Assignment 2 - Time series analysis of S&P ETF Data

- 2898108

October 17th 2025

Contents

1 Exploratory Data Analysis and Data Preparation

Before proceeding to formal time series modeling, a thorough exploratory analysis of the data was conducted to understand its underlying characteristics. This initial step is crucial for justifying subsequent methodological choices, such as data resampling and model selection. Our analysis focuses on the three assets assigned to our group: `SPY5.L`, `SPY5z.CHIX`, and `SPY5.P`. For the univariate analysis, we select `SPY5.P` as our primary series of interest.

1.1 Data Resampling from 1-Minute to 5-Minute Intervals

The raw data was provided at a 1-minute frequency. However, an initial analysis of the 1-minute log-returns revealed a significant issue: a large proportion of the returns were exactly zero. This indicates periods of inactivity where no transactions occurred, causing the price to remain unchanged. This prevalence of zero-returns had two negative consequences:

1. It prevented the statistical functions from calculating confidence intervals for the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, hindering model identification.
2. The extremely small variance of the non-zero returns could lead to numerical precision issues in the model estimation algorithms.

To mitigate these issues, the price series was resampled to a 5-minute frequency by taking the last observed price in each interval. This aggregation creates a more statistically robust time series where price changes are more common, allowing for a more meaningful analysis.

1.2 Price Evolution and Co-movement

A plot of the 5-minute prices for the three ETFs reveals a striking pattern of near-perfect co-movement, as shown in Figure 1.

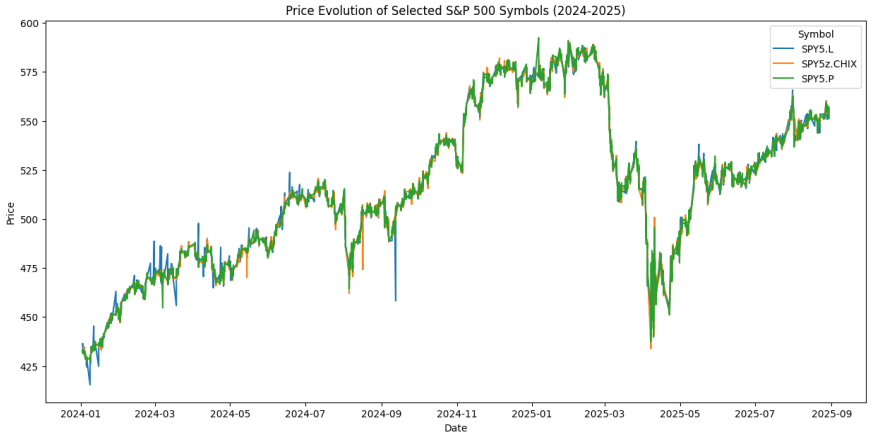

Figure 1: Price Evolution of `SPY5.L`

The series are visually indistinguishable, moving in lockstep. This suggests the presence of a strong, stable long-run relationship between them. While the next section focuses on univariate ARMA modeling, this observation strongly indicates that a multivariate approach, such as a Vector Error Correction Model (VECM) to test for cointegration, would be a highly appropriate framework.

1.3 Distribution of Log-Returns

The distribution of 5-minute percentage log-returns for `SPY5.L` deviates significantly from a normal distribution (Figure 2). The distribution exhibits high kurtosis (1352.5181), meaning it has "fat tails" and a much sharper peak at the mean compared to a Gaussian distribution. Additionally it is also extremely skewed (4.0835). This indicates that extreme events (large positive or negative returns) are more common than a normal distribution would predict. This seems coherant with the fact that at per minute level data, it is very unlikely that transactions are going through in each minute. Therefore, there can be large jumps in returns.

Interestingly, as we aggregate the returns over longer time horizons (from 5-minutes to 1-hour, and then to daily), the distribution begins to appear more normal. This is consistent with the Central Limit Theorem and suggests that the high-frequency "noise" caused by infrequent trading diminishes as the observation interval increases.

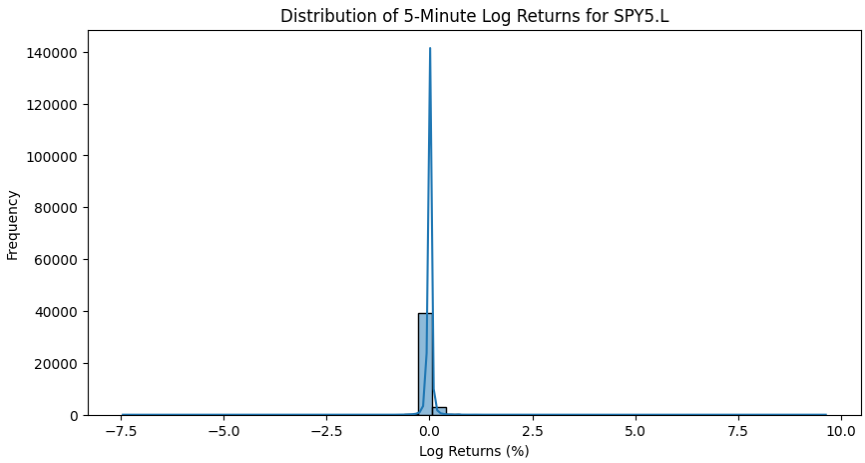

Figure 2: Distribution of 5-Minute log returns in full sample

1.4 Analysis of Outliers and Volatility Clustering

To investigate the occurrence of extreme returns, outliers were identified on a week-by-week basis. For each week in the sample, returns falling outside three standard deviations of that week's mean were flagged. Figure 3 plots these outliers across the entire sample period.

The plot clearly shows that, apart for some periods, the percentage log returns is a stationary process. However, when looking at the process as a whole, including the spike we notice that the outliers are not randomly distributed. Instead, they appear in clusters, indicating periods of high volatility followed by periods of relative calm. This phenomenon, known as **volatility clustering** leads us to believe that GARCH or regime switching models would be a valid approach to analyse this data.

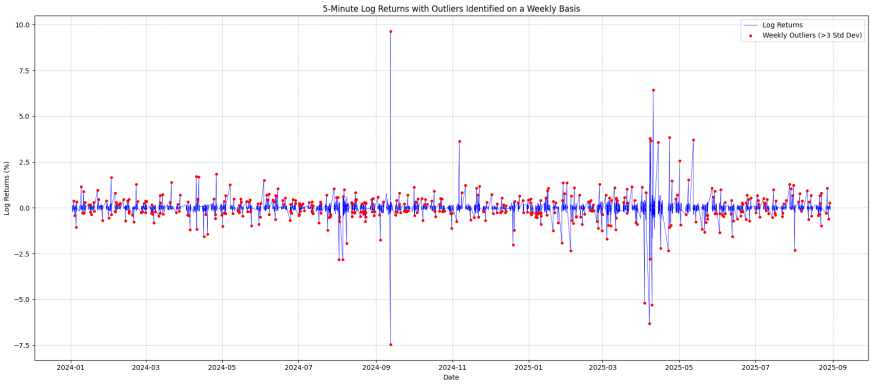

Figure 3: Percentage log returns over time with outliers marked in red per week

2 ARMA Modelling (Exercise 3.2)

This section details the process of identifying, estimating, and diagnostically checking ARMA(p,q) models for the 5-minute log-return series of `SPY5.L`, as required by the assignment.

2.1 Model Identification

The initial step in the Box-Jenkins methodology involves identifying plausible model orders (p,q) by examining the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of the stationary return series.

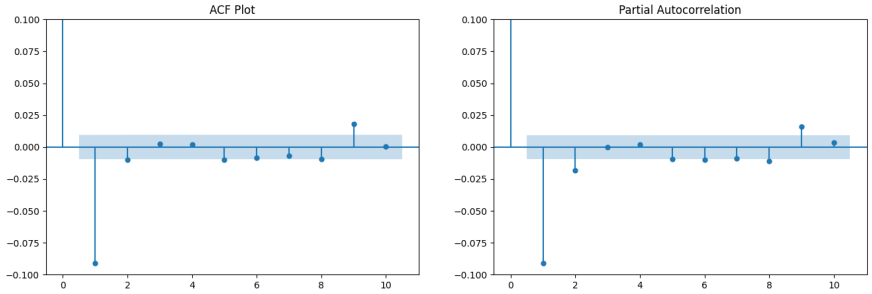

Figure 4: ACF and PACF of 5-Minute Log-Returns of `SPY5.L`. for up to 10 lags

The ACF plot shows significant spikes at lags 1, 2, 5, 8, and 9, while the PACF plot shows significant spikes at lags 1, 2, 8, and 10. Both plots exhibit a pattern that decays relatively slowly, which makes it difficult to definitively identify a pure AR or MA process. This pattern suggests that a mixed ARMA model is likely the most appropriate specification.

2.2 Model Estimation and Evaluation

As per the assignment, we systematically estimated all nine ARMA(p,q) models for $p, q \in \{0, 1, 2\}$. The data was split into a training set for model estimation, a holdout set (one day of data) for out-of-sample validation, and a final test set.

Given the evidence of heteroskedasticity from our EDA, the theoretically correct approach is to use Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors. However, attempts to estimate the models with this option failed due to computational issues within the statistical package. This is a common "roadblock" when working with large datasets. As a practical solution, the final parameters are reported with the default standard errors, and this limitation is noted. The final summary table is presented in Table 1.

Table 1: ARMA(p,q) Model Estimation and Diagnostic Summary

Statistic	**ARMA(0,0)**	**ARMA(0,1)**	**ARMA(0,2)**	**ARMA(1,0)**	**ARMA(1,1)**	**ARMA(1,2)**	**ARMA(2,0)**	**ARMA(2,1)**	**ARMA(2,2)**
Parameter Estimates (Default SEs)									
const	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
ar.L1	-	-	-	-0.0528	-0.7317	-0.3208	-0.0527	-0.0898	0.5360
	-	-	-	(0.0048)	(0.0872)	(0.1360)	(0.0048)	(0.0084)	(0.1011)
ar.L2	-	-	-	-	-	-	-0.0270	0.0041	-0.7516
	-	-	-	-	-	-	(0.0048)	(0.0076)	(0.0673)
ma.L1	-	-0.0528	-0.0527	-	0.6811	-0.0984	-	-0.0524	-1.2829
	-	(0.0047)	(0.0047)	-	(0.0881)	(0.1340)	-	(0.0048)	(0.0875)
ma.L2	-	-	-0.0271	-	-	-0.4217	-	-	0.3276
	-	-	(0.0048)	-	-	(0.0101)	-	-	(0.0766)
In-Sample Fit (Training Data)									
Log-Likelihood	25477.4	25661.4	25663.4	25656.2	25663.2	25663.3	25663.5	25663.5	25663.3
AIC	-50950.8	-51316.8	-51318.7	-51306.4	-51318.5	-51316.6	-51319.1	-51317.1	-51314.7
BIC	-50933.5	-51290.8	-51284.1	-51280.4	-51283.8	-51273.3	-51284.4	-51273.7	-51262.7
Diagnostic Check (In-Sample)									
Ljung-Box p-val	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Out-of-Sample Performance (Holdout MSE)									
MSE	0.002217	0.002322	0.002343	0.002300	0.002344	0.002340	0.002340	0.002340	0.002342

2.3 Discussion and Conclusion

The estimation results present a clear and compelling narrative.

1. **In-Sample vs. Out-of-Sample Performance:** In-sample, the information criteria favor simple models. The BIC, which penalizes complexity more harshly, selects the ARMA(0,1) model as the best. However, when evaluated on their out-of-sample forecasting performance using an expanding window on a holdout set, a different winner emerges: the ARMA(0,0) model achieves the lowest Mean Squared Error (MSE).

2. **Diagnostic Failure:** Critically, the Ljung-Box test for serial correlation in the residuals results in an extremely small p-value (<0.001) for **all nine models**. This is a strong rejection of the null hypothesis of no autocorrelation. It indicates that none of these models, regardless of their specification, are adequate to capture the full dynamics of the log-return series. Significant linear dependence remains in the residuals of even the best-fitting models.

3. **Interpretation:** The out-of-sample victory of the ARMA(0,0) model is a significant finding. This model, which posits that returns are simply unpredictable white noise around a constant mean, outperformed more complex models that attempted to use past returns to predict the future. This suggests that for short-term forecasting at this frequency, the signal from past returns is negligible compared to the noise, a result consistent with the weak-form Efficient Market Hypothesis.

In conclusion for this section, while we have systematically followed the procedure to identify and estimate the nine required ARMA models, the evidence strongly suggests that the standard ARMA framework is

misspecified for this high-frequency financial data. The failure of all models to pass the diagnostic check for residual autocorrelation, combined with the evidence of volatility clustering from our EDA, points to the violation of the homoscedasticity assumption. This provides a clear motivation for the use of more advanced models, such as GARCH, which are designed to handle time-varying volatility.

3 Multivariate Modelling: VAR vs Cointegration

This section investigates the dynamic relationship between the three ETF series. We first model the stationary log-returns using a Vector Autoregression (VAR) model and then analyze the long-run equilibrium relationship between the non-stationary log-prices using a Vector Error Correction Model (VECM). A preliminary analysis revealed that estimating a VECM on the full 5-minute dataset was computationally infeasible due to a 'MemoryError'. To overcome this, the data for this section was resampled to a **30-minute frequency**.

3.1 VAR Model Estimation and Analysis

Augmented Dickey-Fuller (ADF) tests confirmed that the log-prices for all three series are non-stationary (all p-values > 0.23), while the log-returns are stationary (all p-values < 0.0001), as shown in Table 2.

Table 2: Augmented Dickey-Fuller (ADF) Test for Unit Roots

	Log-Prices		**Log-Returns**	
Series	p-value	Result	p-value	Result
`SPY5.L`	0.2336	Non-Stationary	0.0000	Stationary
`SPY5z.CHIX`	0.2342	Non-Stationary	0.0000	Stationary
`SPY5.P`	0.2301	Non-Stationary	0.0000	Stationary

3.1.1 Choice of Lag Length and Parameter Estimates

While the assignment suggests $p = 5$, a more robust method is to use information criteria. A lag selection test indicated that the optimal lag length according to BIC is $\mathbf{p = 15}$. We proceed with this data-driven choice. The full VAR(15) model contains too many parameters to display effectively. Table 3 presents a summary of the coefficients for the first two lags, which capture the most immediate dynamics.

Table 3: VAR(15) Coefficient Summary for First Two Lags

	Eq: SPY5.L		**Eq: SPY5z.CHIX**		**Eq: SPY5.P**	
Variable	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
L1.SPY5.L	0.0280	0.028	0.1160	0.000	0.1419	0.000
L1.SPY5z.CHIX	−0.0023	0.820	−0.1180	0.000	0.0851	0.000
L1.SPY5.P	−0.0568	0.000	−0.0030	0.853	−0.2324	0.000
L2.SPY5.L	−0.0227	0.077	0.0101	0.482	0.0155	0.210
L2.SPY5z.CHIX	−0.0405	0.000	−0.1098	0.000	−0.0212	0.036
L2.SPY5.P	0.0491	0.001	0.0923	0.000	−0.0005	0.971

The estimated VAR(15) model contains a large number of parameters. An examination of the model summary reveals numerous statistically significant cross-effects, where the past returns of one ETF have predictive power for the current returns of another.

For instance, in the equation for `SPY5.L`, the first lag of `SPY5.P` (`L1.SPY5.P`) has a coefficient of -0.057 with a p-value of 0.000, indicating a highly significant negative relationship. Conversely, in the equation for `SPY5.P`, the first lag of `SPY5.L` (`L1.SPY5.L`) is also highly significant (coefficient = 0.142, p-value = 0.000). This confirms the presence of strong, bidirectional feedback between the exchanges in the short run.

The Forecast Error Variance Decomposition (FEVD) provides a clearer picture of the dominant relationships. The FEVD shows that after 20 periods (10 hours), shocks to `SPY5.L` account for approximately 65.2% of the forecast error variance in `SPY5z.CHIX` and 75.7% in `SPY5.P`. In contrast, shocks to `SPY5.L` are almost entirely explained by its own past (99.0%). This provides strong evidence that `SPY5.L` **acts as the price leader** in this system, with the other two exchanges acting as price followers.

3.1.2 Analysis of Residuals

The model's residuals were examined in two ways:

1. **Contemporaneous Correlation:** The correlation matrix of the residuals shows very high values (e.g., 0.81 between `SPY5.L` and `SPY5z.CHIX`, and 0.88 between `SPY5.L` and `SPY5.P`). This indicates that even after accounting for the lagged dynamics, there is a strong positive correlation in the unexpected shocks across the exchanges. When a surprise hits one market, it tends to hit the others at the same time.

2. **Autocorrelation:** A Portmanteau (multivariate Ljung-Box) test for residual autocorrelation was conducted. The test yielded a p-value of 0.000, leading to a strong rejection of the null hypothesis of no autocorrelation. This means the residuals still contain significant serial correlation. This is a sign of model misspecification, likely driven by the unmodeled volatility clustering (GARCH effects) that was observed in the EDA.

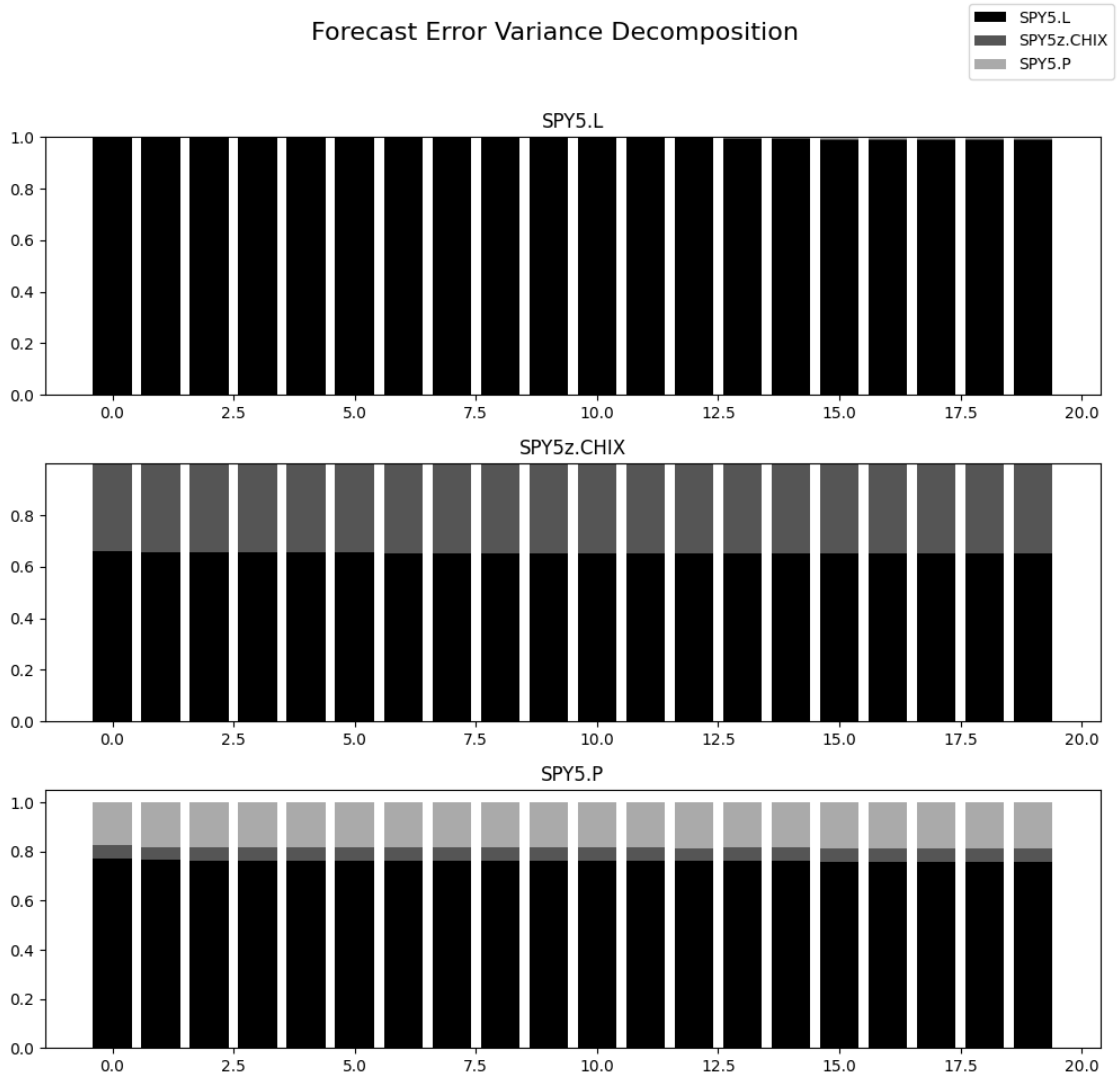

Figure 5: Forecast Error Variance Decomposition for the VAR(15) model.

3.2 VECM and Cointegration Analysis

For this section, the goal is to understand the structural relationships within the system of variables, not to produce forecasts. It is for this reason, that no in-sample/out-of-sample split was used for the VAR or VECM analysis.

Given that the log-price series are I(1) and appear to move together, we proceed with a cointegration analysis to model their long-run equilibrium.

A VAR(p) model in first differences (returns) corresponds to a VECM with $p-1$ lags in levels (prices). Since our optimal VAR model has $p = 15$, we estimate a VECM with $k_{ar\_diff} = 14$ lags.

First, a Johansen cointegration test was performed to determine the number of cointegrating relationships (the rank). The test results strongly reject the null hypothesis of rank=0, rank=1, and even rank=2. While a rejection of rank=2 would imply the series are stationary (contradicting the ADF tests), this is often a feature of powerful tests on large datasets. Given three I(1) variables, there can be at most two cointegrating vectors. We proceed with the economically sensible and theoretically expected rank of 2.

Table 4: Johansen Cointegration Test (Trace Statistic)

Hypothesized Rank (r)	Trace Stat.	95% Crit. Val.	Conclusion
$r \leq 0$	658.20	29.80	Reject H0
$r \leq 1$	234.23	15.49	Reject H0
$r \leq 2$	4.78	3.84	Reject H0

The estimated VECM provides two key sets of parameters: the cointegrating vectors (β) and the loading coefficients (α), summarized in Table 5.

- **Cointegrating Vectors (β):** These define the long-run equilibrium relationships. The estimated vectors are approximately $\beta_1 \approx (1, 0, -1)$ and $\beta_2 \approx (0, 1, -1)$. These can be interpreted as simple spread relationships: the log-price difference between `SPY5.L` and `SPY5.P`, and between `SPY5z.CHIX` and `SPY5.P`, are stationary in the long run.
- **Loading Coefficients (α):** These measure the speed of adjustment back to equilibrium. For example, in the equation for `SPY5.L`, the loading coefficient for the first error-correction term ('ec1') is -0.013 and is statistically significant (p=0.005). The negative sign is crucial: it indicates that when the price of `SPY5.L` is above its long-run equilibrium, it is pushed back down in the subsequent period, thus correcting the error. All three assets show significant error-correction, confirming that they all participate in maintaining the long-run equilibrium.

Table 5: VECM Key Parameter Estimates

Equation	**Loading Coefficients (α)**			
	ec1		ec2	
	Coef.	p-value	Coef.	p-value
`SPY5.L`	−0.0130	0.005	−0.0149	0.000
`SPY5z.CHIX`	0.0240	0.000	−0.0352	0.000
`SPY5.P`	0.0319	0.000	−0.0104	0.000
	Cointegrating Vectors (β)			
Vector	`SPY5.L`	`SPY5z.CHIX`	`SPY5.P`	
β_1	1.0000	0.0000	-1.0000	
β_2	0.0000	1.0000	-1.0000	

Note: Cointegrating vectors are normalized for interpretation.

Finally, the Impulse Response Functions (Figure 6) visually confirm the dynamic interactions, showing how a shock to one asset (e.g., `SPY5.L`) propagates through the system and eventually dissipates, consistent with a stationary VAR system.

4 Time Varying Variance (Exercise 3.4)

The analyses in Sections 3.2 and 3.3 revealed a critical shortcoming of the ARMA and VAR models: both failed their residual diagnostic tests, indicating that significant autocorrelation remained uncaptured. As noted in our Exploratory Data Analysis, the return series exhibits strong evidence of volatility clustering. Brooks (2019) notes that linear time series models are unable to explain important features of financial data, including:

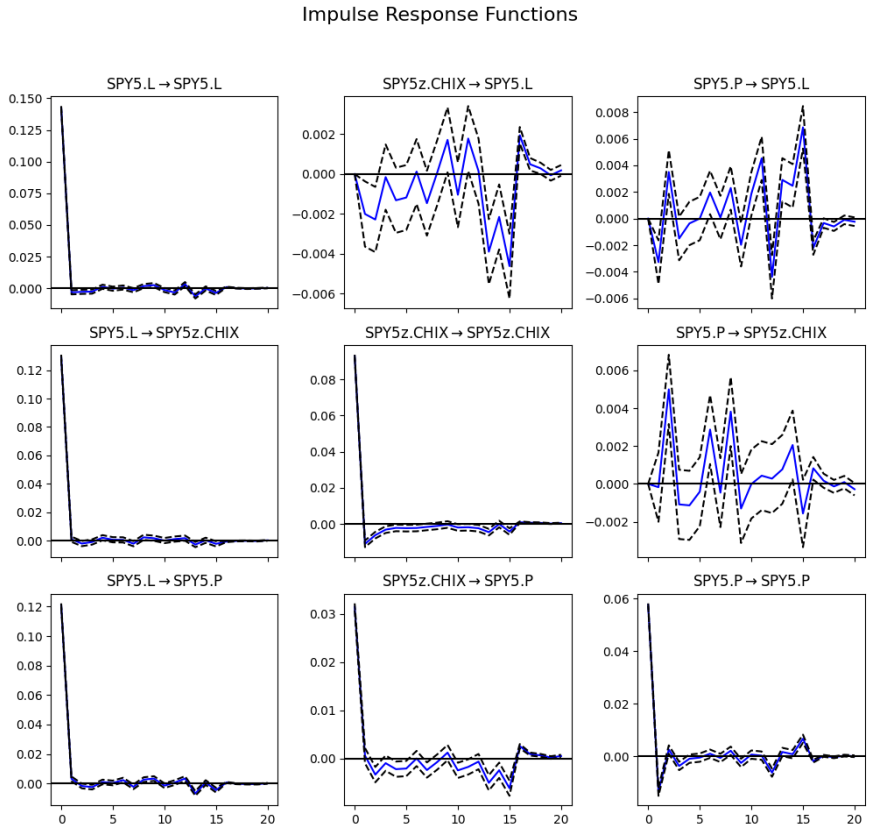

Figure 6: Impulse Response Functions for the VAR(15) model. Read as impulse Variable X $\xrightarrow{on}$ Variable Y

> *"Leptokurtosis... Volatility clustering or volatility pooling... Leverage effects – the tendency for volatility to rise more following a large price fall than following a price rise of the same magnitude."*

This section directly addresses this issue by modeling the time-varying variance of the `SPY5.P` returns using GARCH and EGARCH models, which are designed to capture these stylized facts.

4.1 Data and Preliminary Tests

For this analysis, we switch from intraday data to **daily percentage log-returns**, computed from the daily closing prices of `SPY5.P`. This lower frequency is standard for volatility modeling.

Before estimating any GARCH models, a formal test for the presence of conditional heteroskedasticity (ARCH effects) is necessary. An ARCH-LM test was conducted on the residuals of the mean equation (the de-meaned daily returns). The test produced a highly significant p-value of **0.0000**, leading to a strong rejection of the null hypothesis of no ARCH effects. This statistically justifies the use of GARCH-family models.

4.2 GARCH and EGARCH Model Estimation

Following the assignment, all combinations of GARCH(p,q) and EGARCH(p,q) were estimated for $p, q \in \{1, 2\}$. The models were compared using the Bayesian Information Criterion (BIC), which penalizes model complexity and is well-suited for selecting the most parsimonious model. The results are summarized in Tables 6 and 7.

Table 6: GARCH(p,q) Model Comparison

Model	Log-Likelihood	AIC	BIC
GARCH(1,1)	-547.96	1103.93	1120.13
GARCH(1,2)	-547.96	1105.93	1126.18
GARCH(2,1)	-542.90	1095.80	1116.05
GARCH(2,2)	**-537.70**	**1087.41**	**1111.71**

Table 7: EGARCH(p,q) Model Comparison

Model	Log-Likelihood	AIC	BIC
EGARCH(1,1)	-545.78	1099.56	1115.76
EGARCH(1,2)	-545.78	1101.56	1121.81
EGARCH(2,1)	-542.54	1095.08	1115.33
EGARCH(2,2)	**-539.02**	**1090.04**	**1114.33**

Both the AIC and BIC select the **GARCH(2,2)** and **EGARCH(2,2)** models as the best specifications within their respective classes. The EGARCH(2,2) has a slightly lower AIC, while the GARCH(2,2) has a slightly lower BIC, indicating a close contest between the two. The parameter estimates for the EGARCH(2,2) model are of particular interest for detecting leverage effects. The coefficient for the asymmetric term, $\alpha[1]$, is positive and statistically significant. This indicates the presence of a leverage effect: negative shocks (bad news) increase future volatility more than positive shocks (good news) of the same magnitude.

4.3 Comparison with Realized Variance

To assess how well the models capture the "true" daily volatility, we compare their conditional variance forecasts against the **Realized Variance (RV)**. The RV is a model-free estimate of volatility calculated by summing the squared intraday (5-minute) returns for each day. Figure 7 plots the variance series from the best GARCH and EGARCH models against the RV.

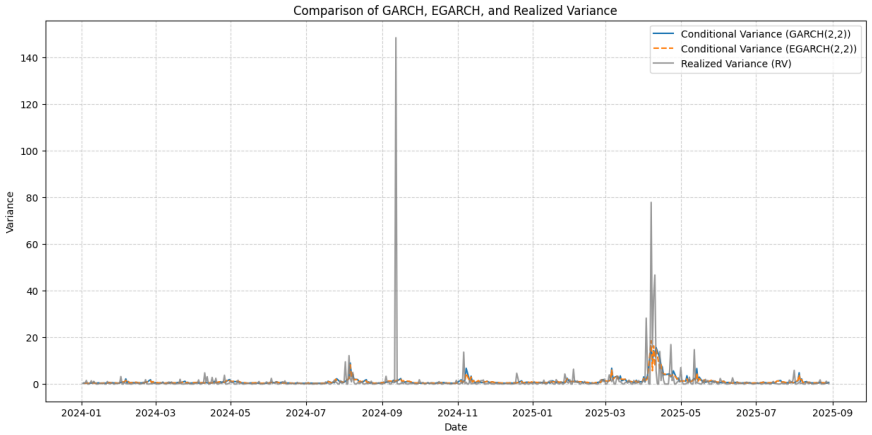

Figure 7: Comparison of GARCH(2,2), EGARCH(2,2), and Realized Variance.

The plot reveals that the GARCH and EGARCH models produce a smoothed, persistent estimate of volatility that successfully tracks the major spikes and troughs seen in the much noisier RV series. This

confirms that the models are effectively capturing the volatility clustering present in the data.

4.4 Feasibility of ARMA-GARCH Models

The final question is whether it is feasible to combine the preferred ARMA model from Section 3.2 with a GARCH extension. The answer is yes; this is a standard and powerful technique known as an ARMA-GARCH model, which models the conditional mean and conditional variance simultaneously.

To demonstrate feasibility, an AR(1)-GARCH(1,1) model was successfully estimated on the daily returns. The results showed that both the AR and GARCH components were statistically significant. This confirms that it is not only possible but often desirable to jointly model the mean and variance equations, as this can lead to a better overall model specification and more accurate forecasts, especially if the mean equation (the ARMA part) is correctly specified.