Time Series and Forecasting

Empirical Project

Riccardo Sturla 13457A, Riccardo Tenuta 26940A

A.Y. 2023/2024

Contents

Introduction

The goal of the project is to implement different models to forecast the US Real Gross Domestic Product. A total of six models has been tested and their performance are summarized in Section 4. The dataset used is the FRED-QD dataset, a large database of US macroeconomic data provided by the Federal Reserve of Economic Data, containing hundreds of economic time series.

1 Section 1

In this section, we plot and shortly describe all the series used for the subsequent empirical analyses: y_t, Δy_t, $\log PCECTPI_t$, π_t, $Tspread_t$.

- $y_t = LogGDP$: logarithm of US Real Gross Domestic Product in Billions of Chained 2012 Dollars.

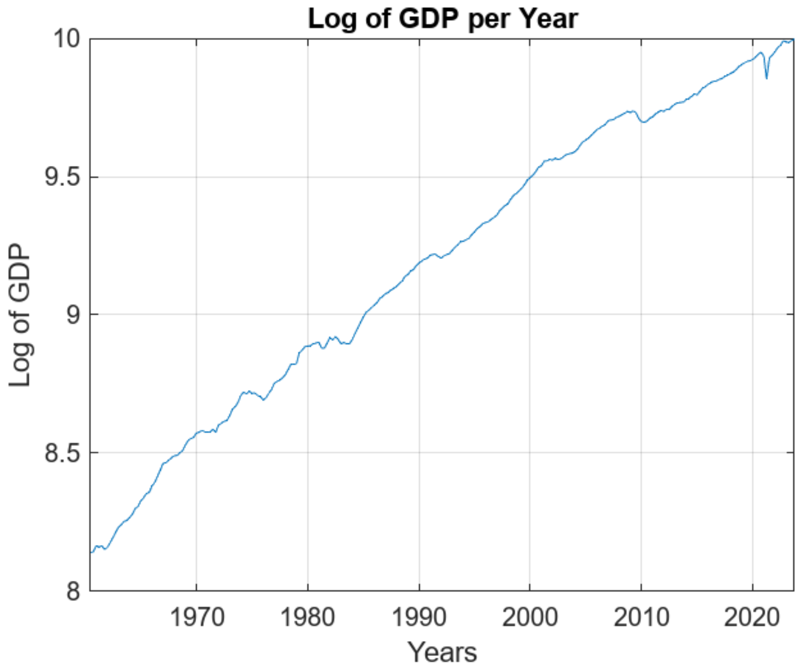

Figure 1: US Log GDP

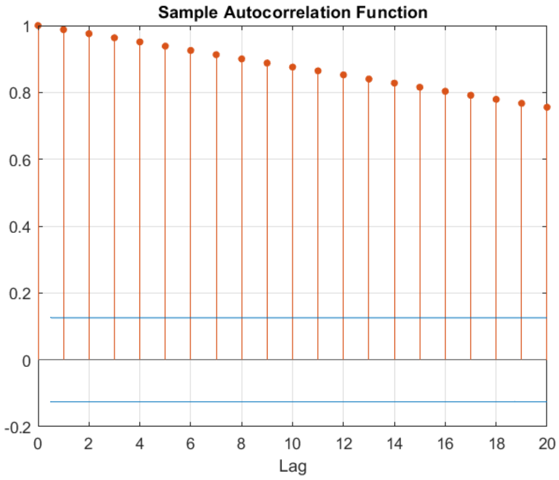

Figure 2: ACF logGDP

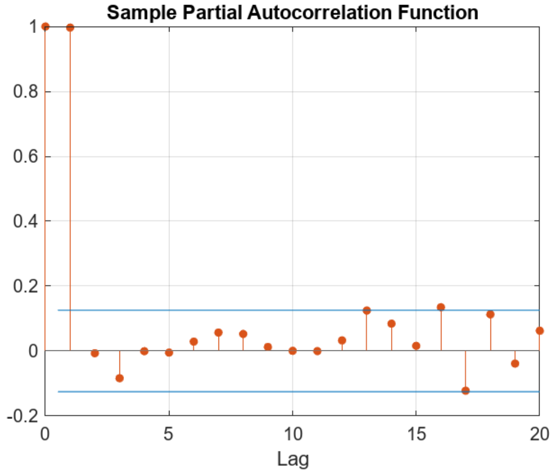

Figure 3: PACF logGDP

The GDP series is obviously very persistent with a strong trend and cycles. The PACF in Figure 3 shows a single significative spike at lag 1.

- Δy_t: first difference of US Real Gross Domestic Product.

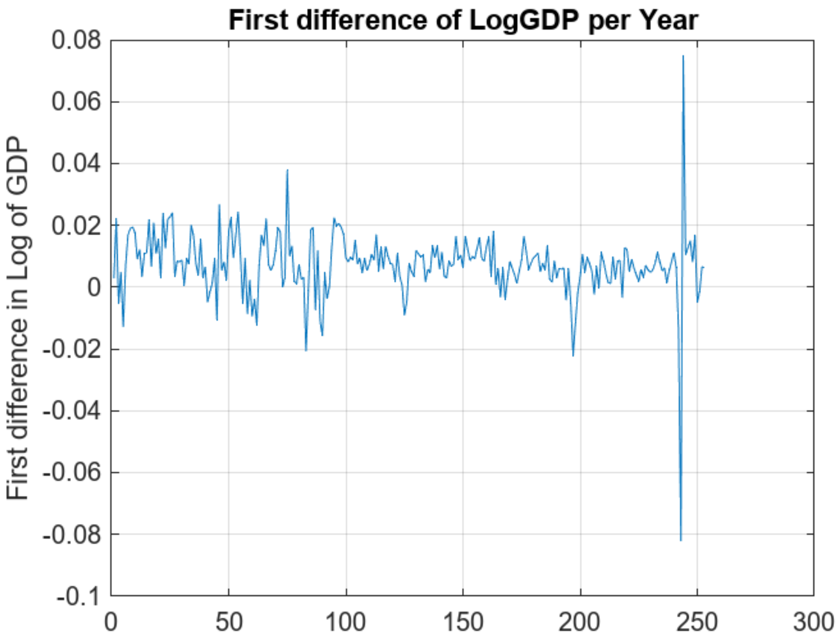

Figure 4: First difference of LogGDP

Taking the first difference of logGDP removes the trend and makes the series mean reverting and close to stationarity.

- Log of PCE: logarithm of Real Personal Consumption Expenditures in Billions of Chained 2012 Dollars.

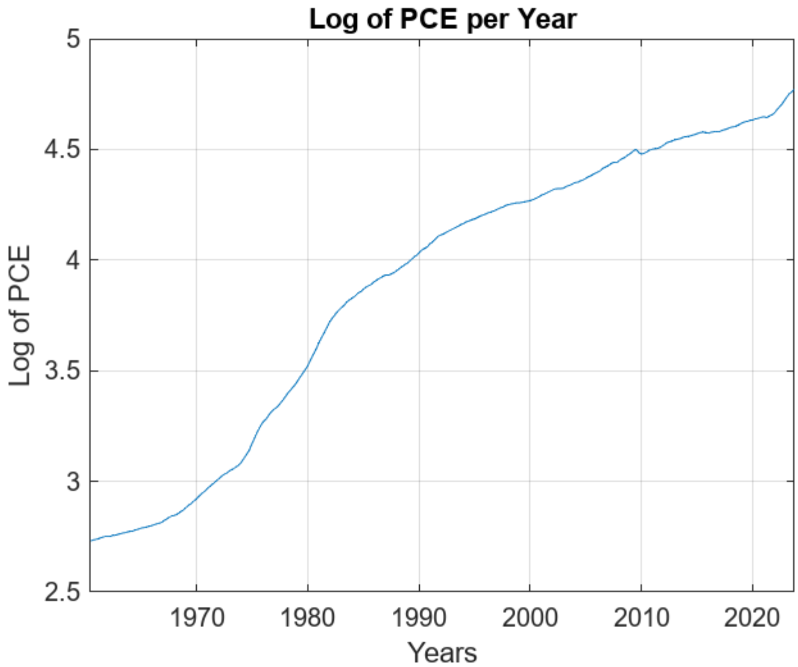

Figure 5: Log of PCE

The series in log displays a clear ascending trend.

- $\Delta LogPCE$: first difference of logarithm of PCE.

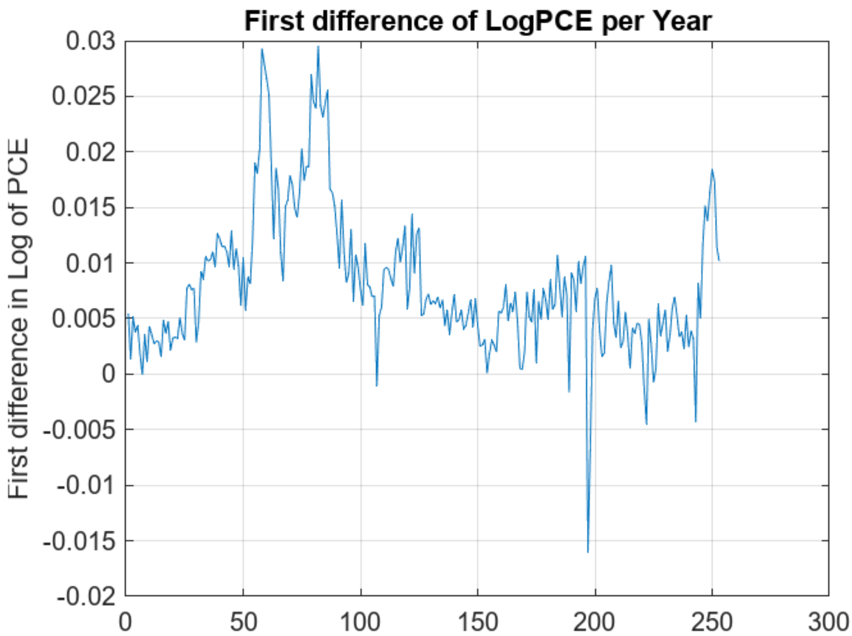

Figure 6: First difference of LogPCE

Taking the first difference of the series removes the trend, but leaves a lot of variability.

- T spread: defined as the difference between 10-Year Treasury Constant Maturity Rate (Percent) and 3-Month Treasury Bill (Secondary Market Rate, Percent).

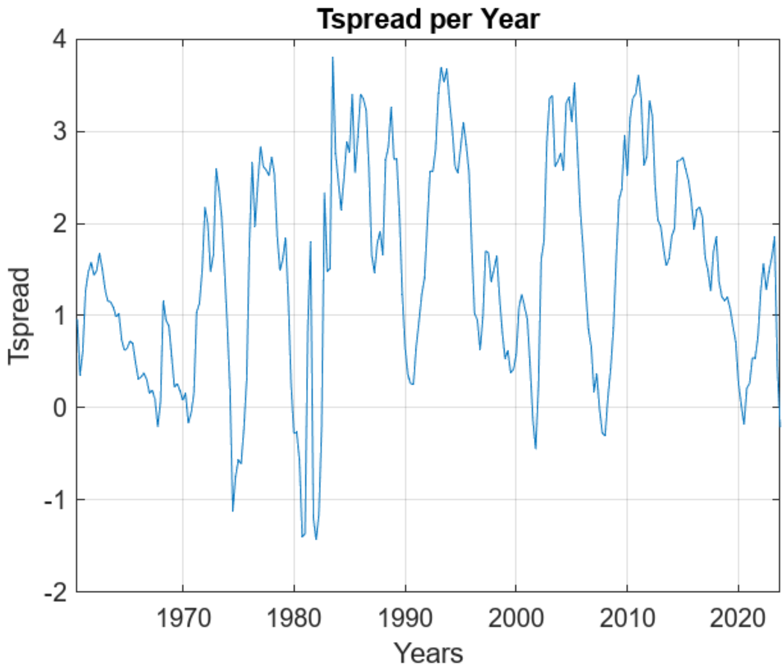

Figure 7: T spread

The series shows some cycles even if it fluctuates around a mean value.

2 Section 2

In this section we consider an hypothetical VAR(p) model with three variables: Δy_t, a measure of inflation π_t, and the term spread define before *T spread*. We analyse the series, plotting the corresponding autocorrelation functions, and we then try to select the best lag order p for the model and estimate it.

2.1 Time Series entering the VAR model

We plot the sample Autocorrelation (ACF) functions of the series entering the VAR(p) model.

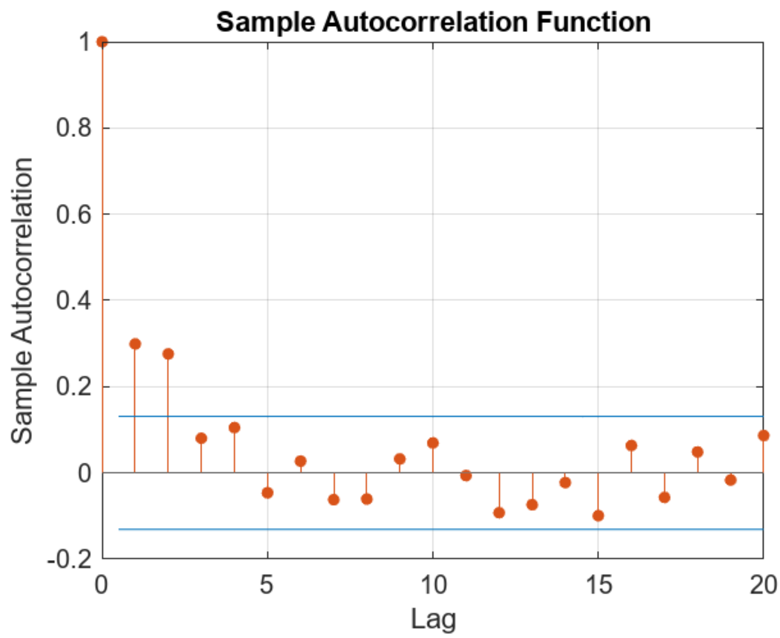

Figure 8: ACF Δy_t

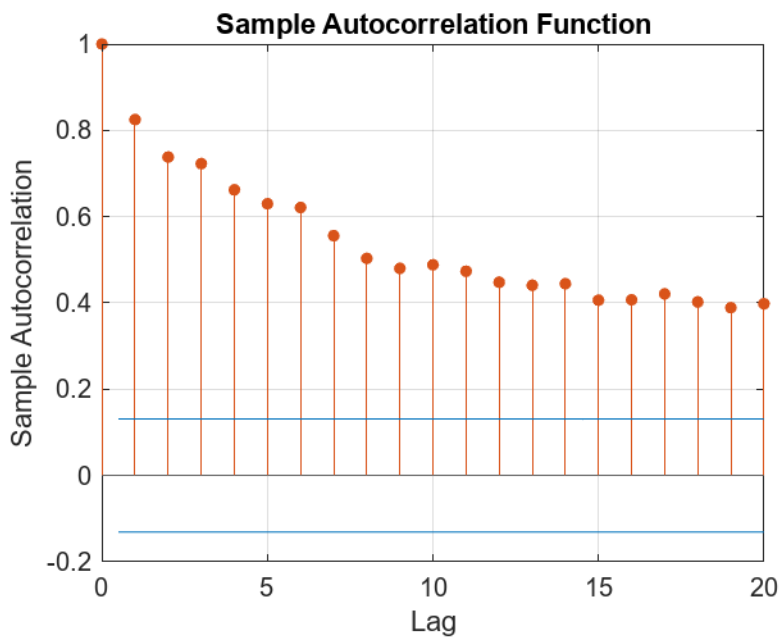

Figure 9: ACF π_t

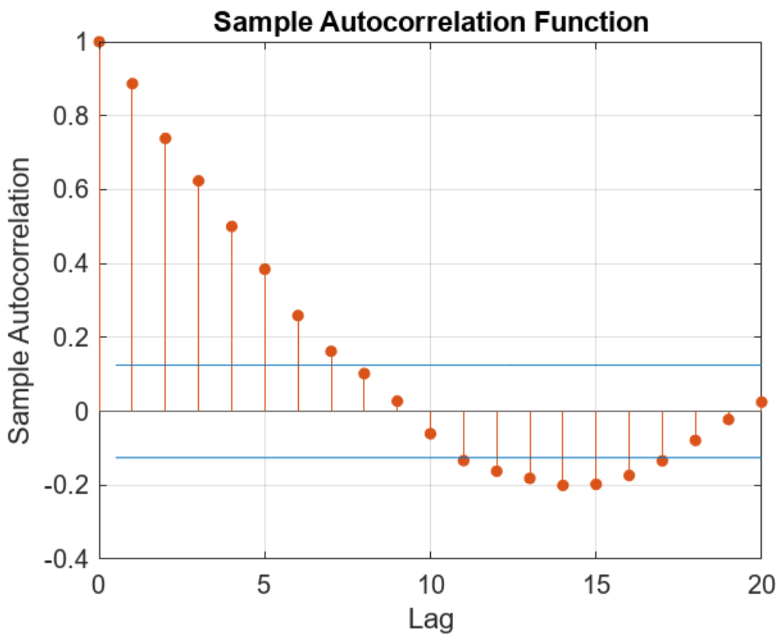

Figure 10: ACF T spread

Running the Augmented Dickey-Fuller test on the three series, the results indicate that the test fails to reject the null hypothesis of a unit root against the autoregressive alternative; in other words there is sufficient evidence to suggest that the series are trend stationary.

2.2 Model selection with AIC

The Akaike Information Criteria is computed estimating 12 VAR(p) models, using lag values (p) from 1 to 12 and looking at the correspondent AIC. The results are showed in the following scree plot.

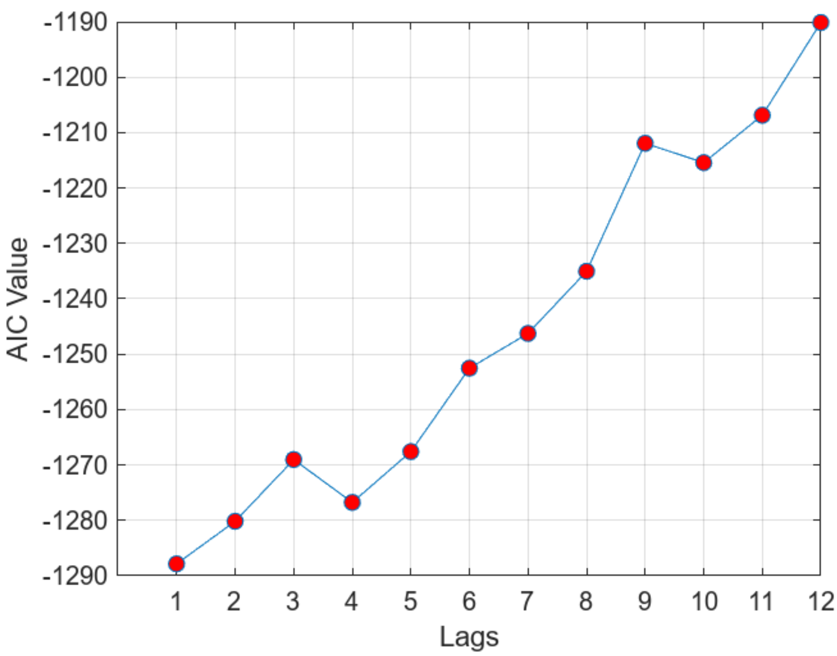

Figure 11: Scree plot for AIC values

The lag order p which minimize the AIC is 1. We can conclude that VAR(1) is the best model to minimize the trade off between fit and complexity according to AIC and should perform better than the suggested VAR(4).

2.3 Model estimation

The resulting VAR(1) model is estimated using the following equation:

$$\begin{bmatrix} \Delta y_t \\ \pi_t \\ T_t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_{1,11} & a_{1,12} & a_{1,13} \\ a_{1,21} & a_{1,22} & a_{1,23} \\ a_{1,31} & a_{1,32} & a_{1,33} \end{bmatrix} \begin{bmatrix} \Delta y_{t-1} \\ \pi_{t-1} \\ T_{t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix}$$

where c_1, c_2, c_3 are the intercepts and $a_{i,j}$ are the coefficients associated with the lagged variables.

2.4 Residuals and their Autocorrelation

We procede to plot the residuals of the three series and their sample ACF and Partial ACF.

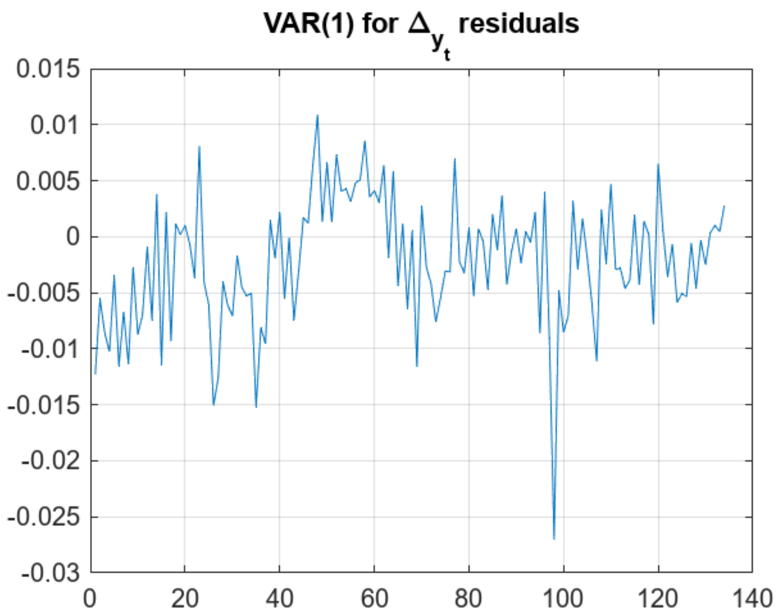

Figure 12: Residuals of Δy_t

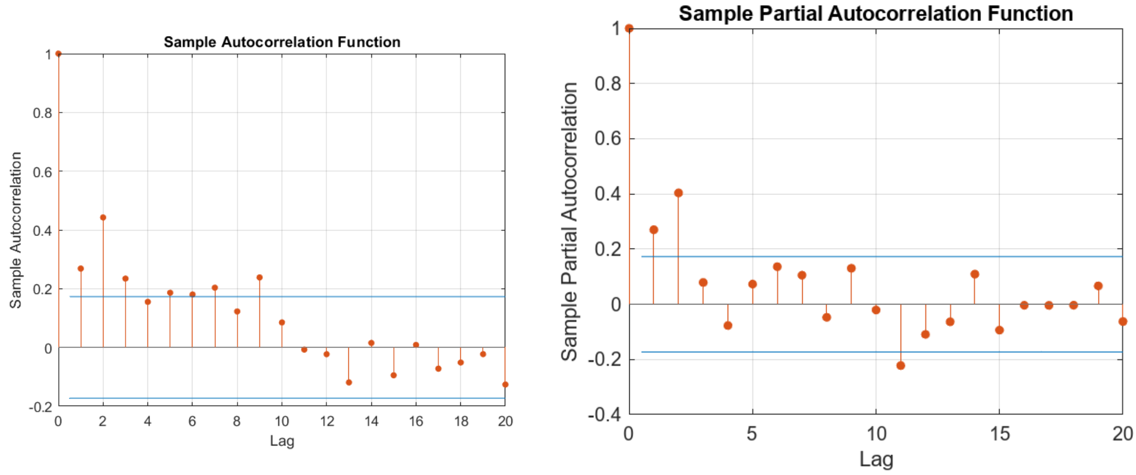

Figure 13: Residuals sample ACF (left) and PACF (right)

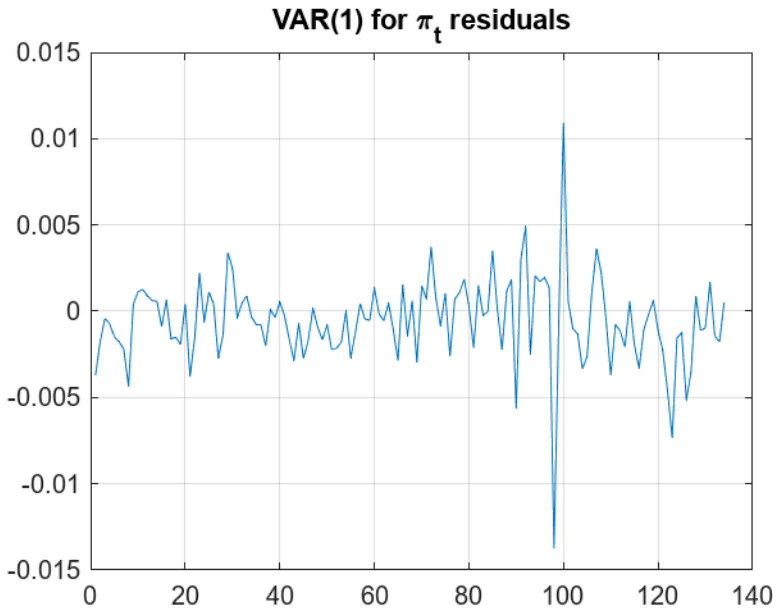

Figure 14: Residuals of π_t

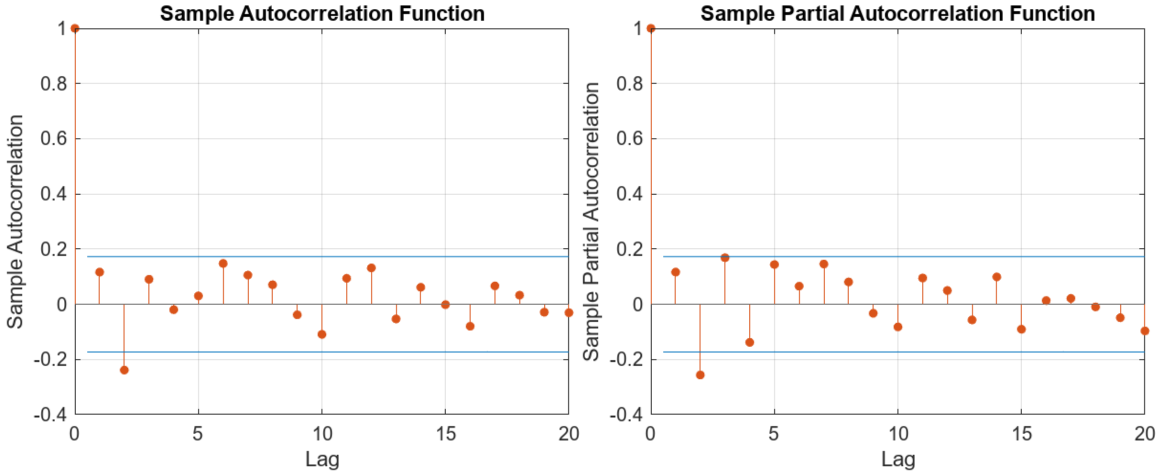

Figure 15: Residuals sample ACF (left) and PACF (right)

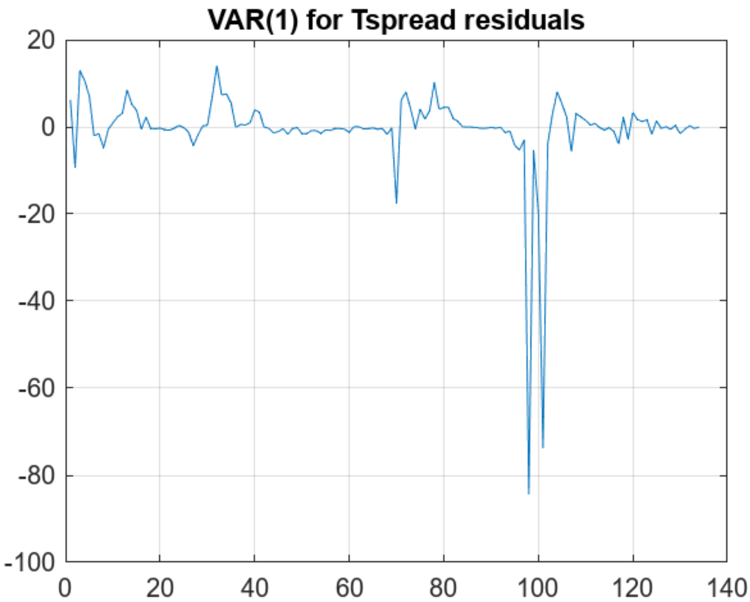

Figure 16: Residuals of T spread

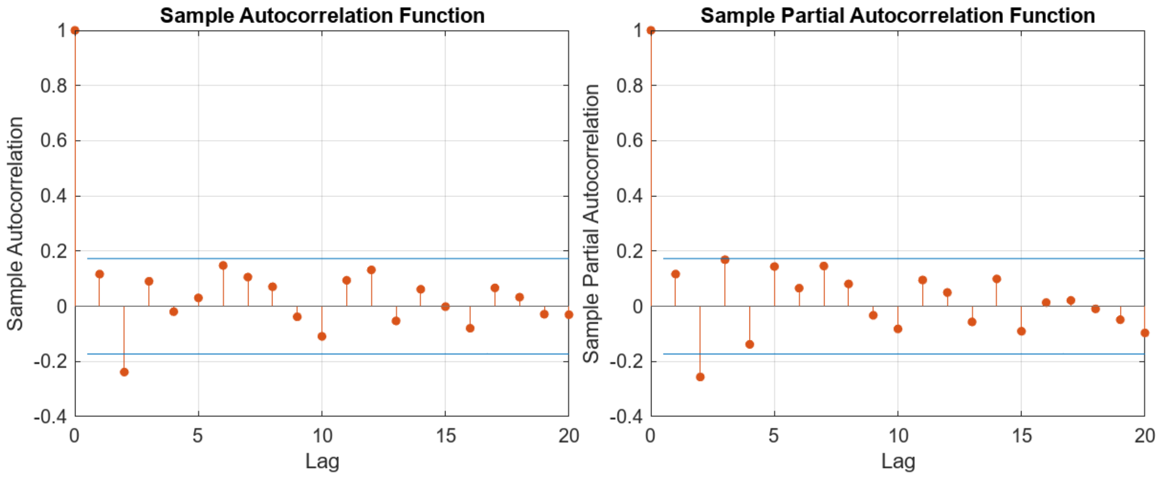

Figure 17: Residuals sample ACF (left) and PACF (right)

The residuals seems to be close to a 0 mean. Moreover, they are uncorrelated, except for the first 2 lags of Δy_t.

3 Section 3

In this section we compute PCA on the "Factors" table in the dataset, in order to use the first one as a Factor F for an AR-X model.

3.1 PCs and explained variance

We estimate the principles components both on the first 100 observation and on the entire set.

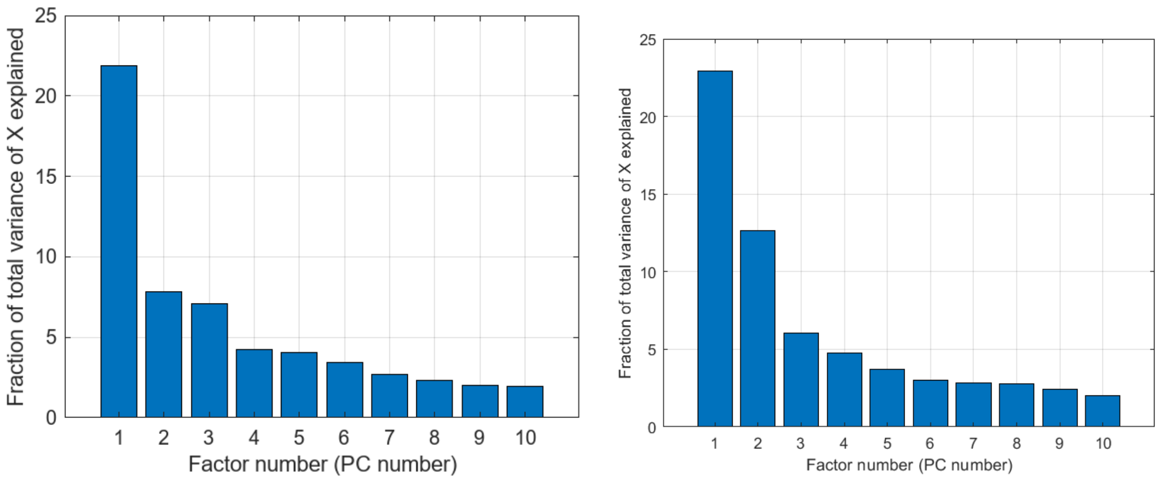

Figure 18: PCs for all the dataset (left) and for the first 100 observations (right)

In both cases the first principal component explains more than 20% of the total variance of X, followed by the second PC with only 12.6% (on first 100 observations) and about 7.5% (on all observations).

3.2 Series in the first PC

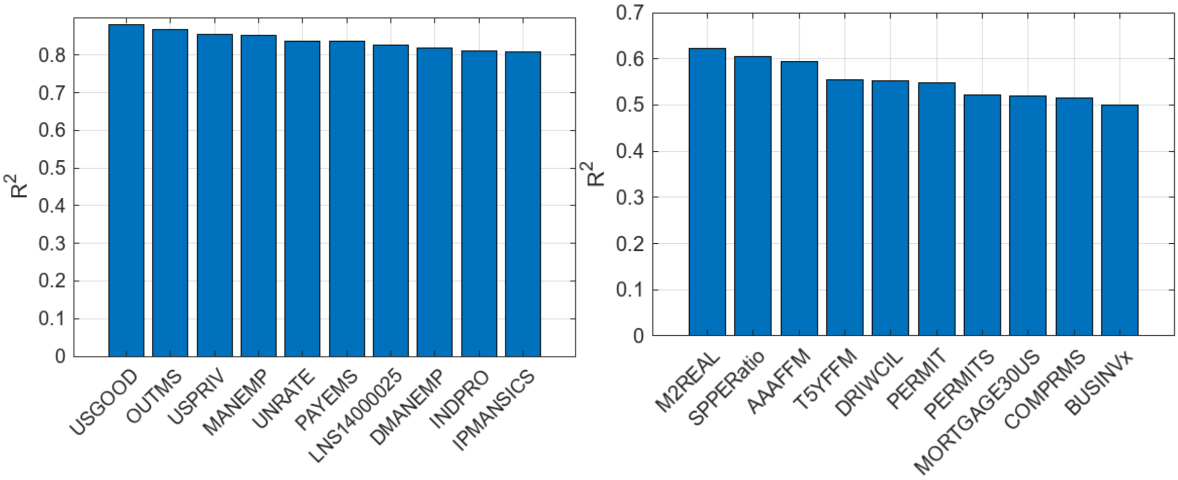

Figure 19: Series in first PC and their R^2

In the two figures, it is possible to look at some of the most influent series in the first Principal Component, ordered by descending R^2 obtained regressing F on each of them.

4 Section 4

In this section we formally define all the models used to forecast real GDP, estimating them and providing a plot for each forecast against real GDP. Finally, we summarize the results in Table 1 with all the models RMSE.

4.1 Models

- Random Walk (without drift):

$$y_t = y_{t-1} + \varepsilon_t \quad \varepsilon_t \sim WN(0, \sigma^2)$$

- AR(4) for Δy_t:

$$\Delta y_t = c + \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \phi_3 \Delta y_{t-3} + \phi_4 \Delta y_{t-4} + \varepsilon_t$$

  Δy_t is the first difference of the time series at time t, while ε_t is the white noise error term at time t. The coefficients $\phi_1, \phi_2, \phi_3, \phi_4$ represent the autoregressive parameters of the model, indicating the influence of the first four lagged differences on the current difference.

- VAR(4) for Δy_t, the measure of inflation $\pi_t = log(PCECTPI)_t - log(PCECTPI)_{t-1}$ and the term spread defined as $Tspread_t = GS10_t - TB3MS_t$.

Since we have 3 time series with 4 lags each, the model is built on this equation of 39 parameters:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_{1,11} & a_{1,12} & a_{1,13} \\ a_{1,21} & a_{1,22} & a_{1,23} \\ a_{1,31} & a_{1,32} & a_{1,33} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{bmatrix} + \begin{bmatrix} a_{2,11} & a_{2,12} & a_{2,13} \\ a_{2,21} & a_{2,22} & a_{2,23} \\ a_{2,31} & a_{2,32} & a_{2,33} \end{bmatrix} \begin{bmatrix} y_{1t-2} \\ y_{2t-2} \\ y_{3t-2} \end{bmatrix} + ... + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix}$$

where y_{1t}, y_{2t} and y_{3t} are respectively Δy_t, π_t and $Tspread_t$.

We omitted the third and fourth lags in the equation to have a clearer view of the model.

- VAR(1):

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_{1,11} & a_{1,12} & a_{1,13} \\ a_{1,21} & a_{1,22} & a_{1,23} \\ a_{1,31} & a_{1,32} & a_{1,33} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix}$$

- AR(4)-X for Δy:

$$\Delta y_t = a + \sum_{j=1}^{4} b_j \Delta y_{t-j} + c\hat{F}_{1t-1} + \varepsilon_t$$

- Bonus model:

  1. AR(2) on Δy_t:

  $$\Delta y_t = c + \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \varepsilon_t$$

  Looking at the GDP partial autocorrelation function in Figure 3, we can see only one significative spike at lag 1. That could be the sign that a simple AR(1) model would fit well to the series. However, the ARIMA function estimate the model on the first difference of y_t, so, observing Figure 8, we can try to fit an AR(2) model on Δy_t using ARIMA(2,1,0).

All models are estimated using a rolling window of 100 observations. The first forecast origin is R = 1985 : Q2, and the last forecast origin is T = 2018 : Q3, for a total of P = 134 forecasts.

4.2 Forecasts

After estimating the models, we produced the forecasts. The following figures show the predictions compared to the real values and growth of US GDP.

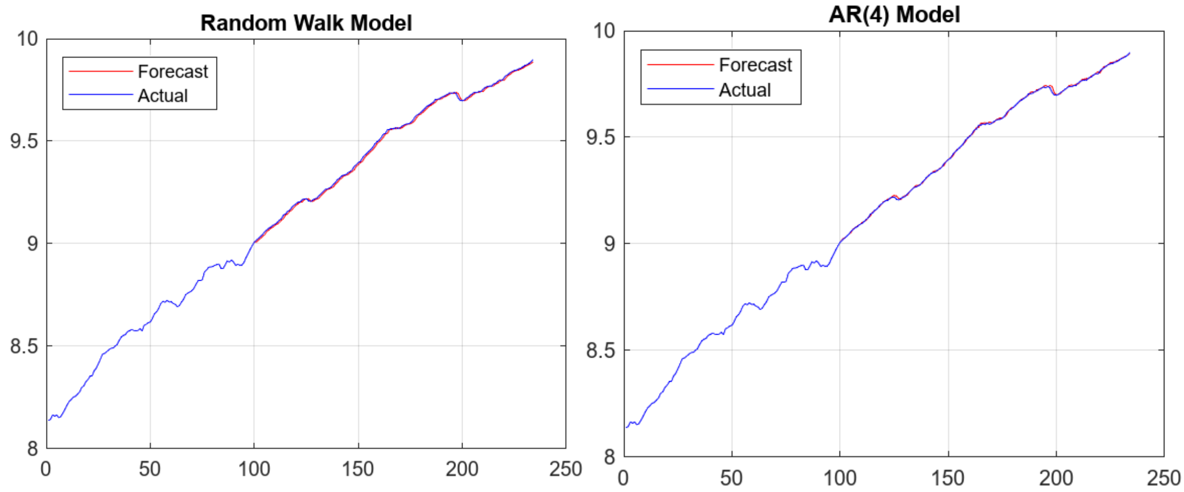

Figure 20: Forecasts VS Actual of GDP for Random Walk and AR(4) models

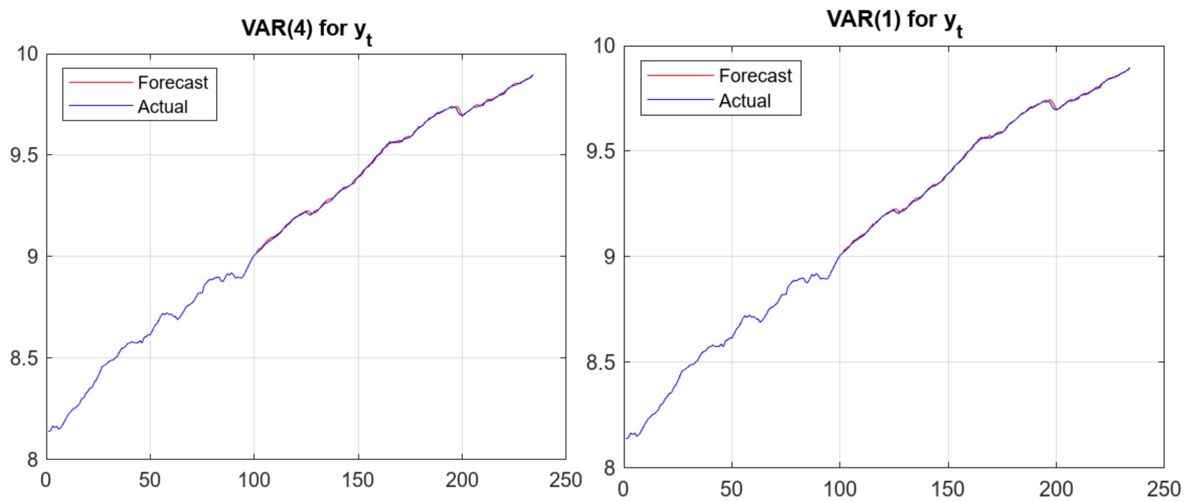

Figure 21: Forecasts VS Actual of GDP for VAR(4) and VAR(1) models

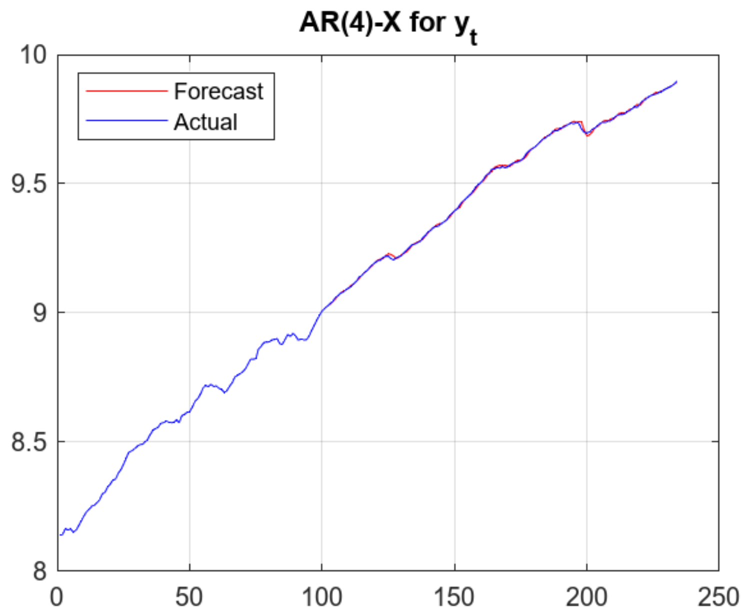

Figure 22: Forecasts VS Actual of GDP for AR(4)-X model

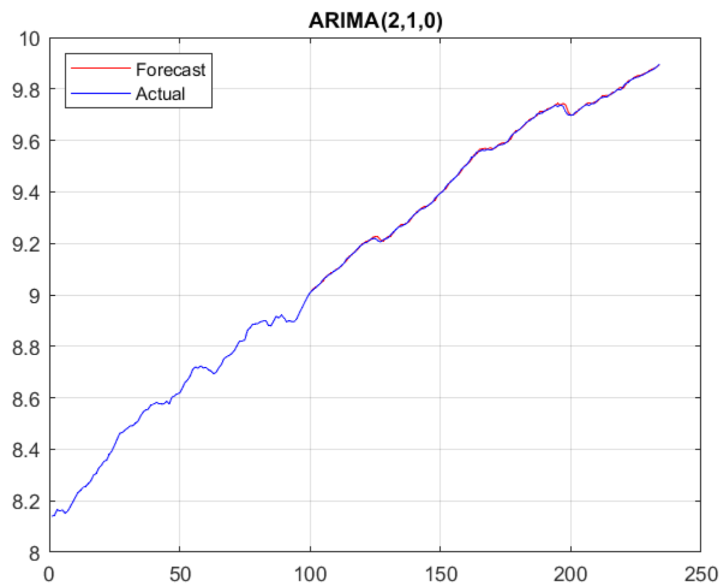

Figure 23: Forecasts VS Actual of GDP for AR(2), the bonus model

4.3 Final results

In conclusion, after estimating, forecasting and computing the residuals for all the models, each Root Mean Squared Error is displayed in the following table:

Model	RMSE
Random Walk	119.44
AR(4)	80.38
VAR(4)	93.34
VAR(1)	84.99
AR(4)-X	84.76
AR(2)	79.57

Table 1: RMSE for all the models

As we can see in Table 1, all the models were able to largely beat the RW benchmark model. VAR(1), as expected, performed better than VAR(4); AR(4)-X has a higher RMSE than the AR(4), probably due to its complexity. However, the best model to forecast the US GDP is the Autoregressive model of order two on Δy_t. As said before, it's the best AR model due to the fact that in the correlogram of the autocorrelation function the first two lags have a significative impact on the value of the real GDP. Furthermore, the AR(2) model is very simple and has only four parameters to estimate, which means that increasing the complexity and the number of coefficients and parameters of the models is not always the best way to reach a high accuracy in predictions.