MSDS 604 Time Series Analysis

Shan Wang

Fall 2023 Module 2

Homework 5 Due Thursday Nov 30 11:59pm

Introduction: submit one .pdf file containing all you answers for the homework, including screenshot of python code, output or plot if asked. The .pdf can be converted from Latex file, pictures of your handwriting solutions, word files, markdown files and .etc (anything that can be converted into .pdf). If there are coding problems, only include the answer of the question in the .pdf, and upload a separate notebook for Python code. This homework requires the submission of both pdf file and python notebook.

1. **In this question, you will acquainted with fitting data with SARIMA.** The data beer.csv were collected in São Paulo, Brazil in 2015, in a university area, where there are some parties with groups of students from 18 to 28 years of age (average). The data recorded the date, median temperature, and daily beer consumption by liters. In this question, our goal is to forecast the daily beer consumption.

   (a) Split the data into history=before 2015-11-30 (included), and the rest as a test. We will first focus on **using history to select and train a model**:

   (b) Use the plots of the original beer consumption in history, combined with ADF test if you wish, to determine if there's trend and/or seasonality in the data.

   (c) Use a combination of plots, differencing, and/or ADF test to give a range of d, D and a number for seasonal freq m for your grid search.

   (d) Based on your forecasting goal: reach a good average performance on predicting the length of the test set, design a 5-fold cross-validation method with MAE using the historical data. Write Python functions based on this design, and select a model from SARIMA((p,d,q)*(P,D,Q,m)). Report the model selected and validation avg-MAE.

   (e) fit the model of chosen orders with all the data from history, then generate the forecast for the test data. Draw the plot with "history", "test" and "test forecast" together and report the MAE for the test forecast.

2. **In this question, you will acquainted with fitting data with SARIMAX.** Use the same train and test from the above question, and consider the column "median temperature" as an exogenous predictor for beer consumption. Use the train set however you like to select and fit a SARIMAX model to forecast beer consumption of the test period. Report the selected model and the test MAE.

3. **In this question, you will acquainted with fitting data with VAR.** In L5, we have use the "microdata" from statsmodels to practice SARIMAX. In this questions, consider "realgdp" and "realcons" as a vector time series and fit a VAR model:

   (a) Aggregate the data into quarterly data the same way as how I did it in L5. Split the data into the train (on or before the end of 2007) and the test (the rest data). Nothing to be submitted for this part.

   (b) Use the train data however you want to select and fit a VAR model and forecast the len(test). The key here is you need to difference the data manually to reach stationary, fit VAR on the stationary data, then reverse-difference to give the forecast of the original scale. The model should provide forecast for both "realgdp" and "realcons". Draw a plot with "train_realgdp", "train_realcons", "test_realgdp", "test_realcons", "test_realgdp_pred", "test_realcons_pred". Attach this plot as the answer of this question.