MATH 191 TOPICS IN DATA SCIENCE: ALGORITHMS AND MATHEMATICAL FOUNDATIONS A COMPARISON OF CORRELATION MEASURES USED FOR PREDICTION OF STOCK RETURNS

LAWRENCE OUYANG

December 12, 2015

Abstract. The prediction of the returns of a stock has been a very popular problem since the very existence of the stock market. Given the day's returns, what will tomorrow's look like? What will it look like next week, or next month? An idea to examine this is the use of correlation measures. The market is not an isolated system; all things affect one another. This work will use a variety of known correlation measures (Pearson's, Spearman's, Hoeffding, maximal correlation, distance correlation, MIC) to compute the most relevant instruments to which we will apply linear regression. The goal is to find an accurate predictor for future returns.

Key words. Correlation; Returns; Linear Regression

1. Introduction.

The idea behind examining logarithmic returns is extremely rational. By using returns instead of raw prices, our values become normalized, which is a requirement for accurate comparisons and measurements. Logarithmic functions are log-normal as well, and as such, our returns our now normally distributed. However, with these benefits also come a price. The logarithmic returns does not have an implied one-to-one relationship with the simple returns, as well as having higher variance which can possibly reduce expected returns[1]. For simplicity, this paper will focus on the use of logarithmic returns to simplify our results.

1.1. Correlation Measures.

Determining our significant predictors using only a standard correlation measure like Pearson or Spearman would be too mundane. We instead examine other measurements, beginning with the maximal correlation. Maximal correlation

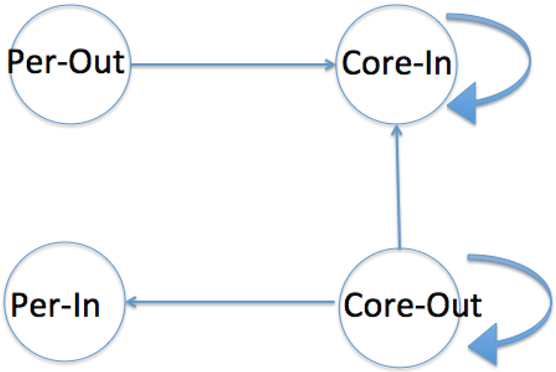

Fig. 1.1. *This is how you add a plot to a Figure. It is always a good idea to add such a caption to each Figure, and explain what the figure is about. Also, if your plot has x and y axis, please always label your graphs (within MATLAB) so that your plots/results can be easily read and understood.*

In Section 3 we ... In Section 4 we test the above algorithms on synthetically generated data sets, while in Section 5 we do so for a real data set. Finally, in Section 6 we conclude with a summary of our results, and discuss future possible research direction.

2. Related work. Questions/Comments/Things that could be done

- What is a good notion of core-periphery structure in directed networks? Would the null model shown in Figure 1.1 be a good model?
- Can one build on or expand some of the above methods and apply them to directed networks?
- As a starting point, perhaps apply simulated annealing to the objective function induced by the above null model
- Apply this to a real network, a good such example might be the migration network between counties in the United States, which we have seen in class in the past.
- This is how you add an url link
  http://people.maths.ox.ac.uk/porterm/papers/prestige_final.pdf

This is how you color tex in red ...
This is how you color tex in blue ...

3. Our work.

3.1. The Data. The data set being used is the logarithmic return prices from date to date for 477 stocks. Since the data contains many not-available entries, we begin by stripping those out of our measurements. Instruments with a high number of NA entries are entirely removed along with missing rows. This leaves us with a dataset containing number instruments with data for number of days.

3.2. The Process. With our primmed and proper data, we begin by calculating their coefficient matrices. Considering the size of our data, the calculation of our coefficient matrices take a very large amount of time. Consider the following R code to compute the correlation coefficients:

```
# Consider the given RETS matrix, remove poor columns and rows:

RETS = na.omit(RETS);
numStock = dim(RETS)[2];

# Calculate the various correlations:
# Pearson, Spearman:
PCOR = cor(RETS, method = "pearson");
SCOR = cor(RETS, method = "spearman");

#Maximal
#install.packages('acepack');
MaxCOR = matrix(, nrow = numStock, ncol = numStock);
colnames(MaxCOR) = colnames(RETS);
rownames(MaxCOR) = colnames(RETS);

for (i in 1:numStock) {
   for(j in 1:numStock) {
      transfVars = ace(RETS[,i],RETS[,j]);
      MaxCOR[i,j] = cor(transfVars$tx,transfVars$ty)[1];
   }
}

#Hoeffding's D
#install.packages('Hmisc');
HCOR = hoeffd(RETS)$D;

#Distance
DCOR = matrix(, nrow = numStock, ncol = numStock);
colnames(DCOR) = colnames(RETS);
rownames(DCOR) = colnames(RETS);
#install.packages('energy');
for (i in 1:numStock) {
   for(j in 1:numStock) {
      DCOR[i,j] = dcor(RETS[,i],RETS[,j]);
   }
}

#MIC
#install.packages('minerva');
MICCOR = mine(RETS,n.cores = 4)$MIC
```

Of course, this is the simple part. Now for the instrument we would like to predict, we need to sort it's correlation values in descending order, and choose the amount to use in our linear regression. Once sorted, we build our linear regression model and evaluate its accuracy.

4. Numerical experiments on synthetic data. Present here the numerical results you obtain on a synthetically generated data set...

5. Numerical experiments on real data. Present here the numerical results you obtain on a real data set...

6. Summary and conclusion. Summarize your work in this section.

REFERENCES

[1] Robert S. Hudson and Andros Gregoriou, *Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns*, SSRN, (2010).

[2] M. P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha, *Core-periphery structure in networks*, SIAM J. Appl. Math., 74 (2014), pp. 167–190.