Homework 9: A Diversified Portfolio

36-402, Advanced Data Analysis

Due at 11:59 pm on Monday, 6 April 2015

> WARNING: Some questions require slow computations. Do not wait until Monday to start this assignment.

Classical financial theory suggests that the log-returns of corporate stocks should be IID Gaussian random variables, but allows for the possibility that different stocks might be correlated with each other. In fact, theory suggests that the returns to any given stock should be the sum of two components: one which is specific to that firm, and one which is common to all firms. (More specifically, the common component is one which couldn't be eliminated even in a perfectly diversified portfolio.) This in turn implies that stock returns should match a one-factor model.

The data file portfolio.csv consists of the log returns for the stocks of 22 selected large US corporations, centered to have mean zero and scaled to have standard deviation 1. Each row is labeled by the relevant date.

1. (10)
   (a) (5) Report the weights of the first principal component. Since this is a vector of length 22, it will be better to report this visually than as a table or list of numbers. Comment on any notable patterns.
   (b) (5) Plot the projection on to the first principal component against date. Comment on any notable patterns.
2. (10) Fit a one-factor model.
   (a) (5) Report the vector of factor loadings. (Again, this will be most easily reported visually.) Comment on any notable patterns, and compare it to the first principal component.
   (b) (5) Plot the factor score against the date. Comment on any notable patterns, and compare to the projection on the first principal component.
3. (10) Use case bootstrapping to provide 90% confidence intervals for the factor loadings of the one-factor model. Report the results as a figure rather than a table.
4. (5) What is the p-value of a goodness of fit test for the hypothesis that one factor is adequate? Explain carefully just what hypothesis is being tested, and what is entailed by rejecting or retaining it.

5. (5) Download the function charles from the class website. Explain carefully what arguments the function takes, what the function does, and exactly what its return value is. (An acceptable answer to this question could be a thoroughly-commented version of the function.)

6. (15) Write a function which finds the cross-validated log-likelihood of a factor model with a given number of factors. That is, it should take a data set and a number of factors as inputs, divide the data randomly into folds, calculate the log-likelihood on a test fold of a model fit on the other folds, and return the average log-likelihood across folds. You are encouraged to re-use existing code from the solutions and notes; `charles` may or may not be useful. Report the five-fold cross-validated log-likelihood of factor models with from 1 to 10 factors for this data. What is the favored number of factors?

7. (10) Using the `mvnormalmixEM` function from the `mixtools` package, fit a two-component Gaussian mixture model to the data.

   (a) (5) Report the parameters of the two mixture components, and their relative weights. Avoid excessive precision.

   (b) (5) Use `posterior` component of the object returned by `mvnormalmixEM` to classify each day as belonging to one mixture component or the other. Plot the mixture components over time, and comment on any patterns.

8. (15) Write a function, `loglike.mvnormalmix`, which takes in a data set and a model returned by `mvnormalmixEM`, and returns a log-likelihood. Check that it works by seeing that it gives the correct value of the log-likelihood when a two-component mixture is fit to the whole data. (*Hint:* read section 21.4.4 of the notes.)

9. (8) Write a function which calculates the log-likelihood of mixture models through cross-validation, as in problem 6. Report the five-fold cross-validated log-likelihood of mixture models with from two to four components for this data. What is the favored number of mixture components?

   Warning: five-fold CV for four mixture components on the full data might take several hours. Start early, and make sure you debug your code on small parts of the data rather than the whole thing.

10. (2) Can you decide whether factor models or mixture models fit this data better?

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the text of the corresponding problems. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate `.R` file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output. For full credit, all code runs, and the Markdown file knits (if applicable).

Company	Abbreviation
Altria (formerly Philip Morris)	MO
Amazon	AMZN
Apple	AAPL
Archer Daniels Midland	ADM
Automatic Data Processing	ADP
Bank of America	BAC
Corrections Corporation of America	CXW
Dow Chemicals	DOW
Equifax	EFX
ExxonMobil	XOM
Ford	F
Halliburton	HAL
General Electric	GE
Goldman Sachs	GS
Graham Holding Companies	GHC
Microsoft	MSFT
Proctor and Gamble	PG
Time Warner	TWX
United States Steel	X
Walmart	WMT
Yahoo!	YHOO
Yum! Brands	YUM

Table 1: Abbreviations for the companies included in the data set.