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

LAWRENCE OUYANG

December 13, 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, distance correlation) 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[2]. For simplicity, this paper will focus on the use of logarithmic returns to simplify our results. Below is a brief description of our correlation measures.

1.1. Pearson Correlation. The Pearson correlation is an extremely basic and simple correlation measure used for many statistical practices. It is defined as:

$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{X})(y_i - \bar{Y})}{n-1} \tag{1.1}$$

$$Cor(X,Y) = \frac{Cov(X,Y)}{Var(X)Var(Y)} \tag{1.2}$$

However, because of it's simplicity, the Pearson correlation faces many shortcomings. As shown in equation 1.1, the covariance is greatly affected if x_i, y_i are extremely large or small values. Since Pearson only measure linear relationships, any nonlinear dependency will not be represented[1].

1.2. Spearman's Rank Correlation. Spearman's correlation is similar to Pearson's except rather than using raw values, Spearman uses ranked values. This better normalizes the data set and accounts for extreme values in the correlation. Consider d to be the distance between rank x_i, y_i, then Spearman's rank correlation can be calculated as[3]:

$$\rho_{Spearman}(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1.3}$$

Spearman also suffers the same shortcomings as Pearson, however, it does handle certain cases more gracefully, and thus is often also considered alongside Pearson.

1.3. Hoeffding's D. 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 Hoeffding's D. Hoeffding's D measures the difference of joint ranks and the product of the marginal ranks. The shortcomings of Spearman and Pearson are their inability to detect nonlinear relationships. However, Hoeffding's D is capable of this, making it more versatile[1].

1.4. Distance Correlation. The distance correlation simply uses the euclidean distance between our data points to calculate the correlation[4]. Similar to Hoeffding's D, distance correlation is able to identify nonlinear relationships, making it more versatile. It is also relatively easy to implement. Consider the distant covariance:

$$dCov(X,Y) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} x_{ij}y_{ij}} \tag{1.4}$$

$$dVar(X) = dCov(X,X) \qquad dVar(Y) = dCov(Y,Y) \tag{1.5}$$

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}} \tag{1.6}$$

With the above measurements considered, in Section 3 we put them to the test by extracting the top values and using it for linear regression. In Section 4 we test the algorithms for a real data set. Finally, in Section 5 we conclude with a summary of our results, and discuss future possible research direction.

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

- Since obtaining such large correlation matrices are calculation intensive, what other methods can be used to trim our selection before creating our matrices?
- What kind of accuracy would using alternating entries give?
- What correlation measure is most affective, and why?
- How many instruments should be considered when building our model? Using more would indeed increase our R-squared, but would it result in over-fitting?

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 436 instruments with data for 2629 days. This presents a problem: our dataset is too large to due a correlation function in a reasonable time frame. To combat this, we will only use the first 500 entries to determine correlation. The remaining values will be used to as our test matrix.

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. Of course, building the matrices is the simple, although lengthy, 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.

With our model, we then use our remaining data points to create a predicted vector $\hat{y}$. We then compare our prediction with the actual values, and calculate the average error ϵ.ϵ is defined as:

$$\epsilon = |\frac{y - \hat{y}}{nrow - 100}| \tag{3.1}$$

We want a minimized error ϵ as that would denote a more accurate prediction. Our results are shown in the next section.

4. Numerical Results. The tables below represents a selection of 5 instruments, and the 20 instruments that had the strongest relationships from left to right with them. This includes the average error and the adjusted R-squared value for the given regression model.

Pearson

SPY	C	JPM	GS	AXP	MS	GE	BBT	TROW	WFC	KEY	EMR	NTRS	PPG	ITW	ETN	LNC	PCAR	BEN	BK	CINF
YHOO	AMZN	EBAY	SPY	NTAP	LLTC	ADI	VRSN	XLNX	JBL	KLAC	TER	JDSU	MOLX	JNPR	AMAT	FISV	BRCM	TMO	LRCX	ADBE
DELL	SPY	LLTC	CSCO	INTC	ADI	AMAT	MOLX	XLNX	MSFT	KLAC	CTAS	ALTR	IBM	PCAR	LRCX	TXN	HPQ	JBL	PBI	AXP
AAPL	SPY	IR	NTAP	PPG	INTC	CSCO	LLTC	MSFT	JCI	DELL	JNPR	SIAL	PCAR	CHRW	PH	PBI	YHOO	FISV	AA	ITW
IBM	SPY	MSFT	LLTC	AXP	INTC	MS	OMC	XLNX	AMAT	CSCO	C	MOLX	BK	ORCL	GS	PCAR	SCHW	TROW	ADI	CINF

	SPY	**YHOO**	**DELL**	**AAPL**	**IBM**
R-Squared	0.8825	0.5807	0.4723	0.2501	0.5363
Error ϵ	0.00414	0.01414	0.01117	0.01454	0.00656

Spearman

SPY	C	AXP	GS	JPM	NTRS	MS	STI	ETN	PCAR	BBT	BEN	TROW	DOV	GE	WFC	EMR	KEY	LM	LNC	BAC
YHOO	EBAY	AMZN	SPY	NTAP	JNPR	VRSN	JDSU	ADI	XLNX	LLTC	TER	JBL	KLAC	ADBE	TXN	INTC	BRCM	QCOM	FFIV	MOLX
DELL	SPY	INTC	CSCO	LLTC	AMAT	MSFT	ADI	LRCX	IBM	XLNX	CTAS	PCAR	HPQ	NTAP	MOLX	ALTR	ORCL	KLAC	TXN	BRCM
AAPL	SPY	NTAP	INTC	JNPR	DELL	MSFT	IR	CSCO	BRCM	PCAR	PPG	SIAL	DOV	LLTC	FISV	JCI	PH	CHRW	MOLX	ORCL
IBM	SPY	MSFT	INTC	CSCO	C	XLNX	DELL	ORCL	LLTC	CTAS	ADBE	MS	AXP	FISV	SCHW	OMC	PCAR	AMAT	CINF	LM

	SPY	**YHOO**	**DELL**	**AAPL**	**IBM**
R-Squared	0.8851	0.5780	0.4741	0.2514	0.5365
Error ϵ	0.00469	0.01422	0.01118	0.01447	0.00658

Hoeffding's D

SPY	C	AXP	NTRS	JPM	GS	MS	BEN	PCAR	DOV	ETN	STI	BBT	TROW	GE	KEY	LM	WFC	MLX	LNC	EMR
YHOO	EBAY	AMZN	SPY	NTAP	VRSN	TER	JNPR	XLNX	JDSU	LLTC	JBL	ADI	KLAC	QCOM	ADBE	TXN	FFIV	INTC	AMAT	ROP
DELL	SPY	INTC	CSCO	LLTC	MSFT	AMAT	IBM	ADI	LRCX	HPQ	MOLX	XLNX	TAS	ORCL	ALTR	BRCM	NTAP	PCAR	JDSU	TXN
AAPL	SPY	NTAP	INTC	DELL	JNPR	MSFT	IR	BRCM	PCAR	CSCO	ORCL	DOV	MOLX	SIAL	LLTC	CTAS	PPG	CHRW	FSIV	JCI
IBM	SPY	MSFT	INTC	CSCO	FISV	C	DELL	ORCL	ADBE	XLNX	AXP	LLTC	CTAS	OMC	HPQ	MOLX	HIG	CSC	PCAR	AMAT

	SPY	**YHOO**	**DELL**	**AAPL**	**IBM**
R-Squared	0.8937	0.5793	0.4756	0.2515	0.5432
Error ϵ	0.00444	0.01423	0.0112	0.01450	0.00673

Distance

SPY	C	AXP	NTRS	JPM	GS	MS	BEN	PCAR	DOV	ETN	STI	BBT	TROW	GE	KEY	LM	WFC	MLX	LNC	EMR
YHOO	EBAY	AMZN	SPY	NTAP	VRSN	TER	JNPR	XLNX	JDSU	LLTC	JBL	ADI	KLAC	QCOM	ADBE	TXN	FFIV	INTC	AMAT	ROP
DELL	SPY	INTC	CSCO	LLTC	MSFT	AMAT	IBM	ADI	LRCX	HPQ	MOLX	XLNX	TAS	ORCL	ALTR	BRCM	NTAP	PCAR	JDSU	TXN
AAPL	SPY	NTAP	INTC	DELL	JNPR	MSFT	IR	BRCM	PCAR	CSCO	ORCL	DOV	MOLX	SIAL	LLTC	CTAS	PPG	CHRW	FSIV	JCI
IBM	SPY	MSFT	INTC	CSCO	FISV	C	DELL	ORCL	ADBE	XLNX	AXP	LLTC	CTAS	OMC	HPQ	MOLX	HIG	CSC	PCAR	AMAT

	SPY	**YHOO**	**DELL**	**AAPL**	**IBM**
R-Squared	0.8937	0.5793	0.4756	0.2515	0.5432
Error ϵ	0.00444	0.01423	0.0112	0.01450	0.00673

5. Summary and conclusion. We can easily note that *SPY* has strong relationships with the other instruments. Other big industry names arise as well, with *AMZN*, *DELL*, *IBM*, *JNPR*, and *AXP* appear quite often. This probably indicates that their presence greatly affects the economy of even smaller industries.

As expected, we see no clear results from our Pearson and Spearman models. They seem to show varying results. Hoeffding's D doesn't give much exciting results either, mirroring very closely to Pearson and Spearman. Hoeffding's D does have slightly lower error values than the standard, which may indicate a more viable prediction. More information can be found by repeated testing with all 477 instruments.

We do note that instruments with strong relationships, such as *SPY*, have high R-squared values and low errors. This may demonstrate that correlation ranking models are effective only on predicting major stock returns.

REFERENCES

[1] Michael Clark, *A comparison of correlation measures*, CSR, (2013).

[2] 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).

[3] Asuka Nakata Suzana de Siqueira Santos, Daniel Yasumasa Takahashi and Andre Fujita, *A comparative study of statistical methods used to identify dependencies between gene expression signals*, Briefings in Bioinformatics, 15 (2013), pp. 906–918.

[4] Bakirov N. Szekely G., Rizzo M., *Measuring and testing independence by correlation of distances*, Ann Stat, 35 (2007).