Technical document

1. Current Implementation Status

The library implementations are at a point such that an end-to-end process can be run: data with training parameters are specified, and a CSCV based PBO is output.

A full list of implementations is as follows:

- Network Types
  - Feedforward Network
  - Restricted Boltzmann Machine
  - Stacked Auto Encoder
- Learning Algorithms
  - Contrastive Divergence-1 for RBM pre-training
  - Online Gradient Descent
  - Stochastic Gradient Descent
- Validation Algorithms
  - CSCV
  - PBO
- Data Processes
  - Log fluctuation implementation to format dataset with rolling windows
  - Library to generate synthetic data
- Optimizations
  - SGD: L1 Regularizations
  - Learning Schedule Epochs
  - Denoising: Gaussian Noise & Feature On-Off Switching
- Neural Network Configurations
  - Activations: Sigmoid, ReLU, Softmax
  - Initialisations: Normal, Hinton, Xavier, He (Normal & Uniform variations)
  - Cost Functions: MSE, CE, Loglikehood
- Output Library
  - Effective graphing of training logs (epoch based performance, weight update rates, output reconstructions etc.)
- Technical Testing
  - Basic hyperparameter search method & outputs
  - Automated Unit Tests for CD1, OGD and
- Database Implementation
  - Recording results and configurations in SQL Lite database
  - Recording networks (SAE) in bson files

2. Iteration Results

Iteration 4: 3rd June

Weight Initializations Weight initializations for SAE were tested using the typical Xavier & He initialization techniques, as well as a customed ("DC") initialization.

Xavier:

$$w_{ij} \sim U(-r, r), r = \sqrt{\frac{6}{n_i + n_j}} \tag{1}$$

The initialization for ReLU is different on accounts of the function being equal to zero for half it's potential input range - in this case it makes sense to double the weight variance, and so the He initialization is used

$$w_{ij} \sim U(-r, r), r = \sqrt{\frac{6}{n_i}} \tag{2}$$

DC:

$$w_{ij} \sim U(-r, r), r = \sqrt{\frac{6}{(n_i + n_j)/2}} \tag{3}$$

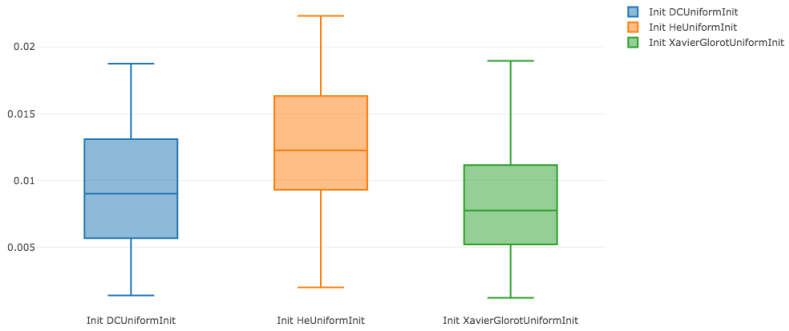

Fig. 1: Weight Initialization for SAE
The box plots show the MSE for a series of SAE networks trained. Surprisingly, the generally used He initialization has the worst performance, whereas the Xavier and DC initializations largely better and on par.

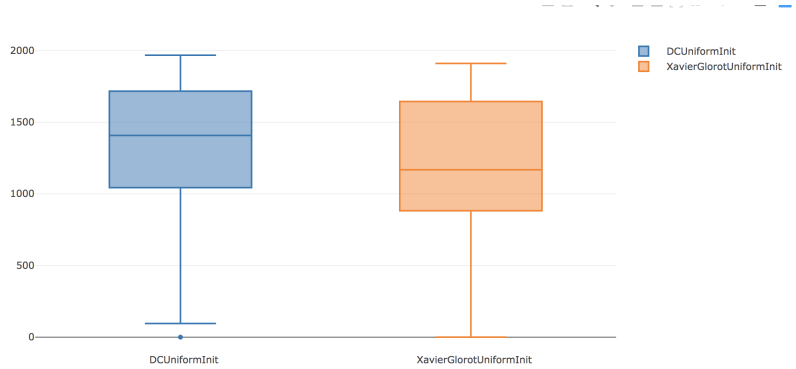

Fig. 2: Weight Initialization for Predictive FFN
The box plots show the profits for a series of FFN networks trained, with the custom DC init showing better performance to Xavier.

Network Structures Networks were trained at various number of layers and layer sizes, as presented below.

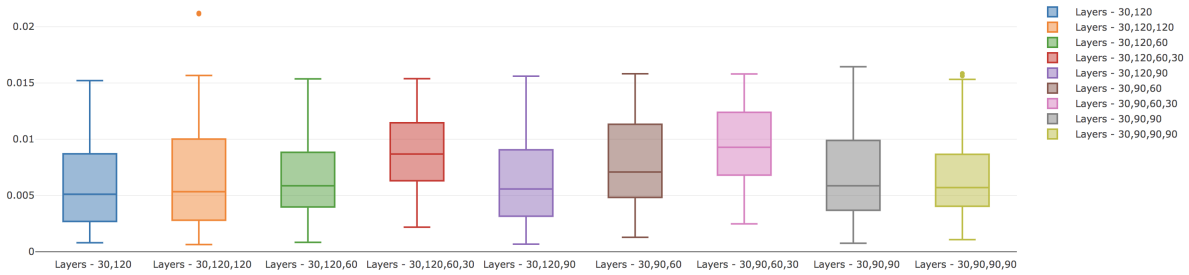

Fig. 3: Network Sizes for SAE
The box plots show the MSE for the series of SAE networks trained. The results show worse performance for the typical descending network sizes used for SAE, and show no improvement form increasing network size and complexity.

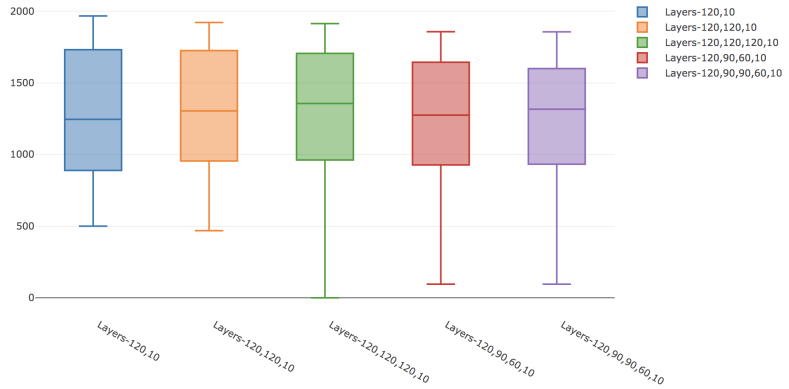

Fig. 4: Network Sizes for Predictive FFN
The box plots show the profits for the series of FFN networks trained, showing small improvments for network size increases.

Data Windows All configurations were tested using data windows of

- (1, 5, 20)
- (5, 20, 60)
- (10, 20, 60)

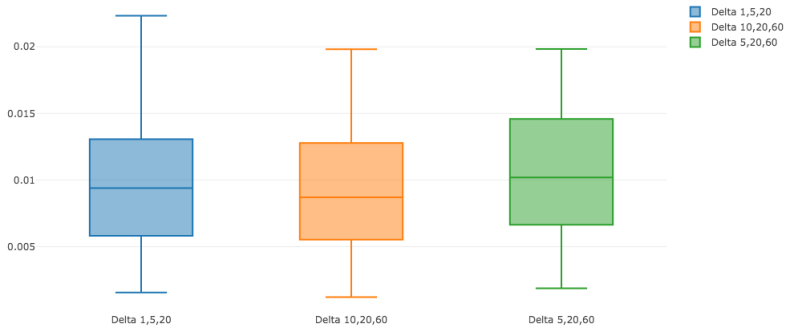

Fig. 5: SAE Data Window Aggregations
The box plots show the MSE for a series of SAE networks trained, grouped by different data window aggregations. The MSE is lower for the higher aggregations, presumably due to reduced noise in the dataset, which may support longer windows in further training.

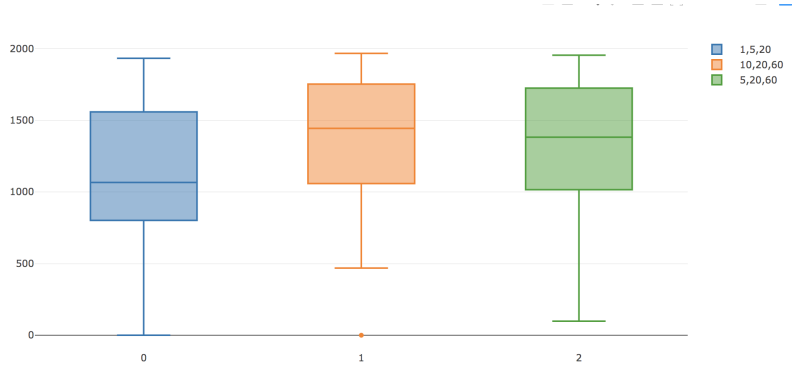

Fig. 6: FFN Data Window Aggregations
The box plots show the profits for the series of FFN networks trained, grouped by different data window aggregations. The profit is higher for the longer aggregations once again supporting longer windows in further training.

Best network performance The stock data generated for this iteration had 10 assets with the following stocks and variances

- (0.9, 0.5)
- (0.7, 0.2)
- (0.05, 0.4)
- (0.05, 0.5)
- (0.04, 0.1)
- (0.02, 0.15)
- (0.01, 0.05)
- (-0.8, 0.55)
- (-0.4, 0.15)
- (-0.1, 0.2)

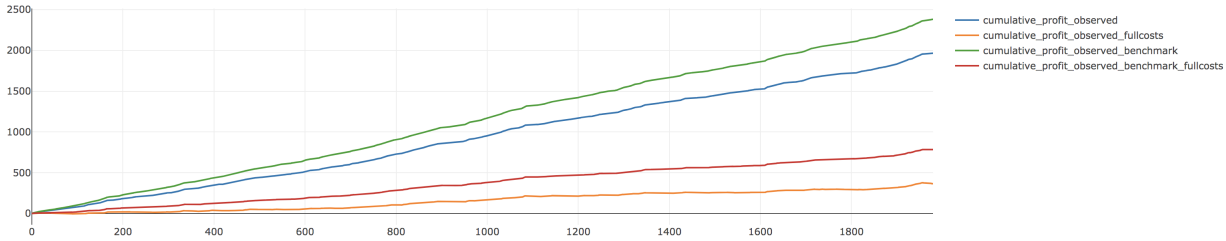

Fig. 7: Best Network Cumulative Profits
The graph here shows cumulative profits for the best network and benchmark, both producing profits with costs attached.

it4_confusion

	model_no_trade	model_trade
benchmark_no_trade	7790	1958
benchmark_trade	2005	8047

Fig. 8: Best Network Confusion Matrix
A confusion matrix showing the match up of trades and no trades on all assets for the best model and benchmark.

Regularization Once again, the regularization shows notable worse performance, now on the predictive network as well. This still stands in contrast to the increased performance from the learning rates schedule.

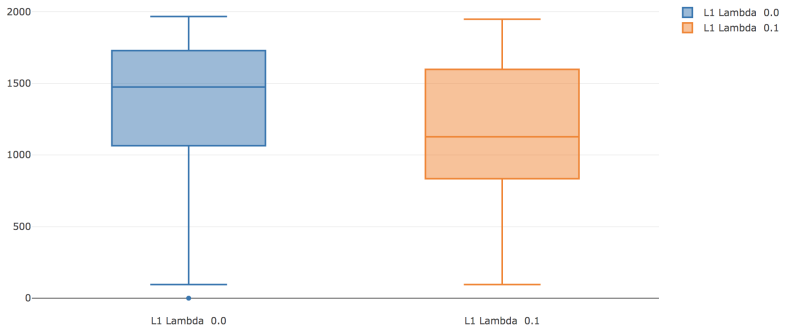

Fig. 9: Predictive Network Regularization
Effects of L1 regularization on predictive network profits.

Iteration 3: 17th May

Linear vs ReLU: Smaller Network Synthetic assets (6) were used on smaller prediction networks to test the performance of Linear vs. ReLU activations, once again showing the higher performance of Linear activations.

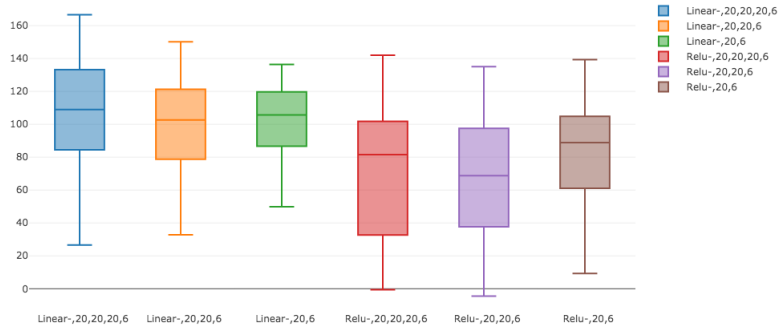

Fig. 10: Linear vs ReLU activation on smaller networks
The boxplots here show the summary of P&L at networks with smaller layer sizes than previously. The networks were trained on 6 synthetic assets with a total of 18 inputs, and box groupings reflect different sized autoencoders and learning rates (720 configurations in total). We once again see the outperformance of ReLU by the linear activations.

Leaky ReLU vs ReLU Leaky ReLU activations were implemented and tested using synthetic data, showing negligible differences on SAE but slighly better performance on the predictive FFN networks. The differences seem to range from none to a marginal gain, which makes sense.

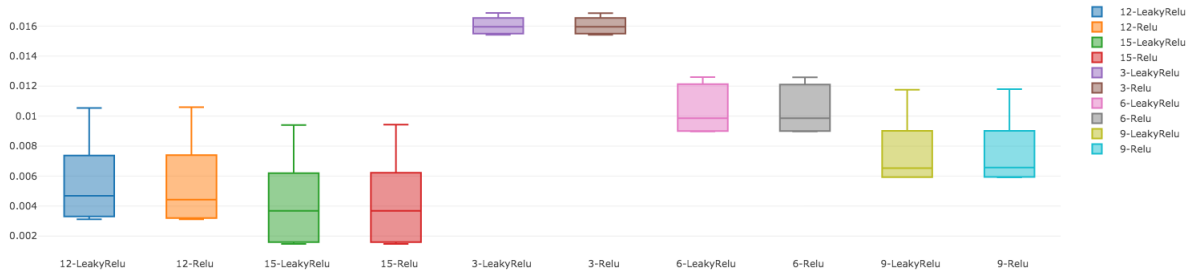

Fig. 11: SAE: Leaky ReLU vs ReLU
The plot above shows the MSE for 120 different SAEs, grouped by encoding size and activations.

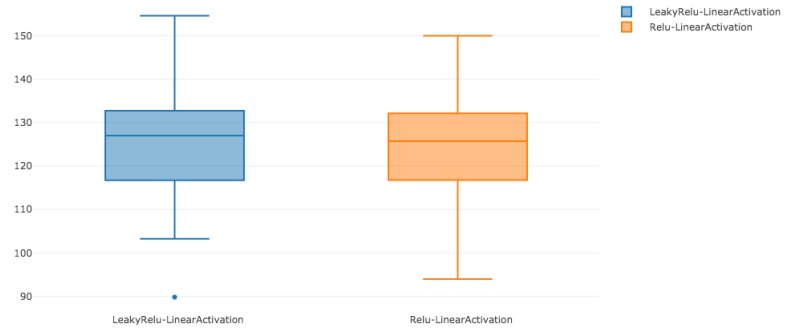

Fig. 12: FFN: Leaky ReLU vs ReLU
The plot above shows the P&L for 80 different predictive networks, grouped by activation, showing some improvements from the Leaky ReLU.

SAE Selection: MSE vs MAPE The P$L was compared for SAEs chosen using MAPE rather than MSE. The performance are fairly similar, with MSE being the better.

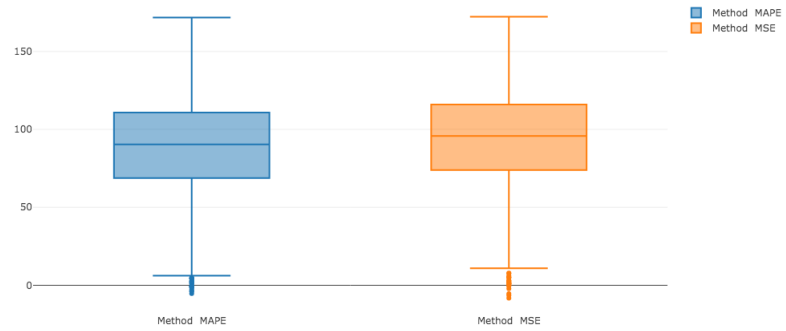

Fig. 13: MSE vs MAPE
The plot above shows the P&L for 1920 different predictive networks, grouped by SAE selection method, showing little difference, but some improvement from MSE.

L1 Regularization L1 Regularization was implemented and tested on a selection of 10 real assets, as per below.

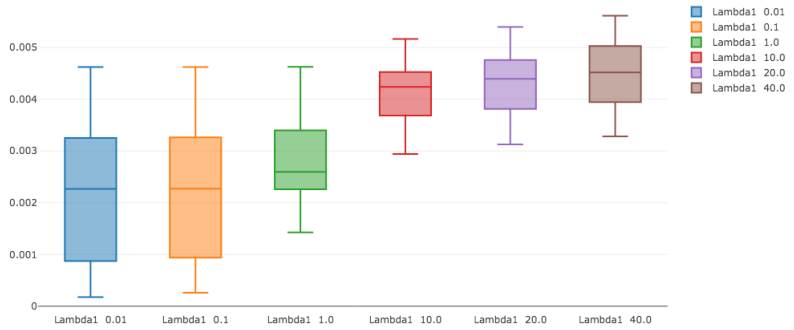

Fig. 14: SAE L1 Regularization - 10 Real Assets
The plot above shows the MSE for 360 different SAE networks, grouped by L1 Lambda, showing notable degradation as regularization is increased. Synthetic data tests showed no performance changes below 0.01.

Learning Rate Schedule Implementation A learning rate schedule, where the learning rate cycles through a sine-like curve every 100/200 epochs was implemented and tested on real asset SAEs as below.

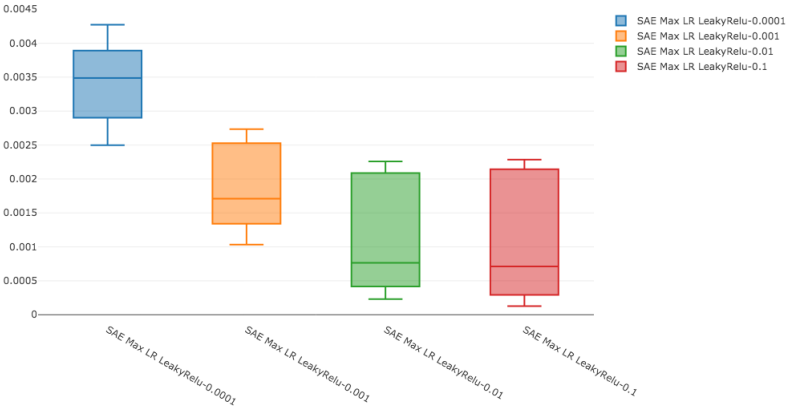

Fig. 15: SAE Learning Rate Schedule - 10 Real Assets
The plot above shows the MSE for 120 different SAE networks, grouped by the maximum learning rate, where all minimum learning rates were 0.00001. The scheduling with high rates shows a clearly better performance and exploration of solution space.

Denoising - On/Off Feature Selection Denoising where random features are switched on or off for each epoch was implemented.

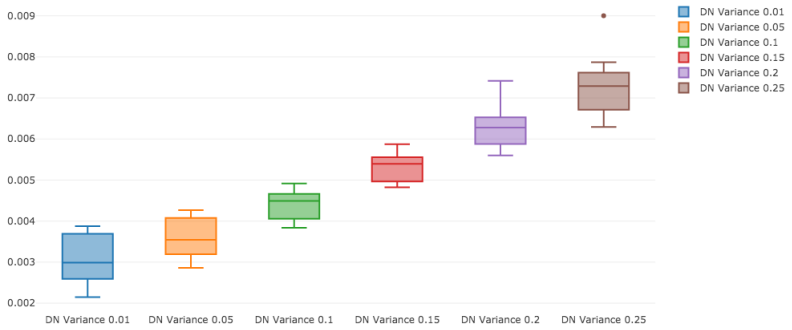

Fig. 16: SAE Denoising - 10 Real Assets
The plot above shows the MSE for 72 different SAE networks, grouped by the percentage of features switch off at random.

Validation Set Percentage A full set of tests were run on synthetic data to determine the effect of excluding data from the SGD training of the predictive network (through the mechanism of the validation set). It should be noted that these predictive networks make use of SAEs that were trained on 80% of the data regardless, which may prime the input of data to the predictive network despite the portion of data being excluded. It would seem that the use of historical data is very beneficial, but not the primary factor in determining P$L. It may even just be the equivalent of pre-training weight initialisation for the OGD phase.

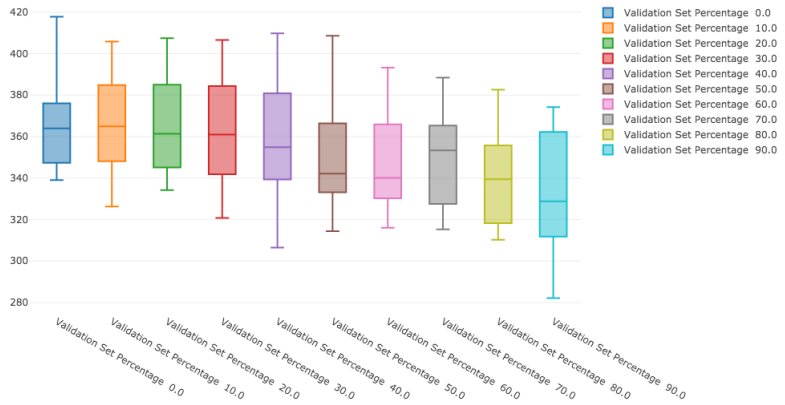

Fig. 17: SAE Denoising - 6 Synthetic Assets
The plot above shows the P&L for 160 different FFN predictive networks, grouped by the percentage of data excluded from the SGD training.

0.1 Limited Synthetic Asset Tests

Tests were run on 1, 2, 3 and 4 synthetic asset datasets using the same mean / same variance approach suggested to determine how well the model is able to learn in a limited (simpler) context.

Tests were run on a 5 day prediction point in order to aim for a profitable strategy, and different past window inputs of [1, 10, 20] days or [10, 20, 60] days to see if the removal of the 1 day price change would change the martingale effect being presented. SAEs were also trained with varying encoding layer sizes to test if the number of assets is typically the optimal number.

1 Asset High Mean, Low Variance test (0.6, 0.04) 32 Predictive Networks trained

Fig. 18: SAE Encoding Size P&L
The plot above shows the P&L for 1 asset according to the different encoding size in the P&L. Maximum P$L is the same for both 1 and 2 encoding.

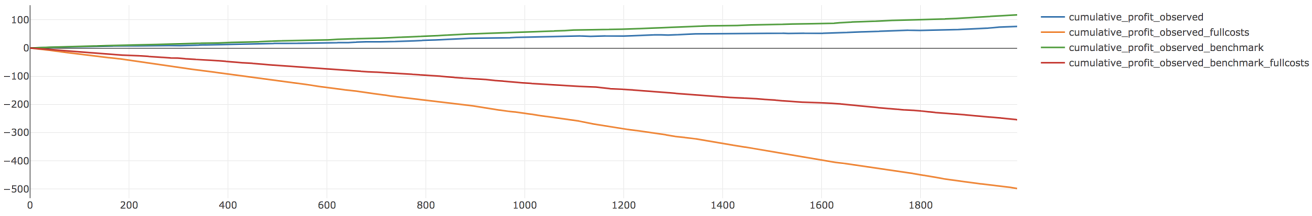

Fig. 19: Cumulative P&L for best 1 Asset network
The plot above shows the P&L for the best 1 asset network.

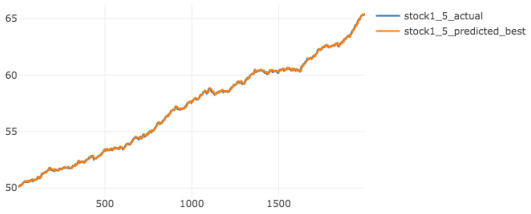

Fig. 20: Price Recreation for best 1 Asset network
The plot above shows the predicted prices for the best 1 asset network.

2 Asset Means = (0.05, 0.6) Variances = (0.02, 0.4)
128 Predictive Networks trained

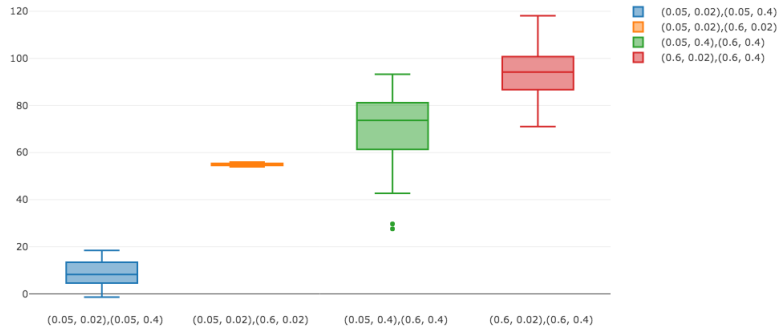

Fig. 21: Variations P&L
The plot above shows the P&L for 2 asset according to the different mean & variance combinations.

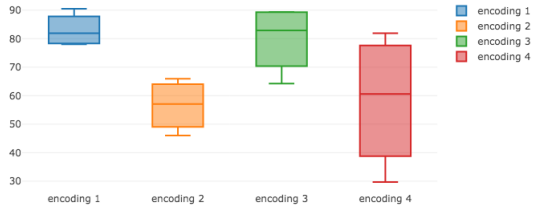

Fig. 22: SAE Encoding Size P&L
The plot above shows the P&L for 2 asset according to the different encoding size in the P&L. This goes against the idea of the number of assets being the best encoding.

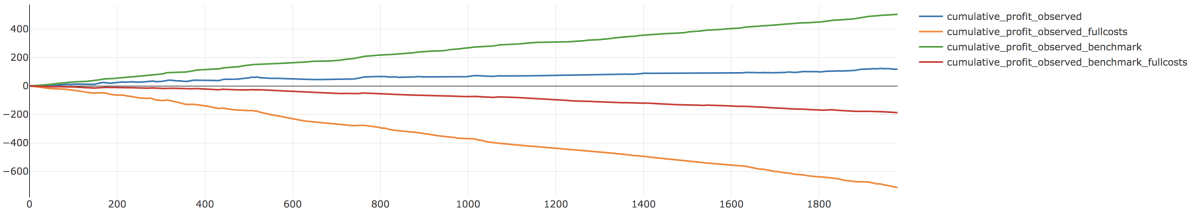

Fig. 23: Cumulative P&L for best 2 Asset network
The plot above shows the P&L for the best 2 asset network.

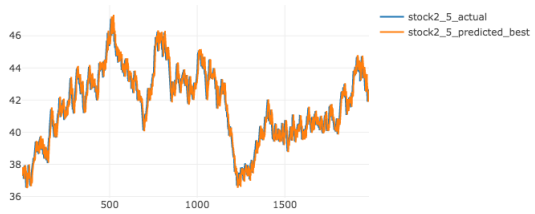

Fig. 24: Price Recreation for best 2 Asset network
The plot above shows one of the predicted prices for the best 2 asset network.

3 Asset Means = (0.05, 0.2, 0.6) Variances = (0.02, 0.1, 0.4)
240 Predictive Networks trained

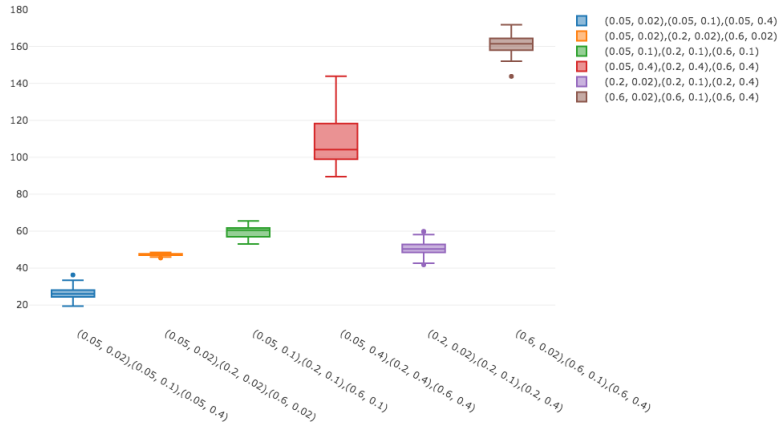

Fig. 25: Variations P&L
The plot above shows the P&L for 3 asset according to the different mean & variance combinations.

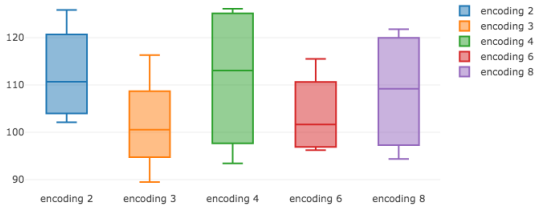

Fig. 26: SAE Encoding Size P&L
The plot above shows the P&L for 3 asset according to the different encoding size in the P&L. This goes against the idea of the number of assets being the best encoding.

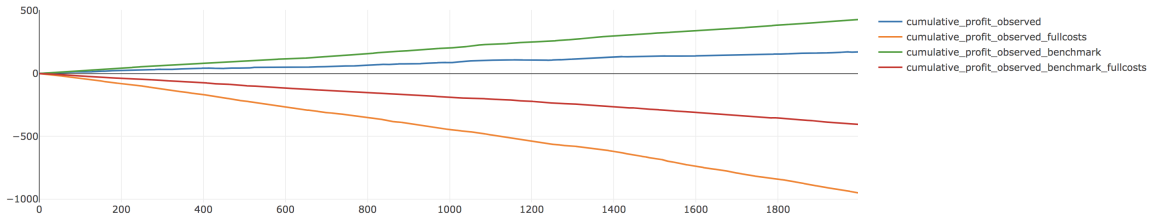

Fig. 27: Cumulative P&L for best 3 Asset network
The plot above shows the P&L for the best 3 asset network.

4 Asset Means = (0.05, 0.2, 0.4, 0.8) Variances = (0.02, 0.1, 0.3, 0.6)
384 Predictive Networks trained

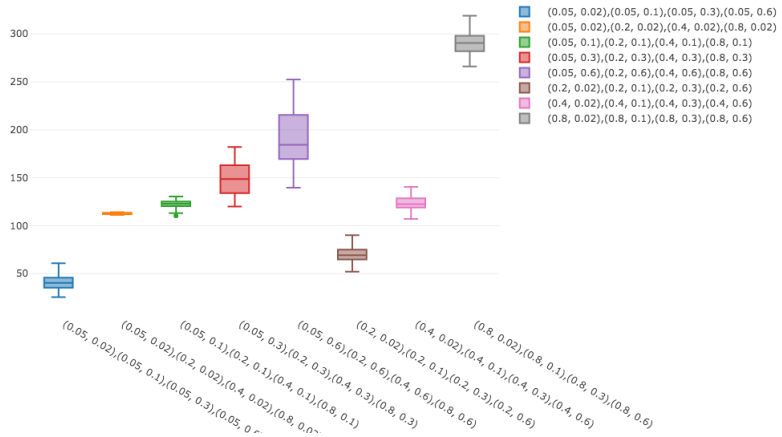

Fig. 28: Variations P&L
The plot above shows the P&L for 4 asset according to the different mean & variance combinations.

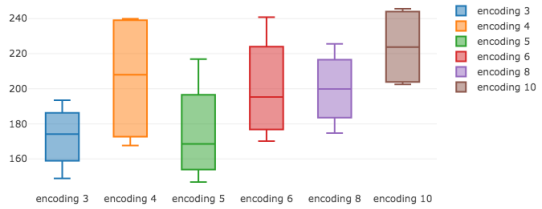

Fig. 29: SAE Encoding Size P&L
The plot above shows the P&L for 4 asset according to the different encoding size in the P&L. This supports the idea of the number of assets being the best encoding.

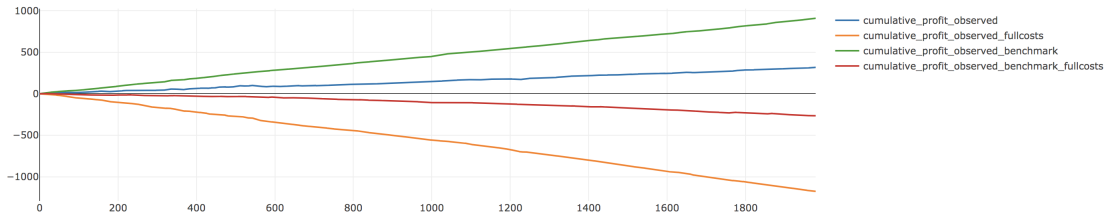

Fig. 30: Cumulative P&L for best 4 Asset network
The plot above shows the P&L for the best 4 asset network.

Effect of Rolling Window Sizes The effects of the [1, 10, 20] and [10, 20, 60] window sizes

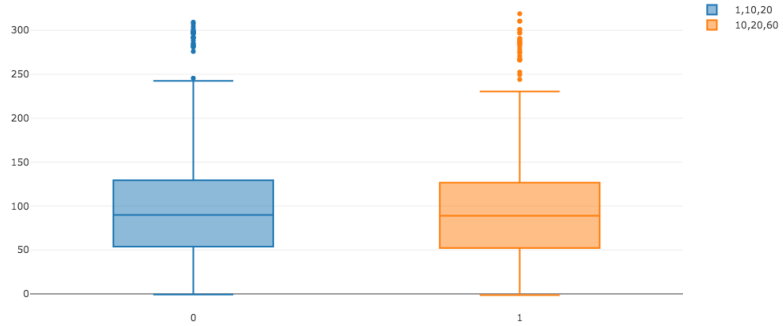

Fig. 31: Effect of Rolling Window Sizes on P$L for 1, 2, 3 and 4 Assets

Iteration 2: 26th April

Sigmoid Pre-training Efficacy The pre-training implementation was tested using the MNIST dataset using 2 validations: the first was RBM training to generate variations of the MNIST data which was inspected manually and considered to be working correctly, the second of which was to use pre-training on a classification network and assess the benefit, as in the graph below.

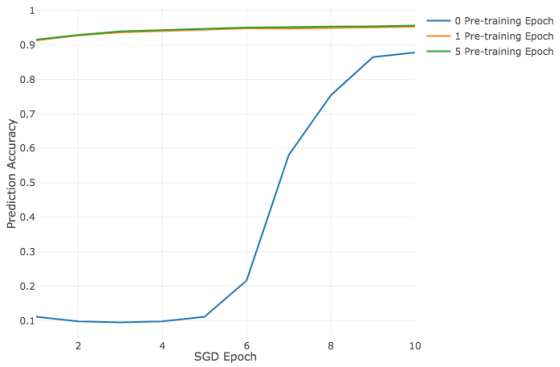

Fig. 32: SAE MSE By Scaling
The three series above show the classification accuracy scores (percentage) by epoch on an AutoEncoder which was used to classify MNIST images. The series were trained with 0, 1 and 5 pre-training epochs, and show a clear improvment in having an epoch of pre-training in the SAE formation (though not much for more than 1).

Having considered the implementation to be correct and effective in the classic use case, it was then trialled on the following SAE configuration, with the results in the graphs below.

- Data: Closing prices for AGL & ACL
- Processing: Time windows of 1, 5 and 20 (thus an input vector of size 6)
- Encoding Layer Activations: Sigmoid and Linear
- Encoding Layer Size: 5
- Hidden Layer Size: 15
- Number of Hidden Layers: 1, 2
- Learning Rates: 0.0000001, 0.00001, 0.001, 0.1, 1.0, 2.0

Fig. 33: Pre-training Effects on financial SAE
The boxplots here show the summary of configuration performances, by minimum MSE achieved, grouped according to the number of pre-training epochs which the network had. There is a clear favour to having no pre-training in this scenario.

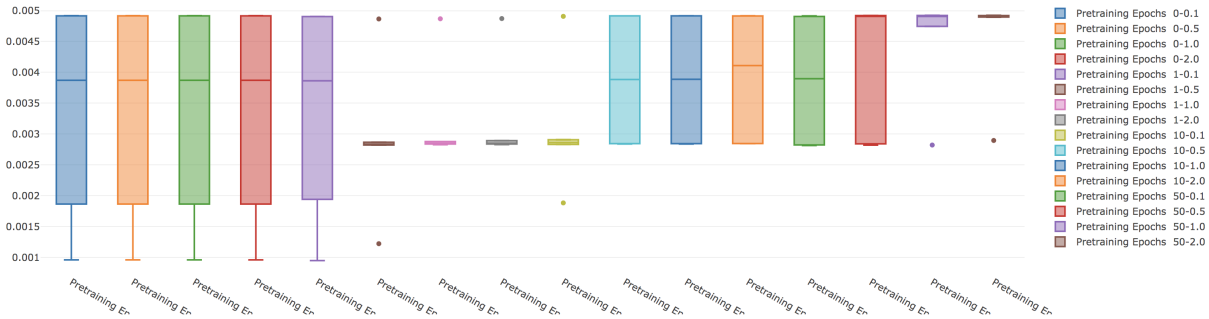

Fig. 34: Pre-training Effects on financial SAE, by learning rate
These boxplots show the same as above, but further grouped by learning rate. We can see the few samples that appears to be benefiting from having 1 epoch of pre-training are simply a result of the learning rate being small enough so as not to have much effect.

SAE Linear Activation Tests In consideration of the previous iterations results, numerous tests were run to assess the efficacy of Linear Activations in the encoding and output layers of the SAE networks. The list of configuration combinations are below.

- Data: Closing Prices for ACL, AGL, AMS, AOD, BAW, BIL, BVT, CFR, CRH, DDT
- Data Windows: 1, 5, 20 (thus input of 30)
- Data Scaling: Normalize, Standardize
- Weight Init,: Xavier Glorot Uniform
- All Layer Encodings: Sigmoid, ReLU, Linear (split by hidden, encoding and output)
- Encoding Layer Size: 5, 15, 25
- Hidden Layer Size: 60, 120
- Number of Hidden Layers: 1, 2, 3
- Learning Rates: Various for each different activation type (chosen by hidden activation used)

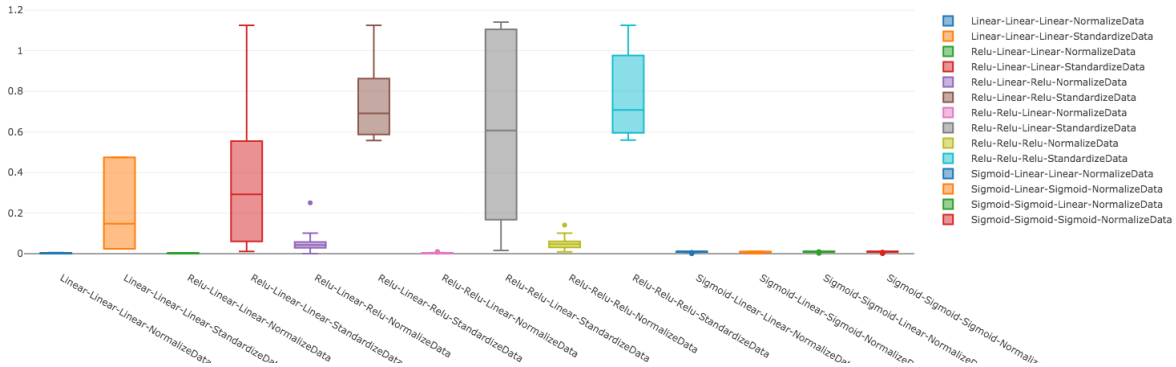

Fig. 35: Effects of Standardizing & ReLU Output
These boxplots show the MSE scores for the combinations run grouped by 'Hidden Activation-Encoding Activation-Output Activation-Scaling Technique. There is significantly poorer performance when Standardizing is used instead of Normalizing or when there is a ReLU output activation. These configurations are excluded from the graphs below.

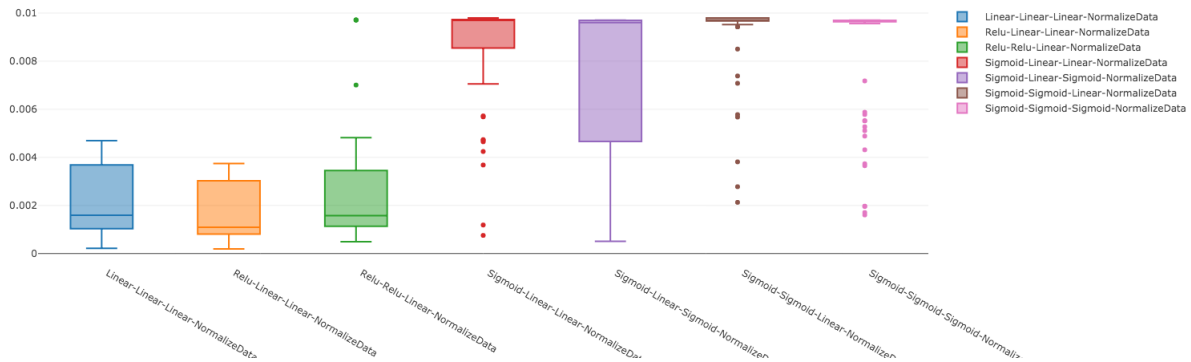

Fig. 36: Effects of Linear Activation
This is the same data as in Figure 4, focusing on the more effective configurations. We can see Sigmoid has largely poor performance unless there is a Linear encoding layer (a widely experienced behaviour), and seems mostly unable to outperform a fully linear network. The best configuration is Relu Hidden layers, with Linear encoding and Output - this makes sense with the non-linear benefit at hidden layers but with less loss of error signal and information at the output and encoding layers.

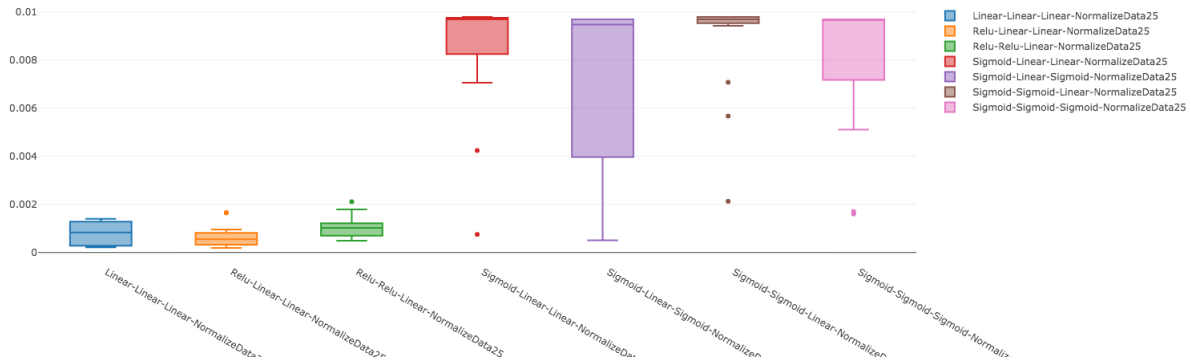

Fig. 37: Encoding Size 25
These plots show the performance for all configurations with an encoding layer size of 25 (input 30). There is once again a surprisingly high performance in the fully linear network.

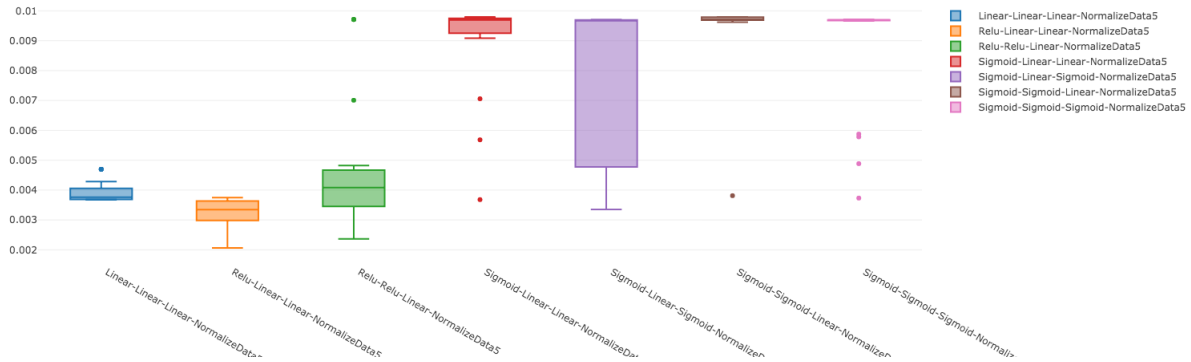

Fig. 38: Encoding Size 5
These plots show the performance for all configurations with an encoding layer size of 5 (input 30). Here we finally see the benefit of non-linear activations in the ReLU based newtorks, which the fully linear is not able to outperform.

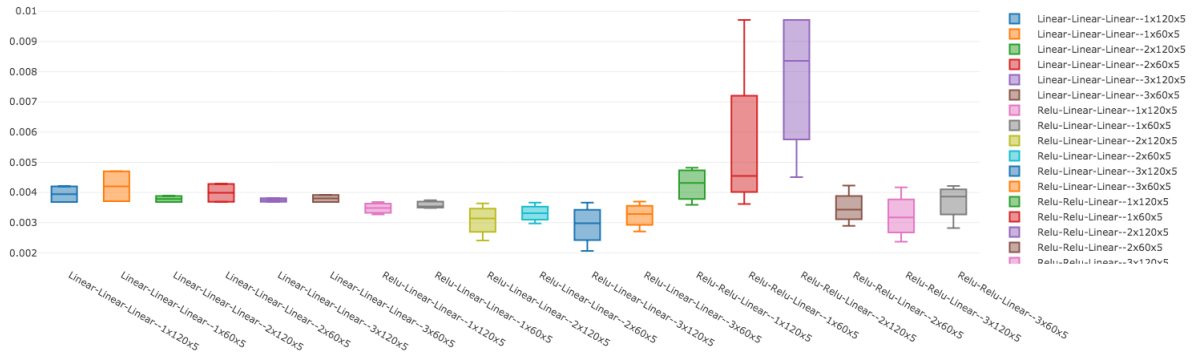

Fig. 39: Performance by Network Size
We can further break these down by network size, and see performance is as one would hope, with the best ReLU configurations corresponding with more layers and of larger sizes.

Denoising SAEs Using the same data and general set up as above, a series of tests were run for DSAE using the best ReLU configurations and a gaussian noise process such that $\tilde{x} = \mathcal{N}(x, \sigma)$

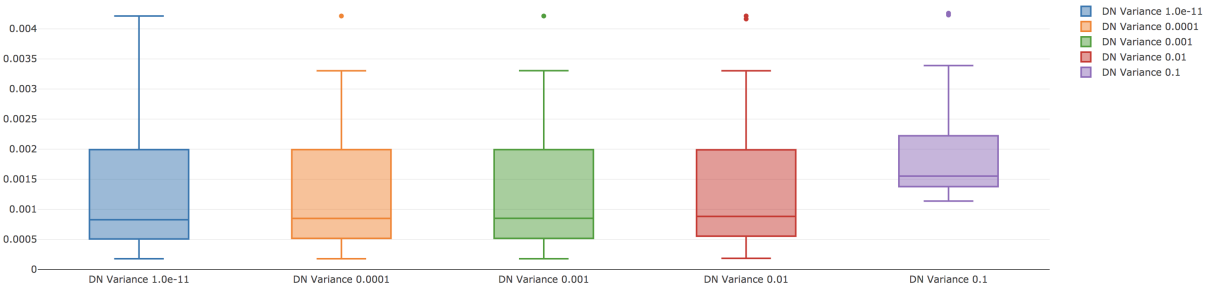

Fig. 40: Performance by Denoising Variation
The above groupings show the decreasing SAE performance as the level of denoising is increased. The grouping with variance 1.0e-11 essentially represents no denoising.

Linear Activations in FFN A series of predictive FFN networks were run, with the following combinations

- SAEs: Best performance (lowest MSE) of encoding sizes 3, 6, 9, 12, 15 (input 18), of both Limited Scaling and typical Scaling methods
- Output Activations: ReLU, Linear
- Hidden Activations: ReLU, Linear
- Layer Sizes: 40, 80
- Number of Layers: 1, 3
- Various Learning rates for SGD & OGD training

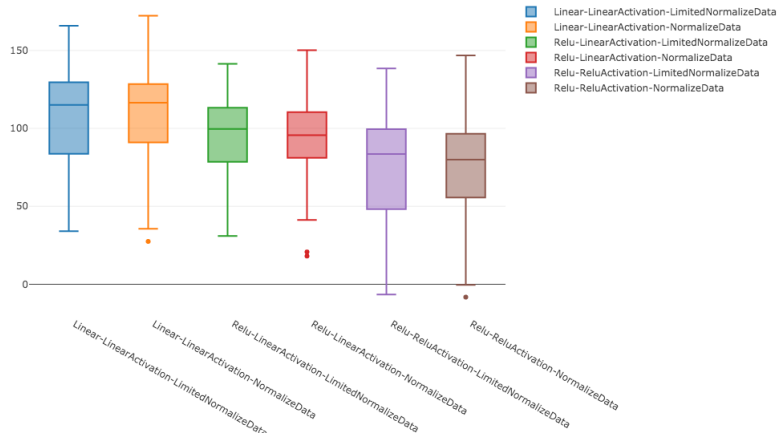

Fig. 41: Linear Activations in FFN
The above groupings show the profits generated for different combinations of Hidden Activation - Output Activation - Scaling Method (which is for both SAE and FFN). There are several takeaways:
1. Higher performance of Linear networks (may be down to network size and amount of input data)
2. The limited scaling technique is having a noticeable impact on profits
3. The decrease in performance with ReLU output persists

SAE Effects The graph below shows an aggregation of the above configuratios by SAE sizes.

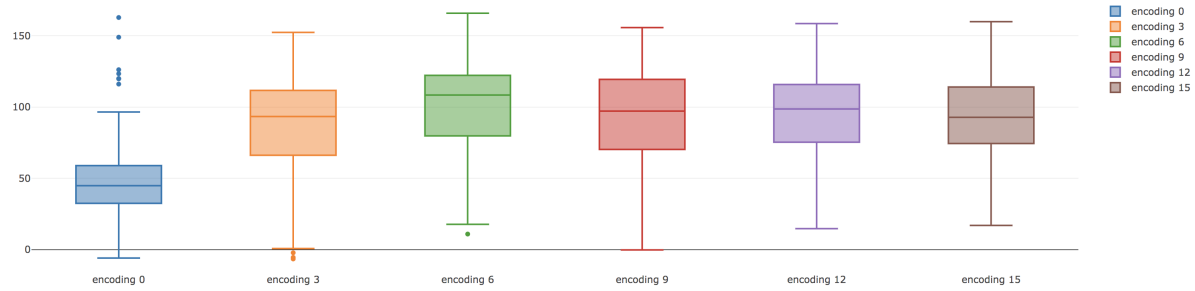

Fig. 42: SAE Effects on P&L
The above groupings show the profits generated for different SAE encoding layers. The results are encouraging, showing that no SAE (encoding 0) has the worst results, and that the best results are found in one of the lower encoding layers (6).

Predictive Network Sizes Similarly, the below graph shows profits for the predictive networks sizes.

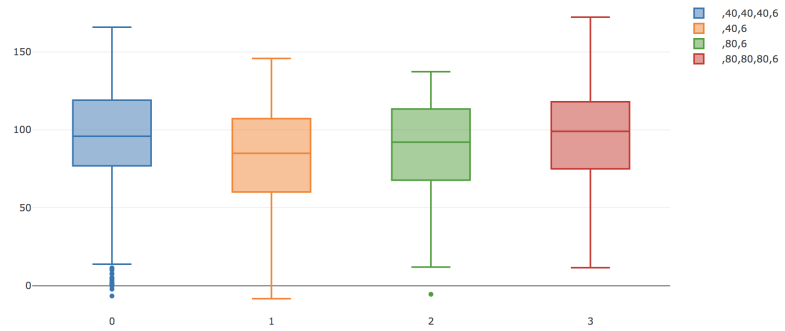

Fig. 43: Network Size Effects on P&L
We can see the effects are largely as expected here, with networks of more layers and larger layer sizes having a better effect on profits.

Cross Validation Effects The same set of configuration were run with a Cross Validation set of 0% and 10%.

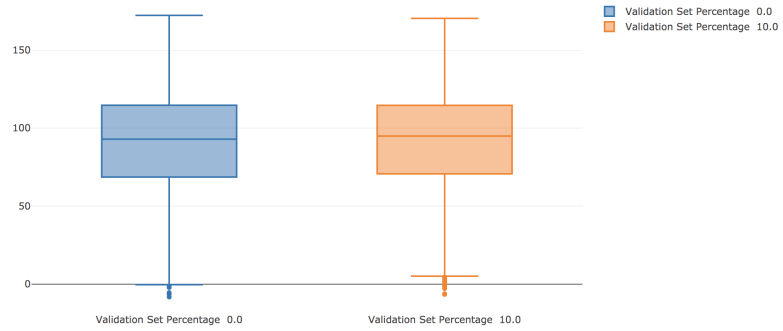

Fig. 44: Validation Set Effects on P&L
This tested the witholding of 10% of the data from the SGD training versus witholding no data. Curiously, there doesn't seem to be a notable effect here, possibly speaking to the much larger general effect of the OGD training in comparison.

Returns Analysis This section will look at the aggregate returns, and the predictions of the best networks by profit and OGD MSE.

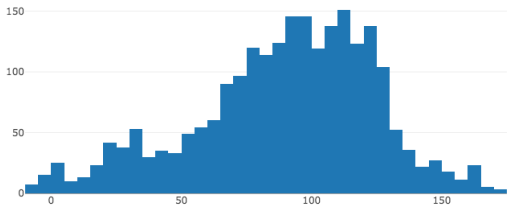

Fig. 45: Histogram of P&L from Predictive MMS
The P&L values seem to be rather low, with few losses. Could warrant a different trading strategy being used.

The returns below are for the best P&L network with Relu hidden activations. This network had an SAE encoding layer of size 3, 3 hidden layers with ReLU activations and a linear output layer, which

produced a profit of 150.128

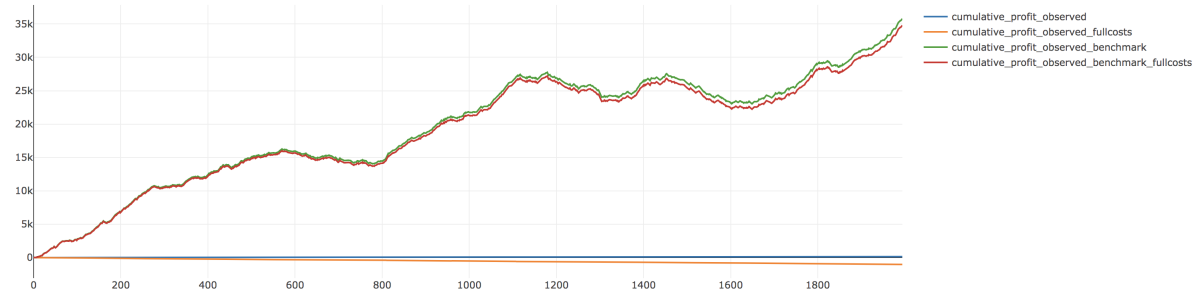

Fig. 46: Cumulative P&L
We can see even the best configuration seems to be delivering very poor results relative to the benchmark, with negative P$L with costs considered.

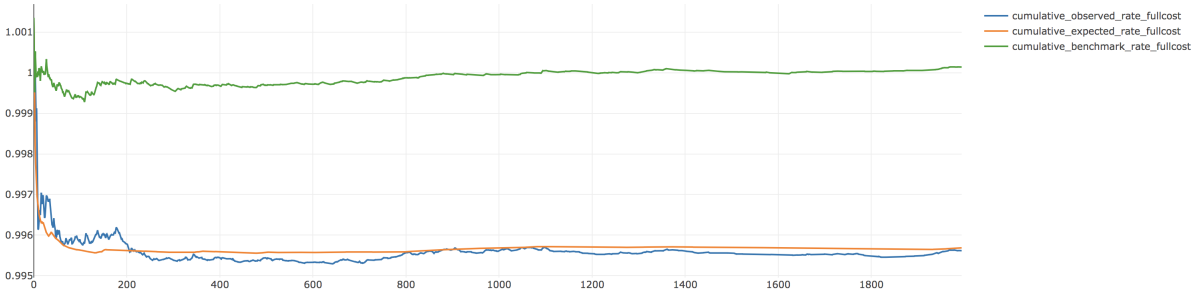

Fig. 47: Cumulative Rates
We can see even the best configuration seems to be delivering very poor results relative to the benchmark, with negative P$L with costs considered.

Fig. 48: Price Predictions
This graphs displays the actual prices for the one stock, as well as the predictions made by the two best networks (chosen by MSE and Profit)

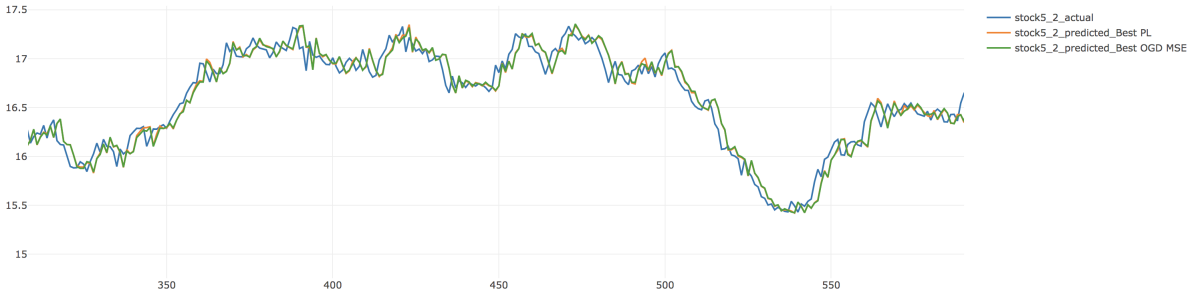

Fig. 49: Price Predictions (Zoomed)
A closer zoom reveals that it looks like the predictions are very much a lagged time series

Iteration 1: 12th April

Scaling Unfortunately a bug resulted in the predictive networks not having a linear activation layer implemented (despite the naming). Current results suggest the limited scaling approach is resulting in worse MSE scores for both SAE and OGD. The OGD profits are also looking potentially worse (a linear activation implementation would need to be tested though).

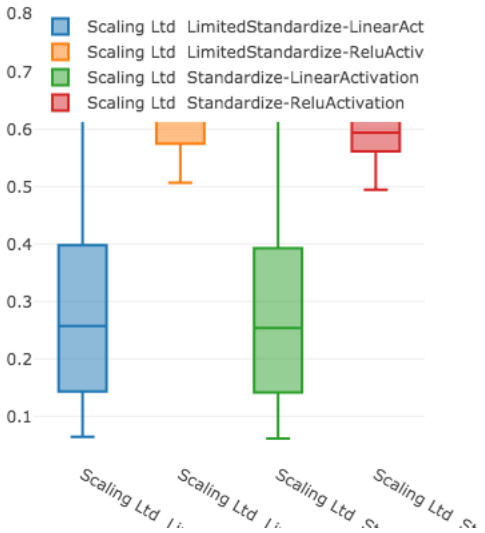

Fig. 50: SAE MSE By Scaling

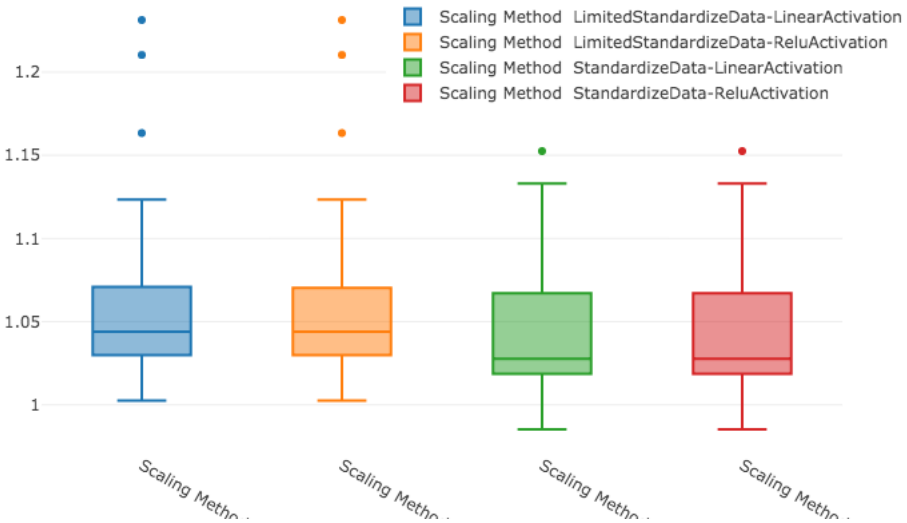

Fig. 51: OGD MSE By Scaling

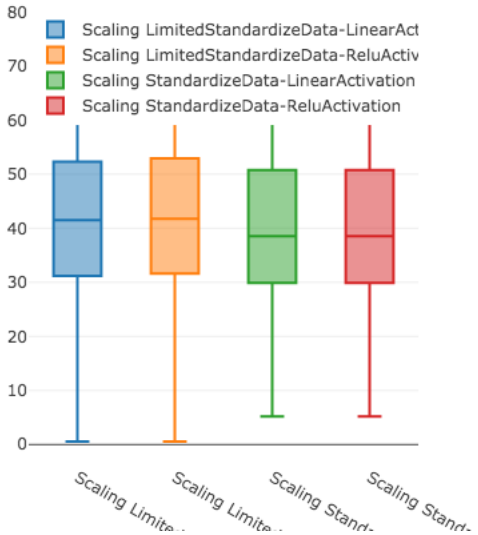

Fig. 52: OGD Profits by Scaling

General Configurations & CV A larger set of configuration were run to test several concepts, as well as the removal of the CV step. The CV methodology needs to be kept for SAE in order to select a best set of SAEs, but is not strictly necessary for the SGD predictive training. In this case the trade off is that without CV a set number of epochs must be chosen (rather than monitoring the changes in OOS errors). All results below are run without CV in the SGD training.

- SAE Configurations:
  - Encoding layers: 3, 6, 9, 12, 15 (input of 18)
  - Layer sizes: 40x1, 40X2, 80X1, 80X2
  - Various learning rates
  - Different scaling functions & output layers
- FFN Configurations:
  - Encoding Layers: 3, 9, 12
  - Different scaling functions & output layers
  - Various learning rates
  - Layer sizes: 40x1, 40X2, 40x3, 80x1, 80x2, 80x3
- Dataset Configurations:
  - Synthetic Data
  - 5000 timesteps
  - Aggregation of 1, 7 and 30 to predict 2
  - 60/40 split between SSGD and OGD

The set of prices tested are shown below

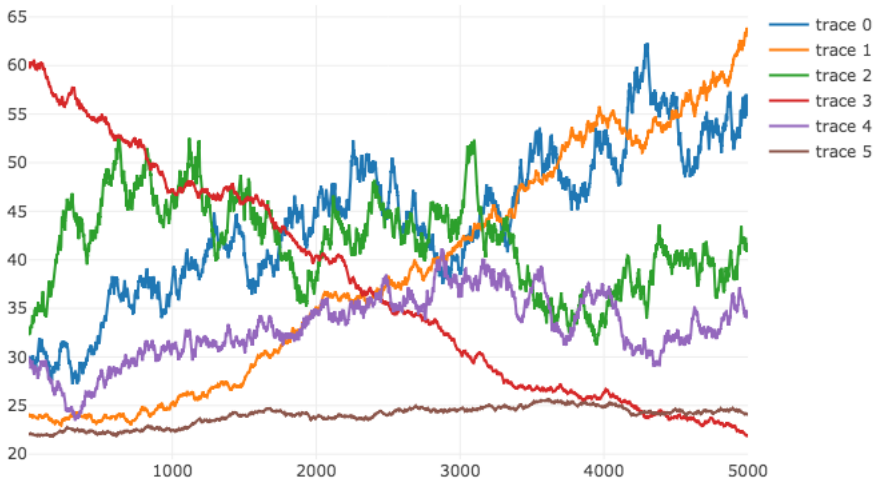

Fig. 53: Price Plot

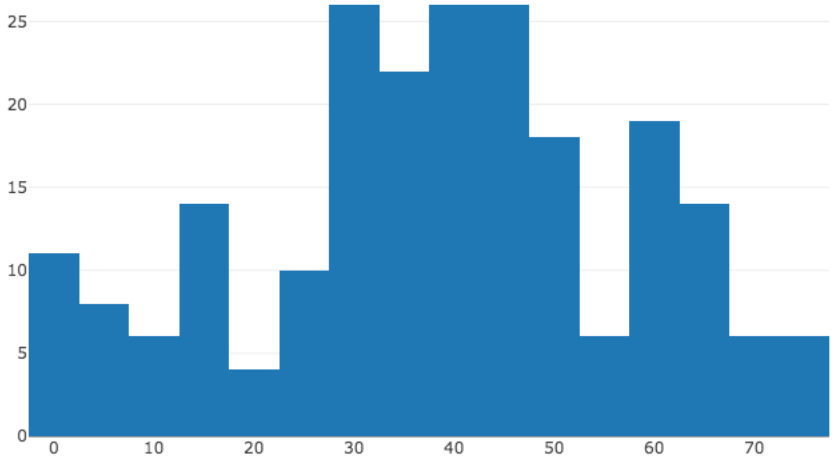

Fig. 54: PDF of all Profits Generated

Profits by the SAEs chosen are below, suggesting that there is some level of feature selection being performed by the SAEs.

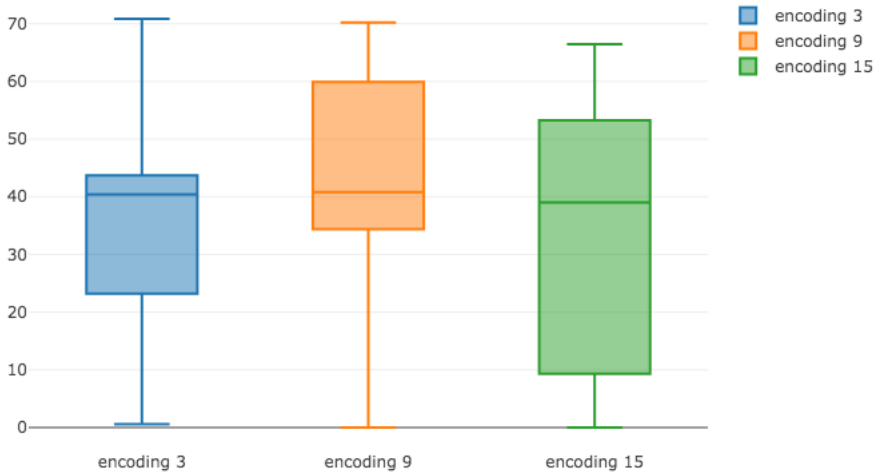

Fig. 55: SAE Profits

Profits by network structure are below, suggesting that simpler networks are performing better. This may point to some leve of overfitting occuring in the larger networks.

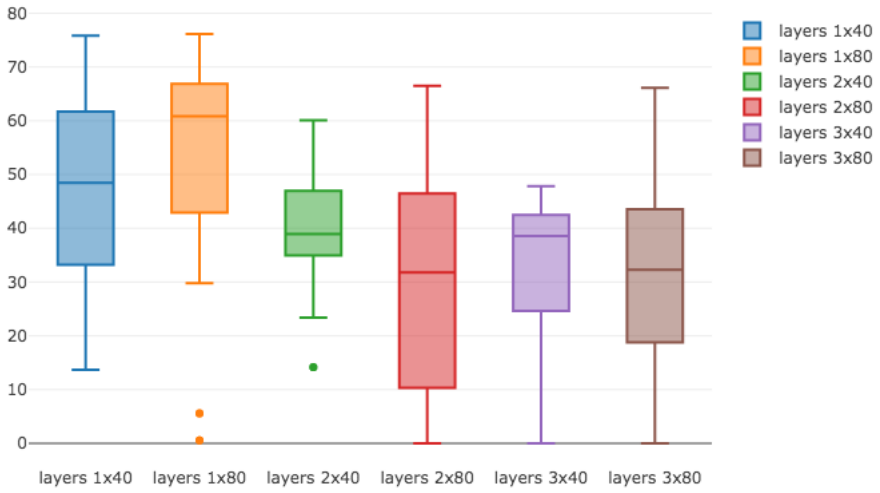

Fig. 56: Network Structure Profits

Return Graphs The graphs below show the cumulative return rates, profits and daily rates of the best network. The best network was chosen by MSE rather than profit, for reasons discussed below.

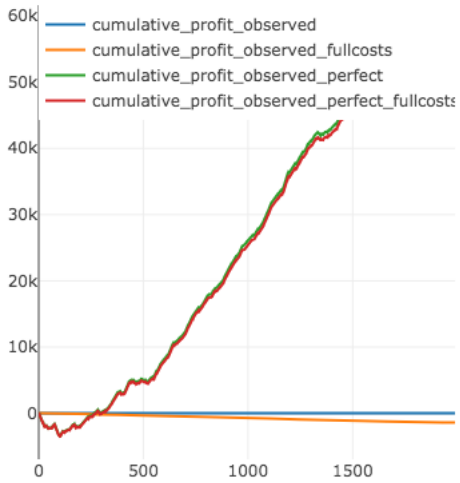

Fig. 57: Cumulative Profits

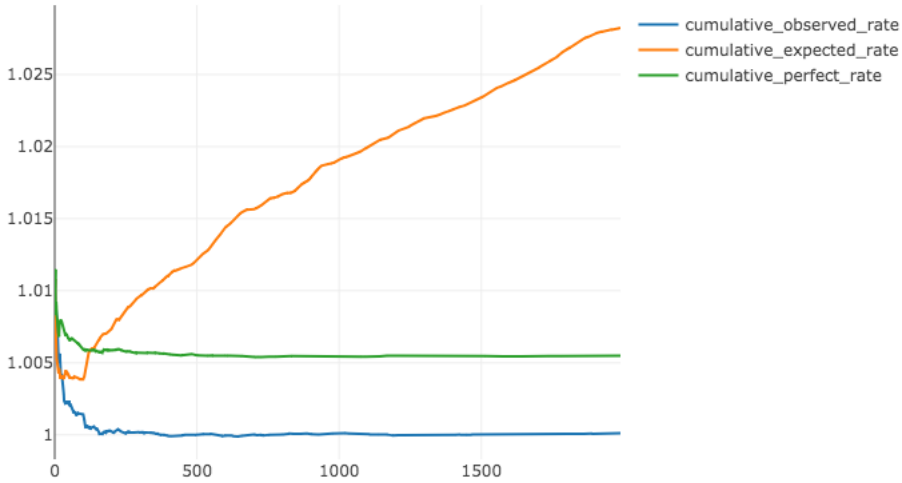

Fig. 58: Cumulative Return Rates

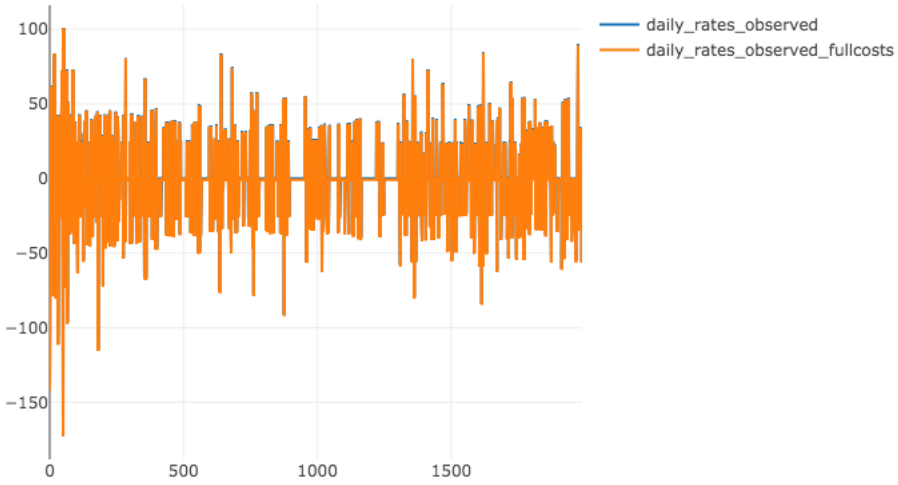

Fig. 59: Daily Rates

Price Predictions The price predictions are showing 2 problems: heaving oscillations occuring at most points, and the most profitable configuration having a relatively poor level of prediction.

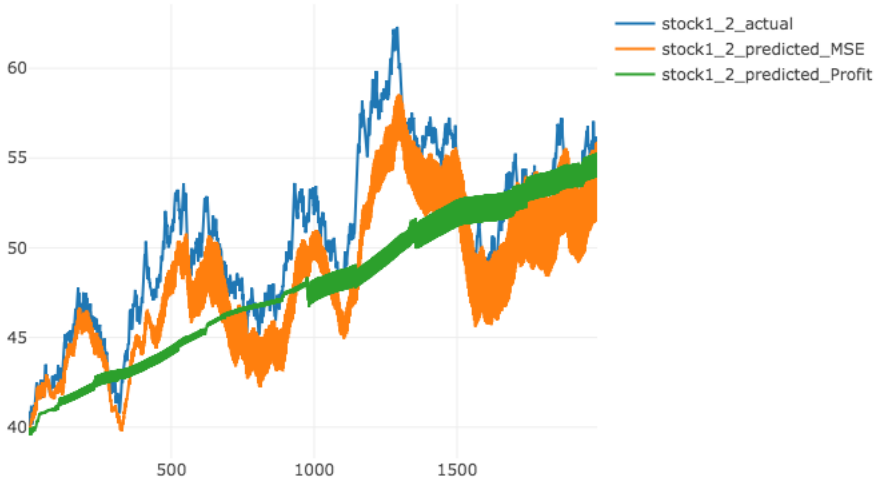

Fig. 60: Stock 1

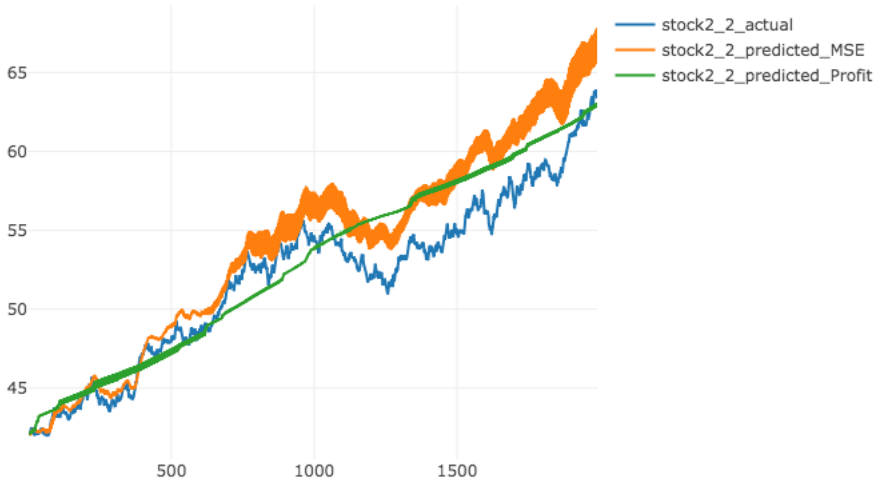

Fig. 61: Stock 2

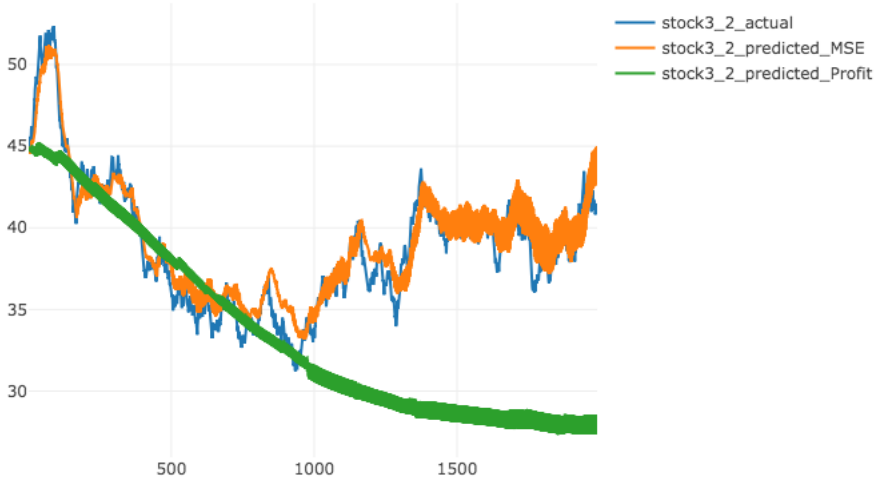

Fig. 62: Stock 3

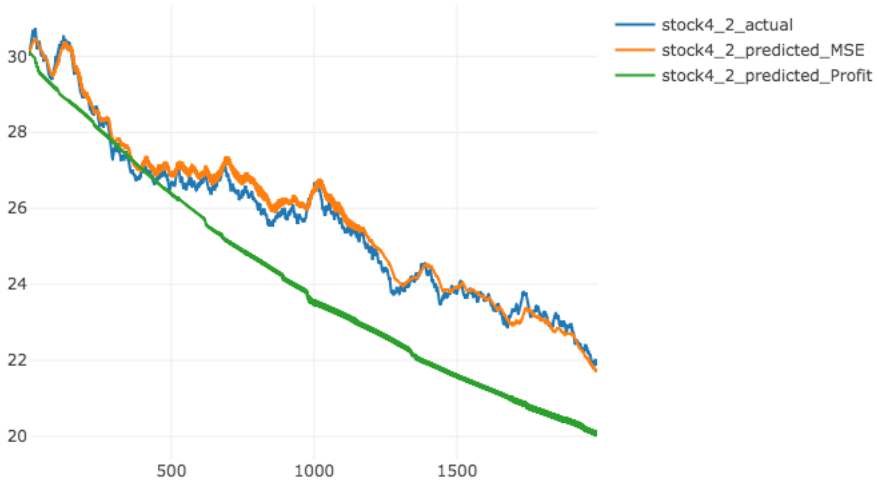

Fig. 63: Stock 4

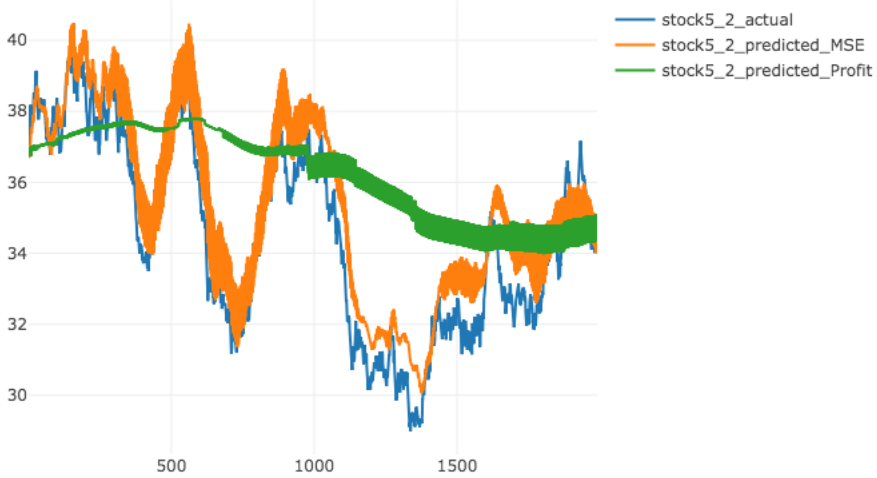

Fig. 64: Stock 5

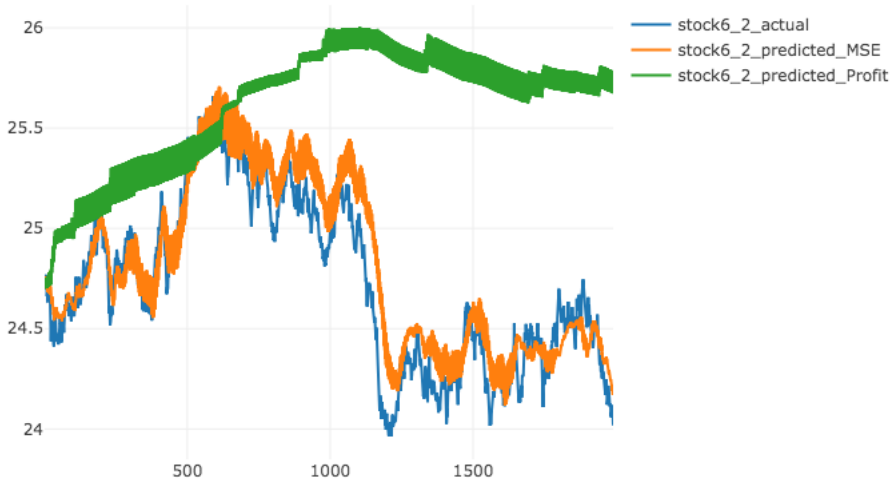

Fig. 65: Stock 6

Sigmoid Pre-training The pre-training for weight initialization in the Sigmoid networks, which is supposed to improve performance, seems to be having an adverse effect:

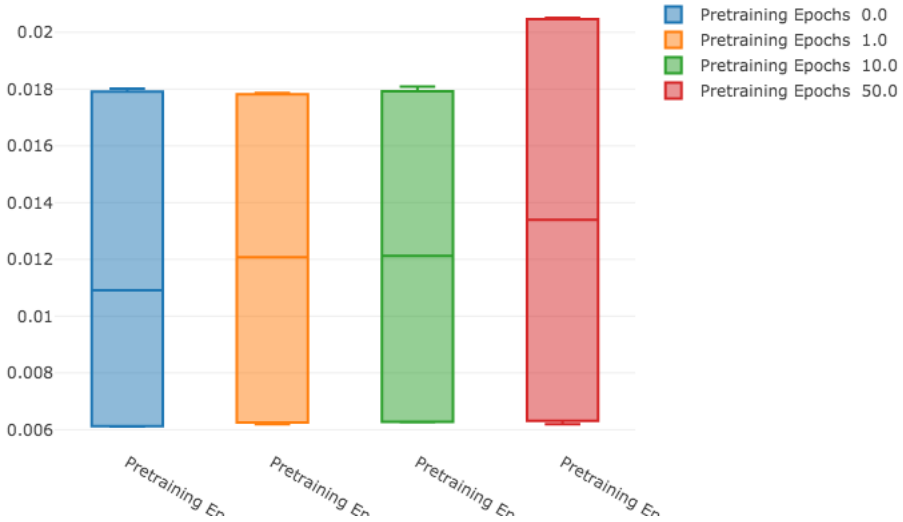

Fig. 66: Stock 6

4. Synthetic Test Plan

Expected Time: 1 Week

General Configuration

- Data: 10 Synthetic Assets
- Rolling Windows: [1, 5, 20] / [5, 20 60] / [10, 20, 60]
- Initializations: Xavier, He
- Activations: Leaky Relu With Linear Encodings and Output

SAE Configurations (945)

- Learning Rate Schedules: 0.005 to (0.01, 0.05, 0.1) by (100) epochs
- Network Structures:
  - 120, 60, 30
  - 90, 60, 30
  - 90, 90, 90
  - 90, 90
  - 120, 90
  - 120,60
  - 120
- Encoding: 25;20;15;10;5

FFN Configurations (2835)

- Best SAEs by encoding size and window aggregations (15 total)
- Learning Rate Schedules: 0.005 to (0.01, 0.05, 0.1) by (100) epochs
- Network Structures:
  - 120;120;120;120
  - 120;90;90;60
  - 120;90;60
  - 120;120;120
  - 120;120
  - 120;60
  - 120
- L1 Lambda: 0; 2 values
- OGD Learning Rates: 0.01; 0.005; 0.05

5. Milestones & Deliverables

Finish Date	Item
Requires Corrections	1 Literature Review
In Progress	2 Data Collection
Done	3 Data Log Fluctuation Processing
Done	4 SAE Implementation
Done	5 Online FFN/LSTM Implementation
Done	6 CSCV Implementation
Done	7 Synthetic Data Generation
Done	8 Finalise Implementation Process
Done	9 Database Implementation
Done	10 Learning Optimizations
	11 Test full implementation on synthetic data
	12 Test full implementation on actual dataset
	13 Statistical Analysis (incl. profitability of trading strategy)
	14 Write up & Revisions

6. Process Implementation

The steps below indicate how everything is expected to fit together at a high level perspective:

1 Data preparation

- The data is processed into day to day log fluctuations
- At each time point, rolling historic summations are calculated (e.g. the past 1, 7 and 30 days)
- At each time point, rolling future summations are calculated (e.g. the next 2 days)

2 Data Segregation

- The processed dataset is split into 2 parts, to be used for the following processes: SAE/SGD training and OGD training (e.g. 60%, 40%)
- Bailey's MinBTL method should be considered to determine this split length

3 SAE Training

- The dataset defined for SAE/SGD is used to train the auto-encoder networks (using either RBM pretraining/weight initialization algorithms + SGD)
- Notably, this SAE/SGD dataset will itself be split into training & testing portions in order to train and validate the SAE network
- The SAE training does not suffer from the same backtesting concerns as for the predictive network. A full set of models for difference encoding layers can be trained to sufficient performance, and then used as a configuration aspect for the predictive training.

4 FFN/SGD Training

- The same SAE/SGD portion of the data will then be used to train the FFN network to predict future prices, using the implemented SGD algorithm & optimizations.

5 OGD Training

- The second portion of data will be used to train and optimize the FNN from 4 for OGD. The purpose of this will be to optimize to train the network and produce returns.

6 Returns generations

- A basic trading strategy will be used to generate returns from the predicted and actual prices in the holdout dataset. These returns for each time period and each configuration of the model defined by steps 3-5 will form the M matrix for the CSCV process.

7 CSCV & PBO

- Using the M matrix from 6, the CSCV process will be run which will then allow a calculation of PBO.

7. Meeting Minutes

16th June 2018

- Literature Review Include: Wilcox & Debbie Hierarchical Paper; Crutchfield & Farmer: Geometry of Time Series; F Takens: Detecting strange attractors in turbulence; Reference Riaz
- Implementation
  - Key implementation point was the data segmentation choices
  - LSTM network would resolve this point for us (CNN convolution could also potentially work)
  - Non LSTM will run risk of correlating times with stocks
  - Representation needs a mixture of time scales as well as the features
  - Will be considering data for the entire index at any time point t
  - Will need to consider how to address the changing constituents
  - Aspect of how often the feature selection is updated (e.g. SDAE in online model)
  - Does it get updated in the online model ?
  - SDAE result should be similar to edge detection which will give set of predictions essentially
  - Use known test cases for CSCV method development
- Data
  - Use simple noise with Gaussian distribution & long trends to develop network models (6 variations of increasing, decreasing and flat with low or high variability)
  - Build backdated tests module with surrogate data afterwards
  - Not going to try intraday data due to different nature of problem
  - Implementation of either a flat file or MongoDB
  - Important to include Volume datav
  - Volume time -¿ might be far more predictable; volume and variance linked
- Other
  - Admin:Single document - table of contents (nice to have) with authors; sections etc. ; full doc with versioning; PDF + latex (always send both)
  - Further Correspondence: * Iterate over FFN/LSTM implementations rather than starting with LSTM * For the meantime, assume the models do not differ fundamentally enough to be considered as separate for the PBO calculations

29th March 2019

- Dissertation
  - Generally frame the idea as a simplest high complexity framework in rational testing framework
  - Need to highlight and sell the SAE separation in terms of reasoning and value
  - SAE training updates: just make comment that not updating
  - Include discussion around decoupling of model components
- Data Considerations
  - 80/20% split for SGD: Should this be used? May not be necessary in full process. Could also consider using all the data for OGD rather than SGD/OGD split.
  - Split for SAE&SGD/OGD split: Baileys MinBTL should at least be discussed, and possibly used.
  - Prediction goal

    1 : A 3/5 days average at end. Buy / sell according to above / below.

    2 : 1-30 days distribution with error variances at each point. Consider how far ahead highest possible forecast is and what the spread on that is.¿ outside scope but need to address to show its been thought carefully about Money management strategy for model (out of scope) but should have some consideration in terms of such a distribution

  - Scaling
    * Paragraph or 2 showing this is a technical problem, and point out that want this learning to be ind. of normalisation
    * Currently choosing to normalise correctly in training set
    * Mention iterative normalisation which will converge: Can benchmark large covariance matrix against noise model for matrix and use PCA and plot eigen values for noise distribution. Would still be removing some level of signal -¿ and signal chosen should be as far outside noise model as possible.
  - Datasets
    * Continue to use surrogate data for testing, and then use close & intraday sets for actual results
    * Daily data have enough to submit
    * Intraday: trade every 5 minutes?
    * Intraday will make a good inclusion and extension

- Results Reporting
  - Nyquest frequency and breakdown on aliasing: need to discuss this to validate the [2] choice, and note that naive prediction could expose us to aliasing and fitting to the band ripple rather than underlying signal in prices
  - Continue with and clearly state how the arithmetic money management strategy is working
  - Processing signals but generating P&L based on the MMS
  - The performance measure is based on these P&L returns
  - Generate daily P&L (same sampling rate as data): includes both trades which are running
  - Split up reporting into the expected and actualised returns for each time point
  - Return calculations need to include costs
    * Capital Costs: 10% p.a. borrowing rate to fund MMS. Payed back every 2 days.
    * Transaction costs: 45bp (as appropriate for Top 40 category shares)
    * Generate results with and without these costs. Can then inform decision making on before and after costs.
    * Articulate: Ideal world would have the MMS inform the learning choices in the model

12th April 2019

- Implementation Choices
  - Scaling Methods: Just expose the decision and impact of using a scaling technique that cant account for future data
  - Cross Validation: Demonstrate payoff either way and make decision from there
  - Denoising Stacked Autoencoders: Try gaussian noise approach and assess results
- Time Window Aggregation
  - Should be 5 for week same for month (should be 1, 5, 20, 60)
  - Should be using a simulation of horizons to create distribution and choose time windows that way
  - Defend through domain knowledge, pose and suggest that it should be learnt - not going to be with primary scope
  - Calendars because thats how traders work - defend this way
  - Know system is hierarchical in feature space and time (articulate this and look at Tims paper)
  - One day: non-tradeable mean reversion, weekend activity for week, mutual funds rebalance monthly
  - Consider inclusion of 90 days
- MMS
  - Highlight the separation of MMS and signal
  - Long only is fine: If theres a strategy / method that consistently loses money - would be a great finding. Right side of MMS, wrong side of pattern recognition.
  - Point out that learning the patterns should not be coupled with making it profitable. Articulate that RL should not be used in the sense that choosing features etc. : highlight modularisation.
  - Daily rates: need to separate pattern recognition signal and MMS in text

  - Benchmark is perfect information world, so correct. Check Can we learn to beat best stock paper ? Paper
  - Shorts: Discuss stop losses and pyramiding stop losses training - way of making MMS more effective (bands implementation). Should be machine learned ideally but dont have that.

- Price Predictions
  - Oscillating predictions: potentially aliasing feedback?
  - Otherwise, MSE vs. Profits finding is interesting: understand why the differences are seen and what the model differences are

- Pre-training Findings
  - Training in a lot of noise may be jumping to pathological corner cases in solution space
  - Should be useful to constrain solution space
  - Data is pathological though, reduced data that can be used for actual algorithm
  - Result is optimising in a subspace - throwing data away in a sense
  - Check IID aspects and requirements for pre-training
  - Important story: dynamic nature of financial market is not the same as some IID distribution
  - Odds are changing through time
  - Ringfence and add diagnostics

- Generally (Dissertation notes)
  - Focus on the modularisation of the framework and how it enables exploring of components and algorithms rather than an explicit solution
  - Perfect -¿ Benchmark
  - Return should be P&L, return is %
  - Text: Times roman not italics for functions and have equation numbers
  - Enumerate rather than bullets
  - Table of tables, figures, variables used & definitions and page number introduced
  - Y tilde :measurement, y U feedback process error, y hat projection, y theoretical
  - Notations: Look at process control by G
  - Trade Long should be teletype

26th April 2019

- General Notes around results
  - SAE encoding results: degrees of freedom argument. Make it clear that it was on synthetic data and same seed etc.
  - Day on day with real data is almost perfectly predictable, but you cant trade it (1 day momentum). Cant trade from the closing auction, but no volume and cant get positions in it. 2 day onwards you get the martingale effect .
  - Prices are nonlinear in breadth, but not in depth of time series
  - Highlight that it is cross sectional depth of data that will give advantage of these models rather than breadth of history - ML is possible, but have to load up with every single current data feed. Patterns are nonlinear but adaptive because of feedback - tech not mature enough to leverage that yet
  - Profit- log scale on graph
  - Network sizes: order by layers and layer size to see pattern
  - Speaks to argument that its cross section of information now rather than temporal dynamics

- Synthetic Tests
  - P&L Found: High means and low means. Do edge case testing
  - 2 stocks: same means, different variances
  - 2 stocks: same variances, different means
  - Then up to 4 and 6 assets and check results for P&L
  - High low medium means etc. Log axes.
  - 2 assets: 0 on means, 4 and 6 after with the different variances. Will pinpoint whether its working or not. Cases with high variance and low means -¿ expect to lose money.
  - Consider test framework on HMM to assess linear activation results

17th May 2019

- General Dissertation & Project Notes
  - Highlight value of framework built, including graphs and diagnostics. Have built a tool that allows exploration of complexity in data.
  - Reason why we have a story is the focus on process that the project has gone through
  - Cumulative Graphs P&L: Highlights markets are predictable but not profitable with costs. Show as table rather than graph: matrix of data, year 1 data, year 2 data
- Results & Synthetic Data
  - Martingale: relief -¿ cant make model any more complicated.
  - In finance: actually data starved due to lack of cross sectional data. Time series effect in comparison to other data.
    - Financial Time Series: Cannot know the probabilities in financial markets. We have evidence in favour of that view
    - They are not organic and stationary - findings argue in that favour
    - Cant predict rare events in finance, and so cannot hedge in finance. Innovation vs hedging. Cant risk manage properly in finance due to lack of predictability of events.
  - L1 Reg is regularising something that isnt overfitting in the first place. If it does help then theres nonlinear signal in the data -¿ ergo framework is fitting relatively simple linear data. Contrasts with the Learning rate schedule results -¿ warrants further testing on different (real) data.

21st June 2019

- Formatting
  - Numbers rather than bullets for all lists
  - Figures: split out and format correctly
- GBM Generated Data
  - Warning that financial time series, ergo all about the time series averages -¿ cant trust ensemble based machine learning
  - Can just reference the Ergodic Economics paper
- Datasets
  - Synthetic Sets are fine
  - Matrix with high to low volatility
  - Represent same was as Synthetic6
- Standardizing Results is probably down to outliers in the real data
  - Discuss the outlier story
  - Read up on Winsorization
- Error Function - will push into Gaussian distribution
  - Fits back into the Standardization issue
  - Good for algorithm which is prone to error maximisation
  - Can then invert back - map weights back
  - Regulariation wont solve the outlier problem, just smoothes over it. Can find a gaussian mapping of the data with error function.
  - Discuss that error function could have been used in the data processing details
- Predictive FFN Activations and Scaling
  - Not necessary to do the linear activation on FFN network - can just note it, can argue that it wont add value
- Summary Table: This is better than that summaries with referencing figures

  - Table for surrogate and real
  - Frame the phenomenology that one gets from financial market data
- SAE Effects
  - Can reference the no SAE encoding in the technical doc & just referenced
  - Note that it was done and what was learnt
- SAE Data Window Aggregations - AGL Data
  - Actual results are showing shrinkage in errors
  - Patterns in features are compressing
  - Dont get caught up on the outliers (Despite the outliers)
  - Ultimately - easier to learn the smoothed version of the graph
  - Can link this back to the outlier issue again
- FFN Data Window Aggregations - AGL Data
  - Emphasise the first 5 days, not just the 1 day mean reversion so gives it traction
- Historical Data and SGD training
  - Makes sense for GBM
  - Do it for real data: 0, 40, 90 percent intervals
- If financial markets are ergodic, more data better -¿ sampling ensembles
- Assessing martingale: Take differences & plot order correlation functions. Order correlation function / power spectrum - see how it peaks
- RBM Pretraining for Sigmoid Networks. Ties back to cross sectional data point as well , where older representations are not of value
- DC init - put paragraph in to emphasise
- Regularization, comes back to ergodicity etc. Ridge or Lasso Works well for IID etc. , not so much for the financial time series. Not moderating the generalization error, just impacting model fidelity.

Online non-linear prediction of financial time-series patterns

Joel da Costa

University of Cape Town

Abstract

We consider pattern prediction of financial time-series data. The algorithm framework and workflow is developed and proved on daily sampled OHLCV (open-high-low-close-volume) time-series data for JSE equity markets. The input patterns are based on input data vectors are equal size data windows pre-processed into a sequence of daily, weekly and monthly sampled feature measurement changes (here log feature fluctuations). The data processing is split into at online batch processed step where data is compressed using a stacked autoencoder (SAE) via unsupervised learning, and then batch supervised learning is carried out using the data-compression algorithm with the output being a pattern sequence of measured time-series feature fluctuations (log differenced data) in the future (ex-post) from the training and validation data. Weight initializations for these networks are implemented with Restricted Boltzmann Machine (RBM) pre-training, and variance based initializations. The historical simulation is then run using an online feedforward neutral network (FNN) initialised with the weights from the online training and validation step. The validity of results is considered under a rigorous assessment of backtest overfitting (Combinatorially Symmetrical Cross Validation), and the results are then considered in terms of test for statistical arbitrage with simple correction for transaction costs.

this wi
to be r
once e
is final

Keywords: online learning, feedforward neural network, restricted boltzmann machine, stacked autoencoder, pattern prediction, JSE, non-linear, financial time series, combinatorially symmetrical cross validation, backtest overfitting

Contents

0.1 Limited Synthetic Asset Tests . . . 9
1 Introduction . . . 4
2 LiteratureReview . . . 5
2.1 Technical Analysis . . . 5
2.2 Neural Networks . . . 5
2.2.1 Training and Backpropagation . . . 6
2.2.2 Activation Functions . . . 6
2.2.3 Deep Learning . . . 6
2.2.4 Backpropagation Improvements . . . 7
2.3 Stacked Autoencoders . . . 7
2.3.1 High Dimensional Data Reduction . . . 7
2.3.2 Deep Belief Networks . . . 8
2.3.3 Stacked Denoising Autoencoders . . . 9
2.3.4 Pre-training . . . 9
2.3.5 Time Series Applications . . . 9
2.3.6 Financial Applications . . . 9
2.4 Data Segmentation . . . 10
2.5 Online Learning Algorithms and Gradient Descent . . . 11
2.6 Backtesting and Model Validation . . . 12
2.6.1 Testing Methodologies . . . 12
2.6.2 Test Data Length . . . 13
2.6.3 Sharpe Ratio . . . 14
3 Data . . . 15
3.1 Data Processing . . . 15
3.1.1 Log Difference Transformation and Aggregation . . . 15
3.1.2 Data Scaling . . . 15
3.1.3 Reverse Data Scaling . . . 15
3.1.4 Price Reconstruction . . . 15
3.2 Synthetic Data Generation . . . 16
3.3 Synthetic Dataset . . . 16
3.4 Real Dataset . . . 16
4 Implementation . . . 17
4.1 Process Overview . . . 17
4.2 Feedforward Neural Networks . . . 17
4.2.1 Notation and Network Representation . . . 17
4.2.2 Activation Functions . . . 17
4.2.3 Backpropagation . . . 18
4.2.4 Gradient Descent Algorithms . . . 19
4.2.5 Regularization . . . 20
4.2.6 Learning Rate Schedule . . . 20
4.3 Restricted Boltzmann Machines . . . 21
4.3.1 Contrastive Divergence . . . 21
4.3.2 CD-1 and SGD . . . 22
4.4 Stacked Autoencoders . . . 22
4.4.1 Sigmoid based Greedy Layerwise SAE Training . . . 22
4.4.2 ReLU based SAE Training . . . 22
4.4.3 Denoising Autoencoders . . . 23
4.5 Variance Based Weight Initializations . . . 23
4.5.1 Initialization Rationale . . . 23
4.5.2 Initializations . . . 23
4.6 CSCV & PBO . . . 24
4.7 Performance Assessment . . . 25
4.8 Money Management Strategy and Returns . . . 25
5 Process Implementation . . . 27
5.1 Data Preparation . . . 27
5.2 Data Segregation . . . 27
5.3 SAE Training . . . 28
5.4 Prediction Network Training . . . 28
5.5 Money Management Strategy . . . 28

6 Results . . . 29
6.1 Datasets Used . . . 29
6.1.1 Synthetic Datasets . . . 29
6.1.2 Actual Datasets . . . 29
6.2 Linearity, Complexity and Structure of Data . . . 30
6.2.1 GBM Generated Data . . . 30
6.2.2 Activations: Linear, Sigmoid, ReLU and Leaky ReLU . . . 30
6.2.3 Effects of Network Size . . . 32
6.3 Feature Selection and Data Aggregation . . . 32
6.4 Weight Initialization Techniques . . . 35
6.4.1 RBM Pretraining for Sigmoid Networks . . . 35
6.4.2 Variance Based Weight Initialization Techniques . . . 36
6.5 Learning Optimizations . . . 37
6.5.1 Learning Rate Schedules . . . 37
6.5.2 Regularization . . . 38
6.5.3 Denoising . . . 39
7 Conclusions . . . 40
8 Appendix . . . 40
8.1 Additional Results . . . 40
8.1.1 Effects of Network Size on SAE and FFN for Synthetic Data . . . 40
8.2 Configuration Sets Used . . . 40
8.2.1 Configuration1 - SAE . . . 40
8.2.2 Config2 - Predictive FFN . . . 41
8.2.3 Configuration3 - SAE . . . 41
8.2.4 Configuration4 - Predictive FFN . . . 41
8.2.5 Configuration5 - SAE . . . 42
8.2.6 Configuration6 - Predictive FFN . . . 42
8.2.7 Configuration7 - SAE . . . 42
9 References . . . 43

1 Introduction

This dissertation explores several ideas for the purposes of effective and valid stock price prediction, and suggests a novel approach to the combination of techniques. The first part of the implementation focuses on the training of deep neural networks, and combines usage of several well known designs for data reduction, deep learning and online learning. The aim here is to produce a deep learning model that is able to effectively produce stock price predictions in an online process, using data reduction of high dimensional finance data in order to achieve better results.

The second part of the paper focuses on whether the model can be trained such that it does not succumb to backtest overfitting, namely by using methods suggested by Bailey et al. [4] in order to calculate the likelihood of overfitting having occurred. As they've noted, and discussed more fully in 2.6, backtesting overfitting for trading strategies have become problematically widespread in financial literature. Here we investigate how more rigorous validation techniques can be applied to deep learning models in order to avoid such overfitting. Further validation takes place in assessing the potential profitability of the model in a live market.

The literature review in chapter 2 has a fuller discussion of work that has been done to precede the various techniques which have been implemented. A brief introduction to technical analysis in the financial sector, with it's perspective and history, is discussed and forms the basis for proposing and using the technical analysis methods throughout this paper 2.1. The literature review then moves onto covering the usage and history of Neural Networks, which has progressed enourmously in recent times. The section covers both the basic foundations, and also discusses the recent work which has resulted in the widespread use of deep learning models 2.2. This is extended to the coverage of Stacked Autoencoders, and their efficacy in data reduction for complex systems, which has lead them to be pivotal tools in deep learning models 2.3. Online learning methods are discussed in 2.5 with a coverage of both the historical basis as well as the more recent devopments which have allowed for further improvment in the algorithms. The chapter finishes off with discussing the impact of backtest overfitting and how some notable works developed by Bailey et al. have provided tools with which this may be avoided 2.6.

Chapter 3 provides the details around the data which has been used in the paper. The chapter starts with describing how the data is processed using log feature fluctuations, and then is expanded to include the fluctuations over rolling window periods. The synthetic data used is generated using Geometric Brownian motion, and is implemented such that each randomly seeded dataset consists of the following stocks:

- 3 stocks with an upward trend, and high, medium and low variance
- 3 stocks with a downward trend, and high medium and low variance
- 3 stocks with a sideways trend, and high medium andlow variance

There is a further summary of the real data used: OHLCV data for the JSE ALSI over a period of 20 years.

Chapter 4 provides more indepth details on the algorithms and structures used to implement the dissertation process. The structure of feedforward neural networks is discussed, as well as how they are trained using the backpropagation algorithm, and how that can be applied in a stochastic descent framework 4.2. The chapter also provides details for how these network weight initializations can impact performance, and how this can be improved with either RBM pre-training (as per [37]) or through newer variance based techniques - this includes the structures and training techniques used for the Stacked Autoencoders 4.4. How the CSCV techniques suggested by [4] are implemented and used to derive a Probability of Backtest Overfitting figure for the full training and testing processes is detailed in 4.6. The chapter ends with a summary of the full end to end process, from data preparation to the output of a PBO figure in **??**.

Chapter 6 is the final chapter, and discusses the results found from the process.

2 LiteratureReview

2.1 Technical Analysis

Technical analysis is a financial analysis practice that makes use of past price data in order to identify market structures, as well as forecast future price movements. The techniques are typically objective methodologies which rely solely on past market data (price and volume). They stand in contrast to fundamental analysis, where experts will consider a companies operations, management and future prospects in order to arrive at an evaluation. The basis of much technical analysis, originally developed through Dow Theory, is the belief that stock market prices will move directionally (upwards, downwards or sideways), and that past movements can be used to determine these trends [65].

One of the primary methods in technical analysis is the use of charts in order to identify price patterns. These charts will be produced using the available market data and a known design, such as the popular candle-bar plot, which can then be compared to historical data to match it to a particular pattern. These patterns are thus indicative that the stock is likely to take on a particular price trend, or is in a particular state [65]. There is a certain amount of controversy around technical analysis, where many argue that it is contradictory to the random walk and weak form efficient market hypotheses, and as such is not valuable or useful [31]. The argument against this, is that technical analysis does not rely on past action to predict the the future, but is rather a measure of current trading, and how the market has reacted after similar patterns have occurred in the past [49]. Further, even if the analysis is unable to effectively forecast future price trends, it can still be useful to exploit trading opportunities in the market [74].

With the advent of processing power becoming cheaply available, there has been an increase in research to adapt computing techniques to technical analysis. The breadth and superhuman speed in which systems are able to perform technical analysis far outstrips what was possible before, and as such they have become the focus of competitive performance for many market participants [48]. To this end, there has been much research to apply machine learning algorithms to perform pattern recognition on stock price movements.

Financial markets have been shown to be complex and adaptive systems, where the effects of interaction between participants can be highly non-linear [3]. Complex and dynamic systems such as these may often exist at the 'order-disorder border' - they will generate certain non-random patterned and internal organisation, which can be assessed and identified, however they will also exhibit a certain amount of randomness in their behaviours, or 'chaos' [18]. As a result, trying to identify these patterns and structures is a simultaneously reasonable and notoriously difficult goal. While it is often clear in hindsight that the patterns exist, the amount of noise and nonlinearity in the system can make prediction challenging. Fittingly then, neural networks have become a popular choice for modelling within the financial markets. Due to their structure, they are able to learn non-linear interactions between their inputs and outputs, with even early research showing their ability to achieve statistically significant results, which lends weight to the argument against the efficient market hypothesis [78].

2.2 Neural Networks

A Neural Network (NN) is a learning model which was originally inspired by the biological mechanisms of neurons in brains. The structure is essentially that of a network system, with connected nodes and edges, or neurons and weights in context. The neurons are based on the same idea as synapses as seen in the brain - where a buildup of input results in a firing of output. The input here is determined by the models input (real numbers typically), and processed through the weights and activation functions of the neuron, which then results in an output value either at an intermediate level, or as the models final output. The system learns by considering input samples sequentially, and adjusting the weights between edges to result in more accurate outputs, which may either be classification or regression values.

Structured neural networks that learn to some extent have been around since the second half of the 21st century [72], though have been through several cycles of popularity. The first versions tended to be very simple with one one layer of hidden neurons [45]. It was only later, through the application of the backpropagation algorithm, that they started to become more practical and popular [88].

With the rise in popularity, many different network formations were developed and suggested. One of the initial suggestions was the conceptually simple Feedforward Neural Network (FNN) as described above - an acyclic graph where inputs are processed in a single direction until the output is reached. The other notable earlier model was the Recurrent Neural Network (RNN), which has a cyclic graph instead - this results in a more powerful computational system than the standard FNN, which was shown to be effective quite early on [77]. The Long Short-Term Memory (LSTM) network was another that used recurrent dynamics, though at a neuron level, in that the neuron is responsible for remembering values for an arbitrary time period [42]. Convolutional Neural Networks (CNN) have a non-recurrent structure, but implement separate pooling layers of neurons which consider the adjacent input values for each feature (e.g. pixels next

to each other). These have been shown to be incredibly effective at tasks such as image recognition, as elaborated on later.

There are three primary learning paradigms used in neural network training - Supervised Learning (SL), where the network is trained on inputs with known outputs; Unsupervised Learning (UL), where the network is trained to identify unknown structures as an output; and Reinforcement Learning (RL), where environmental reactions are used as inputs to train a network for certain outputs [72]. While all of these configurations and paradigms have their benefits and uses, this paper will largely focus on FNN and RNNs, trained through SL and UL.

2.2.1 Training and Backpropagation

Historically, the crux of neural networks popularity has often been based on the development of novel training methodologies, and how they have increased performance. In line with this, the Backpropagation (BP) algorithm (as defined in 4.2.3) has played a pivotal part: while neural network (or perceptron) models were around long before NN popularity, they were largely deemed ineffective, at least in comparison to other available models of the time [64]. It was only during the 1980s that the BP algorithm was applied to NNs, and the field started to gain in popularity again [54, 89].

Rumelhart et al. showed that the BP algorithm as applied in NNs resulted in useful feature representations occuring in hidden layers and the empirical success that resulted thereof [70]. Shortly after, LeCun et al. applied the BP algorithm to CNNs with adaptive connections. They were able to show impressive performance for the time in classifying handwritten images, with the images as a direct input (rather than a feature vector) [55].

While many improvements were made during this time via gradient descent modifications (as expanded on in section **??**), the models were typically of a shallow nature due to problems encountered trying to train deeper networks. Early experiments with deep networks resulted in poor performance due to what is now widely known as the problem of either vanishing or exploding gradients [66]. Essentially, as more layers are added to the network, the backpropagation algorithm (with typical activation function neurons), results in error signals that either shrink or grow out of bounds at an exponential rate. One of first suggested and primary solutions to the problem is to perform pre-training on the network through unsupervised learning [72], which is discussed more fully in 2.3.2.

There were also initial concern that the BP algorithm as applied to high dimensional neural networks would result in the network weights being trapped in local minima if a simple gradient descent was used (e.g. where no small changes to the configuration would reduce the average error rate) [56]. However, empirically, this tends not to so problematic, and large networks usually reach solutions of equitable performance. More recent research has shown that the solution spaces largely consist of many saddle points, each with varying gradients of the features, but which also tend to have similar values of the objective function [19]. Ge et al. have also shown that it is possible to escape saddle points and offer a guaranteed global convergence in certain non-convex problems [27].

2.2.2 Activation Functions

One of the upfront configuration choices necessary is the activation function, which allows the mapping of input to output at the neuron level. There have been many suggestions and experiments with different functions, though there are some common features amongst functions which might make them appropriate: Non-linearity allows for neural networks to operate as universal approximators, as shown in [43], continuous differentiability allows for the use of gradient descent and whether the function is monotonic has been shown to indicate whether the solution can be guaranteed to have a unique periodic solution [90]. Lastly, the range of the function (infinite or finite) can impact both the stability and efficiency of the training.

Some of the most popular functions that have been used are the Sigmoid, TanH, ReLU and Softsign. There have been various studies showing the effectiveness of the different activations under varying initialization (or pre-training) for weights. Glorot and Bengio noted that the typical Sigmoid and TanH functions performed poorly with standard minimization, and result in slower convergence and worse minima, and showed that Softsign with a non-standard initialization resulted in quicker convergence [29]. Further research with Bordes & Bengio found that the rectifier (ReLU) functions were more effective in deep sparse networks compared to the TanH function [30].

2.2.3 Deep Learning

As noted above, most of the earlier work using neural networks relied on shallow models with few layers. However, a resurgence in interest occurred in 2006 after several papers demonstrated the efficacy of unsupervised pre-training of networks prior to supervised training. The effect was substantial enough to allow much deeper layered networks to be

trained than before [10, 36].

The essential point behind the unsupervised learning was to initialize the weights in the network to sensible values in light of the problem context. The methods used trained each layer to be able to reconstruct the model of the features in the layer below (to a varying degree of accuracy). Sequentially pre-training and combining layers like this, the process generated a deep neural network with appropriate weights. Once done, a final output layer was added and the entire network could then be fine-tuned through backpropagation, without suffering such performance degradation through vanishing or exploding gradients [36, 69, 37]. This is expanded up further in 2.3.2. Le Roux and Bengio were able to show that within the DBNs produced by Hinton [36], adding hidden nodes resulted in strictly improved modelling capabilities, and they suggested that increasing the number of layers is likely to result in increased representational ability (subject to efficacy of previous layers), thus establishing the argument for deep networks in theory as well as practice [57].

FNNs had been shown to be effective in modelling high dimensional data even prior to the breakthroughs in deep networks [11], so it fits that the deep networks were shown to be extremely effective in high dimensional data classification. Early implementations shown increased efficacy in handwriting recognition, as well as pedestrian recognition [75]. When it came to data types such as sound and images, CNNs were implemented on several occasions with record breaking model performances in recognition, notably in ImageNet and WaveNet [46, 86].

As more research into deep networks was conducted, it became apparent that with large enough datasets, the layerwise pretraining of networks was not actually necessary to achieve high performance standards [46, 30, 15]. When training for long enough, it was reported that the pre-training offered little to no benefit, though these models were typically using datasets far larger than were attempted before (as a result of hardware improvements enabling as much). While these results did require certain attention was paid to the initialization, as well as the use of nonlinear activation units, it did suggest that pre-training largely acted as a prior which may not be necessary if large enough labelled datasets are available [12]. Naturally, pre-training was still implemented to prevent overfitting in smaller datasets.

2.2.4 Backpropagation Improvements

One of the instrumental improvements which aided the above achievements, was to modify the backpropagation algorithm using the dropout technique, as suggested by Hinton et al [39]. When training of large networks is attempted on small datasets, it often results in overfitting and poor results on out of sample data. Dropout helps resolve this by randomly excluding a certain percentage (usually 0.5) of feature detectors on each training iteration. The effect is to stop co-adaptations of feature detectors, and by rather training each neuron in a wide variety of internal configurations, it forces them to take on more usefully generalizable characteristics (it was noted that this is not a dissimilar technique to ensemble methods, or bagging). The authors were able to show that the method resulting in significant improvements on benchmark data sets (e.g. MNIST, CIFAR-10), and that a simpler model using dropout was able to achieve near comparable performance for the ImageNet dataset.

Goodfellow et al. used the dropout technique as the basis for their maxout activation function technique, which leverages and improves on dropouts fast optimisation and accuracy through averaging characteristics. The maxout model was shown to achieve state of the art performance on benchmark datasets, as well as have a strong theoretical grounding [28]. Further work was done by Wang et al., which improved on the dropout (and potentially maxout) techniques through fast sampling, resulting in an order of magnitude speedup in training [85].

2.3 Stacked Autoencoders

2.3.1 High Dimensional Data Reduction

As noted, machine learning techniques have been shown to be extremely effective at modelling non-linear inputs to outputs - neural networks have even been shown to be universal function approximators in this regard [43]. More traditional statistical models will typically process the available feature data to select the most significant features to be used in the model once its defined - evident in a processes such as subset selection [71]. Machine learning techniques are no different in this regard, and feature data will typically be transformed to smaller observations of more significance prior to being used as input to a model, such as the neural networks described above.

Financial data, in line with the complex and dynamic system that it represents, is often of a very high dimensional nature, which offers opportunities through more sophisticated analysis, but also introduces the curses of dimensionality [21]. The increased dimensionality can result in higher processing complexities when needing to do basic tasks such as estimating a covariance matrix (a commonplace necessity in finance), as well as increase the risk of incorrect assumptions based on spurious variable collinearity [25]. Noise accumulation in high dimensional data can create further problems, resulting

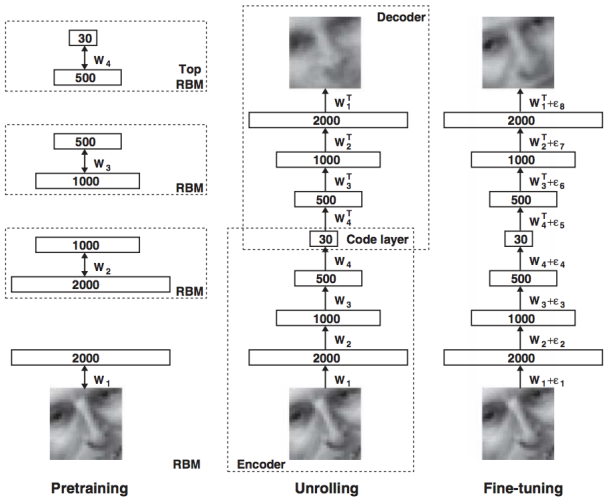

Fig. 67: The Autoencoder training steps [37]

in problems performing variable selection and ultimately having a large impact on classification and regression models [26].

Time series data can introduce its own set of challenges - there is often not enough data available to understand and predict the process [24], the time variable dependence creates complexity in how much past data to consider at any point, and the data is typically non-stationary [52]. Thus, high dimensional time series data (which many financial problems focus on), require careful consideration on how to handle their inputs and analysis.

Deep learning techniques are a natural choice in this context, and much research has been done to show their (varying) efficacy on time series data. The most successful of these models have been ones which modify deep learning techniques to incorporate the temporal aspect of the data (e.g. Conditional Restricted Boltzmann Machines or Recurrent Neural Networks), rather than static, and those which have performed feature selection processes rather than operating on the raw data (e.g. Auto-encoders) [52].

Two of the seminal pieces of research that have lead to the resurgence in machine learning and deep learning were the algorithms for training deep belief networks [36], as well as the usage of stacked auto-encoders [69, 10].

2.3.2 Deep Belief Networks

Autoencoders were suggested by Hinton et. al as a method of transforming high dimensional data to lower dimensional input vectors, in order to alleviate some of the problems detailed above, and increase performance of deep belief networks [37].

One of the more prominent classical techniques for dimension reduction is principal components analysis (PCA), which uses linear algebra to find the directions of greatest variance, and represent the observation samples features along each of these directions, thus maximising the variational representation. Hinton et al. show that autoencoders are a nonlinear generalization of PCA. The structure and training algorithms of the autoencoder show it to be a specialised neural network - there is a multilayer encoder network which is able to transform to a lower dimension, and a symmetrical decoder network to recover the data from the code as represented in Figure 67. As with neural networks, the gradient weights can be trained through the feedforward and backpropagation algorithms.

The primary challenge presented here is the initial weighting of the networks - with large initial weights the autoencoder will often find a poor local minima, and with small initial weights the gradients are too small to effectively train deep layered networks. The critical suggestion by Hinton et al. was to used layered Restricted Boltzmann Machines (RBM) in order to initialise the weights. For each layer of the desired autoencoder, a RBM is formed and trained with the previous layer (or RBM) [38]. Once all the layers have been trained in this way, they are mirrored to form the decoder network. This then forms the initial weights to be fine tuned further, as per the Fine-tuning step in 67. They showed the deep autoencoder networks were significantly more effective than PCA or shallow autoencoders on multiple dataset types.

2.3.3 Stacked Denoising Autoencoders

The second important piece of work was the development of a denoising autoencoder (DAE), by Vincent et al. [82]. One of the problems identified in the DBN model (and those similar), is that if the encoder dimensions were too high, it is likely that the encoder would learn a trivial encoding - essentially creating a copy the input model. The one way of tackling this issue is to constrain the representation with bottlenecks and sparse autoencoder layers, which can be seen in figure 67.

Vincent et al. explore a very different approach to the problem, which was to develop an implementation of autoencoder which focused on partially corrupting the input, and so force the network to denoise it. The theory here is based on two ideas - the first, is that a higher dimensional representation should be robust to partial corruption of the input data; and the second is that the denoising process will force model focus to shift to extracting useful features from the input.

The algorithms and structures are largely the same as described for DBNs above, with the key difference being that the model is trained to reconstruct the original input, but only using a corrupted version of the input (where noise has been added to it), and so is forced to learn smarter feature mappings and extractions. The DAE suggested then is a stochastic variant of the autoencoder, which has the benefit of being able to implement higher dimensional representations without risking training of a trivial identity mapping. Notably, in the Stacked Denoising Autoencoder (SDAE) formation, only the initial input is corrupted (as opposed to the input from layer to layer). It was shown that the SDAE model outperformed previous AE and DBN networks on numerous benchmark datasets [82] .

2.3.4 Pre-training

The methods described above follow a similar approach: greedy layer-wise unsupervised pre-training in order to determine initial weights, followed by supervised fine tuning to arrive at the final model. It is shown numerous times, that the pretraining process results in significant performance gains [82]. However it is not immediately apparent, given the nature of backpropagation algorithms and the like, why this is the case. Erhan et al. performed extensive empirical simulations in order to suggest an explanation to the mechanism of pre-training [23].

While their results were not entirely conclusive, they did lend themselves to a reasonable hypothesis: the unsupervised pre-training results in a form of regularization on the model - variance is minimized, and the bias introduced acts as a prior to direct the model configuration towards a sample space that is effective for the unsupervised learning generalization optimisations.

2.3.5 Time Series Applications

The autoencoder papers reviewed so far in this section derive their results primarily from classification problems, and so do not necessarily account for the problems involved with time series as described in 2.3.1. Due to the inherent difficulties with predictions in the financial system, it can sometimes be unclear if the shortcoming in results is due to this system complexity or if the methodologies used are unsuited for the purpose. In light of this it is worth pointing out that Stacked Autoencoder (SAE) implementations have been shown to be effective in many time series systems.

Lv et al. implemented a deep learning SAE model using the methods described in **??** in order to predict traffic flow at various time intervals (15, 30, 45 and 60 minutes) - a problem not so structurally dissimilar from what will be presented in this paper [61]. They were able to show that the deep SAE was able to offer prediction results which were both objectively good and also persistently outperformed the comparison models used (backpropagation neural network, random walk forecast, support vector machine and a radial basis function neural network).

In a review of unsupervised feature learning and deep learning methods on time series, Langkvist et al. noted that the use of autoencoders, either as a technique in themselves, or as an auxiliary technique to models such as convolutional neural networks, were able to offer performance increases in areas such as video analysis, motion capture data and bacteria identification [52].

2.3.6 Financial Applications

There have of course also been successful applications of stacked autoencoders and deep learning models in finance as well. Takeuchi et al. performed some earlier work showing the use of autoencoders when applied to a momentum trading strategy. They implemented an RBM pre-trained DBN as per 2.3.2, and assessed the networks classification performance for ordinary shares on NYSE, AMEX and Nasdaq. This showed that using a DBN network resulted in significant performance increases compared to the standard momentum strategy [79].

Zhao et al. used SDAEs and combined them with the bootstrap aggregation ensemble method (bagging) in a study of predicting the crude oil price. They compared the proposed model to a variety of benchmarks, including standard SAE, bagged and standard feedforward networks and SVRs. The results indicated that the SAE models were more accurate,

with the bagged SAE model performing the best, though at a significant increase in computational costs in comparison to standard SAE [94].

While much of the financial literature has focused on the use of RBM based models, Autoencoders and SAEs have recently been gaining popularity in performing feature reduction. Troiano et al. specifically investigate the use of different feature reduction models for trend prediction in finance [80]. In line with being primarily interested in the effect of feature reduction techniques, rather than the classification performance itself, only an SVM model was used to test results. Using various periods from historical S&P 500 data, they were able to show that AE outperformed the RBM model significantly in numerous accuracy measures, and was able to do so at a fraction of the training time.

Bao et al. note that the research has been lacking with regards to whether SAEs should be used for financial prediction models or not [8]. They suggest a novel model which combines Wavelet Transformation, SAEs and a Long Short Term Memory (LSTM) network. Using data from several financial exchanges (considering a range of developed and undeveloped markets), they assess the models applicability to OHLC prediction. Comparing the model to configurations without the SAE layers, and a RNN model as benchmark, they showed that the inclusion of SAEs resulted in less volatility and greater accuracy, which in turn offered higher profitabilities in a buy-and-hold trading strategy.

More novel autoencoder applications have also been attempted, with Hsu suggesting the use of a Recurrent Autoencoder for multidimensional time series prediction [44]. There is a clear pattern through the literature that the use of AEs and SAEs both by themselves and when used as an assisting technique result in more accurate prediction results and less computationally expensive training.

2.4 Data Segmentation

One of the aspects of time series not yet discussed is how the data might be segmented for analysis. Financial pattern matching, as discussed in **??**, requires methodologies to decide the length of data to consider when determining whether a subset of data matches a pattern or not.

There are numerous classical approaches to this problem, which were widely applied in machine learning prior to the resurgence in deep learning. One of the earliest and more common approaches was the Sliding Window, where each the model input for each observation is padded by a predetermined number of observations that occurred both prior and after the one in question. This has the advantage of being fairly model agnostic, but fails to capture any correlations in the dependent variable values [16]. The Recurrent Sliding Window method aims to resolve this by including the same number of prior predictions made as part of the input. In this case, input would be $< x_{t-d}, ...x_t, , x_{t+d} > and < y_{t-d}, ..., y_t >$. This was shown to be significantly more effective than the plain sliding window method [7]. Notably, the sliding window approach can only be implemented in an offline model.

These methods can be adapted to take on online formats, as well as incorporate data reduction benefits through algorithms such as Piecewise Linear Approximation (PLA), which adapts a linear representation to the leading portion of the window. Some novel approaches (Feasible Space Window and its stepwise adaption) were suggested by Liu et al. in order to compensate for the computational requirements of reprocessing the entire window at each step for online models [58].

Window based methods represent a fairly static and unsophisticated approach to data segmentation. A suggested solution to this is to segment the data by dynamically identifying significant partition points, known as perceptual important points (PIP) [17]. The computational cost of identifying PIPs was initially rather large, making them unsuitable for quickly changing and dynamic environments such as finance. However, Zhou et al. suggested a novel approach which reduced computational costs through intelligent binary tree traversal. They were able to show this approach was effective in identifying traditional financial stock patterns with the use of a layered neural network [96]. In a similar fashion, input data turning points (TP) were shown to potentially offer lower error rates than PIPs by Yin et al. [91].

More recently, Wan et al. conducted a review on the effect of segmentation on financial time series pattern matching, comparing perceptual important points, piecewise aggregate approximations, piecewise linear approximations and turning point based methods [83]. They use several pattern matching algorithms (template-based, rule-based, hybrid, decision tree, and symbolic aggregate approximation) for each of the segmentation methods in order to determine their effect on a broader level. The analysis was performed on real stock market data, as well as synthetic data generated to display common patterns such as positive/negative head and shoulders (H&S). Measuring accuracy, precision and recall, they showed that PIP based segmentation generally offered superior results to the others (though it is worth noting that there were various performance differences within the segmentation methods depending on the matching algorithm used).

2.5 Online Learning Algorithms and Gradient Descent

Most classic machine learning algorithms operate under the assumption that, for all intents and purposes, the full dataset has been collected and that the amount of training data for the model is both finite and immediately available. However, as the growth of information grows in an exponential fashion, there are numerous areas where the expected training data for the model will continue to grow. In these cases it would be disadvantageous to go through the full training and validation process again in order to incorporate the newly available data.

Online algorithms are designed to offset these issues by adjusting the batch training technique to rather repetitively draw on single samples from the data on which the models parameters can be adjusted. The benefit is that they are able to quickly process a large number of observations and readjust the model, though the downfall is that they are not always able to optimize the cost function to the same extent as offline batch algorithms [1].

Bottou and Cun argue that as the size of the dataset grows significantly, online algorithms advantages result in them outperforming offline models, despite any initial drawbacks [13]. Previous research had shown that online algorithms typically perform as fast as batch algorithms during the search phase of parameter optimization, but that final phase convergence tended to fluctuate around the optima due to the noise present in single sample gradients [53, 14]. Bottou and Cun showed in fact, that it is more practical to consider the convergence towards the parameters of the optima, rather than the optima itself (as defined by the cost function) - the difference between the learning speed and optimization speed, respectively [13]. Theoretical and empirical findings were presented to show that a stochastic online gradient descent (SGD) algorithm **[referenceappendix]** was able to outperform the batch model for parameter estimation, and was able to asymptotically outperform in the number of samples processed in a time period. The stochastic aspect of the algorithm is related to random observations from batch sample groups being used as the gradient basis. Theoretically, this slows down the convergence, but speeds up the processing speed of each batch - a technique which has later been shown to be generally successful [76, 95].

The SGD algorithm has resulted in a fair amount of further research due to its applicability to machine learning and the online benefit, which can largely be group into two categories: improvements affecting gradient learning and convergence rates, and processing improvements through parallelization.

One of the earlier improvements to convergence rates was the Momentum algorithm, as developed by Tseng [81]. As noted, stochastic descent often introduces significant oscillation around an optima, which slows down convergence. Momentum reduces this by decreasing movement in directions of high curvature, and increasing increasing movement towards directions consistent with previous gradients (this is achieved through combining gradient movements in opposite directions) .

There have also been several attempts to introduce effective regularization into the SGD process. Bartlett et al. presented Adaptive Online Gradient Descent, which implements an adaptive step size through a λ penalty on the learning rate, which was shown to be nearly optimal in a strong sense [9]. Langford et al. demonstrated a variation named Truncated Gradient, which introduced an enforced weight sparsity parameter. The weight sparsity is able to achieve equitable effects to L_1 regularization (similar to Lasso Regression). They were able to show that implementation performed effective feature reduction, while having little effect on performance [51]. Other approaches, such as AdaGrad, aim to improve the robustness of gradient training by adjusting the updates to parameters according to frequency - e.g. larger updates to infrequent parameters, and smaller updates to frequent parameters [22, 92].

A parallel implementation of SGD largely rests on the idea of splitting up data to be processed in individual runs of the gradient descent. The results are periodically aggregated, and the new model parameters are distributed to processing nodes for further training [93]. Depending on the configuration, the variance within results can result in poor convergence rates [62]. Mahajan et al. suggest a novel implementation which improves the distribution impacts through the use of better approximating functions in the processing nodes, which in turn improves the efficacy of the algorithms convergence [62]. Povey et al. presented a Natural Gradient SGD, which improves the learning rates through the use of a factor matrix used on the new gradients. They were able to show empirical evidence of this improving performance issues introduced from the parameter averaging typically used in parallelization [68].

There have been some more recent improvements focusing on making the SGD algorithm more dynamic, including the Inconsistent SGD (ISGD) and AdaBatch adaptations. The idea behind ISGD is to treat the training on a particular batch as a stochastic process, and adjust it according to the expected loss identified. Gradient updates from small loss batches are relatively small compared to large loss batches, and so by focusing efforts here ISGD is able to optimize performance. Wang et al. performed careful testing of SGD vs. ISGD and found inconsistent training to consistently outperform in terms of convergence speed and results [84]. It is worth noting that due to its nature, ISGD can be effectively combined with other SGD improvements (e.g. Momentum, as per the authors experiments). AdaBatch, as presented

by Devarakonda et al., introduces dynamism through the adjustment of batch sizes. Small batch sizes in SGD offer the benefit of faster convergence in fewer training epochs, however larger batch sizes are more computationally efficient due to their applicability to parallelization. The algorithm uses a monotonically increasing batch size, which is started small to gain traction in convergence, and later increased to allow the benefits of data parallelism. The effect, as shown, is to offer improvements in training performance of up to 6x, with less than 1% of accuracy when compared to the fixed batch baselines [20]. Both of these algorithms have been shown to be effective and applicable in the context of online neural network training.

2.6 Backtesting and Model Validation

Much of financial academic literature is currently facing a problem in terms of validation and verification of results. The primary method of going about these ends in the past has been to perform historical simulations, or backtests , in order to prove profitability of a trading strategy. The recent advances in both technology and the algorithms available to construct these strategies has resulted in researchers being able to run so many iterations of a model or strategy configuration through these backtests, that its become increasingly difficult to control for spurious results, with some papers suggesting that most published research findings are false [47].

The standard way of implementing backtests is to split the data into two portions: an In Sample (IS) portion which is used to train the model, and an Out of Sample (OOS) portion which is used to test the model and validate results. The problem lies in that millions of different model configurations might be tested, and if more sophisticated test measures are not in place (i.e. not just the standard Neyman-Pearson hypothesis testing framework is implemented), then it is only a matter of time before a false positive result occurs which shows high performance both IS and OOS (i.e. overfitting). The nature of financial data, where there is a low signal-to-noise ratio in a dynamic and adaptive system, and where there is only one true data sequence, makes it difficult to resolve these issues effectively [4, 63].

Overfitting is not a novel issue, and has of course been tackled in various literature areas, including machine learning. However, in that context, the frameworks are often not suited to the buy/sell with random frequency structure of investment strategies. They also do not account for overfitting outside of the output parameters, or take into consideration the number of trials attempted. Other methods, such as hold-out, are arguably still faulty due to researcher knowledge while constructing models [73]. One of the downfalls of the typical IS-OOS set up in the financial context is also that the most recent (and relevant) data will not be able to be used for the model training.

There have been some suggestions to resolve the problem that is occuring in the literature as a result of this - some work suggesting new frameworks, which this section will cover, and others which focus on the review process or how data and replication procedures are made available [67]. While the points made with regard to the review process and so on are certainly important, they don't aid with more effective model training for the researcher up front, and so will not be covered here.

2.6.1 Testing Methodologies

Considering the issues laid out above, there has been much work to develop alternative approaches to backtesting. One of the common approaches to avoid backtest overfitting is the hold-out strategy, where a certain portion of the dataset is reserved for testing true OOS performance. Numerous problems have been pointed out with this approach, including that the data is often used regardless, or that awareness of the movements in the data may, consciously or otherwise, influence strategy and test design by the researchers [73]. For small samples, a hold-out strategy may be too short to be conclusive [87], and even for large samples it results in the most recent data (which is arguably the most pertinent) not being used for model selection [34, 4].

There has been work by several authors to try and lay out techniques to try and avert backtest overfitting. The Model Confidence Set (MCS), as developed by Hansen et al. [32], starts with a collection of models or configurations, and remove models iteratively according to a defined loss function. The confidence set is defined by the remaining models once a non-rejection takes place within the process, and these models are considered to be statistically similar within a certain confidence range. MCS is thus able to facilitate equitable model selection. However, Aparicio et al. [2], showed that while MCS is a potential strategy, in practice is is ineffective due to the inordinate requirement of signal-to-noise necessary to identify true superior models, as well as a lack of penalization over the number of trials attempted.

Bailey et al. [4] have developed a more robust approach to backtesting and how overfitting during strategy selection might be avoided, called Combinatorially Symmetric Cross-validation (CSCV). Their research defines backtest overfitting as having occurred when the strategy selection which maximizes IS performance systematically underperforms median OOS in comparison to the remaining configurations. They use this definition to develop a framework which measures

the probability of such an event occuring, where the sample space is the combined pairs of IS and OOS performance of the available configurations. The probability of backtest overfitting (PBO) is then established as the likelihood of a configuration underperforming the median IS while outperforming IS.

The CSCV methodology provides several important benefits over traditional testing frameworks, including the usual K-fold cross validation used in machine learning. By recombining the slices of available data, both the training and testing sets are of equal size, which is particularly advantageous when comparing financial statistics such as the Sharpe Ratio (SR), which are susceptible to sample size. Additionally, the symmetry of the set combinations in CSCV ensure that performance degradation is only as a result of overfitting, and not arbitrary differences in data sets. It is pointed out that while CSCV and PBO should be used to evaluate the quality of a strategy, they should not be the function on which strategy selection relies, which in itself would result in overfitting.

2.6.2 Test Data Length

The CSCV methodology offers an important but highly generalised framework to assess models and backtest overfitting. It doesnt however indicate which metrics should be used to assess the IS and OOS performance, nor any indication on the amount of data needed to do so effectively. One of the noted limitations of the framework is that a high PBO indicates overfitting within the group of N strategies, which is not necessarily indicative that none of the strategies are skillful - it could be that all of them are. Also, as pointed out, it should not be used as an objective function to avoid overfitting, but rather as an evaluation tool. To this end it helps assess overfitting, but not necessarily avoid it.

A typical measure of evaluation used for financial models is the Sharpe Ratio (SR), which is the ratio of between average excess returns and the returns standard deviation - a measure of the return on risk. In the context of comparing models, SR is typically expressed annually to allow models with different frequencies to be compared. Lo et al. [59] show that annulaized SR can be expressed as

$$SR = \frac{\mu}{\sigma}\sqrt{q} \tag{4}$$

Using sample means and deviations, $\hat{\mu}$ and $\hat{\sigma}$, SR can be shown to converge as follows (as y $\to \infty$)

$$\hat{SR} \to \mathcal{N}[SR, \frac{1 + \frac{SR^2}{2q}}{y}] \tag{5}$$

Thus, when using SR estimations, which follow a Normal distribution, it is possible that where the true SR mean is zero we may still (with enough configurations attempted) find an SR measurement which optimises IS performance. This is shown by Bailey et al. [5], who propose the non-null probability of selecting an IS strategy with null expected performance OOS. Notably, typical methods such as hold-out once again fail, as the number of configurations attempted are not recorded. They add a further derivation, thich is the Minimum Backtest Length (MinBTL), ultimately showing that

$$MinBTL \approx (\frac{(1-\gamma)Z^{-1}[1-\frac{1}{N}] + \gamma Z^{-1}[1-\frac{1}{N}e^{-1}]}{\overline{E[max_N]}})^2 < \frac{2ln[N]}{\overline{E[max_N]}}^2 \tag{6}$$

The statistic highlights the relationships between: selecting a strategy with a higher IS SR than expected OOS, the number of strategies tested (N), and the number of years tested (y). The equation shows that as the number of strategies tested increases, the minimum back test length much also increase in order to contain the likelihood of overfitting to IS SR.

As shown extensively throughout ML literature, increased model complexity and number of parameters is one of the primary causes of overfitting. In context of the MinBTL formula, model complexity affects the number of configurations that are available and which may be tested, which in turn will increase likelihood of overfitting. A lack of consideration, or reporting, of the number of trials makes the potential for overfitting impossible to assess.

Bailey et al. expanded on this view with assessing the impact of presenting overfit models as correct. They were able to show that in lieu of any compensation effects (i.e. a series following a Gaussian random walk), there is no reason for overfitting to result in negative performance. However, where compensation effects apply (e.g. economic/investment cycles, bubble bursts, major corrections etc.), then the inclusion of memory in a strategy is likely to be detrimental to OOS performance if overfitting isnt controlled for [5].

2.6.3 Sharpe Ratio

The use of the Sharpe Ratio in financial backtesting is not just an arbitrary or persistent literature choice. The statistic offers two benefits: the effectively strategy-agnostic financial information contained, as well as being relatable to the t-statistic, and so simple to perform hypothesis testing. The SR ratio (estimate from sample as $\hat{SR}$) is defined as

$$SR = \frac{\mu}{\sigma} \tag{7}$$

The t-ratio is defined as

$$t - ratio = \frac{\hat{\mu}}{\hat{\sigma}/\sqrt{T}} \tag{8}$$

Evidently, the link heree ratio (though it can be generalized to another statistic with a probabilistic interpretation). Additionally, PBO is generally more in line with machine learning literature with the cross validation like approach on time series data.

It should be noted, that the literature detailing usage of the Sharpe ratio for strategy comparison is extensive, with numerous variations and methodologies offered [6]. However, the crux of this paper lies in whether an online neural network is able to make effective enough predictions that a strategy might use the predictions to be profitable. The subtlety here is that we will consider the usage of such forecasting *within* a strategy, rather than *as* a strategy itself. In line with this, statistics such as the Sharpe ratio will be used, but not form a critical consideration of the research here as the comparison of strategies used will be a secondary concern.

3 Data

3.1 Data Processing

The datasets used go through several transformations throughout the usage cycle, each of which are captured below.

3.1.1 Log Difference Transformation and Aggregation

All datasets were transformed into log feature fluctuation values, and were then expanded to include fluctuations over rolling window periods.

$$\texttt{log\_diff}(x_t) = ln(x_t) - ln(x_{t-1}) \tag{9}$$

This log feature fluctuation, is processed for each OHLCV feature and for each time point (from the previous time point). The log fluctuations have the benefit of taking compound effects into account in a systematic way and symmetric in terms of gains and losses.

The datasets are then expanded with the rolling window summations both in the past, for input, and in the future, for predicted output. For example, if the past data aggregation windows were 1, 5 and 20, and the future prediction point was 5 then at each data point t, the following time series are added:

- Log feature fluctuations over the periods $(t-5, t)$ and $(t-20, t)$ (with $(t-1, t)$ already having been processed)
- Log feature fluctuations for the period $(t, t+5)$, to be predicted

Only data points with a full set of features are used for training and prediction later on in the framework.

3.1.2 Data Scaling

Once the log difference data has been processed, the datasets are either standardized or normalized to allow for better learning. This is done in the typical ways:

$$\texttt{standardize}(x_s) = \frac{(x_i - \bar{x})}{\sigma_x} \tag{10}$$

$$\texttt{normalize}(x_n) = \frac{x_i - min(x)}{max(x) - min(x)} \tag{11}$$

Two extensions of these are used later on, dubbed 'Limited Standardization' and 'Limited Normalization', which are used when the data is split up into the SAE training and OGD prediction sets. In this case all the data needs to be scaled, but the information should not be travelling from the OGD dataset to the SAE training set through the use of aggregated measures such as $\bar{x}$. Thus, if the data is split at point t out of n data points, the scaling would be implemented as follows:

$$\texttt{limited standardize}(x_{ls}) = \frac{(x_i - \bar{x_{1:t}})}{\sigma_{x_{1:t}}}, \forall i \in (1, n) \tag{12}$$

$$\texttt{limited normalize}(x_{ln}) = \frac{x_i - min(x_{1:t})}{max(x_{1:t}) - min(x_{1:t})}, \forall i \in (1, n) \tag{13}$$

This log-differenced, aggregated and scaled data is then used as the input for the neural network models.

3.1.3 Reverse Data Scaling

The predicted data points are transformed back into actual prices for returns analysis. The first step is to reverse the scaling that is done in 3.1.2 in order to retrieve the log difference fluctuations.

$$\texttt{reverse standardize}(x_i) = x_s * \sigma_x + \bar{x} \tag{14}$$

$$\texttt{reverse normalize}(x_i) = x_n * (max(x) - min(x)) + min(x) \tag{15}$$

These are applied similarly for the limited variations.

3.1.4 Price Reconstruction

The predicted log fluctuations produced in 3.1.3 are used to reconstruct the predicted prices. For original prices o, stocks s, time window predicted t, predicted log fluctuation p and time steps $i \in (0, n)$i:

$$\texttt{reconstructed price}_{s,i} = o_{s,i-t} * e^{p_{s,i}} \tag{16}$$

These reconstructed prices are then used to assess model returns later.

3.2 Synthetic Data Generation

Synthetic data generation is implemented using Geometric Brownian Motion as described in 1, which allows the stocks simulations to be implemented with a trend (σ) and variance (μ) for each stock. Each dataset was generated with a random seed (using a Mersenne Twister pseudorandom number generator).

Algorithm 1: Geometric Brownian Motion Simulation

Input: σ, μ, S_0, *steps*
t = 1/steps;
prices = $[S_0]$;
foreach *i in 1:t* **do**
  z = random() $\sim N(0,1)$;
  $S_t = S_{t-1} * e^{(\mu - \frac{\sigma^2}{2})t + \sigma\sqrt{t}z)}$;
  prices = [prices; S_t];
end
Result: prices

3.3 Synthetic Dataset

This section will detail the chosen σ and μ values for the dataset. finish

Synthetic datasets were generated to represent known trends with some variance in prices. Geometric Brownian motion was used to simulate stock price movements, and had a structure of 6 stocks:

- 2 stocks with an upward trend, and high and low variance levels
- 2 stocks with a downward trend, and high and low variance levels
- 2 stocks with a sideways trend, and high and low variance levels

3.4 Real Dataset

This section will detail the use of real data - namely the JSE ALSI data collected, and the time periods collected for. Detail will be added once this dataset has been finalised. finish

4 Implementation

4.1 Process Overview

The implementation focuses on bringing together several ideas: data reduction, deep learning with pre-training/weight initialization, online learning and backtest overfitting validation for the purposes of stock price prediction. The process implementation is discussed fully in 5.1, but can be summarised as the following:

1 The dataset is split into 2 subsets: the Training portion, and the Prediction portion.

2 The Training set is used to train the SAE and deep FFN using the SGD algorithm. These networks are implemented with pre-training or weight initialization techniques.

3 The Prediction sett of data is used to continue training the network in an online manner using OGD.

4 The returns generated from OGD can then be used in the CSCV process, to estimate the probability of backtest overfitting.

The rest of this chapter will detail the algorithms used to train the relevant FFN, RBM and SAE networks, as well as the trading strategy and CSCV & PBO testing procedures.

4.2 Feedforward Neural Networks

Feedforward Neural Networks (FFN) in the form of multilayer perceptrons is a well established network technique, providing effective nonlinear representations for both shallow and deep structures [72]. Specifically, a FFN is made up of several non-cyclical layers: the first and last are the input and output layers, respectively, and any inbetween are referred to as 'hidden' layers. Each layer is made up of nodes which are fully connected to the nodes in the previous and following layers, but does not have connections to nodes within the layer - information only travels forward. Each node has an activation function, which acts on the weighted input from the previous layers' nodes.

4.2.1 Notation and Network Representation

For the purposes of this chapter, the following notation will be used:

- A network N will be contructed with L layers, each with e nodes
- The weights from the j^{th} node in the l^{th} layer from the k^{th} node in the $(l-1)^{th}$ layer are represented by w^l_{jk}
- The bias for the j^{th} node in layer l is represented by $w^l_0 j$
- The output for the j^{th} node is defined as $a = o(z_j)$, for an activation function o and weighted input z
- The weighted input for the j^{th} neuron in a layer is $z^l_j = \sum_{i=1}^{e^{l-1}} a^{l-1}_i w_{ij} + w_{0j}$

These definitions allow an input into the network to be propagated through it, having the original values processed through the weights and activations functions, and have an output in the form of the network's last layer.

4.2.2 Activation Functions

As noted in 2.2.2, there are 3 primary characteristics of concern for activation functions: non-linearity, continuous differentiability and monotonicity. While many different functions have been suggested and used, 2 of the most used were implemented here.

Sigmoid The sigmoid, or logistic, function is one of the most popular and widely used activation functions historically, and is defined as

$$f(x) = \frac{1}{1+e^{-x}} \tag{17}$$

The Sigmoid function is in the range [0,1], making it a suitable choice for problems requiring a probabilistic output. The slope of the function curve is both a boon and a drawback: it allows for fast learning initially, but results in learning slowdown later (often casuing what is referred to as node 'saturation'). The exponent calculation is also computationally expensive, relatively speaking.

ReLU The Rectified Linuear Unit (ReLU) is, as discussed in , a newer activation function which has been shown to be effective in deep learning networks. It is defined as

needs section

$$f(x) = max(0, x) \tag{18}$$

The function has the benefits of quick learning which doesn't saturate, as well as being computationally cheap. The downside is that the non-gradient for the negative range of the function can result in 'dead' nodes, which stop updating with the learning process.

Leaky ReLU The Leaky ReLU resolves the dying ReLU problem by adding a small gradient to the negative range of the function. This results in a slow learning being applied to 'dead' ReLU's, which may in turn result in them being used again if necessary.

(19)

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0.01x, & \text{otherwise} \end{cases}$$

Linear Activation Linear activations don't transform the input, and have a constant gradient, which can be useful for key layers where no loss in error signal is desired, such as the output or encoding layers.

$$f(x) = x \tag{20}$$

4.2.3 Backpropagation

The backpropagation algorithm, as discussed in 2.2.1, has allowed for effective training of FNNs for given data. The algorithm relies on incremental improvements of the model, as defined by decreasing the cost function. A common choice for cost is Mean Squared Errors (MSE):

$$C = \frac{1}{2}\|y - a^L\|^2 \tag{21}$$

A technical definition of the backpropagation algrithm is given in Algorithm 2, though the three primary steps of the algorithm are:

Forward Pass

- The samples are propagated through the network, in order to generate the output $\hat{y}_s$

Calculate Cost

- The cost between the training output y_s and the model output $\hat{y}_s$ is calculated, as per 21

Backward Pass

- Propagate the activation values back through the network to calculate the delta values based on the cost value
- Each weights output delta and input activation are multiplied to find the weights gradient and this gradient is reduced by a factor of the learning rate η, which is then subtracted from the network weights to update them

Algorithm 2: Backpropagation

Input: Neural Network N, with randomly initialized weights w, L layers, and activation functions o
Data: Testing set with inputs x, and outputs y

repeat
  Select a sample x_s from x;
  `// Perform the Forward Pass, to calculate the network output,` y_s`, for input sample` x_s
  $a_0 = x_s$
  foreach *l in 1:L* **do**
    foreach *j in 1:*e^l **do**
      $z_j^l = \sum_{i=1}^{e^{l-1}} a_i^{l-1} w_{ij} + w_{0j}$
      $a_j^l = o(z_j^l)$
    end
  end
  `// Calculate the error term (aka cost), as`

$$\lambda^L = \nabla_a C \otimes o'(z^L) = (a^L - y) \otimes o'(z^L) \tag{22}$$

  `// Perform the Backward Pass, to propagate the errors back and update the network accordingly`
  foreach *l in (L-1):1* **do**
    `// Calculate the delta values`

$$\lambda^l = ((w^{l+1})^T \lambda^{l+1}) \otimes o'(z^l) \tag{23}$$

    `// Update the weights`

$$w^l \to w^l - \eta \lambda^l (a^{l-1})^T \tag{24}$$

  end
until *no new samples can be drawn from x*;
Result: Updated Network N

4.2.4 Gradient Descent Algorithms

The backpropagation algorithm is defined at a single sample level, and the learning from a dataset is usually repeated for a number of 'epochs'. Implementation however is usually done using a vectorized version of the algorithm which either samples the entire dataset at once (batch), or a subset of samples (minibatch). The latter is the Stochastic Gradient Descent (SGD), as discussed in **??**, which has been shown to increase the speed at which the backpropagation algorithm can converge a minima in terms of cost. The algorithm runs backpropagation over the entire dataset for the number of epochs, and updates the network incrementally through the epoch. The stopping condition for the algorithm is usually defined as either a particular number of epochs being reached, or cost no longer decreasing for some number of epochs (i.e. a minima has been reached).

Algorithm 3: Stochastic Gradient Descent

Input: Neural Network N, with randomly initialized weights w, L layers, and activation functions o
Data: Testing set with inputs x, and outputs y

repeat
  repeat
    Select m samples x_s from x which have not yet been sampled in this epoch;
    Use backpropgation on the network using the m samples to update the weights, using the aggregated weight update:

$$w^l \to w^l - \frac{\eta}{m} \sum_x \lambda^{x,l} (a^{x,l-1})^T \tag{25}$$

  until *no new samples can be drawn from x*;
until *For each epoch, or until a stopping condition is reached*;
Result: Updated Network N

Online Gradient Descent Where SGD is appropriate and effective for scenarios where the entire dataset is available, Online Gradient Descent (OGD) is applicable for when the model is learning in an online fashion. In this case, the backpropagation is run as defined above in Algorithm 2 but with no repetition (i.e. only 1 epoch).

4.2.5 Regularization

Regularization is a commonly used technique in machin learning used to reduce a model's capacity to overfit to the data, and and so doing reduces the variance in the model results. One of the typical methods, $L2$ (or "weight decay"), is implemented by adding an extra term to the cost function which is itself a function of the weights in the model. This extra term forces the learning process to favour smaller weights, and only allows large weights to occur if they are able to offer an appropriate increase in performance. The modified cost function is displayed below in 26.

$$C = \frac{1}{2n} \sum_x \left\| y - a^L \right\|^2 + \frac{\lambda}{2n} \sum_w w^2 \tag{26}$$

The additional term is scaled by the configurable regularization parameter, λ, which if small approximates the original cost function or if large will increase the degree of regularization used.

In order to implement this in backpropagation, the weight update rule is changed to the following (while biases remain the same):

$$w \to w - \eta \frac{\partial C_0}{\partial w} - \frac{\eta \lambda}{n} w \tag{27}$$

For SGD, this can be simplified to the following:

$$w \to \left(1 - \frac{\eta \lambda}{n}\right) w - \frac{\eta}{m} \sum_x \frac{\partial C_x}{\partial w} \tag{28}$$

4.2.6 Learning Rate Schedule

Learning rates schedules are implemented to allow for a more dynamic exploration of the possible solution space, and allow fine tuning around an optima. With a constant learning rate implementation, one has to choose between a larger learning rate which allows faster progress but less optimisation, or a smaller learning rate which learns slower but is more effective at optimising (but avoiding the learning algorihtm from bouncing around a minima valley). Learning rate schedules aim to get the best of both scenarios by cycling through a minimum and maximum range across epochs.

The learning rate is specified to run schedule is specified to cycle through from min to max every i epochs, from $\eta_{\min}$ to $\eta_{\max}$. Thus for the current epoch T, the learning rate is calculated using 29 below:

$$\eta_t = \eta^i_{\min} + \frac{1}{2} \left(\eta^i_{\max} - \eta^i_{\min}\right) \left(1 + \cos\left(\frac{T_{\text{current}}}{T_i} \pi\right)\right) \tag{29}$$

This approach is known as Stochastic Gradient Descent with Warm Restarts (SGDR) [60]. For a minimum learning rate of 0.1, a maximum of 1.0 and a epoch cycle of 100, it will produce learning rates as displayed in 68 below.

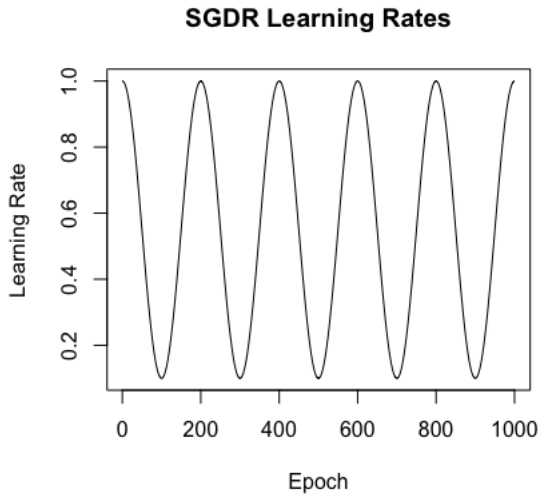

Fig. 68: Learning rates calculated over 1000 epochs with $\eta_{\min} = 0.1$ to $\eta_{\max} = 1.0$ and $i = 100$

4.3 Restricted Boltzmann Machines

Restricted Boltzmann Machines (RBM) are generative networks which can be trained to learn probability distributions over a dataset. They are structurally different from a FFN in that they have a recurrent weight function - a typical RBM has one visible layer (input/output), and one hidden layer. A sample will be processed from the input layer to the hidden layer, and the activation values from the hidden layer will be used by the input layer to provide a recontruction. The hidden units thus correspond to feature detection of the visible unit data structures, and the learning process of the network results in effetive parameter estimation.

One of the primary differences from an FFN lies in the stochastic unit determination - the values in a hidden layer will typically be implemented such that the take on a binary value with a probablistic likelihood. Thus, the input and output have the same structure, and the processing from the hidden layer creates the generative process learned by the RBM. The joint configuration (v, h) of the visible and hidden units has an energy given by:

$$E(v,h) = -\sum_{i \in visible} a_i v_i - \sum_{j \in hidden} b_j h_j - \sum_{i,j} v_i h_j w_i j \tag{30}$$

where v, h correspond to the binary states of the visible and hidden units, with biases a, b, and weights w [40]. It can be shown, that network weights can be adjusted to change the probabilities assigned to a particular training sample [40]. The derivations of 30 show that performing a stochastic asent for the log probability of the data can be implemented through the following weight adjustment:

$$\Delta w_{ij} = \eta(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}) \tag{31}$$

The angular brackets here indicate expectations under the distribution specified by the following subscript. Due to the probabilistic nature of RBMs, the [0,1] ranged Sigmoid activation function is typically used. Thus, for hidden nodes and a data sample v, the it is easy to get unbiased sampling of $\langle v_i h_j \rangle_{data}$

$$p(h_j = 1) = \sigma(w_{0j} + \sum_{i=1} v_i w_{ij}) \tag{32}$$

Similarly, the visible states (or reconstruction), can then be calculated as

$$p(v_i = 1) = \sigma(w_{0i} + \sum_{j=1} h_j w_{ij}) \tag{33}$$

Getting an unbiased sample of $\langle v_i h_j \rangle$ proves more problematic, and so the model reconstructioms via Gibbs sampling are used instead (this is explained below), resulting in the weight updates of

$$\Delta w_{ij} = \eta(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{reconstruction}) \tag{34}$$

While it's important for the hidden units to take on a binary value (and so avoid communicating real values rather than learning structure), the visible units may be chosen to take on the probability values, rather than the stochastic samples, particularly if real valued output is necessary.

4.3.1 Contrastive Divergence

The process of sampling and resampling may be run for many iterations between the two layers before finishing on an output - this potentially long running and stochastic process results in the generative aspect of the network, and constitutes a Gibbs sampling chain. Multiple sampling steps in this chain is known as Contrastive Divergence, or CD-n, where n represents the number of steps, and which allows for effective parameter estimation. Thus, CD-1 is the following:

Algorithm 4: CD-1

Input: RBM Network N, with visible nodes N and hidden nodes H and distribution $\langle vh \rangle$
Data: Training sample dataset V'

For a random training sample V', sample H' from $P(V|H)$;
Sample V from $P(V|H')$;
Sample H from $P(H|V)$;
Return $(V, H) \sim R$, where R is the reconstructed distribution for $\langle vh \rangle$;

Using the reconstructed distribution $R(V, H)$, the weight update for CD-1 is, as above, then

$$\Delta CD(W) = \eta(E_{P(H|V)D(V)}[VH^T] - E_{R(V,H)}[VH^T]) \tag{35}$$

Result: Reconstructed distribution for $\langle vh \rangle$

There's no upper bound on the iterations used for CD, and running for many can prove more effective for certain purposes. In this case however, where RBMs are used for the purposes of weight initialization, CD-1 is usually deemed sufficient.

4.3.2 CD-1 and SGD

In the same way that the backpropagation learning algorithm in 4.2.3 can be implemented in an SGD process for FFNs, so can the CD learning algorithm for RBMs. The framework is kept the same, with the implementation of epochs and weight updates based on stochastically chosen minibatches, but the calculation used to update the weights is CD-1 instead.

4.4 Stacked Autoencoders

As noted in 2.3, the use of Stacked Autoencoders (SAE) have resulted in significant improvements in deep learning networks, and allowed effective reduction of high dimensional data. A single autoencoder is a specialized type of FNN with 3 layers: one input, one hidden, and one output. The network is trained (using backpropagation as per 4.2.3) to reconstruct the input, so the input and output layers have the same structure, and the hidden layer needs to have fewer nodes than the input. This forces the hidden layer to learn effective features of the data, and reduce the dimenstional representation.

Stacked autoencoders follow a similar structure, but with multiple hidden layers. The only strict requirement of the hidden layers is that the middle one, which will be used as the encoder layer, still has fewer nodes than the input. This structure can still be trained using backpropagation, but with more layers, it is likely to begin suffering from the vanishing or exploding gradient problem. As noted, the work by Hinton et al. for initialization of weights helps resolve this.

4.4.1 Sigmoid based Greedy Layerwise SAE Training

The steps for implementing the SAE training suggested by Hinton et al [37] is as follows

1 Define a network structure which conforms to the requirements of an SAE, with L layers

2 For the first hidden layer, train as you would for an RBM with 2 layers - with the input layer, and the first layer as the hidden layer, using CD-1 and SGD

3 For each layer l in $(2, L/2)$:

- Process the data through the previously trained layers using the forward propagation as defined in 4.2.1
- This processed data then forms the input to the l^th layer, which can be trained once again using CD-1 and SGD as if it were two layers

4 Once all the layers up until the encoder layer have been trained in this greedy layerwise fashion, mirror the weights and layers structures after the encoder to create a fully L layered FFN with pre-trained weights

5 This network can then be trained using the backpropagation and SGD algorithms, where cost of reproducing the network input is minimised

6 Once a minima or acceptable level of reconstruction has been reached, the network can be truncated as the encoder layer, and so the first $L/2$ layers are used as the SAE

Notably, this weight initialization will only be effective if the RBM and SAE networks use the same activation function, which due to the RBM implementation, needs to be a function that can output a probabilistic value in [0,1].

4.4.2 ReLU based SAE Training

ReLU activations differ from Sigmoid in that they are not fitting for probability estimations, which makes the algorithm suggest by Hinton et al. unsuitable. The process used here relies on effective weight initialization, and is a simplification of the above.

1 Define a network structure which conforms to the requirements of an SAE, with L layers

2 Use an effective weight initialization, such as Xavier or He

3 This network can then be trained using the backpropagation and SGD algorithms, where cost of reproducing the network input is minimised

4 Once a minima or acceptable level of reconstruction has been reached, the network can be truncated as the encoder layer, and so the first $L/2$ layers are used as the SAE

4.4.3 Denoising Autoencoders

As noted in 2.3.3, denoising can be used as an optimization technique for autoencoders in order to improve general performance and reconstruction. The methdology works by corrupting the input data for training, but using the non-corrupted data as the expected output sample. In doing so, this forces the autoencoder to learn more fundamental representations of the data rather then fitting to sample noise, hence "denoising". Two of the more typical techniques for achieving this have been implemented, as detailed below. In both cases, the noise is reapplied each time training samples are chosen.

Additive Gaussian Noise With Gaussian noise, samples are corrupted such that a degree of variance is added to the input according to a parameterized Gaussian distribution. In this case, the distribution variance is a configurable training parameter.

$$\tilde{x}|x \sim \mathcal{N}\left(x, \sigma^2 I\right) \tag{36}$$

Masking Noise With masking noise, a fraction of the features in the data are set to 0, thus masking them for that sample. The percentage of features chosen is a configurable training parameter.

4.5 Variance Based Weight Initializations

Recent research, as noted in [35], has shown that weights can be initialized to maintain expected variance between the input and output layers. These methodologies have the immediate advantages of simpler implementation, as well as faster computation (as no pre-training is required). The also allow for effective weight initialization of non-probabilistic activation functions, such as ReLU. Whether they result in better reconstructions or predictions is less clear (especially as the linearity assumption would prove faulty), and so the methods are tested here as well.

A common initialization heuristic is to use a uniform, and but not layer agnostic, initialization such as 37 below. While simple enough, it's been shown that this does not always lead to the best training results, and can be outperformed by variance based initializations [29].

$$W_{ij} \sim U\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right] \tag{37}$$

4.5.1 Initialization Rationale

The variance balancing methodology is based on balancing the variance of a linear network. For input X, with n components and linear neurons with weights W, and output Y,

$$Y = W_1X_1 + W_2X_2 + ... + X_nW_n \tag{38}$$

It can thus be shown, that for $i.i.id$ samples with mean 0, then

$$Var(Y) = nVar(W_i)Var(X_i) \tag{39}$$

For the variance of both input X and output Y to be balanced on the forward and backward propagation, then it is necessary for

$$Var(W_i) = \frac{1}{n_{in}} = \frac{1}{n_{out}} \tag{40}$$

In the instance where there are not an equal number of nodes in the two layers, the average can be taken, such that

$$Var(W_i) = \frac{2}{n_{in} + n_{out}} \tag{41}$$

4.5.2 Initializations

Xavier Using this as the weight values expectation function provides us with the Xavier Glorot initialization [29] which can be used for Sigmoid activations, and is defined as follows

$$w_{ij} \sim U(-r, r), r = \sqrt{\frac{6}{n_i + n_j}} \tag{42}$$

where n is the number of nodes in the i^{th} or j^{th} layer.

He The initialization for ReLU is different on accounts of the function being equal to zero for half it's potential input range - in this case it makes sense to double the weight variance, and so the He [35] initialization is used

$$w_{ij} \sim U(-r, r), r = \sqrt{\frac{6}{n_i}} \tag{43}$$

He-Adjusted An adjusted initialization is presented here, dubbed 'He-Adjusted', where the ReLU initialization uses a mean of the input and output layers in order to scale the weight variance. For networks with constant layer sizes, the intialization is the same as He, though for SAE networks which have layer size changes by definition, the He-Adjusted initialization will result in more appropriately sized weights.

$$w_{ij} \sim U(-r, r), r = \sqrt{\frac{6}{(n_i + n_j)/2}} \tag{44}$$

4.6 CSCV & PBO

The combinatorially symmetric cross-validation (CSCV) method developed by Baily et al., as discussed in 2.6.1, can be used to assess the likelihood of backtest overfitting through comparison of IS and OOS return metrics. Formulaically, the definition of backtest overfitting is given by

$$\sum_{n=1}^{N} E[\overline{r_n} | r \in \Omega_n^*] Prob[r \in \Omega_n^*] \leq N/2 \tag{45}$$

Where the search space Ω consists of the N ranked strategies, and their ranked IS performance r and OOS performance $\bar{r}$. This allows the PBO, using the bayesian formula, to be defined as

$$PBO = \sum_{n=1}^{N} Prob[\overline{r} < N/2 | r \in \Omega_n^*] Prob[r \in \Omega_n^*] \tag{46}$$

Notably, the above definitions consider IS as the data made available to the strategy selection, rather than the models calibration (e.g. the full IS dataset, rather than, by was of example, the number of days used in a moving average). This allows the model-free and non-parametric nature of the definition.

They further developed the CSCV framework as a methodology to reliably estimate the probability used in PBO, which allows a concrete application of the concept. The CSCV framework does not require using the typical hold-out strategy (and thus avoids credibility issues), and is ultimately able to provide a bootstrapped distribution of OOS performance. The full methodology is shown in 5 below [4].

Algorithm 5: CSCV

Input: N configuration trials over T time periods with P&L

1 Generate a TxN performance series matrix, M, representing the profits and losses by the N configuration trials over T time periods

2 Partition the M matrix by rows into S submatrices, each of even size (T/SxN)

3 Generate the combinations C_S of M_S, in groups of size $S/2$, for total $\binom{S}{S/2}$ of combinations

4 For each combination in $c \in C_S$:

a Form a training set J by joinined $S/2$ M_S submatrices, in their original order. J is a matrix of order $(T/S)(S/2) * N$

b Form the test set $\bar{J}$ as the complement of J in M, once again in the original order

c Form a vector of R^c of performance statistics of order N, where the Nth component R^c_n of R^c reports the performance associated with the nth column of J

d Repeat [c] for $\bar{J}$ to derive $\bar{R}^c$ and $\bar{r}^c$

e Determine the element n^* such that $R^c_n \in \Omega^*_n$ - i.e. n^* is the best performing strategy IS

f Define the relative rank of $\bar{r}^c_{n^*}$ by $\bar{\omega}_c := \bar{r}^c_{n^*}/(N+1) \in (0,1)$. This is the relative rank of the OOS performance associated with the strategy chosen IS, which should systematically outperform OOS if no backtest overfitting has taken place

g Define the logit $\lambda_c = ln\frac{\bar{\omega}_c}{(1-\bar{\omega}_c)}$. High values here indicate consistency between IS and OOS performances (and so low ovefitting)

5 The logit values can now be used to compute the distribution ranks of OOS, by collecting all λ_c for $c \in C_S$. The relative frequency for λ occurring across all C_S is

$$f(\lambda) = \sum_{c \in C_S} \frac{\chi_\lambda(\lambda_c)}{\#(C_S)} \tag{47}$$

where χ is the characterization function and $\#(C_S)$ is the number of elements in C_S, and so $\int_{-\infty}^{\infty} f(\lambda)d\lambda = 1$

Result: Reconstructed distribution for $\langle vh \rangle$

The CSCV framework and results thus allows the consideration of several notable statistics. First and foremost, the PBO may now be estimated using the CSCV method and using an integral over the $f(\lambda)$ function as defined above which offers a rate at which the best IS strategies underperform the median of OOS trials. The PBO is estimated using

$$\Phi = \int_{-\infty}^{0} f(\lambda)d\lambda \tag{48}$$

If $\Phi \approx 0$, it is evidence of no significant overfitting (inversely, $\Phi \approx 1$ would be a sign of probable overfitting). Critically then, a PBO measure may be used in a standard hypothesis test to determine if a model should be rejected or not. This can be extended, as shown by Bailey et al., to show the relationship between overfitting and performance degradation of a strategy. It becomes clear that with models overfitting to backtest data noise, there comes a point where seeking increased IS performance is detrimental to the goal of improving OOS performance.

4.7 Performance Assessment

This section may detail any choices and implementation of the performance assessment to be used in CSCV if it's not just the Sharpe Ratio.

4.8 Money Management Strategy and Returns

In line with the general approach here, the Money Management Strategy (MMS) has been designed to be relatively simple, such that the results are more indicative of the underlying modelling rather than intricate trading strategies. The MMS follows an arithmetic long strategy of buying any stock for which the predicted $t + D$ price is above the current price, and

selling the stock at $t + D$. Trading costs have been included at 10% capital costs per annum for borrowing to purchase, and 0.45% for transaction costs. Results are compared to a "perfect" model which has exact knowledge of the price in D days, which represents an upper bound on performance.
Thus, for each stock at time t and prediction point X, the following values are stored or calculated:

1 $\texttt{y}_t$, the closing price for t

2 $\texttt{y}_{t+D}$, the closing price in D days

3 $\hat{\texttt{y}}_{t+D}$, the predicted closing price in D days

4 $\texttt{trade\_long}_t = bool(\hat{y}_{t+5} > y_t) \in (0, 1)$, a flag indicating whether a trade will take place on t

5 $\texttt{trade\_long}_{p,t} = bool(y_{t+5} > y_t) \in (0, 1)$, a flag indicating whether a trade will take place on t in the perfect model

6 $\texttt{r}_{o,t+D} = y_{t+D} * trade\_long_t$, the observed return at $t + D$

7 $\texttt{r}_{e,t+D} = \hat{y}_{t+D} * trade\_long_t$, the expected return at $t + D$

8 $\texttt{r}_{p,t+D} = y_{t+D} * trade\_long_{p,t}$, the perfect return at $t + D$

9 $\texttt{cost}_t = y_t * trade\_long_t$, the cost incurred in buying the stock at time t

10 $\texttt{cost}_{p,t} = y_t * trade\_long_{p,t}$, the cost incurred in the perfect model at t

11 $\texttt{trading\_cost} = y_t * (0.1/365 * 2) + y_t * (0.45/100)$, the possible capital and transaction costs

12 $\texttt{fullcost}_t = (y_t + trading\_cost_t) * trade\_long_t$, the full cost of trading the stock at t in the predictive model

13 $\texttt{fullcost}_{p,t} = (y_t + trading\_cost_t) * trade\_long_{p,t}$, the full cost of trading the stock at t in the perfect model

The MMS is then implemented using these calculated values which are aggregated at each time point t for all stocks, s, as follows:

14 $\texttt{R}_{e,t+D} = \sum_{s \in stocks} r_{s,e,t+D}$, the expected return for $t + D$ from all stocks traded at t

15 $\texttt{R}_{o,t+D} = \sum_{s \in stocks} r_{s,o,t+D}$, the observed return for $t + D$ from all stocks traded at t

16 $\texttt{R}_{p,t+D} = \sum_{s \in stocks} r_{s,p,t+D}$, the perfect return for $t + D$ from all stocks traded at t

17 $\texttt{C}_t = \sum_{s \in stocks} cost_{s,t}$ the cost of all trades at t

18 $\texttt{C}_{p,t} = \sum_{s \in stocks} cost_{s,p,t}$ the cost of all trades at t in the perfect model

19 $\texttt{FC}_t = \sum_{s \in stocks} fullcost_{s,t}$ the cost of all trades at t including transaction costs

20 $\texttt{FC}_{p,t} = \sum_{s \in stocks} fullcost_{s,p,t}$ the cost of all trades at t, including transation costs, in the perfect model

The final model return is thus established as $\sum_t R_t / \sum_t C_t$, for any variations of observed, expected or perfect returns as well as base and full costs. These model returns are used to establish the performance of the different predictive models which were trained.

An additional consideration of the MMS is the daily rates generated by the strategy, which incorporate the costs of any trades incurred, as well as any returns seen at t, such that $\texttt{dailyrate}_t = R_{t-D} - C_t$.

5 Process Implementation

Having covered the technical implementations in Chapters 3 and 4, a higher level overview of the end to end experimental process can be detailed. The process here rests on two key principals: implementing a generalised version of a system which could offer exploration of more complex techniques, and ensuring an effective modularisation of steps such that the process can be reconfigured accordingly while maintaining its integrity. In doing so, a separable system is created which brings together data reduction, deep learning with pre-training/weight initialization, online learning and back test overfitting validation.

5.1 Data Preparation

The fulldataset goes through several steps of pre-processing - these are covered in more detail in 3.1, though the steps are included here:

1 The data is processed into day to day log fluctuations

2 At each time point, rolling historic fluctuation summations are calculated (e.g. the past 1, 7 and 30 days)

3 At each time point, rolling future fluctuation summations are calculated (e.g. the next 2 days)

4 The dataset is truncated to only include points with all aggregations or predictions available

5 The dataset undergoes scaling - the different functions are detailed in 3.1.2, though `limited normalize` was used for most of the results produced

Various configurations were tested in terms of the historic summations used, while the fluctuation predictions were either of 2 or 5 days. These are detailed more extensively in . It's worth noting that the prediction horizon is configurable, and in the case of implementing a more complex trading strategy, one would likely choose to predict multiple points across the horizon in order to to develop a distribution. In this case, the MMS was kept purposefully simple, and so didn't warrant the more intensive implementation.

add re

The time aggregations were chosen according to domain knowledge - it's expected that prices may move in daily, weekly, fortnightly, monthly and quarterly patterns as trading is actioned by day traders to institutional investors . Taking this into account, variations of 1, 5, 10, 20 and 60 working days were considered throughout the configuration tests, in the following cominations:

add hi space

1 1,5,20

2 5,20,60

3 10,20,60

When testing the two limited scaling methodologies, standardization and normalization, it was found that the results for standardizing tended to be notably worse than normalization. It's possible that outliers in price fluctuations and changing variance through time (which is not captured effectively through the limited variation) resulted in a less informative representation and worse performance of the FNN models which can be prone to error maximisation. As a result of this, the limited normalization process was used for most of the tests run. It's worth noting that an implementation which incorporates processing the data through the Error Function in order to reduce the impact of outliers would likely prove a worthwhile endevour, and could be easily swapped in due to the systems modularity.

5.2 Data Segregation

Once processed, the dataset is split into 2 portions: the first set is used for the SAE training, as well as the initial SGD training on the predictive FFN network; the second set is used only for the OGD training in order to generate returns from the predictive network. In this sense, the first set is used as what would typically be the historical training set, and the second is used as the testing/validation set (the use of Bailey's CSCV technique negates the need for a hold-out portion of the dataset [4]). For the sake of clarity, these two subsets of the dataset will be referred to as the 'Training' and 'Prediction' datasets.

In this case, a simple split point of 60%/40% was chosen for the Training and Prediction sets respectively. In a more restrictive case, where there is less data to work with, it would be advised to consider a method such as MinBTL as developed by Bailey et al. [5] and the expected number of configurations to be tested in order to decide an appropriate split point.

5.3 SAE Training

The Training dataset is used to train the auto-encoder network using either RBM pretraining or weight initialization algorithms and SGD training, as defined in 4.4. The Training portion of the dataset, as noted in 5.2, is itself split up into a training and validation set for the SAE. The efficacy of the SAE on the validation set is required in order to determine which networks to choose for subsequent steps in the process. The full testing process here will rely on a generated set of best SAE networks at each encoding layer size, to be used for FFN training and prediction (chosen by a minimum MSE score). The benefit of the modularised system is emphasised here, as the SAE training will not suffer from limitations due to backtesting considerations: any amount of configurations or processes can be tested for feature extraction without concern.

Once the SAE networks have been defined and chosen, they can be used to reprocess both the Training and Prediction datasets such that the input is encoded, and the output is as before. These encoded datasets can then be used for the following steps in the process.

In a productionized system, where data is updated daily, there would be a set period after which the SAE network would be retrained to reflect more recent data. While the system detailed here aims to emulate as much, it wasn't necessary to incorporate this step with the static dataset.

5.4 Prediction Network Training

Typically, the SAE and predictive network might be presented as one and the same, however in the process presented we are able to separate the 2 and optimise for feature extraction prior to training for predictions. Having done so, the datasets can have their input encoded, but retain the same output. The predictive network is then trained using the Training set, as detailed in 4.2.3. There's no requirement to validate this part of the training, so the dataset is not split into subportions as it was for SAE training.

Once the predictive network is trained, the OGD process is run through the encoded Prediction dataset in order to generate the predictions for the assets prices that the model produces - thus emulating what would have occured in a live environment.

5.5 Money Management Strategy

6 Returns generations

- The trading strategy, as per **??**, will be used to generate returns from the predicted and actual prices in the OGD dataset. These returns for each time period and each configuration of the model defined by steps 3-5 will form the M matrix for the CSCV process.

7 CSCV & PBO

- Using the M matrix from 6, the CSCV process will be run which will then allow a calculation of PBO.

6 Results

An introduction to the results, highlighting the important findings and sections.

6.1 Datasets Used

6.1.1 Synthetic Datasets

Data was generated and scaled, as per the methods detail in 3.2 and 3.1.2. The sets were generated with the following configurations, each for 5000 timesteps and with a 60/40 split on the SAE&SGD/OGD training.

Synthetic6

- Upward Trend, High variance: $\mu = 0.9$ and $\sigma = 0.5$
- Upward Trend, Low variance: $\mu = 0.9$ and $\sigma = 0.2$
- Downward Trend, High variance: μ = -0.8 and $\sigma = 0.55$
- Downward Trend, Low variance: μ = -0.8 and $\sigma = 0.15$
- Sideways Trend, High variance: $\mu = 0.05$ and $\sigma = 0.4$
- Sideways Trend, Low variance: $\mu = 0.05$ and $\sigma = 0.1$

Synthetic10

- ($\mu = 0.9$, $\sigma = 0.5$)
- ($\mu = 0.7$, $\sigma = 0.2$)
- ($\mu = 0.05$, $\sigma = 0.4$)
- ($\mu = 0.05$, $\sigma = 0.5$)
- ($\mu = 0.04$, $\sigma = 0.1$)
- ($\mu = 0.02$, $\sigma = 0.15$)
- ($\mu = 0.01$, $\sigma = 0.05$)
- (μ = -0.8, $\sigma = 0.55$)
- (μ = -0.4, $\sigma = 0.15$)
- (μ = -0.1, $\sigma = 0.2$)

This set was configured erroneously with the excessively high means.

6.1.2 Actual Datasets

Several datasets have been used using JSE closing price relative data for 2003-2018 . They were processed the same way as the Synthetic sets, following the steps set out in 3 and with a 60/40 split on the SAE&SGD/OGD training.

add re

AGL Using AGL data from JSE close price relatives.

AGL&ACL Using AGL and ACL data from JSE close price relatives.

Actual10 Using the following assets from JSE close price relatives:

- ACL, AGL, AMS, AOD, BAW, BIL, BVT, CFR, CRH, DDT

6.2 Linearity, Complexity and Structure of Data

6.2.1 GBM Generated Data

A general discussion around the following aspects of GBM data in comparison to actual stock data, and how it has affected the test results which follow. Notably,

- Stationarity of GBM data
- Ergodicity and Lognormality

I'm not entirely sure what should be included here, and to what extent it needs to be discussed.

6.2.2 Activations: Linear, Sigmoid, ReLU and Leaky ReLU

In line with the characteristics of the synthetic data noted in 6.2.1, results for networks using a Linear Activation often outperformed the networks using non-linear activations such as ReLU or Leaky ReLU (activation functions are detailed in 4.2.2).

While these configuration were run to generate a comparison of Linear and ReLU activations for the SAE networks, it is typical to use a Linear Activation in the encoding layer of an SAE [37], and the linear output activation has been implemented here to assist with the shortcomings of the scaling methodology, as noted in 3.1.2.

SAE Activations and Scaling In figures 69 and 70 below, the effects of different activations and scaling methods can be seen when applied to the SAE using actual stock data. The MSE results offer several takeaways: using standardizing scaling instead of normalizing results in far worse performance; Sigmoid activations are notably worse than the linear or ReLU activations and linear output layers offer better results than ReLU output layers.

Networks which use only linear activations show very competitive performance overall, though the ReLU activation shows much better performance than linear activations when the encoding layer size is decreased (71 and 72 below), thus forcing the network to take advantage of non-linear effects. Sigmoid acitvations, while also able to take advantage of non-linear effects, show much worse performance. Sigmoid activations are known to suffer from learning slow down and so increased sensitivity to vanishing gradients . SAE networks are deep by nature, and so it is not surprising that the Sigmoid activations are resulting in worse performance over the same training period.

add re

Results were generated using Configuration1 (8.2.1) for the Actual10 dataset (6.1.2).

There is not a clear reason why Standardizing results in such a massive performance decrease.

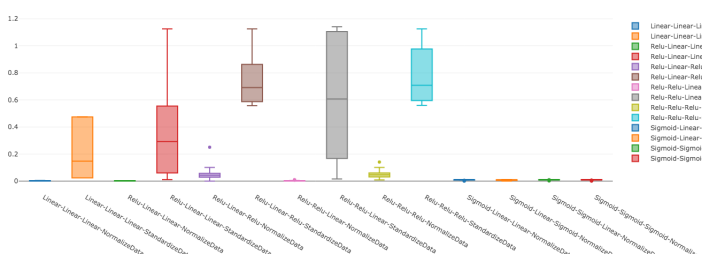

Fig. 69: **Effects of Scaling and Activation Functions on SAE MSE Score (Figure 1/2)** These boxplots show the MSE scores for the Actual10 dataset (8.2.1). Each grouped is labelled in the following format: 'Hidden Activation-Encoding Activation-Output Activation-Scaling Technique. The MSE scores show significantly poorer performance when Standardizing is used instead of Normalizing or when there is a ReLU output activation. These configurations are excluded from the next figure.

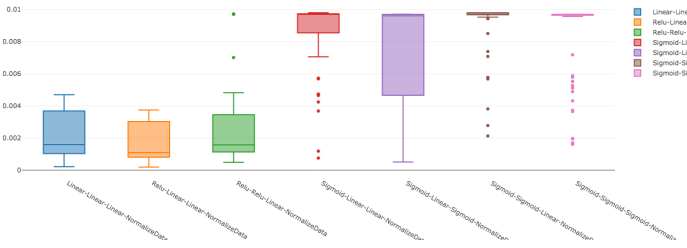

Fig. 70: **Effects of Scaling and Activation Functions on SAE MSE Score (Figure 2/2)** Excluding the Standardizing scaling combinations, we can now see that Sigmoid has largely poor performance unless there is a Linear encoding layer (a widely experienced behaviour), and seems mostly unable to outperform a fully linear network. The best configuration is Relu Hidden layers, with Linear encoding and Output - this makes sense with the non-linear benefit at hidden layers but with less loss of error signal and information at the output and encoding layers.

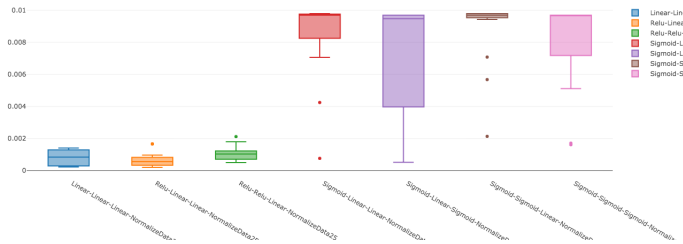

Fig. 71: **Non-linear Benefit in Encoding Layer (Figure 1/2)**
These plots show the performance for all configurations with an encoding layer size of 25 (input 30). The fully linear networks show good performance here, where there is greater scope for linear representation in the encoding.

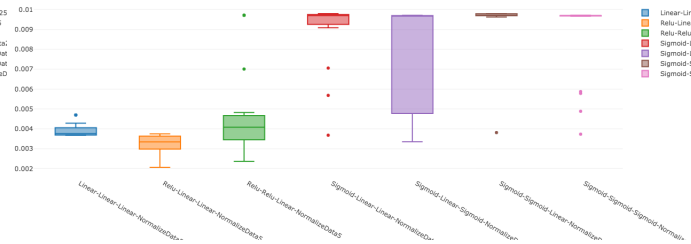

Fig. 72: **Non-linear Benefit in Encoding Layer (Figure 2/2)**
These plots show the performance for all configurations with an encoding layer size of 5 (input 30). Here we now see the benefit of non-linear activations in the ReLU based networks, which the fully linear is not able to outperform.

Predictive FFN Activations and Scaling In figures 73 and 74, a comparison of linear and ReLU activations are made for the Synthetic6 dataset in both smaller and larger sized networks. The linear activations result in notably better P&L when compared to ReLU activations, though unlike the SAE networks, this persists even when the network size is decreased - this difference highlights the effects of GBM data and that it can be represented linearly, as per 6.2.1, whereas the actual stock data benefits from a non-linear representation.

Results were generated using Configuration2 (8.2.2) for the Synthetic6 (6.1.1) dataset.

add sa

As yet, there's no plan to run linear activations for the actual data on the FFN network. Should this be included?

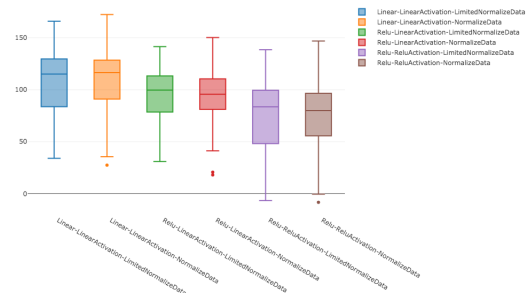

Fig. 73: **Activations in Large FFN Predictive Networks**
The above groupings show the P&L generated for different combinations of Hidden Activation - Output Activation - Scaling Method (which is for both SAE and FFN). For an input of 18, and network sizes of 40 and 80 with hidden layers of 1 an 3, we see the linear activations showing notably better performance than the Relu activations.

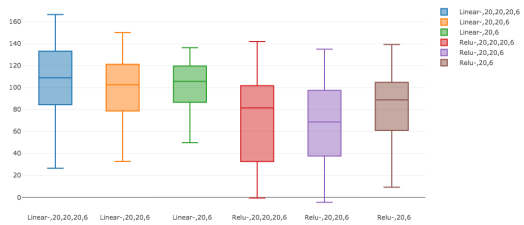

Fig. 74: **Activations in Smaller FFN Predictive Networks**
The boxplots here show the P&L generated for networks with smaller layer sizes, as indicated by the group naming (each number represents a hidden layer and its node size). The networks were trained on the Synthetic6 dataset with 18 inputs, different sized autoencoders and learning rates (720 configurations in total). We once again see the outperformance of ReLU by the linear activations.

Leaky ReLU vs ReLU Leaky ReLU activations were implemented, and as can be expected, do show a small improvement on both the SAE and predictive FFN networks when compared to ReLU. These effects are show for the Synthetic6 dataset in in figures 75 and 76 below.

add co
tion re

add co
tion re
and de

Reorde
ing gra

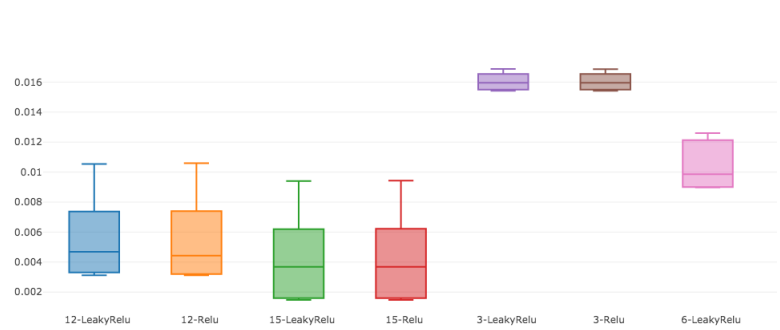

Fig. 75: **SAE: Leaky ReLU vs ReLU**
The plot above shows the MSE for 120 different SAEs, grouped by encoding size and activations.

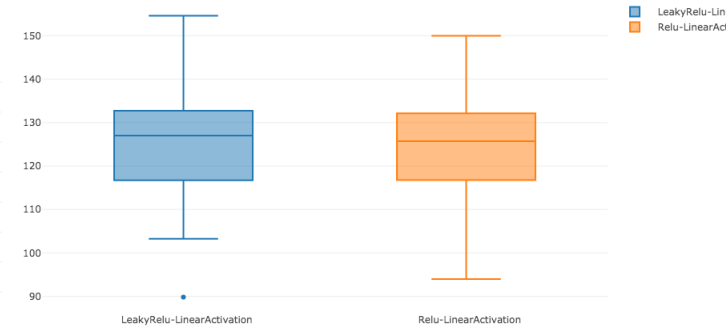

Fig. 76: **FFN: Leaky ReLU vs ReLU**
The plot above shows the P&L for 80 different predictive networks, grouped by activation, showing some improvements from the Leaky ReLU.

6.2.3 Effects of Network Size

The effects of networks size on SAE MSE performance and the P&L predictive FFN performance are shown in figures 77 and 78 below. They show a general increase in performance for both, with improvements from increases in the sizes and numbers of layers in the networks. The classical SAE structure of decreasing layer sizes results in worse performance - it's possible that the move away from greedy layerwise training nullifies the approach of iterative feature reduction, and that with the variance based weight initializations there is a more typical increase in performance with larger networks. Similar trends were seen in the results for the Synthetic10 dataset (6.1.1), which can be seen in the appendix (8.1.1).

Results were generated using the AGL dataset (6.1.2) and Configurations 3 and 4 (8.2.3, 8.2.4) for the SAEs and predictive FFNs respectively.

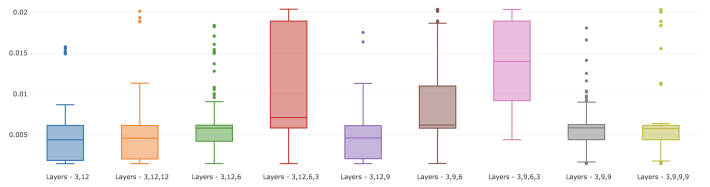

Fig. 77: **Performance of AGL SAE networks by size**
This figure shows the MSE performance of the SAE networks for AGL. The scores show improved results for larger networks (i.e. the '9,9,9' and '12,12' sized networks have the lowest scores). The results also show much worse results for the networks in the networks with gradually decreasing layer sizes.

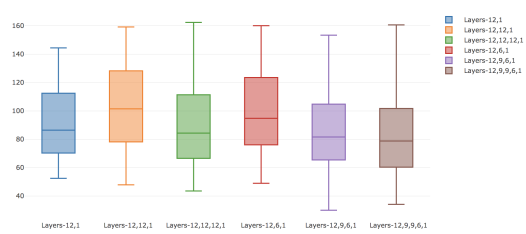

Fig. 78: **Performance of AGL predictive FFN networks by size**
The plot above shows the P&L for AGL predictive FFN networks. Similarly, there is increased preformance for networks of larger sizes, and no notable benefits of a decreasing layer size structure.

6.3 Feature Selection and Data Aggregation

Feature Selection The figures below show the efficacy of the different encoding layer sizes from the SAE and how they impact P&L in the predictive FFN network. It's apparent in the synthetic dataset that feature selection is occuring, with the optimal size being 10, around the degrees of freedom (or number of assets). This pattern is less clear in the AGL results where there were only 2 encoding sizes that could be used appropriately.

Results were generated using the AGL dataset (6.1.2) with Configurations 3 and 4 (8.2.3, 8.2.4) for the SAEs and predictive FFNs respectively, and the Synthetic10 dataset (6.1.1) with Configurations 5 and 6 (8.2.5, 8.2.6) for the SAEs

and predictive FFNs respectively.

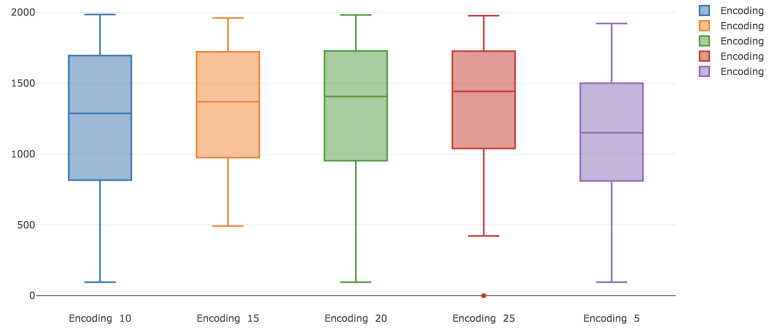

Fig. 79: **Encoding Size Effects on P&L for Synthetic Data**
This figure shows the effecst of the difference encoding layer sizes on the P&L when the relevant SAE is used for the predictive FFN, as for the Synthetic10 dataset (with an input of 30 features - 3 per asset). There is a clear effect of the SAE being able to perform effective feature reduction, with the best performance being at encoding layer size 10.

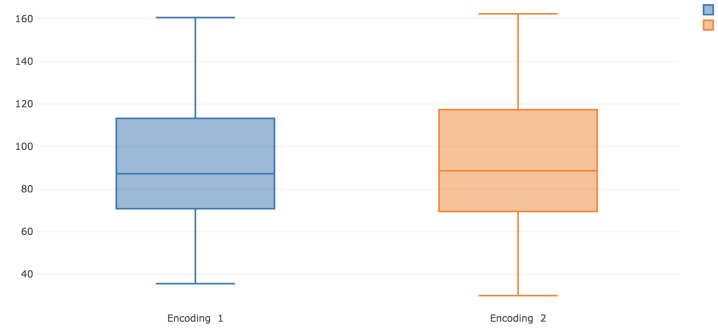

Fig. 80: **Encoding Size Effects on P&L for AGL Data**
This figure shows the effecst of the difference encoding layer sizes on the P&L when the relevant SAE is used for the predictive FFN, as for the AGL dataset (with an input of 3 features). Performance for both sizes are similar, though increasing the number of assets may result in a different effect being seen.

Would it be worthwhile to go back and run the configurations with no SAE to see the effect? This was done in a previous it

Data Aggregation Input data was scaled to 3 different configurations across a series of tests in order to assess the effect of shorter and longer configurations on both SAE MSE performance, as well as the predictive FFN P&L results. The configurations tested (in trading day window periods) were:

- 1, 5, 20
- 5, 20, 60
- 10, 20, 60

Results highlighted the differences between the synthetic and actual data in terms of stationarity, and the benefit of more real-time data when it comes to actual stocks.

Results were generated using the AGL dataset (6.1.2) with Configurations 3 and 4 (8.2.3, 8.2.4) for the SAEs and predictive FFNs respectively, and the Synthetic10 dataset (6.1.1) with Configurations 5 and 6 (8.2.5, 8.2.6) for the SAEs and predictive FFNs respectively.

The results for SAE MSE are not totally clear at the moment. It may be more apparent with a larger number of actual stoc

Fig. 81: **SAE Data Window Aggregations - Synthetic Data**
The box plots show the MSE for a series of SAE networks trained, grouped by different data window aggregations. The minimum and median MSE is lower for the higher aggregations, suggesting that the reduced noise in the stationary feature set may support longer windows in further training.

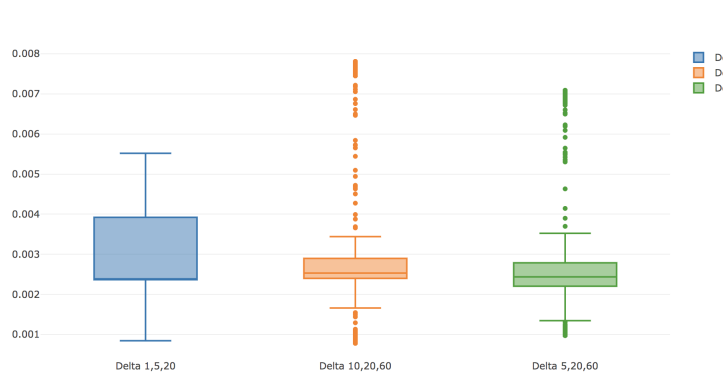

Fig. 82: **SAE Data Window Aggregations - AGL Data**
For the AGL dataset, the minimum MSE score was found in the '10,20,60' aggregated dataset, though this configuration also results in the worst MSE scores and highest median, suggesting that there is more difficultly in a non-stationary dataset.

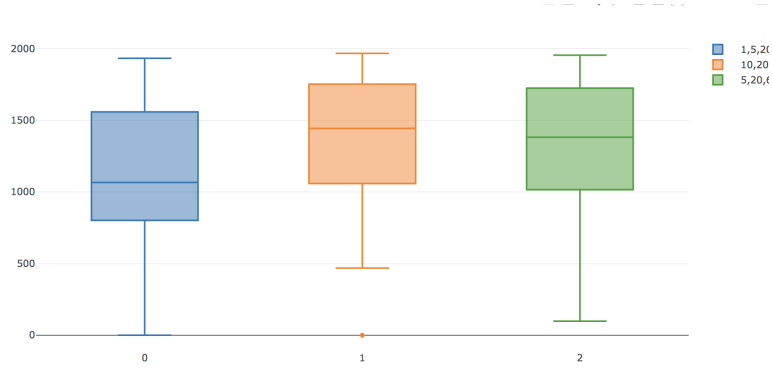

Fig. 83: **FFN Data Window Aggregations - Synthetic Data**
The figure here shows the P&L for the series of FFN networks trained, grouped by different data window aggregations. The profit is higher for the longer aggregations, which is in line with expectations, where the network is able to learn a mean for the synthetic GBM data.

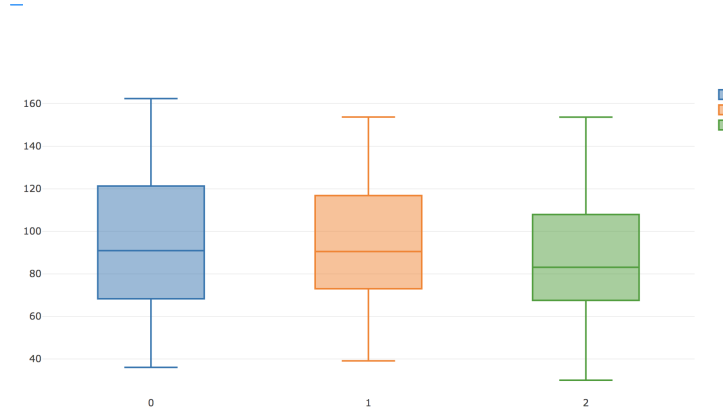

Fig. 84: **FFN Data Window Aggregations - AGL Data**
The figure here shows a notably higher P&L for datasets with a 1 day aggregation, highlighting again the non-stationarity of the real data, and thus the benefit of more accurate recent data.

Historical Data and SGD training Tests were run to determine the effect of excluding data from the SGD training of the predictive network. It should be noted that these predictive networks make use of SAEs that were trained on 80% of the data regardless, which may prime the input of data to the predictive network despite the portion of data being excluded from SGD training.

The number of epochs run in the SGD training were not adjusted according to the amount of data used, so ultimately the smaller dataset configurations were trained on fewer samples. It would seem that the use of historical data or the more extensive training is beneficial, but not the primary factor in determining P&L, as decreases in profit were certainly not proportional to the decrease in the period of data.

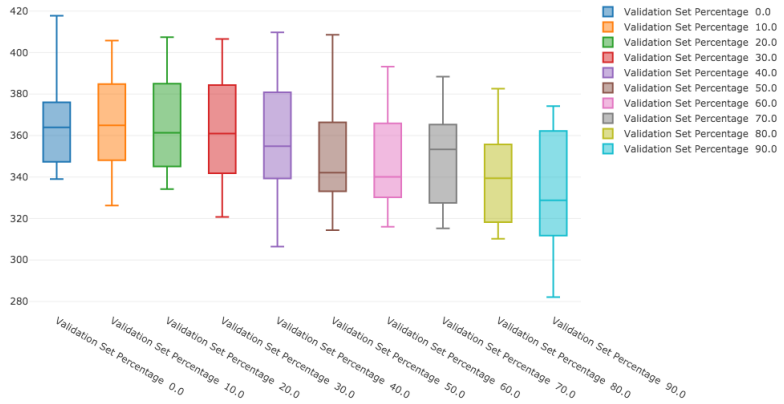

Fig. 85: **SGD Training Dataset Size - 6 Synthetic Assets**
The plot above shows the P&L for FFN predictive networks, grouped by the percentage of data excluded from the SGD training. While there is a decrease in P&L, it is much less than the decrease in data, and may only be a result of the reduced samples used to train (due to a constant number of SGD epochs)

Results were generated using Configuration2 (8.2.2) for the Synthetic6 dataset (6.1.1).

add SA

The trainig for this also used the first X% of data, rather than the last X% in the SAE/SGD set, which is slightly counterint

The tests for this are quite extensive - not sure how to balance that with PBO for the actual data.

Data aggregation results could be down to martingale effect.

6.4 Weight Initialization Techniques

6.4.1 RBM Pretraining for Sigmoid Networks

While previously considered the best approach for training deep neural networks, the methodology of greedy layerwise RBM pre-training for Sigmoid SAE networks (as described in [37]) had detrimental effects on network performance. Experiments assessed the effects of running the pre-training for 0, 1, 10 and 50 epochs, where there is a clear decrease in performance as the number of epochs increase (the low value outliers that can be seen for the 1 epoch configuration were with learning rates which were low enough to approximate no epochs).

The RBM pre-training technique does assume that data is Independent and Identically Distributed (IID), and does ultimately traverse a different solution space and loss function. While it may be effective in some contexts, it was shown to result in a counterproductive exploration of the weight space for this non-IID dataset.

Results were generated using Configuration7 (8.2.7) for the AGL&ACL dataset (6.1.2).

The CD vs Backprop differences don't appear to have a theoretical justification (that I could find). Not sure how much I ne

Fig. 86: **Pre-training Effects on financial SAE MSE Scores**
The boxplots here show the summary of configuration performances, by minimum MSE achieved, grouped according to the number of pre-training epochs which the network had. There is a clear favour to having no pre-training in this scenario.

Sigmoid Activation Functions Due to the poor performance seen in Sigmoid function based SAE's here, as well as the poorer results when compared to ReLU activations (noted in 6.2.2), Sigmoid functions were largely excluded from further configuration testing.

6.4.2 Variance Based Weight Initialization Techniques

More recent research has focused on the use of weight initialization using variance base methodologies, as discussed fully in 4.5. As suggested, the DC initialization for Leaky ReLU activations outperforms both Xavier and He when used for networks with varying layer sizes. When considering the predictive FFN networks, with more consistent layer sizes, the performance between He and DC becomes comparable, with both outperforming Xavier (as expected, with ReLU activations).

Results were generated using the AGL dataset (6.1.2) with Configurations 3 and 4 (8.2.3, 8.2.4) for the SAEs and predictive FFNs respectively, and using the Synthetic10 dataset (6.1.1) with Configurations 5 and 6 (8.2.5, 8.2.6) for the SAEs and predictive FFNs respectively.

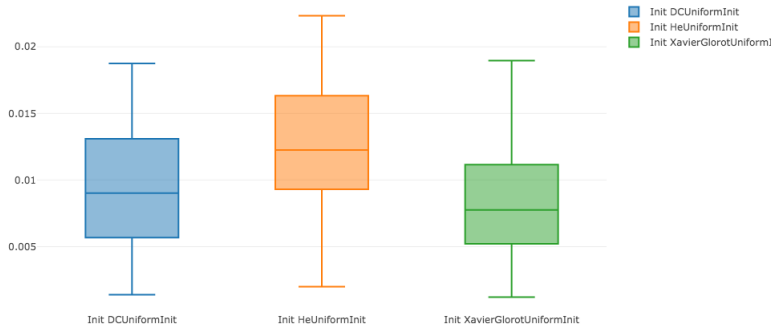

Fig. 87: **Weight Initialization for SAE MSE - Synthetic10**
The box plots show the MSE for a series of SAE networks trained. As suggested in 4.5, the He initialization results in poorer fitting for networks with varying network layer sizes. The Xavier and DC initializations were largely on par, thought with DC resulting in better performance for smaller encoding layers.

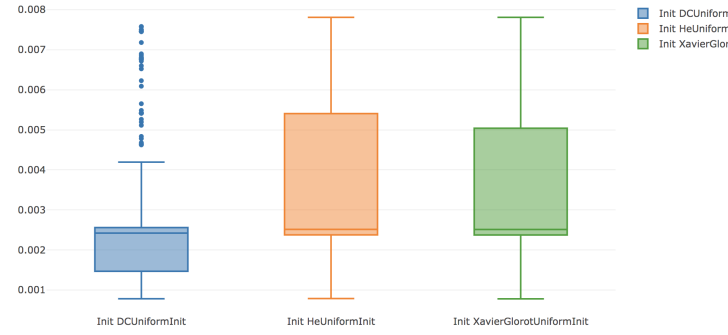

Fig. 88: **Weight Initialization for SAE MSE - AGL**
The configurations run for AGL show notably better performance using the DC based initialization in comparison to both He and Xavier. The emphasis seen here can be attributed to increased pressure on suitable weight initizliations with 1 and 2 node encoding layers used.

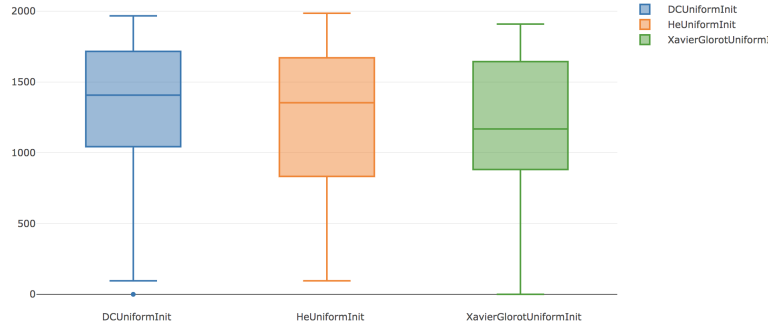

Fig. 89: **Weight Initialization for Predictive FFN P&L- Synthetic10**
The figure here shows the best network P&L performance for DC and He being higher than Xavier, as to be expected with more balanced network layers and ReLU activations. The better performance in the lower end of the DC configurations when compared to He can be attributed to the inclusion of inconsistently sized layers in the configuration set, where DC would have performed better.

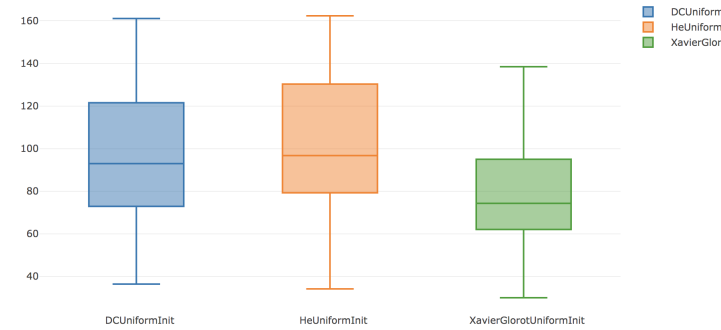

Fig. 90: **Weight Initialization for Predictive FFN P&L- AGL**
The configurations run for AGL show notably better performance using the DC and He based initialization in comparison Xavier, thought with less of a difference between the 2 in comparison to the Synthetic set.

The DC init here is 'semi-novel' so to speak (afaik). I'm not sure how far this should be emphasised/addressed.

6.5 Learning Optimizations

6.5.1 Learning Rate Schedules

Learning rate schedules were implemented such that the learning rate would rise and fall through a set of values during a set number of epochs, as discussed in **??**. Conceptually, this should allow a more effective traversal of the solution space, where larger learning rates are able to dislodge the configuration from saddle points, and smaller learning rates are able to optimize at a minima. The schedules presented here have a constant minimum learning rate of 0.00001, with an adjusted maximum learning rate as presented.

Results were generated using the AGL dataset (6.1.2) with Configurations 3 and 4 (8.2.3, 8.2.4) for the SAEs and predictive FFNs respectively.

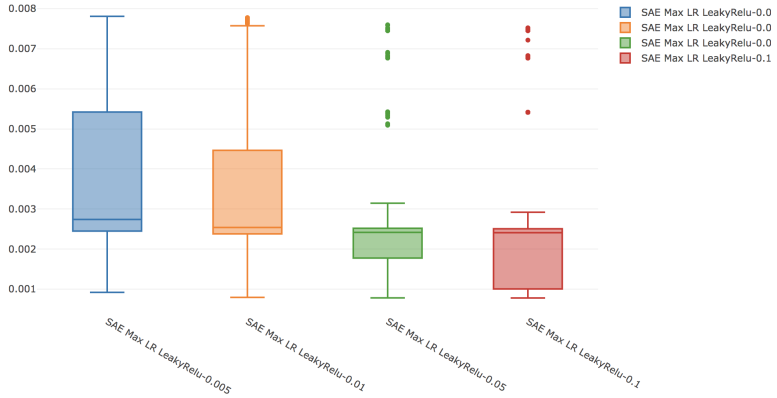

Fig. 91: **Effects of Learning Rate Schedule on SAE MSE**
The figure here shows the improvements in MSE scores gained from increasing the upper bound on the learning rate schedule for SAE configurations, with a notably better performance from the largest upper bound.

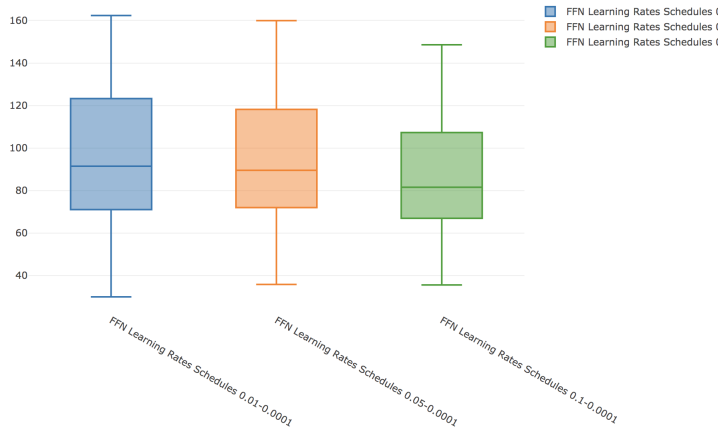

Fig. 92: **Effects of Learning Rate Schedule on Predictive FFN P&L**
Unlike the SAE configurations, the predictive FFN performance deteriorates as the upper bound is increased, with the largest upper bound having significantly worse P&L results.

The results here highlight the different nature of the SAE and predictive networks and the solution space they inhabit. The feature reduction performed by the SAE is expectedly complex, and so benefits from a more dynamic learning process. Conversely, the predictive network seems to take on martingale properties which would not require as complex searching through the solution space.

Need to address the validity of this claim. Also, is there any need to include synthetic results here? (they would need to be

6.5.2 Regularization

L1 Regularization was implemented and tested on a selection of 10 real assets, as per below.

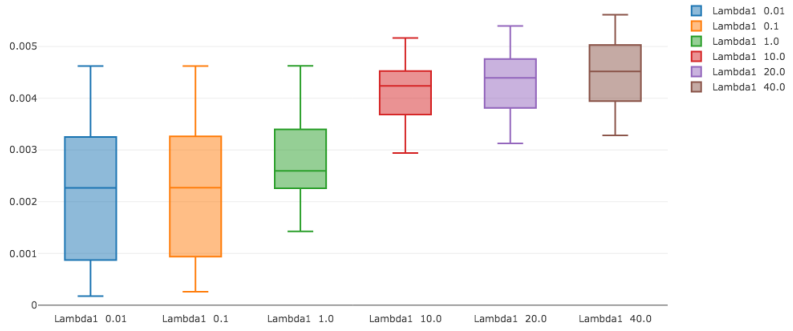

Fig. 93: **SAE L1 Regularization - 10 Real Asset**
The plot above shows the MSE for 360 different SAE networks, grouped by L1 Lambda, showing notable degradation as regularization is increased. Synthetic data tests showed no performance changes below 0.01.

This still contrasts with the learning rate schedule. It could perhaps be the nature of L1 vs L2 regularization though.

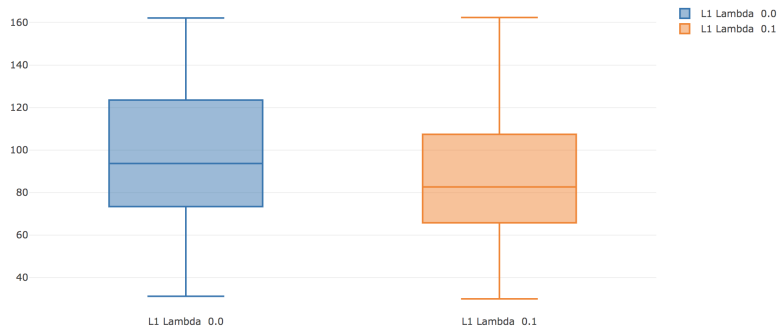

Fig. 94: **L1 Regularization - AGL Predictive FFN**
Similarly, performance decreases for the AGL predictive FFN as regularization increases.

6.5.3 Denoising

Denoising tests so far don't seem worth including, though should be considered for future experiments.

7 Conclusions

8 Appendix

8.1 Additional Results

8.1.1 Effects of Network Size on SAE and FFN for Synthetic Data

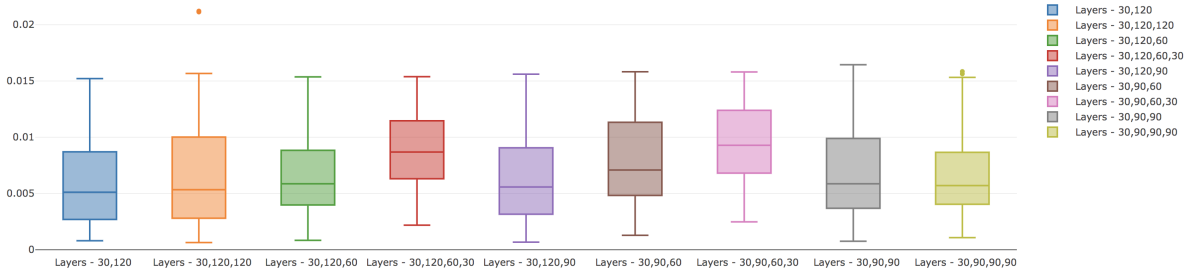

Fig. 95: Network Sizes for SAE on the Synthetic10 dataset
The box plots show the MSE for the series of SAE networks trained. The results show worse performance for the typical descending network sizes used for SAE, and show no improvement form increasing network size and complexity.

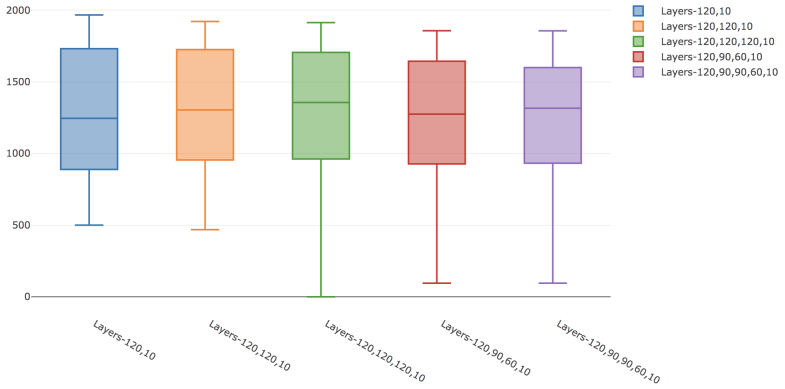

Fig. 96: Network Sizes for Predictive FFN on the Synthetic10 dataset
The box plots show the profits for the series of FFN networks trained, showing small improvments for network size increases.

8.2 Configuration Sets Used

8.2.1 Configuration1 - SAE

- Data: Closing Prices for ACL, AGL, AMS, AOD, BAW, BIL, BVT, CFR, CRH, DDT
- Data Windows: 1, 5, 20 (thus input of 30)
- Data Scaling: Normalize, Standardize
- Weight Init,: Xavier Glorot Uniform
- All Layer Encodings: Sigmoid, ReLU, Linear (split by hidden, encoding and output)
- Encoding Layer Size: 5, 15, 25
- Hidden Layer Size: 60, 120
- Number of Hidden Layers: 1, 2, 3
- Learning Rates: Various for each different activation type (chosen by hidden activation used)

8.2.2 Config2 - Predictive FFN

- SAEs: Best performance (lowest MSE) of encoding sizes 3, 6, 9, 12, 15 (input 18) from Config1, of both Limited Scaling and typical Scaling methods
- Output Activations: ReLU, Linear
- Hidden Activations: ReLU, Linear
- Layer Sizes: 40, 80
- Number of Layers: 1, 3
- Various Learning rates for SGD & OGD training

8.2.3 Configuration3 - SAE

- Data Windows: [1,5,20], [5,20,60], [10,20,60]
- Data Scaling: Normalize
- Weight Init: Xavier Glorot Uniform, He, DC
- Layer Encodings: Leaky ReLU with Linear encoding and output
- Encoding Layer Size: 1, 2
- Network Sizes: (12,6), (12), (12,12), (12,9), (9,6), (9,6,3),(12,6,3), (9,9,9), (9,9)
- Learning Rates: Scheduled from (0.005, 0.01, 0.05, 0.1) to 0.0001
- Learning Rate Epoch Schedules: (100,300)
- SGD Epochs: 400
- Minibatch Size: 20
- SAE Validation Split: (0.8)

8.2.4 Configuration4 - Predictive FFN

- SAEs: Best networks by encoding size and data aggegations from Configuration3.
- Data Windows: [1,5,20], [5,20,60], [10,20,60]
- Data Scaling: Normalize
- Weight Init: Xavier Glorot Uniform, He, DC
- Layer Encodings: Leaky ReLU with Linear output
- Network Sizes: (12, 12), (12, 12, 12), (12, 9, 9, 6), (12, 9, 6), (12)
- Learning Rates: Scheduled from (0.01, 0.05, 0.1) to 0.0001
- Learning Rate Epoch Schedules: (100)
- SGD Epochs: 400
- Minibatch Size: 20
- Training Validation Split: (1.0) (i.e. none)
- OGD Learning Rates: (0.01, 0.05)
- L1 Regularization: (0, 0.1)

8.2.5 Configuration5 - SAE

- Data Windows: [1,5,20], [5,20,60], [10,20,60]
- Learning Rate Schedules: 0.005 to (0.01, 0.05, 0.1) by (100) epochs
- Network Sizes:(120, 60, 30), (90, 60, 30), (90, 90, 90), (90, 90), (120, 90), (120,60), (120)
- Encoding: 25;20;15;10;5
- SGD Epochs: 400
- Minibatch Size: 20
- SAE Validation Split: (0.8)
- Weight Init: Xavier Glorot Uniform, He, DC
- Layer Encodings: Leaky ReLU with Linear encoding and output

8.2.6 Configuration6 - Predictive FFN

- Best SAEs by encoding size and window aggregations (15 total) from Configuration5
- Learning Rate Schedules: 0.005 to (0.01, 0.05, 0.1) by (100) epochs
- Network Structures: (120;120;120;120), (120,90,90,60), (120,90,60),(120,120,120),(120,120),(120,60),(120)
- OGD Learning Rates: (0.01, 0.05)
- Data Windows: [1,5,20], [5,20,60], [10,20,60]
- Data Scaling: Normalize
- Weight Init: Xavier Glorot Uniform, He, DC
- Layer Encodings: Leaky ReLU with Linear output
- Learning Rate Epoch Schedules: (100)
- SGD Epochs: 400
- Minibatch Size: 20
- Training Validation Split: (1.0) (i.e. none)
- L1 Regularization: (0, 0.1)

8.2.7 Configuration7 - SAE

- Data: Closing prices for AGL & ACL
- Processing: Time windows of 1, 5 and 20 (thus an input vector of size 6)
- Encoding Layer Activations: Sigmoid and Linear
- Encoding Layer Size: 5
- Hidden Layer Size: 15
- Number of Hidden Layers: 1, 2
- Learning Rates: 0.0000001, 0.00001, 0.001, 0.1, 1.0, 2.0

9 References

[1] Albers S. Online Algorithms: a survey. Mathematics Subject Classification (1991): 68W25, 68W40. Available at: https://link.springer.com/content/pdf/10.1007

[2] Aparicio, Diego and Lopez de Prado, Marcos, How Hard Is It to Pick the Right Model? (December 2017). Available at SSRN: https://ssrn.com/abstract=3044740 or http://dx.doi.org/10.2139/ssrn.3044740

[3] B. Arthur (1995) Complexity in Economics and Financial Markets, Complexity 1 (1), 20???25.

[4] Bailey, David H. and Borwein, Jonathan and Lopez de Prado, Marcos and Zhu, Qiji Jim, The Probability of Backtest Overfitting (February 27, 2015). Journal of Computational Finance (Risk Journals), 2015, Forthcoming. Available at SSRN: https://ssrn.com/abstract=2326253 or http://dx.doi.org/10.2139/ssrn.2326253

[5] Bailey, David H. and Borwein, Jonathan and Lopez de Prado, Marcos and Zhu, Qiji Jim, Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance (April 1, 2014). Notices of the American Mathematical Society, 61(5), May 2014, pp.458-471. Available at SSRN: https://ssrn.com/abstract=2308659 or http://dx.doi.org/10.2139/ssrn.2308659

[6] Bailey, David H. and Lopez de Prado, Marcos, The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality (July 31, 2014). Journal of Portfolio Management, 40 (5), pp. 94-107. 2014 (40th Anniversary Special Issue).. Available at SSRN: https://ssrn.com/abstract=2460551 or http://dx.doi.org/10.2139/ssrn.2460551

[7] G. Bakiri and T. G. Dietterich. Achieving high-accuracy text-to-speech with machine learning. In R. I. Damper, editor, Data Mining Techniques in Speech Synthesis. Chapman and Hall, New York, NY, 2002. 22

[8] Bao W, Yue J, Rao Y (2017) A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE 12(7): e0180944. https://doi.org/10.1371/journal.pone.0180944

[9] Bartlett P., Hazan E. Adaptive Online Gradient Descent. Available at: http://papers.nips.cc/paper/3319-adaptive-online-gradient-descent.pdf

[10] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In Bernhard Scholkopf, John Platt, and Thomas Hoffman, editors, Advances in Neural Information Processing Systems 19 (NIPS06), pages 153160. MIT Press, 2007. Available at: http://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf

[11] Bengio, Yoshua, and Samy Bengio. "Modeling high-dimensional discrete data with multi-layer neural networks." Advances in Neural Information Processing Systems. 2000. Available at: http://papers.nips.cc/paper/1679-modeling-high-dimensional-discrete-data-with-multi-layer-neural-networks.pdf

[12] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6472238&tag=1

[13] Bottou L., Le Cun Y. Large Scale Online Learning. Available at: http://papers.nips.cc/paper/2365-large-scale-online-learning.pdf

[14] Bottou, L. and Murata, N. (2002). Stochastic Approximations and Efficient Learning. In Arbib, M. A., editor, The Handbook of Brain Theory and Neural Networks, Second edition,. The MIT Press, Cambridge,

[15] Cirean, Dan Claudiu, et al. "Deep, big, simple neural nets for handwritten digit recognition." Neural computation 22.12 (2010): 3207-3220. Available at: `https://www.mitpressjournals.org/doi/abs/10.1162/NECO_a_00052`

[16] Chu C., Time series segmentation: A sliding window approach. Information Sciences, Volume 85, Issues 13, July 1995, Pages 147-173. Available at: `ttps://www.sciencedirect.com/science/article/pii/002002559500021G#!`

[17] F. L. Chung, T. C. Fu, R. Luk, and V. Ng, Flexible time series pattern matching based on perceptually important points, in Proc. Int. Joint Conf. Artificial Intelligence Workshop Learning from Temporal and Spatial Data in International Joint Conference on Artificial Intelligence (IJCAI'01), Seattle, Washington, USA, 4-10 August, 2001, p. 1-7. Available at: http://hdl.handle.net/10397/48578

[18] J. P. Crutchfield, (2011), Between order and chaos, Nature Physics, Vol. 8, January 2012, 17-23

[19] Dauphin, Y. et al. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Proc. Advances in Neural Information Processing Systems 27 29332941 (2014) Available at: http://papers.nips.cc/paper/5486-identifying-and-attacking-the-saddle-point-problem-in-high-dimensional-non-convex-optimization

[20] Aditya Devarakonda, Maxim Naumov & Michael Garland. ADABATCH: ADAPTIVE BATCH SIZES FOR TRAINING DEEP NEURAL NETWORKS. Available at: https://arxiv.org/pdf/1712.02029.pdf

[21] David L. Donoho, High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality (August 8, 2000). Available at: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.329.3392&rep=rep1&type=pdf`

[22] John Duchi, Elad Hazan and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research 12 (2011) 2121-2159. Available at: http://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf

[23] Erghan D., Bengio Y., Courville A, Manzagol P., Vincent P. Why Does Unsupervised Pre-training Help Deep Learning?, Journal of Machine Learning Research 11 (2010) 625-660 . Available at: http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf

[24] E.F. Fama, The behavior of stock-market prices, J. Bus., 1 (1965), pp. 34-105. Available at: https://doi.org/10.1086/294743

[25] Jianqing Fan, Runze Li, Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery (7 Feb, 2006). Available at: https://arxiv.org/abs/math/0602133

[26] Fan, J., & Fan, Y. (2008). High Dimensional Classification Using Features Annealed Independence Rules. Annals of Statistics, 36(6), 26052637. http://doi.org/10.1214/07-AOS504

[27] Ge, Rong, et al. "Escaping from saddle pointsonline stochastic gradient for tensor decomposition." Conference on Learning Theory. 2015. Available at: http://proceedings.mlr.press/v40/Ge15.pdf

[28] Goodfellow, Ian J., et al. "Maxout networks." arXiv preprint arXiv:1302.4389 (2013). Available at: http://proceedings.mlr.press/v28/goodfellow13.pdf

[29] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010. Available at: http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

[30] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks." Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. 2011. Available at: http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf

[31] Griffioen, Gerwin A. W., Technical Analysis in Financial Markets (March 3, 2003).

[32] Hansen, Peter Reinhard and Lunde, Asger and Nason, James M., The Model Confidence Set (March 18, 2010). Available at SSRN: https://ssrn.com/abstract=522382 or http://dx.doi.org/10.2139/ssrn.522382

[33] Harvey, Campbell R. and Liu, Yan, Backtesting (July 28, 2015). Available at SSRN: https://ssrn.com/abstract=2345489 or http://dx.doi.org/10.2139/ssrn.2345489

[34] Hawkins, Douglas. (2004). The Problem of Overfitting. Journal of chemical information and computer sciences. 44. 1-12. 10.1021/ci0342472.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Available at: https://arxiv.org/pdf/1502.01852.pdf

[36] Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. Neural Computation, 18:15271554, 2006. Available at: https://www.mitpressjournals.org/doi/abs/10.1162/neco.2006.18.7.1527

[37] G. E. Hinton, R. R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks. Science 28 Jul 2006: Vol. 313, Issue 5786, pp. 504-507 DOI: 10.1126/science.1127647. Available at: http://science.sciencemag.org/content/313/5786/504/tab-pdf

[38] Geoffrey E. Hinton, Training Products of Experts by Minimizing Contrastive Divergence,Neural Computation Volume 14 — Issue 8 — August 2002 p.1771-1800 . Available at: https://www.mitpressjournals.org/doi/pdf/10.1162/089976602760128018

[39] Hinton, Geoffrey E., et al. "Improving neural networks by preventing co-adaptation of feature detectors." arXiv preprint arXiv:1207.0580 (2012). Available at: https://arxiv.org/pdf/1207.0580.pdf

[40] Geoffrey E. Hinton, A Practical Guide to Training Restricted Boltzmann Machines, August 2, 2010. Available at: http://learning.cs.toronto.edu

[41] Campbell R. Harvey & Yan Liu & Heqing Zhu, 2016. " and the Cross-Section of Expected Returns," Review of Financial Studies, vol 29(1), pages 5-68.

[42] Sepp Hochreiter, Jrgen Schmidhuber. Long Short-Term Memory. Neural Computation Volume 9 — Issue 8 — November 15, 1997 p.1735-1780. Available at https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735.

[43] K. Hornik, Multilayer feed-forward networks are universal approximators, Neural Networks, vol 2, 1989

[44] Hsu D. Time Series Compression Based on Adaptive Piecewise Recurrent Autoencoder (17 August 2017). Available at: https://arxiv.org/pdf/1707.07961.pdf

[45] Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. IEEE Transactions on Systems, Man and Cybernetics, (4), 364378. Available at: https://ieeexplore.ieee.org/document/4308320/

[46] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012. Available at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[47] Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

[48] N. Johnson, G. Zhao, E. Hunsader, H. Qi, N. Johnson, J. Meng & Brian Tivnan (2013). Abrupt rise of new machine ecology beyond human response time. Sceintific Reports 3(2627). DOI: 10.1038/srep02627

[49] Kahn, Michael N. . Technical Analysis Plain and Simple: Charting the Markets in Your Language, Financial Times Press, Upper Saddle River, New Jersey, p. 9. ISBN 0-13-134597-4.(2006)

[50] Stefan Knerr, Handwritten Digit Recognition by Neural Networks with Single-Layer Training. 962 IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 3, NO. 6, NOVEMBER 1992. Available at: `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=165597`

[51] John Langford, Lihong Li and Tong Zhang. Sparse Online Learning via Truncated Gradient. Journal of Machine Learning Research 10 (2009) 777-801 Submitted 6/08; Revised 11/08; Published 3/09. Available at: http://www.jmlr.org/papers/volume10/langford09a/langford09a.pdf

[52] Martin Lngkvist, Lars Karlsson, Amy Loutfi, A review of unsupervised feature learning and deep learning for time-series modeling, Applied Autonomous Sensor Systems, School of Science and Technology, rebro University, SE-701 82 rebro, Sweden. Available at: `https://www.sciencedirect.com/science/article/pii/S0167865514000221#bi005`

[53] LeCun Y., Bottou L, Orr G, Muller K. Efficient Backprop. Available at: http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf

[54] LeCun, Y. (1988). A theoretical framework for back-propagation. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), Proceedings of the 1988 connectionist models summer school (pp. 2128). `Availableat:https://www.researchgate.net/profile/Yann_Lecun/publication/2360531_A_Theoretical_Framework_for_Back-Propagation/links/0deec519dfa297eac1000000/A-Theoretical-Framework-for-Back-Propagation.pdf`

[55] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4), 541551. Available at: http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf

[56] Yann LeCun, Yoshua Bengio & Geoffrey Hinton. Deep learning, Nature volume 521, pages 436444 (28 May 2015). Available at: https://www.nature.com/articles/nature14539

[57] Le Roux, Nicolas, and Yoshua Bengio. "Representational power of restricted Boltzmann machines and deep belief networks." Neural computation 20.6 (2008): 1631-1649. Available at: http://www.iro.umontreal.ca/ lisa/publications2/index.php/attachments/single/22

[58] Liu X., Lin Z., Wang H. Novel Online Methods for Time Series Segmentation. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 12, DECEMBER 2008. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4445667&tag=1

[59] Lo, Andrew W., The Statistics of Sharpe Ratios. Financial Analysts Journal, Vol. 58, No. 4, July/August 2002. Available at SSRN: https://ssrn.com/abstract=377260

[60] Ilya Loshchilov, Frank Hutter, SGDR: Stochastic Gradient Descent with Warm Restarts, ICLR 2017 conference paper. Available at: https://arxiv.org/abs/1608.03983

[61] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16, NO. 2, APRIL 2015. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6894591&tag=1

[62] Mahajan D., Keerthi S., Sundararajan S., Bottou L. A Parallel SGD method with Strong Convergence. Available at: https://arxiv.org/pdf/1311.0636.pdf

[63] McLean, R. David and Pontiff, Jeffrey, Does Academic Research Destroy Stock Return Predictability? (January 7, 2015). Journal of Finance, Forthcoming. Available at SSRN: https://ssrn.com/abstract=2156623 or http://dx.doi.org/10.2139/ssrn.2156623

[64] M Minsky, SA Papert, L Bottou (1988). Perceptrons: An introduction to computational geometry. Available at: https://books.google.co.za/books?hl=en&lr=&id=PLQ5DwAAQBAJ&oi=fnd&pg=PR5&dq=Minsky+Papert+1969+%E2%80%94+Perceptrons&ots=zyEDwMvmXX&sig=aVXF7DBJKAOxW066S4UCyzhNARw#v=onepage&q=Minsky%20Papert%201969%20%E2%80%94%20Perceptrons&f=false

[65] Murphy, John J. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Penguin, 1999.

[66] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." International Conference on Machine Learning. 2013. Available at: http://proceedings.mlr.press/v28/pascanu13.pdf

[67] Lopez de Prado, Marcos, The Future of Empirical Finance (May 31, 2015). Journal of Portfolio Management, 41(4). Summer 2015. Forthcoming.. Available at SSRN: https://ssrn.com/abstract=2609734 or http://dx.doi.org/10.2139/ssrn.2609734

[68] Povey D., Zhang X., Khudanpur S. Parallel training of Deep Neural Networks with Natural Gradient and Parameter Averaging. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.745.6995&rep=rep1&type=pdf

[69] MarcAurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann LeCun. Efficient learning of sparse representations with an energy-based model. In B. Scholkopf, J. Platt, and T. Hoffman, editors, Advances in Neural Information Processing Systems 19 (NIPS06), pages 11371144. MIT Press, 2007. Aavilable at: http://papers.nips.cc/paper/3112-efficient-learning-of-sparse-representations-with-an-energy-based-model.pdf

[70] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, & J. L. McClelland (Eds.), Parallel distributed processing, vol. 1 (pp. 318362). MIT Press.

[71] Robert L. Schaefer (2012) Subset Selection in Regression, Technometrics, 34:2, 229, DOI: 10.1080/00401706.1992.10484917

[72] J Schmidhuber (2015). Deep learning in neural networks: An overview. Neural Networks Volume 61, January 2015, Pages 85-117. Available at https://www.sciencedirect.com/science/article/pii/S0893608014002135.

[73] Schorfheide, Frank, and Kenneth I. Wolpin. 2012. "On the Use of Holdout Samples for Model Selection." American Economic Review, 102 (3): 477-81.

[74] Jack D. Schwager. Getting Started in Technical Analysis, Page 2 (1999)

[75] Sermanet, Pierre, et al. "Pedestrian detection with unsupervised multi-stage feature learning." Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013. Available at: https://arxiv.org/abs/1212.0142

[76] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In Proceedings of the Twenty-Fourth International Conference on Machine Learning (ICML-07), 2007.

[77] Siegelmann, H. (1992). Theoretical foundations of recurrent neural networks (Ph.D. thesis), New Brunswick Rutgers, The State of New Jersey: Rutgers. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.920&rep=rep1&type=pdf

[78] Skabar, Cloete, Networks, Financial Trading and the Efficient Markets Hypothesis (http://crpit.com/confpapers/CRPITV4Skabar.pdf)

[79] Takeuchi L, Lee Y. Applying Deep Learning to Enhance Momentum Trading Strategies in Stocks. Technical Report, 2013. Available at: http://www.smallake.kr/wp-content/uploads/2017/04/TakeuchiLee-ApplyingDeepLearningToEnhanceMomentumTradingStrategiesInStocks.pdf

[80] Troiano L., Mejuto E., Kriplani P.On Feature Reduction using Deep Learning for Trend Prediction in Finance (11 Apr 2017). Available at: https://arxiv.org/abs/1704.03205.

[81] Tseng P., AN INCREMENTAL GRADIENT(-PROJECTION) METHOD WITH MOMENTUM TERM AND ADAPTIVE STEPSIZE RULE. SIAM J. OPTIM. Vol. 8, No. 2, pp. 506531, May 1998. Available at: https://pdfs.semanticscholar.org/1a29/6a1577478654a54a9f801f93f71b7d853c53.pdf

[82] Vincent P., Larochell H., Lajoie I., Bengio Y., Manzagol P., Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. Journal of Machine Learning Research 11 (2010) 3371-3408. Available at: http://www.jmlr.org/papers/volume11/vincent10a/vincent10a.pdf

[83] Wan Y., Gong X., Si Y. Effect of segmentation on financial time series pattern matching. Applied Soft Computing Volume 38, January 2016, Pages 346-359. Available at: https://www.sciencedirect.com/science/article/pii/S1568494615006341

[84] Linnan Wang, Yi Yang, Renqiang Min, Srimat Chakradhar. Accelerating deep neural network training with inconsistent stochastic gradient descent. Neural Networks Volume 93, September 2017, Pages 219-229. Available at: https://www.sciencedirect.com/science/article/pii/S0893608017301399

[85] Wang, Sida, and Christopher Manning. "Fast dropout training." international conference on machine learning. 2013. Available at: http://proceedings.mlr.press/v28/wang13a.pdf

[86] Van Den Oord, Aaron, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016). Available at: https://arxiv.org/abs/1609.03499

[87] Weiss, S. M, & Kulikowski, C. A. (1991). Computer systems that learn : classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Mateo (Calif.): Kaufmann.

[88] Werbos, P. J. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences (Ph.D. thesis), Harvard University. Available at: https://www.researchgate.net/publication/244947191_1974_Beyond_regression_New_tools_for_predicting_and_analysis_in_the_behavioral_sciences

[89] Paul J. Werbos. Applications of advances in nonlinear sensitivity analysis. System Modeling and Optimization pp 762-770. Available at: https://link.springer.com/chapter/10.1007/BFb0006203

[90] Wu, Huaiqin. "Global stability analysis of a general class of discontinuous neural networks with linear growth activation functions." Information Sciences 179.19 (2009): 3432-3441. Available at: https://ac.els-cdn.com/S0020025509002539/1-s2.0-S0020025509002539-main.pdf?_tid=fc9d5d20-c0fe-4107-947c-6a4d9e472aab&acdnat=1526319168_bc401e01e4ab496b50b8d568dcafb0b3

[91] Yin J., Si Y., Gong Z. Financial Time Series Segmentation Based On Turning Points. Proceedings of 2011 International Conference on System Science and Engineering, Macau, China - June 2011. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5961935

[92] Zeiler M. ADADELTA: An Adaptive Learning Rate Method. Available at: https://arxiv.org/abs/1212.5701

[93] Zinkevich M, Weimer M, Smola A, Li L. Parallelized Stochastic Gradient Descent. Available at: http://papers.nips.cc/paper/4006-parallelized-stochastic-gradient-descent.pdf

[94] Zhao Y., Li J., Yu L. A deep learning ensemble approach for crude oil price forecasting. Available at: https://doi.org/10.1016/j.eneco.2017.05.023

[95] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of the Twenty-First International Conference on Machine Learning (ICML-04), pages 919926, 2004

[96] Zhou B., Hu J. A Dynamic Pattern Recognition Approach Based on Neural Network for Stock Time-Series. 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC 2009). Available at:`https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5393674`