Hi Thomas, PJ

Hope you are well.

Following the email I sent you previously which answered some of your questions from our last meeting and also provided a few reference papers to give a glimpse of the current literature on research topics similar to mine, I have conducted two more kinds of experiments with regard to predictions that start after earnings reports have been released:

1. I have been trying to ***derive*** the 1-day-to-30-day forecasts by using the model generated 0-day-to-30-day forecasted returns and the actual 0-day-to-1-day returns (which will be known by the time we are at the end of the first day after earning release).

I do this both by naïve subtraction of the 0-day-to-30-day predicted return by the known 0-day-to-1day return, and also by firstly aligning the underlying distributions of all these actual and predicted returns (by aligning their z-scores) before deriving the 1-day-to-30-day returns.

Unfortunately, these estimated the 1-day-to-30-day returns don't bear much relevance to the actual 1-day-to-30-day returns.

I also have studied the empirical relevance of these time series. Below is the correlation stats for Q3 2018 with respect to movement direction:

904 stocks from Q3 2018

Correlations

actual 0d-to-30d / actual 0d-to-1d

0.313059249

actual 0d-to-30d / actual 1d-to-30d

0.631915144

actual 0d-to-1d / actual 1d-to-30d

-0.052509809

predicted 0d-to-30d / actual 0d-to-30d

0.128358244

predicted 0d-to-30d / actual 1d-to-30d

0.03508886

We can see that the actual 0d-to-1d return is uncorrelated to the actual 1d-to-30d and in fact I've counted that nearly half of the time they show opposite signs with each other. That explains why I can't simply *estimate* 1-day-to-30-day return with the help of the actual 0d-to-1d return.

Although the predicted 0d-to-30d shows correlation to the actual 0d-to-30d, it shows no correlation with the actual 1d-to-30d.

2. I have also carried out experiments by accounting for the volatility of stock returns. For that matter I have changed my system to predict the z score of any post-earnings stock return with respect to that stock's own historical return distribution. It means I am effectively predicting where a stock's return will sit on its historical return distribution. A predicted z-score will then get transformed back to an equivalent (predicted) stock return. I am achieving similar results compared to forecasting stock returns directly, but again this methodology doesn't generate the same kind of result when the dependent variable is a stock return after earnings release.

I have to concede I have almost exhausted my experiment methods in this case. I am afraid there is always going to be certain limitations in machine learning results when we work with a limited set of *numerical data*, and it's also particularly challenging when the training data is composed of data from a large number of companies each of which follows a different distribution of its own. Data normalization can't eliminate those contradicting characteristics of these companies.

I want to ask for your permission to move on from here to explore other aspects of the phd thesis which may end up helping this current project further down the road. I do however think there is a number of genuine contributions and new research angels I will have made to the literature (partially outlined in the answer document I sent you in the last email) and would like to attempt a publication. This will be on its own a way to get feedback from the academia. ***I am more than happy to re-scope the key topics of the paper and re-position the research angel, should you have any opinions***.

In terms of the next step of the research, I can think of three directions:

1. Use GAN (generative adversarial network) to generate quansi-realistic data for an individual company and perform forecast on that company alone using a mixture of real (a smaller set) and generated data (a much bigger set). This is so as to bypass the (huge) interference by other companies' data. I have not attempted this so can't comment the effectiveness of the generated samples.
See https://arxiv.org/ftp/arxiv/papers/1904/1904.11419.pdf

2. Explore a number of financial applications using reinforcement learning as per my previous proposal (Please see attached)

3. Work with PJ on a project he proposed at the last meeting

I would like to seek your opinions about the number of points I raised in this email. Please kindly let me know when we can have our meeting.

Many Thanks

Joseph