MSc Mathematical and Computational Finance

Statistics and Financial Data Analysis - Problem Sheet 4

PART A

1. (a) Download the following file `Data_Q2.csv` and save it as dataframe `df`. This contains the input data x and output y for your data analysis task.

   (b) Create a family of linear regression models by using as predictors polynomial transformations of x, varying the degrees p of x, for $p \in [1, ..., 10]$. That is, for polynomial degree 4 and 10, the corresponding input design matrices should look like this:

$$X_{(p=4)}= \begin{bmatrix} 1 & x_1 & x_1^2 & ... & x_1^4 \\ 1 & x_2 & x_2^2 & ... & x_2^4 \\ \vdots & & \ddots & & \vdots \\ 1 & x_n & x_n^2 & ... & x_n^4 \end{bmatrix} \quad \text{and} \quad X_{(p=10)}= \begin{bmatrix} 1 & x_1 & x_1^2 & ... & x_1^{10} \\ 1 & x_2 & x_2^2 & ... & x_2^{10} \\ \vdots & & \ddots & & \vdots \\ 1 & x_n & x_n^2 & ... & x_n^{10} \end{bmatrix}$$

   (c) Create a family of polynomial regressions by taking the matrices above as predictors, using `OLS` from `statsmodels`. You should have 10 model outputs. What is the best model according the BIC criteria? Show graphs and tables to support your results and display the coefficients of the best model under BIC.

   (d) What would be the best model if the AIC criteria was chosen instead? Plot both AIC and BIC in the same graph for comparison.

   (e) Using the design matrix $X_{(p=10)}$ created above as predictors (11 inputs, including constant 1), fit a Ridge model to the data. Using a 10-fold Cross-Validation, select the optimal value of λ using a grid search.

   The following code may be helpful, with appropriate replacements for `***'MSE of Test'***`, `***'FITAMODEL'***`, `*** 'TRANSFORM YOUR TRAIN DATA'***`:

```
import sklearn as sk
from sklearn.model_selection import KFold

kf= KFold(n_splits = 10)
err = []

for train_index, test_index in kf.split(X):
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

    X_train_transform = *** 'TRANSFORM YOUR TRAIN DATA'***
    X_test_transform = *** 'TRANSFORM YOUR TRAIN DATA'***

    # Fit model on Training Data
    modelResults = ***'FITAMODEL'***.fit(X_train_transform,y_train)

    # Assess model using MSE of Test data
    y_predTest = modelResults.predict(X_test_transform)
    err.append((***'MSE of Test'***))
```

Hint 1: The code above would run a Cross-Validation for a λ that has been already pre-specified. You would need to loop through the CV for each λ.
Hint 2: Assume the split is done on the original data x downloaded from the csv file. Hence the training data and test data need to be accordingly transformed to their polynomial powers, to get $X_{(p=10)}^{train\_data}$ and $X_{(p=10)}^{test\_data}$.
Hint 3: Choosing the λ range to run Cross-Validation can be tricky: see how they can be proportional to $||\hat{\beta}^{OLS}||_2^2$ for Ridge and $||\hat{\beta}^{OLS}||_1$ for LASSO. Alternatively, start by creating very small ranges $[10^{-6}, 1]$ and increase limits or shift limits accordingly.

(f) Display a graph of the Cross-Validation error against λ. Recall, for a given λ_i, the CV mean error (CV_{λ_i}) is the average of all test data sets MSE for that λ. Find the λ for which the CV error is minimum by doing a simple grid search.

(g) Repeat the exercise above for LASSO. In addition, display a table of the coefficients for the optimal LASSO regression (using λ_{min}^{lasso} chosen via the CV criteria). Compare this with the coefficients for the optimal OLS model in part (a) using BIC criteria.