Modeling of High-Frequency Financial Data

Andrés Mireles, Luca Ibra

December 12, 2024

1 Introduction

Understanding the behavior of financial markets is essential for effective risk management and portfolio optimization. This study investigates high-frequency financial data through three key lenses: Value-at-Risk (VaR) estimation, GARCH modeling, and Functional Data Analysis (FDA). VaR, a fundamental concept in financial risk management, measures potential losses within a specified confidence interval. First, we examine our data under the assumption of independence and identical distribution (i.i.d.), evaluating its validity before moving to conditional volatility modeling using GARCH. Given the limitations of traditional VaR models in capturing volatility clustering and heavy-tailed distributions, we fit GARCH(1,1) models under both Normal and Student's t-distributions, analyzing their efficacy for volatility estimation. The final section employs FDA to analyze high-frequency intraday cumulative returns, capturing intraday dynamics that are often discarded in daily aggregations. By decomposing returns into their principal modes of variability using Functional Principal Component Analysis (FPCA), we explore their main trends and patterns. Moreover, we test whether modeling these modes with Functional Time Series (FTS) techniques provides additional predictive insights.

2 Data

2.1 Data preprocessing

The data used for the analysis represents intraday movements in the share price of a given stock, observed over a three-year period. Each observation in the dataset represents the stock price at the associated tick. Every day is comprised of multiple ticks, each of which represents an event recorded in the stock market, such as the occurrence of a trade. Multiple trades can be executed at the same price, which is why the same price may be associated with different ticks. Before conducting our research, we first cleaned our dataset as it consisted of observations that were irregularly spaced. To do so we firstly defined an equally spaced tick grid using the number of minutes per day and secondly, we interpolated the log prices across the tick grid. Therefore, our high-frequency data now consisted of a price per minute, for a period of 3 years, resulting in 1,085,732 observations. The last part of the data cleaning process dealt with zero values (no price change) which were in the first place filled with "NaN values" and were later replaced with linearly interpolated values. Regarding the estimation of the VaR, we aggregated the data to a lower frequency, retrieving the daily closing prices in order to reduce the impact of micro-structure noise. Moreover, the same process was applied in the estimation of the GARCH model in order to avoid the mentioned noise, which is highly undesirable as it might distort the estimates of our GARCH models. However, note that as it was shown in (1), high-frequency

data can be incorporated as a proxy in the estimation of GARCH and conditional volatility, but this task goes beyond the scope of this project.

2.2 Data analysis

Figure 1 shows the dynamics of intraday log prices and log returns over time. The image on the left displays an upward trend, whereas the right-hand side image illustrates that returns seem to be concentrated around zero with large spikes being rare in both positive and negative regions of the y-axis. Note that the largest values appearing in our graph correspond to the difference between the closing price of a trading day and the opening price of the following day. These outliers would be filtered out in the FDA section where we treat intraday cumulative returns.

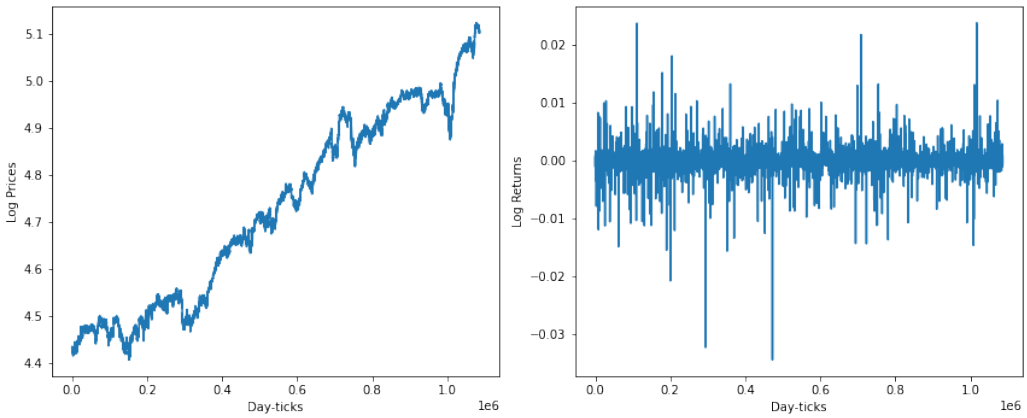

Figure 1: Intraday log prices (left) and log returns (right) over time

Following the resampling of our data to a daily frequency, we plotted the histogram of daily log returns (from now on daily returns). We also fitted both the normal and the Student's t-distribution (from now on, t distribution) to our data using the Maximum likelihood estimation (MLE) method (2). Then, we computed some relevant statistics in order to make accurate assumptions during our analysis.

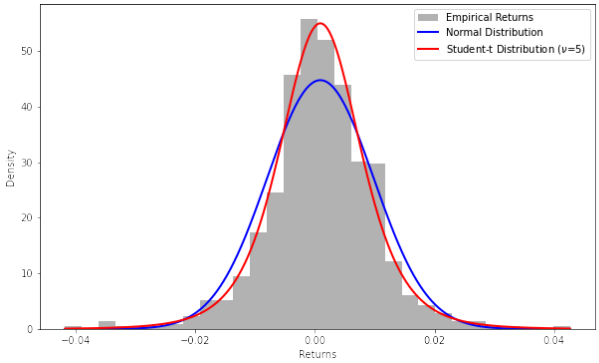

Figure 2: Daily returns distribution

Statistic	Value
Mean	0.000900
Standard Deviation	0.008930
Skewness	-0.215630
Excess Kurtosis	2.548622

Table 1: Daily returns statistics

Figure 2 shows the distribution of daily returns, with both the normal and the t distribution fitted to our data. The table on the right instead, displays the main statistics of the data. Visually it is clear that the t distribution better describes our data, and this conjecture is supported by the key statistics contained in the table. By analyzing the two, we were able to conclude that the normal distribution does not suit log-price data. The positive excess kurtosis implies that the distribution is fat-tailed and the negative skewness, which is common in financial

returns, indicates negative values occur more frequently than positive values do. In summary, the t distribution (with five degrees of freedom) is more appropriate to model our data as it gives more weight, in terms of probability, to extreme events, unlike the normal distribution. Moreover, as it is required for our assumptions, we tested stationarity in our data by employing the ADF test (3), showing that indeed our data is stationary.

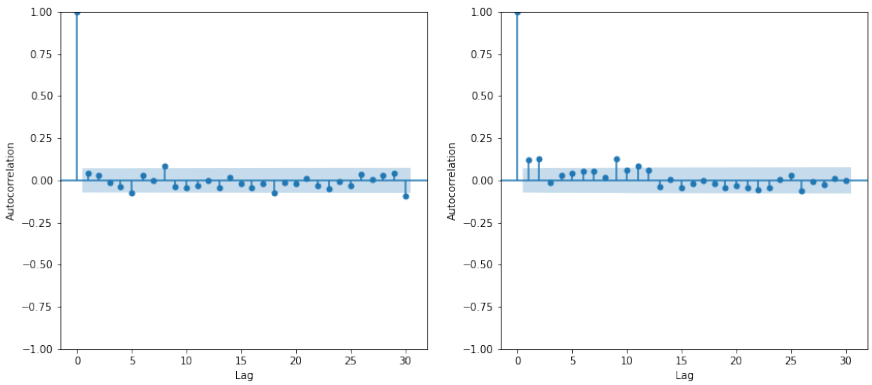

Figure 3: ACF of daily simple (left) and squared (right) returns

We proceed by analyzing whether our data has what is commonly known as volatility clustering. Figure 3 displays the autocorrelation function (ACF) computed for both simple and squared daily returns. The ACF reveals whether there exists some degree of relationship between our observations and their lagged values for a given fixed level. In both cases, we selected the same number of lags (40) and a level of 0.05. Concerning the ACF of daily returns, it seems that the data is correlated, however, the visual inspection is not as informative as the one performed on squared returns. The graph on the right, clearly shows multiple correlation values that exceed the confidence bands. Computing the ACF of squared returns is of essential importance, as it detects whether the observations are independent. Actually, the ACF plot for daily returns fails to do so as it neglects relationships of degrees higher than one. This can be further tested using the Ljung–Box test (4), which was performed in our research, and the corresponding results can be found in the attached Python code. From the results of this exploratory analysis, we can conclude that our daily data is stationary but neither i.i.d nor white noise.

3 Methodology

In this section we present the theoretical framework under which we have conducted our analysis.

3.1 Value-at-Risk

Estimating Value-at-risk amounts to computing the quantile of the underlying variable's distribution:

$$V_p(X) := \inf \{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq p\} \tag{1}$$

Equation 1 states that the quantile associated to a p-level, with $p \in [0, 1]$, is given by the greatest lower bound, that is a real value, such that the probability that the underlying variable is less than the quantile, is greater or equal to p (5). Initially we worked with a fixed level $p = 0.05$, whereas later in our research we computed estimates for $p \in [0, 0.3]$. We adopted both the parametric and the non-parametric approach, i.e. we made assumptions about the underlying

distribution of returns and we entirely relied on the historical data, respectively. Concerning the parametric approach, we fitted a normal and a t distribution to our data. During this section, we assumed to be working in a homoscedastic setting, except for an analysis we performed by computing a daily moving VaR. For this, we obtained the filtered series, which served as the basis for the daily VaR by applying:

$$Z_t = \frac{X_t - \mathbb{E}(X_t)}{\mathrm{SD}(X_t)} \tag{2}$$

where we used local varying measures for both the mean and the standard deviation (with a window of 60 trading days). The daily VaR was obtained by computing the quantile of the filtered series which was then scaled by the local volatility and then shifted by the moving average. Finally, we compared the different VaR approaches in the accuracy of the estimates. The comparison was performed by counting the number of exceedances of each estimated VaR with respect to the actual number of violations. The test we used is the Kupiec's Proportion of Failures (KPF) likelihood ratio (6), which is given by Equation 3.

$$KPF = 2\ln\left(\left(\frac{\frac{x}{T}}{\alpha}\right)^{x}\left(\frac{1-\frac{x}{T}}{1-\alpha}\right)^{T-x}\right) \tag{3}$$

The statistic follows a chi-square distribution and indicates whether the estimated VaR is breached more or less frequently than the expected $100\% \times \alpha$ of the time.

3.2 GARCH and Volatility Estimation

As we have seen in previous sections, our daily returns data is stationary but exhibits volatility clustering and a fat-tailed distribution. For treating this kind of data, we make use of the GARCH(1,1) model, which provides a robust framework for modeling volatility in financial time series (7). The model achieves this through a combination of equations governing the mean and variance dynamics of returns.

The mean equation models the expected return of a financial asset:

$$r_t = \mu + \epsilon_t \tag{4}$$

where r_t represents the return at time t , μ is the constant mean of the returns, and ϵ_t is the residual term so that deviations from the mean are random and governed by the conditional variance. The evolution of conditional variance σ_t^2 is given by:

$$\sigma_{t+1}^2 = \omega + \alpha\epsilon_t^2 + \beta\sigma_t^2 \tag{5}$$

This equation decomposes volatility into three components: ω representing the long-run average variance, α measuring the impact of past shocks (ϵ_t^2) on current volatility and β, which quantifies the persistence of volatility over time. Note that for estimating the three parameters for the conditional variance equation, we make use of the MLE method under two different distributions, the Normal and Student's t.

Once our GARCH(1,1) model is fitted to our data, we can use Equation 5 to estimate the one-step ahead volatility and therefore estimate the VaR under the GARCH modeling:

$$\mathrm{VaR} = \mu + z_\alpha \cdot \sigma_t \tag{6}$$

3.3 Functional Data Analysis for High-Frequency Returns

In the previous approaches, we transformed high-frequency prices into daily data by using the closing prices. However, this method discards valuable intraday information, such as volatility patterns and short-term trends. To address this limitation, we employed Functional Data Analysis (FDA) to analyze high-frequency data while preserving the continuous dynamics of intraday cumulative returns. By representing intraday data as smooth functions over time, FDA captures the underlying trends and variability of price movements throughout the trading day (8).

Under the FDA framework, the high-frequency cumulative log returns are modeled as:

$$Y_n(t_{nj}) = X_n(t_{nj}) + \varepsilon_{nj}, \quad n = 1, \ldots, N, \quad j = 1, \ldots, J_n, \tag{7}$$

where $\varepsilon_{nj} \sim (0, \sigma^2)$ represents the noise term, and $X_n(t)$ is the smooth functional representation of the returns, defined as:

$$X_n(t) = \mu(t) + \sum_{k=1}^{p} \xi_{nk}\psi_k(t). \tag{8}$$

Here, $\mu(t)$ is the mean function capturing the average behavior of intraday returns, $\psi_k(t)$ represents the k-th eigenfunction, representing dominant modes of variability, ξ_{nk} are the principal component scores for day n , and p represents the number of principal components retained.

The principal components $\psi_k(t)$ are orthogonal functions that explain decreasing proportions of variability in the data. Although in theory there are infinitely many eigenfunctions, we can retain only the first p components and achieve a low-dimensional yet informative representation.

To represent $X_n(t)$ as a smooth function, we utilized B-spline basis functions due to their computational efficiency, flexibility, and smooth transition across time points (8). These piecewise polynomial functions are defined over a sequence of knots, which divide the time domain into intervals. For our analysis, we selected the Cubic B-splines (degree 3) and placed the knots uniformly across ticks.

Each trading day's cumulative log returns are expressed as a linear combination of the B-spline basis functions:

$$X_n(t) = \sum_{j=1}^{p} c_{nj}\phi_j(t), \tag{9}$$

where $\phi_j(t)$ are the B-spline basis functions and c_{nj} are the coefficients estimated by minimizing the least-squares error between observed and reconstructed data (8).

Once we had this setup ready, we employed the same approach to model the evolution of our scores. For developing these forecasting methods, we split the data into training and testing sets: the first two years of data were used for training, while the third year was reserved for evaluating prediction performance.

For each principal component score sequence, we employed AR(1), or FAR(1), models (9), defined as:

$$\xi_{n+1,k} = \phi_k\xi_{n,k} + \epsilon_{n,k}, \tag{10}$$

where ϕ_k represents the autoregressive coefficient for the k-th component, and $\epsilon_{n,k}$ denotes the residual noise term.

Using the FAR(1) models fitted on the training data, we generated h-step-ahead forecasts for the scores $\xi_{N+h,k}$. These forecasts were then combined with the mean function $\mu(t)$ and the eigenfunctions $\psi_k(t)$ to reconstruct the forecasted trajectories of the intraday log returns. Finally, we evaluated the forecasting performance on the test set using the RMSE.

4 Results

In the present section, we cover the results of the methodology explained in the previous sections.

4.1 Value-at-Risk

Figure 4 shows the daily return series and the various VaR estimates, with the associated exceedances marked with a different color whenever a violation of the VaR occurs. The first thing we noticed is that assuming a normal distribution of returns led to a higher (in absolute value) estimate of VaR, denoted by the orange line, than for the Student's t VaR, colored in yellow. This is because the t-distribution assigns a higher probability to more extreme values, resulting in a given quantile (in our case, the 95%) being closer to the mean compared to the Normal distribution. The green line is the VaR associated to this confidence level by using the non-parametrical historical distribution. We observe how the Student's t VaR is closer to this estimation. Lastly, the moving VaR outperforms the Normal Var by effectively capturing volatility clustering, although it fails to improve Student's t VaR. The results for the KPF ratio will be discussed in Section 4.3.

It is important to note, however, that this analysis was conducted on the same dataset used for both estimation and evaluation, which could introduce bias; a separate training and testing approach would have been more robust.

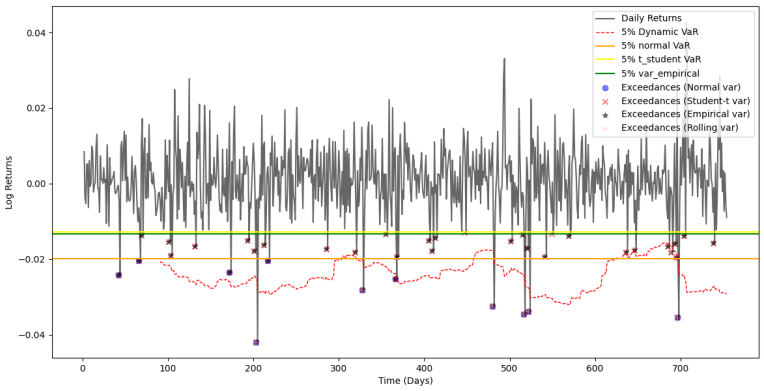

Figure 4: Comparing VaR estimates in homoscedatic setting

4.2 GARCH and Volatility Estimation

In the previous subsection, we explored risk management under the i.i.d. assumption (in addition to a rolling VaR that disregards this assumption). However, in Section 2.2 we observed that our returns exhibited autocorrelation and fat-tails, suggesting that the i.i.d. assumption is not valid. Therefore, for modelling these characteristics, we have applied a GARCH(1,1) model following the procedure detailed in Section 3.2.

First, we fit our data to a GARCH(1,1) model using the Normal distribution by maximizing the log-likelihood function to fit the model parameters, achieving the results presented in Table 2. There, the results of fitting the GARCH model under the t distribution are also shown.

Once the parameters were estimated, we assessed whether our GARCH models successfully captured the autocorrelation in daily returns. To evaluate this, we examined the autocorrelation in the residuals using the Ljung-Box test, as applied previously. The results indicate that neither

Parameter	Normal Distribution	Student's t-Distribution
μ	9.00×10^{-4}	9.00×10^{-4}
ω	8.65×10^{-6}	7.08×10^{-6}
α	9.89×10^{-2}	6.84×10^{-2}
β	7.91×10^{-1}	7.57×10^{-1}
$\alpha + \beta$	8.90×10^{-1}	8.26×10^{-1}

Table 2: GARCH(1,1) Model Fit Summary for Normal and Student's t-Distributions

the Normal nor the Student's t GARCH(1,1) models exhibit significant autocorrelation in their residuals, demonstrating that the models effectively captured all the correlation in the data.

With these results for the parameters of the GARCH model under both distributions, Equation 5 was used to compute the one-step ahead conditional variance for each day. The resulting volatility can be observed along the log-prices evolution in Figure 5

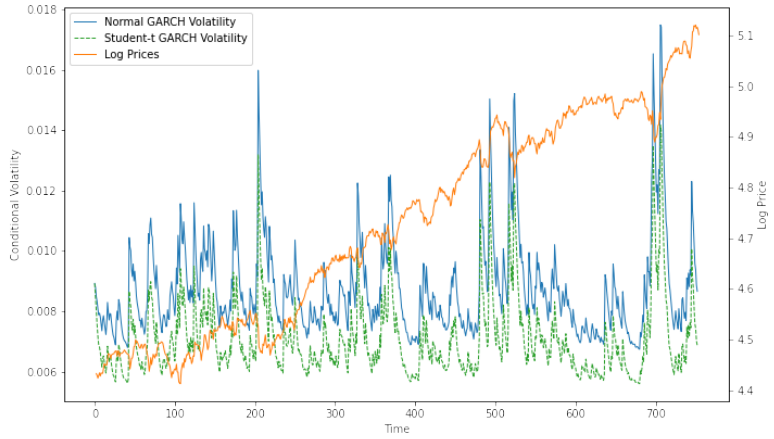

Figure 5: GARCH(1,1) Estimated Conditional Volatility (Normal and Student's t)

This representation allows us to observe two important facts. First of all, both the normal and the t distribution estimated conditional volatility increase whenever the prices jump (either up our down). Second, the estimated conditional volatility under the t distribution is smaller than the normal distribution. This is because the t-distribution accounts for the heavy tails observed in financial returns, leading to lower sensitivity to extreme price movements compared to the normal distribution, which tends to overestimate volatility in such cases (7).

Finally, we can use the estimated volatility to compute a moving VaR model in our returns using Equation 6. The results for a confidence interval of 95% are shown in Figure 6. For this specific interval, both Normal and t-distribution models work similarly (they count 39 and 40 exceedances respectively, with the theoretical being 38). However, in the next subsection, we will see how the t-distribution GARCH VaR outperforms the Normal one.

4.3 VaR comparison

In this section we compared all the VaR estimates for difference confidence intervals, using the number of violations of the VaR, calculated based on historical data, as the benchmark. As expected, the number of exceedances is inversely proportional to the p level.

As previously discussed, the Student-t distribution provides a better fit for financial returns. Consequently, we anticipated the two corresponding VaR estimates to outperform the others.

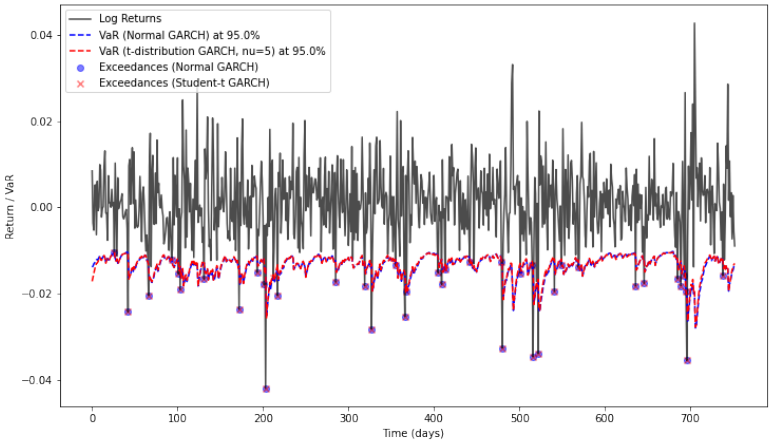

Figure 6: Moving Daily GARCH VaR (Normal and Student's t)

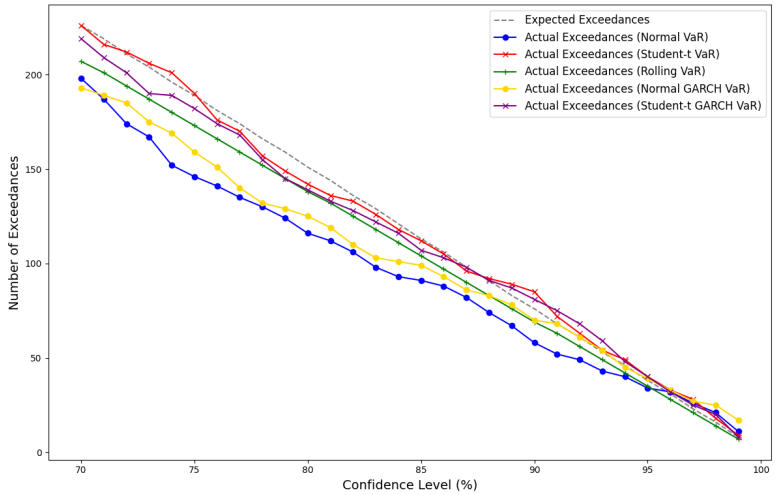

Figure 7: Comparison of VaR estimates relative to actual violations

Figure 7 confirms this expectation, showing that the number of violations for the Student-t VaR estimates closely aligns with the actual number of exceedances.

Notably, the VaR estimate that accounts for non-constant volatility proves more accurate than the one derived under the homoscedastic setting. This outcome is consistent with our expectations, as the assumption of i.i.d. returns fails to capture the characteristics of our data. On the other hand, the normal distribution underestimates the number of exceedances due to its thinner tails, leading to less reliable predictions.

Model	Average KPF Ratio	Root Mean Squared Error (RMSE)
Student's t-distribution	0.2759	4.5680
Normal distribution	5.9294	26.7058
Normal GARCH	3.7183	20.9117
Student's t GARCH	0.4789	7.2961
Rolling	0.8744	11.2086

Table 3: Average KPF ratio and RMSE for each VaR model

For comparing the results numerically, Table 4.3 presents the average test statistic for each model, as outlined in Equation 3, across all $p \in [0, 0.3]$. The t-distribution assumption consistently proves to be the most suitable for our data, as its test statistics are close to zero, indicating that the difference between the estimated VaR and the one obtained from historical data is small. Additionally, we calculated the Root Mean Squared Error (RMSE) for each model, which measures the difference between the predicted and observed exceedances. The results confirm that the most accurate estimates are provided by assuming a t-distribution. However, note that as mentioned before, this result might be biased since the training and test data is the same.

4.4 Functional Data Analysis for High-Frequency Returns

In contrast to the previous sections, where we used daily data to develop different risk assessment strategies, in this section, we use intraday returns to try to model them and understand their main modes of variability.

We first performed exploratory analysis where we observed that the intraday log returns, similar to the daily data, were stationary and followed a distribution with fatter tails than a Normal one. Then, we employed the methodology presented in Section 3.3 to decompose the intraday cumulative log returns into their principal modes of variability using Functional Principal Component Analysis (FPCA) with B-spline basis functions. The intraday mean function as well as the first three eigenfunctions are presented in Figure 8.

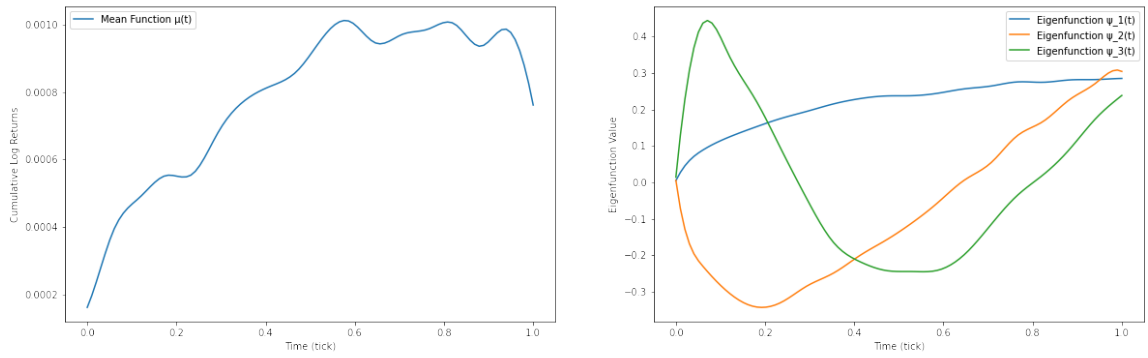

Figure 8: Estimated Mean Function $\mu(t)$ (left) and first three eigenfunctions $\Psi(t)$ (right)

The left plot shows an upward trend in cumulative log returns with a small shift at the end of the day, reflecting a general positive drift in stock prices. The first eigenfunction captures overall shifts in return levels throughout the trading day, with a slowly increasing behaviour. The second eigenfunction highlights midday deviations, contrasting return behaviours during the middle of the session with those at the opening and closing periods, indicating variability in intraday momentum. The third eigenfunction reveals more complicated intraday fluctuations, such as early-morning rallies followed by midday declines and late-day recoveries.

We now answer to whether using the first three eigenfunctions is enough for representing our daily intraday cumulative returns. For this task, we make use of the cumulative explained variance as a function of the number of eigenfunctions used (p). Also, we measure the reconstruction error (or noise variance) by computing the RMSE between the p-reconstrued trajectory and the realised trajectory of our stock. Both results are presented in Figure 9. In this plot, we discover that with the first three eigenfunctions, we are able to explain more than 90% of the variance, achieving the elbow point. Moreover, including more eigenfunctions does not reduce

significantly the reconstruction error. Therefore, our choice of using 3 eigenfunctions is justified.

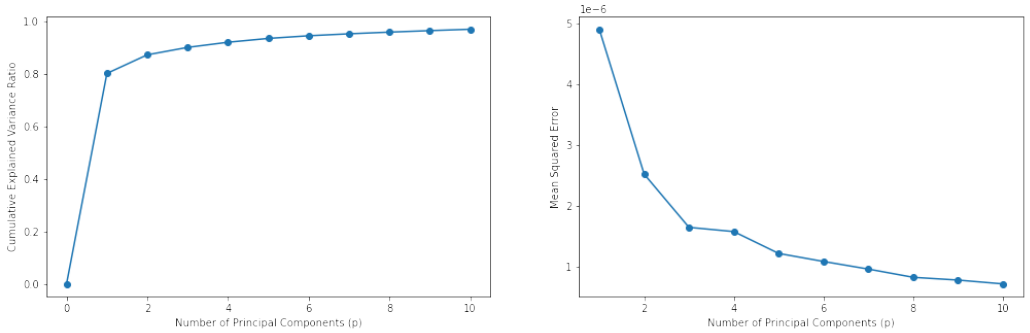

Figure 9: Cumulative Explained Variance (left) and Reconstruction Error (right) as a function of the number of eigenfunctions (p)

Just for illustration purposes, in Figure 10 we see how "well" we are able to reproduce the intraday movements over 4 randomly selected days with the reconstructed function using 20 B-Splines and with the 3 first eigenfunctions. We observe that, in general, we are able to capture the global movements with just 3 eigenfunctions without including the erratic noise.

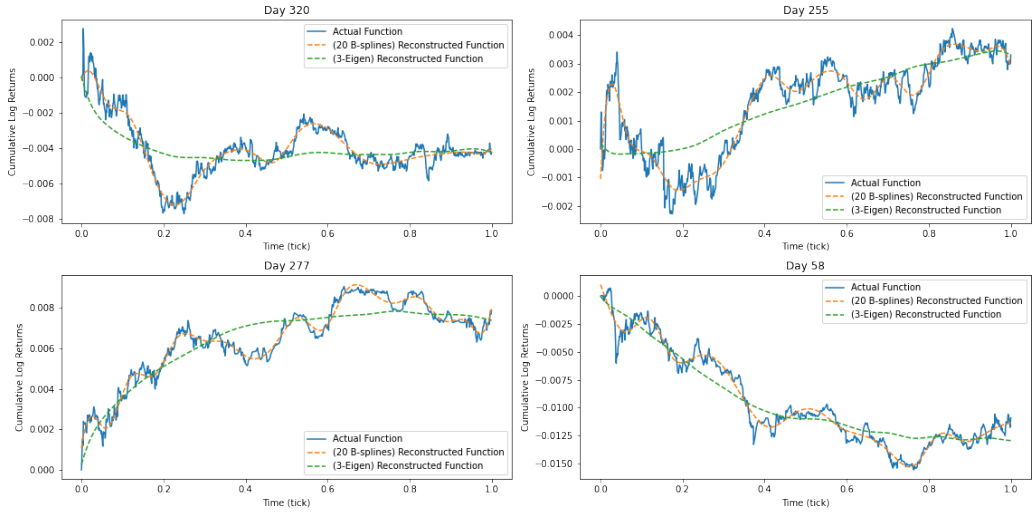

Figure 10: Actual vs. Reconstructed Intraday Trajectories for 4 random days

Once we have chosen the number of eigenfunctions, we can analyze the corresponding scores (eigenvalues). For this, in Figure 11 we show the distribution of them, as well as the corresponding Q-Q plot. With these results, we conclude that scores align with a normal distribution, especially in the central region, with some deviations in the tails indicating slight departures from perfect normality, possibly due to heavier tails in returns. Moreover, we have computed the ACF function and applied the Ljung-Box test to each of the scores, and we have been able to discard any autocorrelation in them. Therefore, we say that the scores are a white noise process.

4.5 Functional Time Series

In the previous subsection, we applied FDA techniques to our data to decompose it into the principal modes of variability. These modes have an associated score that determines the percentage of total variance explained. We have also discovered that, in our case, these scores

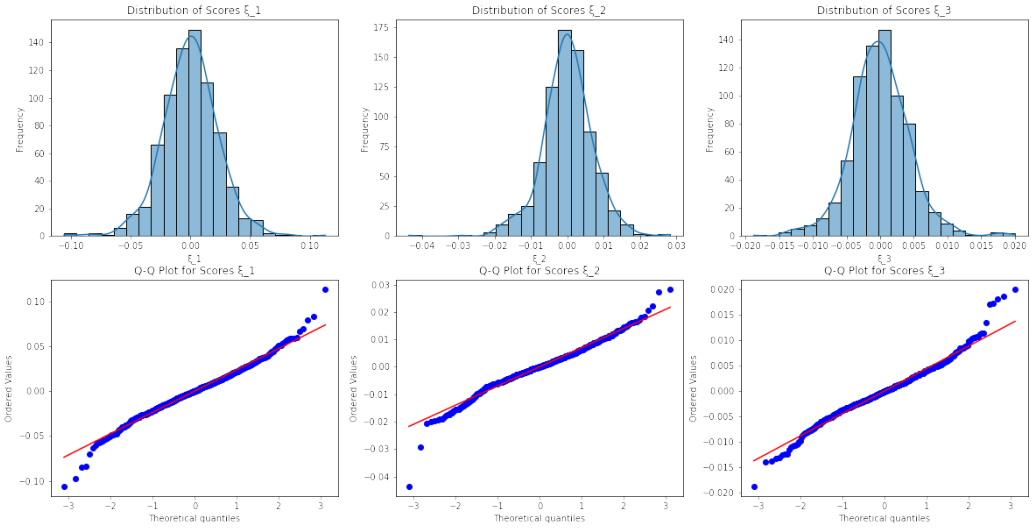

Figure 11: Distributions and Q-Q plots for the first three scores (eigenvalues) across days

follow a white noise process, i.e. stationarity and no autocorrelation. Consequently, applying time-series models such as AR(1) to these scores yields no improvement over simply using the mean function as a predictor. However, we need to test this hypothesis.

In order to fit a Functional Time Series model to our data, we split the data points into a training and test set. Then, we trained an AR(1) (or FAR(1)) model in the training set.

Score	ϕ_k
ξ_1	-0.0344
ξ_2	0.0033
ξ_3	0.0031

Table 4: AR(1) Model Coefficients for the First Three Principal Components

The corresponding ϕ coefficients are shown in Table 4. These coefficients are smaller than 1, indicating stationarity, but are also very close to 0, confirming our hypothesis that the AR(1) model lacks predictive capability for this data. To validate this, we applied the trained AR(1) model to the test set and evaluated the RMSE of the predictions. Comparing this to a simple benchmark with the sample mean of the training set, we observed similar RMSE values (around 6×10^{-3}), further reinforcing the limited predictive utility of the AR(1) model in this context.

5 Conclusions

This study explored high-frequency financial data using Value-at-Risk (VaR) estimation, GARCH modeling, and Functional Data Analysis (FDA). We found that the normal distribution assumption for returns was inadequate due to the data's heavy tails and negative skewness. The Student's t-distribution provided more accurate VaR estimates, closely matching the actual number of exceedances. GARCH(1,1) models effectively captured volatility clustering inherent in financial time series. The model under the Student's t-distribution outperformed the normal distribution variant, highlighting the importance of accounting for heavy-tailed distributions

and time-varying volatility in risk assessment. With FDA, we were able to capture over 90% of the variability in intraday cumulative returns using just the first three principal components. However, our approach to model these components with Functional Time Series models did not improve our predictive capabilities, as the scores followed white noise processes.

Future work could address some limitations of the present work by employing separate training and testing datasets; incorporating intraday returns as a proxy for volatility or utilizing more advanced models for conditional volatility such as EGARCH; and exploring non-linear models or machine learning techniques (such as RRNN or LSTM (10)) for forecasting the FDA scores, considering even including exogenous variables (such as macroeconomic variables) into the feature space.

References

[1] C. Deng, X. Zhang, Y. Li, and Q. Xiong, *GARCH Model Test Using High-Frequency Data*, Mathematics, vol. 8, no. 11, p. 1922, Nov. 2020. 10.3390/math8111922.

[2] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed., Duxbury Press, Pacific Grove, CA, 2002.

[3] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a), 427–431.

[4] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Wiley, Hoboken, NJ, 2015.

[5] Nicolas Privault, *Notes on Financial Risk and Analytics*, 2023, Lecture Notes, Nanyang Technological University, Singapore. `https://www.math.nus.edu.sg/~matnp/notes_financial_risk_analytics.pdf`

[6] Kupiec, P. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. Journal of Derivatives

[7] Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. 10.1016/0304-4076(86)90063-1.

[8] Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed.). Springer.

[9] Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Springer.

[10] W. Dai, Y. An, and W. Long, *Price change prediction of ultra high frequency financial data based on temporal convolutional network*, ITQM 2020&2021. arXiv:2107.00261.

$$\mathbb{E}[\epsilon_{t+1}^2|\mathcal{F}_t] = \sigma_{t+1}^2 = \omega + \alpha\epsilon_t^2 + \beta\sigma_t^2$$

Declaration

We hereby declare that this report is our own work. All sources of information and assistance have been acknowledged and referenced appropriately.