Maggie Nolan
Paper 3 Proposal

**Forecasting Market Volatility with Neural Networks**

*ARCH type layers, volatile forecasting periods, comprehensive parameter selection*

The basis of this third paper is an extension upon the second paper, which focused primarily on comparing the relative success of complex ARCH models (specifically GARCH, PARCH, and TARCH models) against varying neural networks in forecasting S&P 500 volatility. For the purposes of the second paper, the sample data was limited to the period between 10/4/2010 and 3/2/2020. This date range was selected in order to update the data originally analyzed by the Alam et al. study\(^1\) referenced in that paper, which considered various ARCH models as tools for forecasting volatility in the DSE general index returns from 2001 to 2011. The general conclusion of the second paper was that the neural networks, regardless of parameter selections, outperformed all ARCH models suggested. However, this selection of data from 2010 up through 2020 (ending before the recent COVID-19 related fluctuations in the S&P 500) failed to include any significantly volatile periods, such as those that were tested in the Alam et al. study (i.e. the 2008 recession). While this study did not consider neural networks, there was slightly more discrepancy between ARCH models in forecasting performance than was found in the second paper. As such, one extension of this paper will be to consider a longer sample period, perhaps training on data that includes the 2008 recession and testing on data that includes the recent COVID-19 fluctuations. Therefore, the data will be the same source and type as the second paper (using S&P 500 returns) just on a different time frame.

Additionally, I will extend upon the existing neural networks to incorporate some of the ARCH model results into the neural nets themselves, as is similarly proposed by Li et al. (2005)\(^2\) with a different kind of neural network (Li used generalized regression neural net as opposed to the backward propagation model in the second paper). Li et al. pursues an interesting hybrid model that combines a GARCH model with a generalized regression neural network (GRNN) – in this model, one of the inputs into the neural network is information obtained through a prior GARCH model fit. This information is entered into the model after its first layers’ coefficients are calculated. Given the success of the hybrid model over both a regular GARCH and GRNN separately, this paper will look to extrapolate beyond that conclusion to see if the trend persists with backward propagation models. Additionally, I’ll incorporate more than just the GARCH model when considering ARCH type inputs into neural networks. One result of the Li et al. paper that was particularly interesting was the discrepancy in improvement between the GRNN-GARCH


model versus the GRNN and the GRNN-GARCH model versus the GARCH. The hybrid model performed significantly better than a simple GARCH, but did not add much information beyond the plain GRNN model. Because of this, it will be interesting to see if the backward propagation methods perform differently, or if different variants on the ARCH models (beyond just a GARCH model) show larger discrepancies from the basic neural net when included in the hybrids. Generally, the question from this modification then becomes how much new information can GARCH type results add to a neural network beyond its underlying volatility forecasts?

Given the superior performance of the backward propagation models in general from the second paper, I hypothesize that we will see a higher accuracy when creating hybrid models using the ARCH model results, but I don’t anticipate the hybrid with the TARCH or PARCH will necessarily perform better than a GARCH hybrid. This is based on the conclusion from the second paper that the GARCH (in absence of neural network hybrids) outperformed the TARCH and PARCH models (although only slightly). Although, it is possible that the new information that each model provides will have different effects, as the neural networks don’t have any specific kinds of information they’re designed to pick up on. So, for example, if the neural network can pick up standard volatility patterns used in the GARCH model, but not the asymmetric information coefficient included in the TARCH model, we could see the TARCH hybrid outperforming the GARCH hybrid. Additionally, getting back to the idea of sample volatility, it’s possible that the TARCH or PARCH does better in periods of high volatility, which would explain their equal (if not slightly lesser) performance compared to the GARCH models in the second paper.

As an added enhancement, I’m going to perform a more rigorous parameter selection procedure for the neural networks in terms of determining the number of layers and nodes. In the second paper, I looked at trends rather informally when different parameters were changed (i.e. looked at how higher versus lower parameters affected fit), but this could have missed an optimum network structure. For instance, if higher layer parameters did better with a certain range of node parameters, this might not have been noticed. To improve upon this, I’ll build out a few functions that test a variety of parameters simultaneously to see which performs best on the out of sample data. Because of the added model hybrid component, you effectively introduce a new parameter as well, which is when/where the information with GARCH results is injected into the model. It’s possible that it could be an original data point (with the forecast up until that point linked to each prior years’ realized volatility) used in the training data, or it could be that these values are added at a later point in the model (as was done in Li et al.). It may be worth considering both data entry methods, as (depending on the other model parameters regarding nodes and layers), one could be superior to the other.
In terms of model comparisons, I’ll use the same method that was used in the second paper, which was the usage of out of sample root mean squared error metrics for evaluating forecasting success. However, given that I’m looking at several different alterations (sample period, ARCH inputs into the neural nets, neural network parameters), it will be slightly more complicated in determining where these comparisons occur. Simultaneous parameter and input variation should not be too complicated, as each ARCH input configuration could be tested using the aforementioned parameter variation function. These models, however, should in theory have two separate train-test splits. Meaning, regardless of how many neural networks are created using these different parameters and inputs, all of them will be trained on two data sets – one including a volatile period (likely the 2008 recession) and one trained on a period of equal length without a highly volatile period.

Once the pairs of models are fitted, they will be tested on out of sample data. Then, we will see (overall) whether training on volatile data matters in the eventual effectiveness of the model. While it is common to test neural networks on a variety of data splits, the intention here is to see whether the results of the second paper were at all misleading in the ability of the neural networks to outperform ARCH models. For instance, if neural networks perform extremely well in periods of low volatility, but fail in periods of higher volatility compared to GARCH models, then the largely nonvolatile data tested in the second paper may have had misleading results. By training (and testing) the data on nonvolatile, volatile, and mixed periods, a better assessment can be made by comparing the fitted models’ performance out of sample. Additionally, it may be interesting to consider how these models compare when you don’t have continuous data, so to speak. When you train the model on one period and test it on a period beginning immediately after, you may benefit from having a model fitted to recent data. It would be interesting to see how it performs with gaps in between the training and testing data, as constant refitting of the model to new data may not be practical (or wise) in its actual implementation and usage.

This question of gaps in data, or rather the ability of a model to be extrapolated to different time periods, raises one potential issue for the paper. If structural changes in the economy in recent decades affect the performance of the model, then we may run into an issue of not having enough data – for example, if the model doesn’t work as well on data far in the past (prior to the turn of the century), then you’re realistically only going to be working with more recent data. In the second paper, the data was fitted on a 7 year period of daily observations. If I’m hoping to do all of these different train-test splits with different gaps between the training and testing data, it might mean that I’d either have to shorten the training period or use more historical data, both of which could affect the accuracy of the volatility forecasts.