Assignment 2: Time series data
Econometrics for Quantitative Finance

C.S. Bos

2025/9/21

1 Introduction

With high frequency data, oftentimes there are few if any explanatory variables which could help in explaining the movements of e.g. stock prices. Instead, most of the information for predicting time series data oftentimes comes from the time series itself, or from 'similar' time series which are available at the same frequency.

To investigate this, you are asked in this assignment to take a closer look at high frequency prices of Exchange Traded Funds, in case the SPDR S&P 500 UCITS ETF as provided by State Street. This ETF trades at a number of different exchanges. Price data is made available for you at Canvas, courtesy of ETFbook.

2 Data and setup

Data provided at Canvas consists of the minute by minute average bid/ask midprice for a range of European exchanges. You are asked to use 3 of the symbols listed in the datasets, disregarding the other data.

Data is combined into monthly files, with timing indicated in Central European Standard Time, and all prices translated into Euro, for easier comparison. In total, data is available from all trading days from 2024-01 until 2025-08.

3 Exercise

3.1 Data preparation

The data download comes with an outline of a program `prepdata_sp500_outline.py`. Adapt this outline to ensure that you read the data for your group, combine all monthly data files together for only your symbols, and store the output.

At time of handing in your report + programs, ensure that you hand in this `prepdata` file, do **not** hand in the underlying nor extracted data (as I should be able to recreate your exact data easily with your prepdata program).

Alternatively, only if you prepare your data by hand/in some other language, do hand in the prepared data + preparation code.

3.2 ARMA modelling

Read your prepared ETF prices, and transform a single series of your own choice into percentage log-returns. On the log-returns, perform an investigation into the autocorrelation of the series: What shape of the autocorrelation function do you see? What ARMA models would you deem reasonable?

Afterwards, confirm your findings, by estimating an ARMA(p, q) for each combination of $p = 0, 1, 2$ and $q = 0, 1, 2$, summarising the output into a single table. The table should contain:

1. columns for the 9 different models
2. with parameter estimates in the rows, including standard errors, estimated over the first year of the data only
3. below the estimates, a list of (at least) the mean squared error (MSE), the log-likelihood, the AIC, and e.g. the Ljung-Box statistic of the residuals; give these statistics both for the in-sample and the out-of-sample period

In the discussion, compare the results from the table with your ACF analysis. Discuss as well your choice for the standard errors reported: Did you use the 'standard' standard errors, or some other more robust version? Why, what is the difference?

Note that this exercise, on purpose, again leaves you to solve some open ends and surprises. You may run into convergence issues, you may find possibly disappointing results, it is possible you have trouble dealing with missing data or outliers. This is to be expected; your task is to show in the report how you address such problems.

3.3 Multivariate modelling: VAR vs Cointegration

The SPDR S&P 500 is traded at multiple exchanges. Clearly, the returns on those exchanges should be related, as the prices at the exchanges should be relatively similar. Hence, take your series and estimate a VAR(p) model, of lag-length $p = 5$. Create a clear table for the parameter estimates, with standard errors and further statistics of interest, and discuss, e.g.

1. Is this value of $p = 5$ a good choice? What could be better in your case?
2. What parameters seem to be significant? Do you see any cross-effects in the VAR model? Where, in what direction, what exchange seems to be driving what other exchange?
3. What can you say of the residuals? Do they look clean? Is there a way you can quickly look if the correlation between the two series of residuals seems zero?
4. What in-sample/out-of-sample split did you use? Is there evidence of a change of behaviour?

Possibly there are other topics to discuss: Feel free to add to the list.

But of course, while modelling the log-returns brings a first set of insights, in the end one would prefer to model the log-prices of ETFs together. Hence, take the logarithm of the prices instead: What VECM or cointegration model on log-prices would correspond to the VAR(5) you have estimated? Can you obtain results here? Provide a table of estimates of the VECM model, and discuss the relationship to the parameters found for the VAR model.

3.4 Time varying variance

In Section 3.2, you estimated an ARMA model, assuming that the variance over time is constant. Of course, this assumption is wrong for most financial data.

For this part of the exercise, extract the daily closing prices of your asset of choice, and compute the *daily* percentage log returns. On these returns, estimate both the GARCH(p, q) and EGARCH(p, q), for all combinations of $p = 1, 2; q = 1, 2$, and compare the 8 sets of parameter estimates. As your sample is now relatively small, feel free to take the full sample for estimation (or download a longer series from e.g. Yahoo Finance).

Discuss the results, and extract the variance estimate for the best GARCH and the best EGARCH model. Also, using the intraday log-returns, construct the daily realized variation

$$\hat{\sigma}^2_{RV,t} = \sum_i r^2_{it},$$

(that is: For each day, sum the squared intraday returns). Plot together the $\hat{\sigma}^2_{\text{GARCH}}, \hat{\sigma}^2_{\text{EGARCH}}, \hat{\sigma}^2_{\text{RV}}$. What do you see?

In 3.2, you may have ended up with an AR(1) or MA(1), or even a combination of the two. Is it feasible to estimate your preferred model of Section 3.2 with a GARCH or EGARCH extension? Why, why not, what results do you see/would you expect?

4 On the report

Hints:

- Write a report with your findings; try to find the right equilibrium between detail in the matters of interest, and a succinct writing style for the remainder. However, do not let me guess as to the decisions you have had to make.
- As mentioned before: On purpose there are loose ends in this assignment, for which you have to find a practical solution yourself. It may not be possible to solve each problem, then describe where you got stuck and why.
- Do NOT use screenshots, translate your results to true LaTeX/Word tables
- Think a bit about the presentation: What results should be combined with what other results?

- You are free to use whatever language/environment/packages you prefer, as long as I am able to check your results. If you use Python, send me code + data. If you use another language/package (e.g. Stata or Eviews), send me the code (if there is any) plus (separate from the report) sufficient screenshots to see the results.
- Submit a PDF (not `.docx`!) of the report, plus a zip-file (or similar) with working code + data/screenshots, as a group. Exclude the underlying raw data, but do include your prepdata routine, and the outcome of it.
- Note that you earn points by discussing sensibly what you found, showing that you compute purposefully the interesting elements of the question.