})$$

where  $Q$  follows asymptotically a chi-square distribution with  $K$  degrees of freedom. Ljung and Box proposed a small-sample correction to improve the approximation :

$$Q_1 = n(n+2) \sum_{k=1}^K \frac{\hat{\rho}^2(k)}{n-k} \xrightarrow{\mathcal{L}} \chi^2(K) \quad (\text{A.8})$$

This adjusted statistic  $Q_1$  is typically used in applied work. The interpretation is the following : if the  $p$ -value associated with  $Q_1$  is small, we reject  $H_0$  and conclude that the residuals of the model exhibit significant autocorrelation up to lag  $K$ . Otherwise, we fail to reject  $H_0$ , indicating no evidence of serial correlation.

## A.4 Stationarity and Augmented Dickey-Fuller test

Let  $\{X_t\}_{t \in \mathbb{Z}}$  be a stochastic process.

**Definition A.4.1** (Strong Stationarity). *The process  $\{X_t\}$  is said to be strongly stationary if the joint distribution of  $(X_{t_1}, X_{t_2}, \dots, X_{t_k})$  is identical to that of  $(X_{t_1+h}, X_{t_2+h}, \dots, X_{t_k+h})$  for all integers  $t_1, \dots, t_k$  and any time shift  $h$ .*

$$F_{X_{t_1}, \dots, X_{t_k}}(x_1, \dots, x_k) = F_{X_{t_1+h}, \dots, X_{t_k+h}}(x_1, \dots, x_k) \quad \forall k, \forall h$$

In words, all joint distributions are invariant under time translation.

**Definition A.4.2** (Weak Stationarity). *The process  $\{X_t\}$  is said to be weakly stationary if the following conditions hold :*

1.  $\mathbb{E}[X_t] = \mu$  is constant and independent of  $t$ ;
2.  $\text{Var}(X_t) = \gamma(0) < \infty$ ;
3.  $\text{Cov}(X_t, X_{t+h}) = \gamma(h)$  depends only on the lag  $h$ , not on the specific time  $t$ .

Weak stationarity implies that the process has time-invariant mean, variance, and autocovariance structure.

The Augmented Dickey-Fuller (ADF) test is used to detect the presence of a unit root in a time series, i.e. to test whether a process is non-stationary. The test is based on the following regression model :

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{i=1}^p \delta_i \Delta X_{t-i} + \varepsilon_t \tag{A.9}$$

where:

- $\Delta X_t = X_t - X_{t-1}$  is the first difference;
- $\alpha$  is an intercept (optional);
- $\beta t$  is a deterministic time trend (optional);
- the lagged differences  $\Delta X_{t-i}$  account for autocorrelation in the residuals;
- $p$  is the number of augmenting lags.

The hypothesis are :

$$H_0 : \gamma = 0 \text{ (unit root, non-stationarity)} \quad \text{vs} \quad H_1 : \gamma < 0 \text{ (stationarity)} \tag{A.10}$$

The ADF test statistic is the  $t$ -statistic associated with  $\hat{\gamma}$  :

$$\tau_{\text{ADF}} = \frac{\hat{\gamma}}{\text{SE}(\hat{\gamma})} \tag{A.11}$$

Under the null hypothesis of a unit root, the statistic  $\tau_{\text{ADF}}$  does not follow a standard  $t$ -distribution. Instead, its asymptotic distribution is given by (see chapter 4 of [8] for details)

$$\tau_{\text{ADF}} \xrightarrow{\mathcal{L}} \frac{\int_0^1 W(r) dW(r)}{\left[ \int_0^1 W(r)^2 dr \right]^{1/2}} \tag{A.12}$$

where  $W$  is a Wiener process. This limiting distribution is non-standard and defines the Dickey-Fuller distribution. Let  $\tau_{\text{ADF,crit}}$  denote the critical value from this distribution at significance level  $\alpha$ . We reject  $H_0$  if  $\tau_{\text{ADF}} < \tau_{\text{ADF,crit}}$ . That is, if the test statistic is more negative than the critical value, we reject the null hypothesis of a unit root and conclude that the series is stationary.

## A.5 Log-likelihood and Exponential-GARCH

Let  $\{y_1, y_2, \dots, y_T\}$  denote a sample of observations drawn from a statistical model with density (or probability mass) function  $f(y_t | \theta)$ , where  $\theta \in \Theta \subseteq \mathbb{R}^k$  is a vector of unknown parameters. The likelihood function of  $\theta$  given the observed sample is defined as

$$\mathcal{L}(\theta | y_1, \dots, y_T) = \prod_{t=1}^T f(y_t | \theta) \quad (\text{A.13})$$

which represents the joint probability (or density) of observing the data  $\{y_t\}$  as a function of the parameter vector  $\theta$ . In estimation,  $\mathcal{L}(\theta)$  is viewed as a function of  $\theta$ , with the data fixed. Because the product of densities may result in extremely small numerical values, it is common to work with the natural logarithm of the likelihood function, known as the log-likelihood function, defined as

$$\ell(\theta | y_1, \dots, y_T) = \log \mathcal{L}(\theta | y_1, \dots, y_T) = \sum_{t=1}^T \log f(y_t | \theta) \quad (\text{A.14})$$

The log-likelihood function is typically easier to handle both analytically and numerically. The parameter estimate  $\hat{\theta}_{ML}$  that maximizes  $\ell(\theta)$  is called the maximum likelihood estimator (MLE) and satisfies

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} \ell(\theta) \quad (\text{A.15})$$

Now, consider a zero-mean GARCH(1,1) process defined as

$$\epsilon_t = \sigma_t Z_t, \quad \sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (\text{A.16})$$

where  $Z_t \sim \text{i.i.d. } \mathcal{N}(0, 1)$ , and the parameter vector is  $\theta = (\omega, \alpha, \beta)'$  with  $\omega > 0$ ,  $\alpha \geq 0$ , and  $\beta \geq 0$ . Given  $\mathcal{F}_{t-1} = \{\epsilon_{t-1}, \epsilon_{t-2}, \dots\}$ , the conditional distribution of  $\epsilon_t$  is

$$f(\epsilon_t | \mathcal{F}_{t-1}; \theta) = \frac{1}{\sigma_t} \phi\left(\frac{\epsilon_t}{\sigma_t}\right) \quad (\text{A.17})$$

where  $\phi(\cdot)$  denotes the standard normal density function. Assuming a sample  $\{\epsilon_t\}_{t=1}^T$ , the conditional likelihood function is given by

$$\mathcal{L}(\theta | \epsilon_1, \dots, \epsilon_T) = \prod_{t=1}^T f(\epsilon_t | \mathcal{F}_{t-1}; \theta) = \prod_{t=1}^T \frac{1}{\sigma_t} \phi\left(\frac{\epsilon_t}{\sigma_t}\right) \quad (\text{A.18})$$

Taking logarithms, the log-likelihood function becomes

$$\ell(\theta) = \sum_{t=1}^T \log f(\epsilon_t | \mathcal{F}_{t-1}; \theta) = -\frac{1}{2} \sum_{t=1}^T \left[ \log(2\pi) + \log(\sigma_t^2) + \frac{\epsilon_t^2}{\sigma_t^2} \right] \quad (\text{A.19})$$

This is the same formula used in the computation of the log-likelihood for derived GARCH models such as the TGARCH, which is employed in the LR test discussed in Chapter 2, as well as for the EGARCH model. The latter is defined as follows:

$$\log(\sigma_t^2) = \omega + \beta \log(\sigma_{t-1}^2) + \alpha \left( \frac{|\epsilon_{t-1}|}{\sigma_{t-1}} - \sqrt{\frac{2}{\pi}} \right) + \gamma \frac{\epsilon_{t-1}}{\sigma_{t-1}} \quad (\text{A.20})$$

where  $\epsilon_t = \sigma_t Z_t$  and  $Z_t \sim \text{i.i.d. } \mathcal{N}(0, 1)$ . Similarly to the TGARCH specification, the EGARCH(1,1) model allows for the presence of leverage effects in the volatility dynamics. To verify the statistical

significance of asymmetric effects in the data, we perform an LR test comparing the (unrestricted) EGARCH(1,1) model with its restricted counterpart in which  $\gamma = 0$  is imposed. The null hypothesis of the test is thus

$$H_0 : \gamma = 0 \quad (\text{A.21})$$

Under  $H_0$ , and following the same derivation as in Chapter 2, the test statistic, as  $n \rightarrow \infty$ ,

$$2(\ell_{\text{unrestricted}} - \ell_{\text{restricted}}) \xrightarrow{\mathcal{L}} \chi^2(1) \quad (\text{A.22})$$

The results for the log-return series of the Tesla stock are shown in the Figure A.3. We observe that the evidence of asymmetric effects is confirmed even more strongly than in the TGARCH case.

```
==== EGARCH(1,1) (restricted: gamma = 0) ====
{'omega': -8.956641, 'alpha': 0.150006, 'beta': 0.92, 'gamma': 0.0}
logLik: -431707307.090, success=True, nit=4

==== EGARCH(1,1) (unrestricted: gamma free) ====
{'omega': -0.08646, 'alpha': 0.109556, 'gamma': -0.002574, 'beta': 0.985926}
logLik: 4964.325, success=True, nit=33

==== LR test H0: gamma = 0 (no asymmetry) ====
LR statistic: 863424542.831, df=1, p-value: 0.000000
```

Figure A.3: Results LR test

## A.6 Brownian motion and Jarque-Bera test

**Definition A.6.1** (Brownian Motion). *Let  $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$  be a filtered probability space. A stochastic process  $\{B_t\}_{t \geq 0}$  adapted to  $\{\mathcal{F}_t\}$  is called a Brownian motion (or Wiener process) if the following properties hold :*

1.  $B_0 = 0$  almost surely;
2. the sample paths  $t \mapsto B_t$  are continuous almost surely;
3. for every  $0 \leq s < t$ , the increment  $B_t - B_s$  satisfies :
  - (a)  $B_t - B_s \sim \mathcal{N}(0, t - s)$ ;
  - (b)  $B_t - B_s$  is independent of  $\mathcal{F}_s$ .

The eigenfunctions and eigenvalues of the covariance operator of a Brownian motion on  $[0, 1]$  are given by

$$\psi_j(t) = \sqrt{2} \sin\left(\left(j - \frac{1}{2}\right)\pi t\right), \quad \lambda_j = \frac{1}{\left(j - \frac{1}{2}\right)^2 \pi^2}, \quad j = 1, 2, \dots \quad (\text{A.23})$$

and therefore the Karhunen-Loève expansion takes the form

$$B(t) = \sum_{j=1}^{\infty} \frac{Z_j}{\left(j - \frac{1}{2}\right)\pi} \sqrt{2} \sin\left(\left(j - \frac{1}{2}\right)\pi t\right), \quad Z_j \sim i.i.d. \mathcal{N}(0, 1) \quad (\text{A.24})$$

This representation also suggests one of the possible approaches to simulate a Brownian motion. Moreover, if one plots the shape of these eigenfunctions, they appear remarkably similar to those obtained for the cumulative log returns in Figure 3.3 of Chapter 3.

The Jarque-Bera test is used to assess whether a sample comes from a normal distribution. The test is based on the fact that for a normally distributed random variable, the skewness is equal to 0 and the kurtosis is equal to 3. The hypotheses of the test are

$$H_0 : \text{the data are normally distributed} \quad (\text{A.25})$$

Let  $\hat{S}$  denote the sample skewness and  $\hat{K}$  the sample kurtosis. The Jarque-Bera test statistic is defined as

$$JB = n \left( \frac{\hat{S}^2}{6} + \frac{(\hat{K} - 3)^2}{24} \right) \quad (\text{A.26})$$

where  $n$  is the sample size. Under  $H_0$ , as  $n \rightarrow \infty$ , this statistic satisfies the asymptotic distribution

$$JB \xrightarrow{\mathcal{L}} \chi^2(2) \quad (\text{A.27})$$

that is, a chi-square distribution with 2 degrees of freedom. The corresponding  $p$ -value is therefore computed as

$$p\text{-value} = 1 - \chi_2^2(JB_{\text{obs}}) \quad (\text{A.28})$$

where  $\chi_2^2$  denotes the cumulative distribution function of the  $\chi^2(2)$  distribution.

# References

- [1] Alexander Gerber, Christoph Hanck, Martin Arnold, and Martin Schmelzer. *Introduction to Econometrics with R*. University of Duisburg-Essen, 2024.
- [2] Ahmed Kebaier. Lecture notes on time series. Université d'Évry, 2025.
- [3] Whitney K. Newey and Kenneth D. West. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, pages 631–653, 1994.
- [4] Juhyun Park. Lecture notes and slides on financial econometrics. ENSIIE, 2025.
- [5] James H. Stock and Mark W. Watson. *Introduction to Econometrics*. Pearson Education, 4th global edition, 2019.
- [6] Marie-Luce Taupin. Lecture notes, slides and code on linear models. Université d'Évry, 2024.
- [7] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition, 2005.
- [8] Eric Zivot and Jiahui Wang. *Modeling Financial Time Series with S-PLUS*. Springer, 2003.