MSc Mathematical and Computational Finance

Introduction to Statistics - Problem Sheet 'Week 2'

1. Data Analysis Report: restrict answers for this session to no more than 2 pages, clearly stating the problem you are examining, describing the data, labelling your graphs and backing up your conclusions with evidence. Your research may be extensive, however, the report should focus on describing the models considered, summarising your key findings with relevant data. The report should not be reduced to a 'cut-and-paste' exercise of the outputs of all your code.

   (a) The following `csv` files `IR.csv, Repossession.csv, Unemployment.csv` were created using data available at:

   `https://www.gov.uk/government/statistical-data-sets/`
   `live-tables-on-repossession-activity`
   `https://www.bankofengland.co.uk/boeapps/database/Bank-Rate.asp`
   `http://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork`

   (b) Investigate the historical relationship in the UK between the bank base interest rates, the rate of unemployment, and the rate of mortgage repossession. Write a brief report of your findings, including the main models considered and their assumptions and how they were verified.

   (c) Create a single dataframe `df` in Python, aligning the dates for the three datasets. Display the first three and last three elements of your dataframe using the commands `df.head(3)` and `df.tail(3)`. Display this output in your report (insert the picture of the output generated in Python).

   (d) Note: you should try simple regression models, polynomial regression and piecewise polynomial regression (no need for kernel regression). Consider which features to include and if normalising the data improves the model. Also use appropriate model selection criteria to support your conclusions.

2. We seek to fit a model of the form $Y_i = f(x_i) + \sigma\epsilon_i$. Write a Python script/notebook which searches for the optimal spline approximation to this model (with knots given by the quantiles of x), using the AIC as an optimality criterion. Your script should take vectors y and x as inputs, and return the number of knots in the optimal model. You may assume that the optimal number of knots is less than `sqrt(length(x))`, and that the vector x has at least this many distinct quantiles.

3. (a) Consider a model of the form:

   $$y_i = x_i\beta + \sigma\epsilon_i$$

   where x_i is a row vector of predictors, β is a vector of parameters to be estimated and σ is known. The ϵ are normally distributed with mean 0 and variance 1, however, for every i, j, ϵ_i and ϵ_j have known correlation $\rho^{|i-j|} \in (-1, 1)$. Find the maximum likelihood estimator of β as a matrix-vector equation.

(b) Consider a model of the form

$$y_i = \beta_0 + \beta_1(x_i + \sigma_1\eta_i) + \sigma_2\epsilon_i$$

where η_i and ϵ_i are independent normally distributed random variables, and y_i and x_i are scalar observations. Given a sample of pairs $(y_1, x_1), ..., (y_n, x_n)$, find the maximum likelihood estimator of the pair (β_0, β_1), as a function of the observations and the known quantities σ_1 and σ_2. What happens in the extreme cases $\sigma_1 = 0$ and $\sigma_2 = 0$?

(Hint: Consider the likelihood as a function of β_0, β_1 and the (unknown) variable $x_i^* = x_i + \sigma_1\eta$. Maximizing with respect to all three of these variables gives a system of equations to solve. The quantity $\lambda = \sigma_2^2/\sigma_1^2$ will be useful. It is possible to get β_0 and x^* as simple formulae and β_1 as the solution of a certain quadratic equation.)