International Stock Market Prediction using Artificial Neural Networks
AC 299r Independent Study

Chang Liu
Advisor: Neil Shephard

7 May, 2017

1 Introduction

According to the Efficient Market Hypothesis, past information can not be used to predict the future prices of a financial asset. A body of literature in financial economics suggests that international equity markets can have cross-market momentum, where one market in a region influences another market in the same region or in a different region. Profitable trading strategies are devised to exploit the predictability of future prices using cross-market momentum. The literature has investigated this factor within a single stock market and across international markets. Within the US market, Lo and Mackinlay (1990) [2] found positive cross-autocorrelation in the CRSP stock data that past weekly returns of small stocks lag those of large stocks. They argue that contrarian profits can not be due to overreaction of investors (implying negative autocorrelation) but significant cross-sectional autocorrelations. With the more recent availability of high-frequency data, Chen et al (2017) [3] designed a hierarchical neural network that uses cross-sectional 5-minute price and volume information of S&P 500 constituents to forecast the next 5-minute direction of price movement of the 100 most liquid constituents. Their method yields higher accuracy rates compared with ARMA-GARCH and other benchmark models. Outside the US market, Burgess (1990) [1] exploited the cointegration relationship between UK FTSE index and a basket of indices in the US and Europe, which operate in a relatively similar time zone, and designed a portfolio of models hybridizing neural network and genetic algorithms that reaped consistent out-of-sample profits. Looking at completely asynchronously traded markets, Leung et al (2016) [5] observed co-integration of daily returns between SPY index and Asia-country ADRs and backtested a pair-trading strategy that earned excess returns. Other studies [4, 7, 8, 9] also created statistical trading systems to exploit cross-market momentum signals in international equity markets.

Previous studies primarily focus on case studies that explore the dynamics of a single or a few selected major country index such as the US, UK, Japan, and China, while ignoring the dynamics with other countries in the same region or other regions. On the other hand, the majority of existing literature primarily uses only past cross-sectional returns as predictors or features up to certain lags. In this paper, we seek to expand the coverage of assets and investigate the price predictability of the regional indices as well as major country indices that cover the world equity markets according to the MSCI world index classification. Our approach is a prediction pipeline of machine learning models including multi-layer feed-forward neural networks to predict returns. In addition to lagged cross-sectional returns, we distill past price and volume information into a number of momentum technical indicators as input features to our models.

2 Methodology

In this section, we describe the data, problem statement, feature engineering, neural network architecture, and forecasting method.

2.1 Data

It is a common issue that spurious cross-autocorrelation can be a result of thinly traded markets, asynchronous trading, or both. To minimize these impacts, we have used large, liquid iShares Exchange-Traded-Funds (ETF) trade data for international equity indices downloaded from Yahoo! Finance. Most of the international ETFs track the MSCI equity regional or country indices and trade in the US market hours. For the US market, we use SPDR S&P 500 ETF data downloaded from Yahoo! Finance. The coverage of ETFs is similar to the MSCI world equity index classification, except that we have both major country indices and the regional indices excluding countries (which usually make up a substantial percentage of the regional index capitalization), in order to better separate the effect of the two. The comprehensive coverage also represents three major market time zones: Asia and Pacific, Europe, and America time zones.

For the purpose of prediction, we break daily returns of a region or country into intraday and overnight returns of day t that are defined as follows:

$$R_{overnight,t} = \frac{O_t - C_{t-1}}{C_{t-1}}$$

$$R_{intraday,t} = \frac{C_t - O_t}{O_t}$$

where C_t, O_t represent the close and open price of day t, respectively.

The returns are driven by fundamentally different drivers, depending on the overlapping time zone. For example, Asia overnight returns will be driven primarily by market sentiments in Asia, while Asia intraday returns primarily by those in the US and other regions. Table 1 gives liquidity information and summarizes the difference between intraday and overnight returns of these ETFs. We will refer to the ETF by the region it represents from now on.

	Asia ex Japan (AAXJ)	**China (FXI)**	**Pacific ex Japan (EPP)**	**Japan (EWJ)**	**Eurozone (EZU)**	**UK (EWU)**	**S&P 500 (SPY)**	**Canada (EWC)**	**Latam (ILF)**
Intraday Mean	-0.02%	0.02%	0.03%	0.16%	0.04%	0.03%	0.003%	0.01%	0.004%
Intraday Volatility	1.20%	1.3%	1.0%	0.9%	1.1%	1.1%	1.0%	1.20%	1.5%
Overnight Mean	0.05%	0.01%	-0.01%	-0.01%	-0.04%	0.01%	0.03%	0.02%	0.03%
Overnight Volatility	1.20%	2.0%	1.6%	1.3%	1.5%	2.0%	0.6%	1.1%	1.9%
Avg. Daily Volume* (Million shares)	1.0	19.5	0.6	7.8	5.7	2.2	87	2.4	1.7
Net Asset* ($Billion)	3.2	3.1	2.9	15	11	2.6	237	1.5	1.2

Table 1: Summary statistics and liquidity information of regional iShare ETFs and SPDR S&P 500 ETF during 2004-10-08 - 2017-4-7. *Average daily volumes are calculated from the last 250 days (as of 4-7-2017) of volume data from Yahoo! Finance. Net asset values (NAV) of iShares ETFs are reported from iShares' official site, and NAV for SPY from Yahoo! Finance.

Outside North America, we observe that the overnight volatility is higher for all regions , in which case the fundamental drivers of overnight and intraday returns are different. Overnight index prices should respond to local market sentiments outside North America while the intraday prices respond to the sentiments within North America, depending how much they overlap with the North American market hours. For this reason, the overnight and intraday returns should be separated for the prediction task.

2.2 Problem Statement

Given the past cross-sectional price information of equity indices of day t across regions, our goal is to predict the next-day overnight or intraday return of one index:

$$\hat{R}_{overnight,t+1} = f(R_{intraday,t}, C_t, \text{Technical indicators}_t)$$

$$\hat{R}_{intraday,t+1} = f(R_{overnight,t+1}, O_{t+1}, \text{Technical indicators}_t)$$

where R_t denotes cross-sectional intraday or overnight returns of day t; O_t, C_t denotes the cross-sectional last available open or close price of day t. On any day t, overnight returns come first, then come intraday and daily returns. Technical indicators of day t are calculated based on last available close price of the index we are predicting (see the feature engineering section). f is a model that generates a prediction given the inputs.

2.3 Feature engineering

Technical Indicators For each target variable of a regional ETF, we will transform its lagged price and volume information into a set of technical indicators. The following momentum indicators are selected from the literature [6, 3] that are reported to have the more predictive power than other technical indicators. They are all based on price and volume information of or before day t:

- Moving Averages (MA) over 5 days, $MA_t(5) = \frac{1}{5}\sum_{i=0}^{4} C_{t-i}$
- AD: number of advancing stocks of day t minus that of declining stocks
- ADV: volume of advancing stocks of day t minus that of declining stocks
- Past 5-day momentum, $C_t - C_{t-4}$
- Past 5-day stochastic oscillator, $\%K = \frac{C_t - L_5}{H_5 - L_5}$
- Past 5-day stochastic oscillator, $\%D = \frac{1}{5}\sum_{i=0}^{4} \%K_{t-i}$
- Past 5-day rate of change, $(\frac{C_t}{C_{t-5}})$
- Larry William's %R in 5 days, $\%R = \frac{H_5 - C_t}{H_5 - L_5}$
- Disparity in 5 days, $\frac{C_t}{MA_t(5)}$
- Past k-day volatility of daily return
- Past k-day log return, $\log C_t - \log C_{t-k}$
- Exponential Moving Averages over k periods, $EMA_t(k) = (C_t - EMA_{t-1}(k))\frac{2}{k+1} + EMA_{t-1}(k)$
- Day of week, 0,1,2...6

where O_t, H_t, L_t, C_t represent the open, high, low, and close price of day t, respectively; H_5, L_5 are the highest high price and lowest low price in the past 5 days. The parameter $k = 1, 2, 3, 4, 5$ days except for exponential moving averages where $k = 1, 2, 3..., 10, 15, 20, 25$ days.

Feature Selection With the above technical indicators, and 3 types of (lagged intraday/overnight/daily) cross-sectional returns, plus the last available price of the target region, we have more than 60 input features with much redundant information. From practice, it adds noise that the neural network confuses with signals so that the results are not satisfactory. We use Random Forest Regressors with Mean Square Error as the criterion for selecting the top 20 optimal features in the total in-sample period (i.e. the samples that model is allowed to see in during training and validation).

Importance Rank	Feature Importance	Features
1	0.120	Canada open price
2	0.108	ADV
3	0.082	UK 5-day volatility
4	0.068	UK 5-day momentum
5	0.033	China open price
6	0.021	Latin Am. last intraday return
7	0.020	UK last 2-da return
8	0.019	Canada last daily return
9	0.018	UK last 9-day return
10	0.016	Latin Am. open price

Table 2: Top 10 feature importance scores of features selected by Random Forest Regressor in Python's sklearn package by 1000 estimators. The target variable to predict is intraday return of UK. Importance score of a feature for a decision tree is defined as the normalized total reduction of mean squared error brought by that feature. For a forest of trees, the final score is the average scores of individual trees.

As an example, we only report the top 10 feature selected to predict intraday return for UK in Table 2 due to space constraint. Depending on intraday or overnight returns as the target variable, the most commonly selected features are the last available cross-sectional price, returns, and Price Trend Indicators, as well as past returns and volatility of that region to be predicted. The selected features verify the hypothesis that some of the most relevant features will depend on the type of returns we want to predict. For example, the cross-sectional overnight returns and open prices are relevant for predicting intraday returns of the same day, whereas the cross-sectional intraday returns and close prices are relevant for predicting overnight returns of the next day.

2.4 Neural Network Architecture

A relatively simple architecture is chosen for the following reasons: 1) with limited training samples we have (1577 for Asia ex Japan, 2450 for the rest) for each rolling window (see the forecast method below), the number of parameters should be always less than of the training set to avoid over-fitting; 2) the first hidden layers should have units at least as many as input layer and 2 hidden layers can help to capture complexity in the inputs; 3) from the literature and from practice, larger networks (e.g. with more than 2 hidden layers and more hidden units) tend to over-fit quickly, which suggests that a network is able to capture the dependence structure inherent in the data. Therefore, several regularization techniques are used to prevent over-fitting of individual models.

- A 2-layer fully-connected network: 20 input units - 24 hidden units - 3 hidden units - 1 output unit
- Mean Square Loss function
- ADAM optimizer
- ReLu activation for all nodes except linear activation for output node
- L2 penalty; max norm constraints of weights; dropout layers

- Monitor training process – early stopping if validation not improving
- Normalize inputs by z-score transformation to improve the convergence and stability

2.5 Ensemble Forecast Method

We employ a walk-forward method that is suited for back-testing financial time series. For each rolling window, we train the neural network on 1000 bootstrapped samples and compute the accuracy on validation set. We then select the top 50% neural networks according to the directional accuracy rate (which we define in the evaluation section) to form a committee and predict on the test set. The average of the committee predictions will be the final prediction of the ensemble. Due to the computational intensity, each ensemble is used to next 200-day forecast and will not be retrained until a good portion (10 20%) of the training set is updated. The main benefit of the model averaging technique is to reduce the instability of a single neural network due to 1) random initialization of the parameters, and 2) performance of different local optima on the test set, in addition to the variability of bootstrap sampling.

In all regions except Asia ex Japan, for each rolling window, the training set has 2450 samples, validation set 272 samples, and test set 200 samples; we use 4 windows, totaling 800 test samples. However, since Asia ex Japan has only 2152 observations in total (with an inception date in 2008), we use only 2 windows. Therefore, to preserve training samples, we exclude Asia ex Japan as a feature when we predict other regions' returns. Nevertheless, China, which makes up a large percentage of the Asia index, is included in all experiments.

The entire prediction pipeline is summarized in Figure 1.

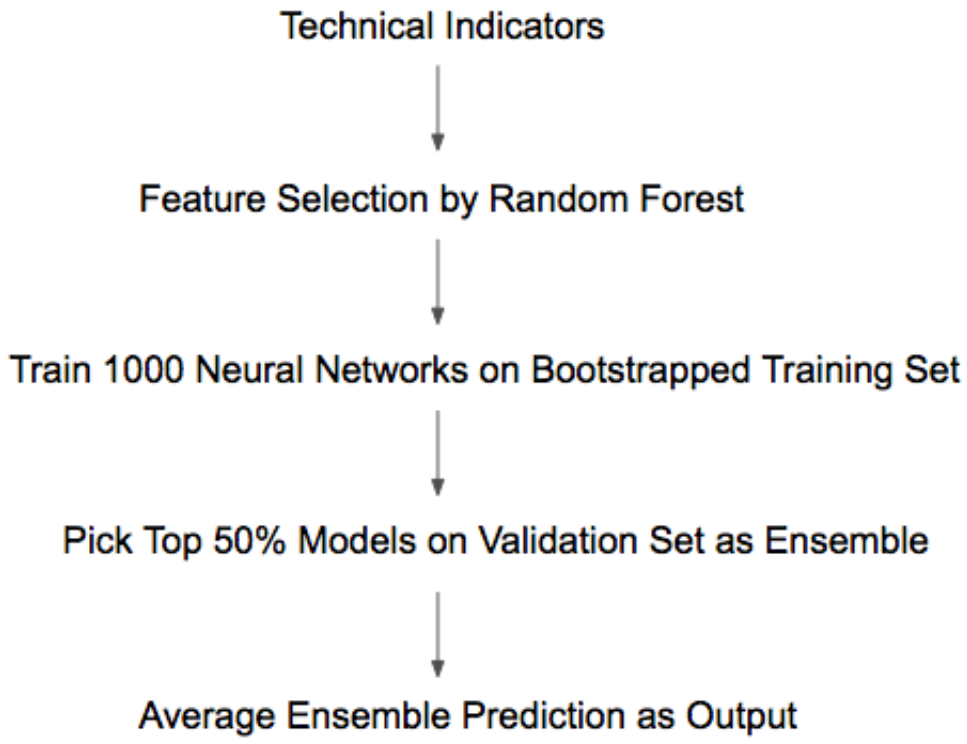

Figure 1: Prediction Pipeline

In this study, regularized linear methods Ridge and Lasso are used to compare with our ensemble neural network. We apply the same two evaluation criteria to evaluate the final predictions.

2.6 Evaluation Metrics

Accuracy To measure how close our predictions are to the true returns, we use directional accuracy, the percentage of times of our predictions are in the same direction (i.e., positive or negative returns) as the true returns. This metric is only used to measure the validation accuracy for model selection of the ensemble neural net. As for the final prediction, we want to measure its ability to predict large values and the model performance as a trading strategy. For an incorrect prediction in terms of direction, a loss is incurred; otherwise, a profit is gained. Thus we use a new accuracy metric according to Chen et al (2017) on the test set.

Let R_t be the true return at time t of an asset and $\hat{R}_t$ our prediction based on a model in our test set. Then the model's real return is $R'_t = sign(\hat{R}_t R_t)|R_t|$. We then threshold our predictions at the p percentile value $\hat{R}^{(p)}$ in the test set and obtain adjusted return $R^*_t = 1\{|\hat{R}_t| > \hat{R}^{(p)}\}R'_t$ where $1\{\cdot\}$ is an indicator function. The adjusted returns measures our performance if we only transacted $100 \cdot (1-p)\%$ of the largest absolute predictions. Lastly, we use the adjusted returns to measure accuracy:

$$Accuracy^* = \frac{\sum_t 1\{R^*_t > 0\}}{\sum_t 1\{R^*_t \neq 0\}}$$

Sharpe Ratio To evaluate the risk-reward ratio for our trading models, Sharpe ratio is commonly used a standard and calculate Sharpe ratio based on the adjusted returns for our models accordingly:

$$\text{Sharpe ratio} = \frac{\bar{R}^*}{\sigma^*}$$

where $\bar{R}^*$ is the mean and σ^* the standard deviation of the adjusted return.

3 Experimental Results

We test our models from 2015-09-08 to 2017-04-07 over 400 market days for Asia ex Japan which fewer data, and from 2015-02-05 to 2017-04-07 over 800 market days for all other regions. Results are shown in Figures 2, 3, 4, and 5. For comparison, a baseline is calculated as the fraction of positive returns in the test set, which does not vary with the proportion of transaction $100 \cdot (1-p)\%$.

We observe several patterns:

- For overnight returns, on average, all the models perform slightly better than the baseline, except for Asia ex Japan. For intraday returns, on average, the models perform on par with the baselines. This suggests that there is predictive information in the features of overnight returns across all regions. One reason might be that the trading outside the exchange hours does not happen as often as within them, so that the market does not react as timely as information arrives. In addition, in a less efficient after-hour market, the bid-ask spreads are probably higher in trading; therefore, to access the execution feasibility of the trading strategy, it would be helpful to compute the break-even trading costs based on our model returns to compare with real-world trading costs.

- A prediction accuracy higher than the baseline is not always equivalent to a good Sharpe ratio or feasible strategy; nor does a feasible strategy require a high prediction accuracy. For example, our ensemble neural net has below baseline accuracy in predicting overnight returns in Asia ex Japan. Yet, its Sharpe ratio is significantly above 1.0 and outperforms the benchmark models. This suggests a small proportion of good predictions that turn out to earn large returns can outweigh a large proportion of bad predictions with smaller losses. On the other hand, the benchmark models with higher-than-baseline accuracy, for example in predicting Canada overnight returns, can have zero or even negative

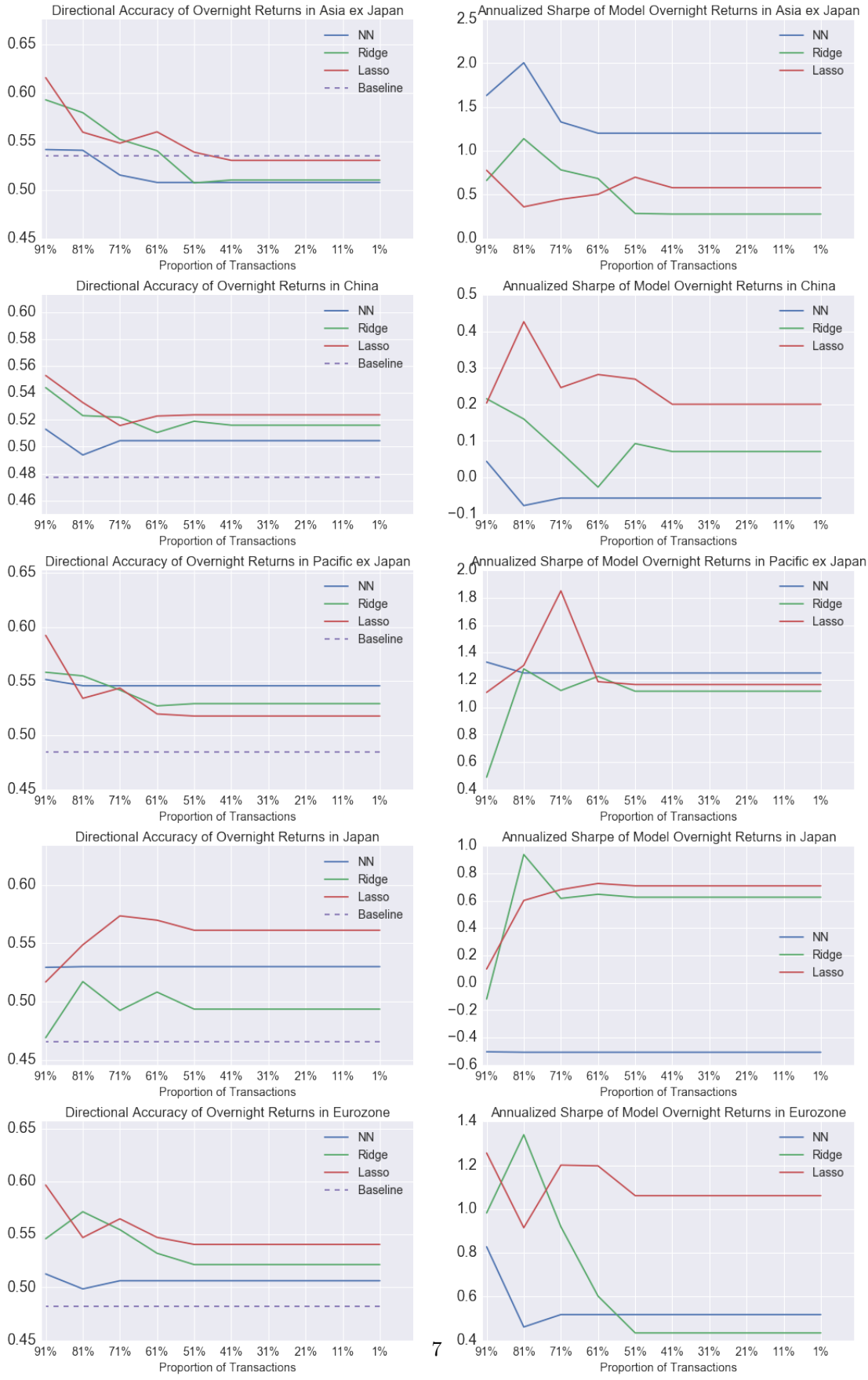

Figure 2: Accuracy and annualized Sharpe ratios of overnight returns. The proportion of transactions is set to 1%,11%,...91%.

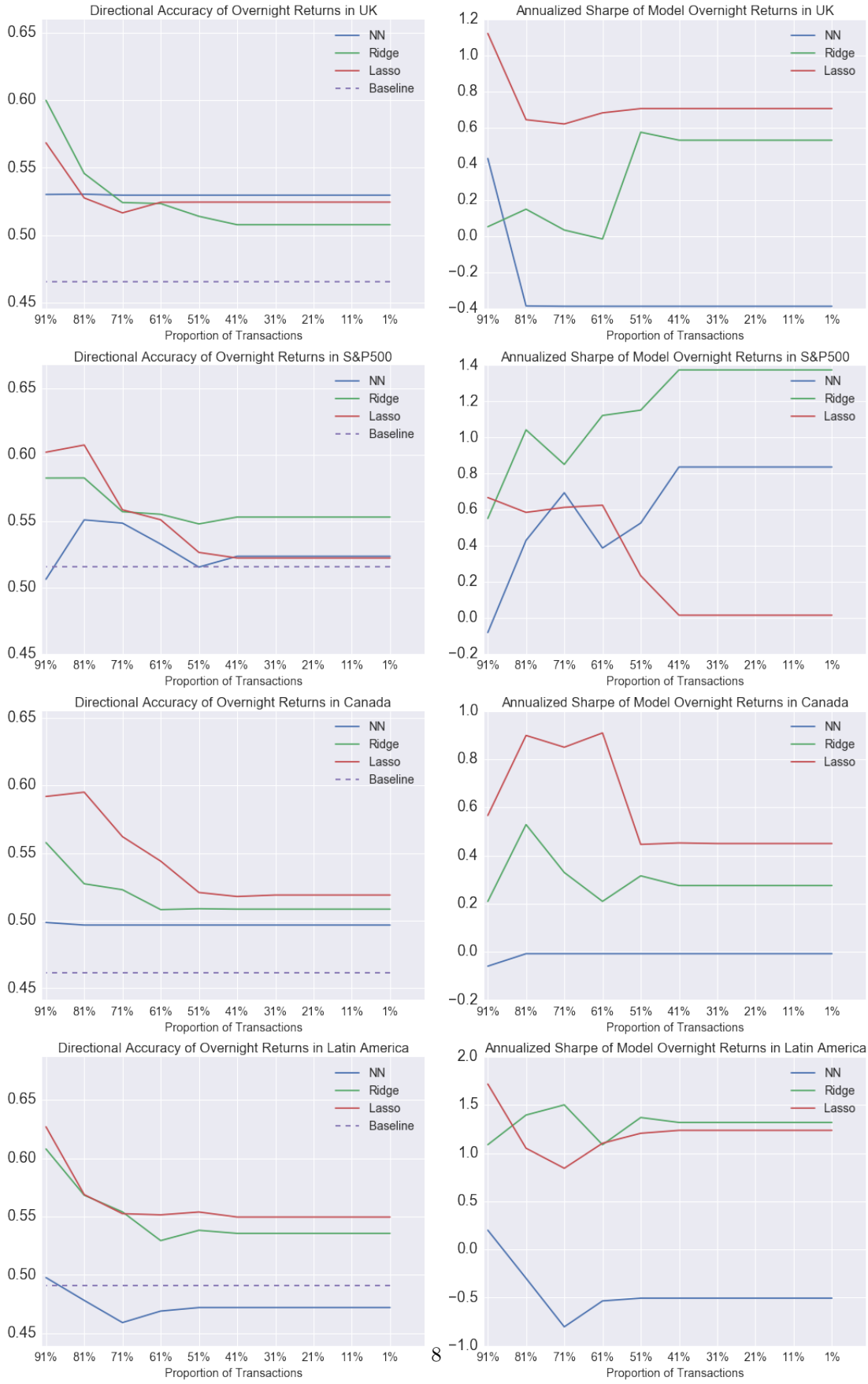

Figure 3: Accuracy and annualized Sharpe ratios of overnight returns. The proportion of transactions is set to 1%,11%,...91%.

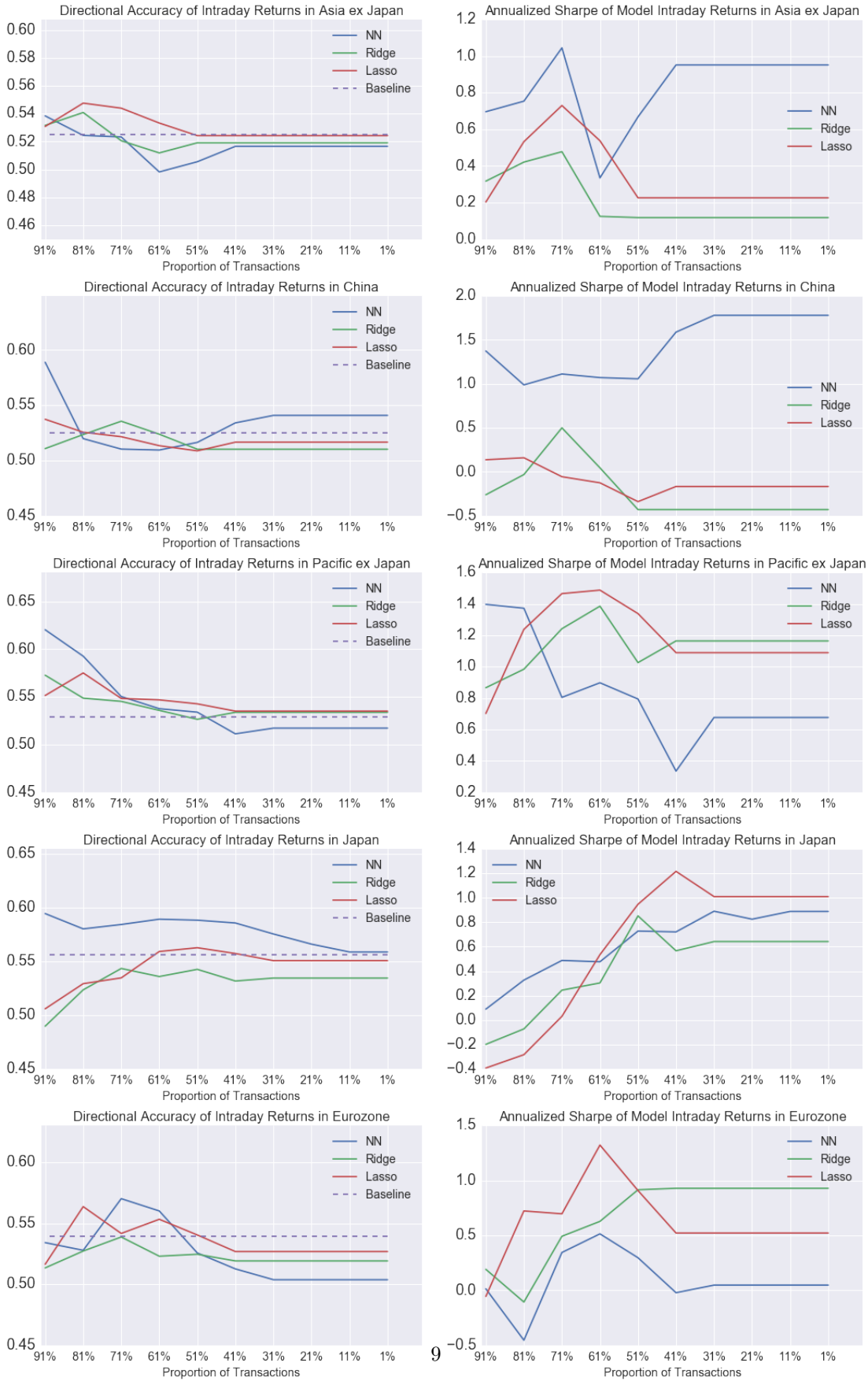

Figure 4: Accuracy and annualized Sharpe ratios of intraday returns. The proportion of transactions is set to 1%,11%,...91%.

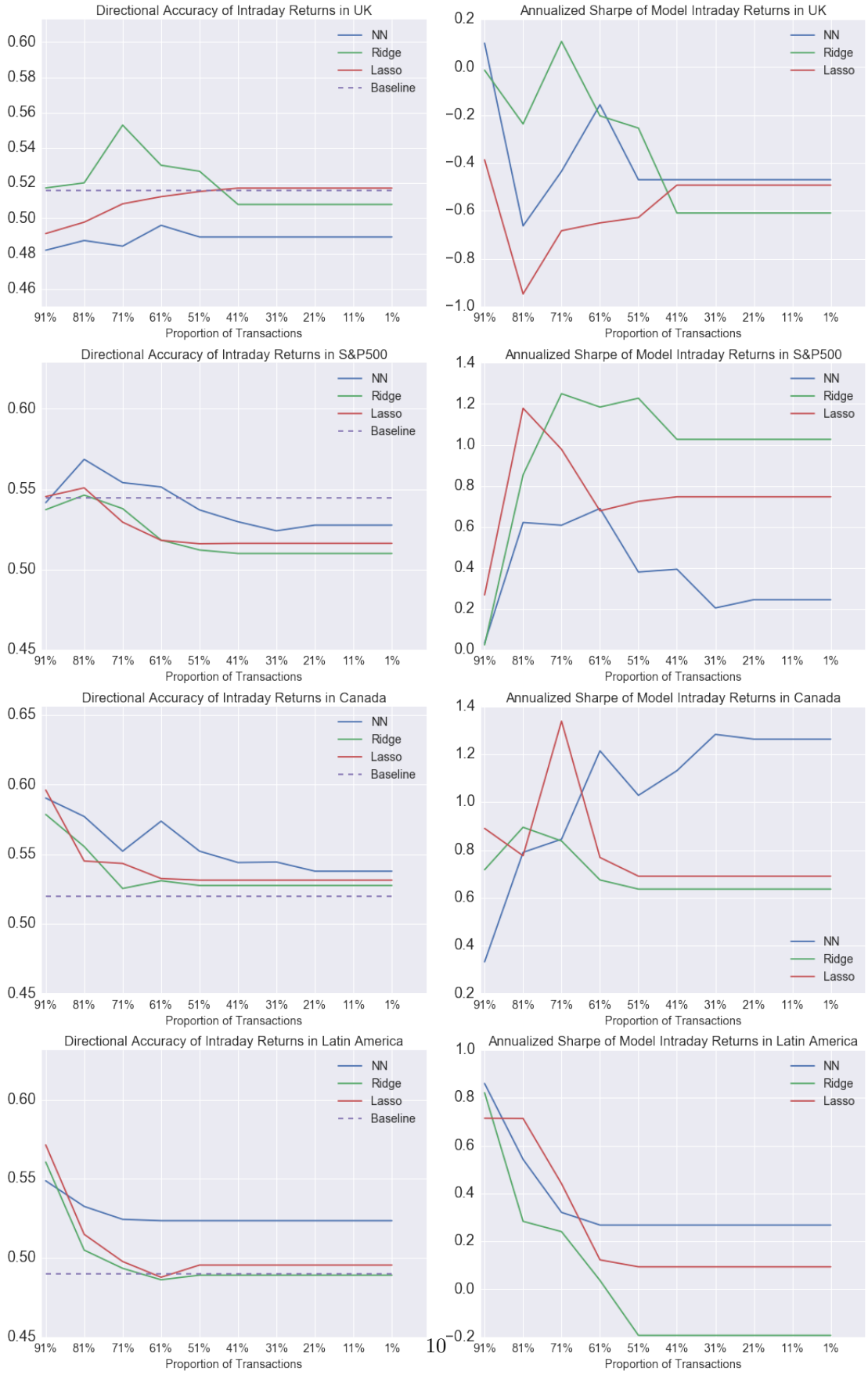

Figure 5: Accuracy and annualized Sharpe ratios of intraday returns. The proportion of transactions is set to 1%,11%,...91%.

Sharpe ratios. Therefore, it is important to observe consistency in both accuracy and Sharpe ratio curves to confirm performance stability of a model.

- In terms of Sharpe ratio, our ensemble neural net outperforms the benchmarks in predicting intraday returns in China and overnight returns in Asia ex Japan. In all other cases, its performance is on average at par or even worse than the benchmark models. The sub-par performance highlights one potential issue with the neural net that the ensemble after model averaging can still have variation that cause performance instability, because the neural nets with powerful learning capacity can mistake noises as true patterns in the data. However, given the large set of neural network hyper-parameters (precise architecture, learning rate, convergence criteria, regularization, etc), we have minimal amount of fine-tuning the algorithm, striking a balance between an overly-tuned model and simplicity. It is certainly possible to select optimal parameters using other machine learning algorithms such as particle swarm optimization and genetic algorithm. Another explanation for the performance instability is that we have given very similar input features for all experiments; however, returns of different types and regions may have different dynamics and require specialized features.

4 Concluding Remarks

In this study, we propose to predict the intraday and overnight returns of international equity indices that cover most of the world equity markets. To minimize spurious correlations due to low liquid or asynchronous trading, we use large, liquid Exchange-Traded-Funds trade data that closely track MSCI regional indexes or the S&P 500. We aim to predict the next intraday or overnight returns of an ETF based on the last available price and volume information. We computed the technical indicators of the price and volume information and selected the most important features as inputs. Then we train an ensemble of neural networks and benchmark it with Lasso and Ridge, in terms of adjusted accuracy rates and Sharpe ratio. The results show that in all regions except for Asia ex Japan, all the models outperform the baseline in predicting overnight returns in terms of directional accuracy, which suggests predictability of cross-market momentum. However, quantifying to what extent the result can be attributed to thin markets for overnight returns is outside the scope of this study.

In addition, our ensemble of 2-layer models is on average on par with the regularized linear models as benchmarks. This highlight some issues regarding application of neural networks to financial time series: 1) much of the nuisance the neural net learns (better than a simple linear model) in financial time series can be just noise, which is random or inherently unpredictable. 2) to fine-tuning an ensemble to beat linear methods, the computational cost and risks of over-fitting are high; and 3) to improve the overall performance, no matter what method to use, one could focus on engineering the appropriate features in addition to tuning the model.

5 Acknowledgement

I would like to express my sincere gratitude to my advisor Professor Neil Shephard for providing stimulating thoughts throughout the study. My thanks also goes to Professor David Parkes who offers practical comments in machine learning.

References

[1] A. N. Burgess. *Modelling relationships between international equity markets using computational intelligence*, Knowledge-Based Intelligent Electronic Systems, 1998. Proceedings KES '98. 1998 Second International Conference on, Adelaide, SA, 1998, pp. 13-22 vol.3. doi: 10.1109/KES.1998.725946

[2] Andrew W. Lo, A. Craig MacKinlay. *When Are Contrarian Profits Due to Stock Market Overreaction?*. Rev Financ Stud 1990; 3 (2): 175-205. doi: 10.1093/rfs/3.2.175

[3] Hao Chen, Keli Xiao, Jinwen Sun, and Song Wu. 2017. *A Double-Layer Neural Network Framework for High-Frequency Forecasting*. ACM Trans. Manage. Inf. Syst. 7, 4, Article 11 (January 2017), 17 pages. DOI: https://doi.org/10.1145/3021380

[4] Hargreaves, Carol; Yi Hao. *Prediction of Stock Performance Using Analytical Techniques*. Journal of Emerging Technologies in Web Intelligence. May 2013, Vol. 5 Issue 2, p136-142. 7p.

[5] Leung, Tim and Kang, Jamie Juhee. *Asynchronous ADRs: Overnight vs Intraday Returns and Trading Strategies* (October 23, 2016). Studies in Economics Finance, 2016, Forthcoming. Available at SSRN: https://ssrn.com/abstract=2858048

[6] Qiu M, Song Y. *Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model*. PLoS ONE 11(5): e0155133 (2016). https://doi.org/10.1371/journal.pone.0155133

[7] Selmi, N., Chaabene, S Hachicha, N. *Forecasting returns on a stock market using Artificial Neural Networks and GARCH family models: Evidence of stock market S P 500.*Decision Science Letters, 4(2), 203-210 (2015).

[8] Y. Yetis, H. Kaplan and M. Jamshidi. *Stock market prediction by using artificial neural network*. 2014 World Automation Congress (WAC), Waikoloa, HI, 2014, pp. 718-722. doi: 10.1109/WAC.2014.6936118

[9] Yanshan Wang. 2014. *Stock price direction prediction by directly using prices data: an empirical study on the KOSPI and HSI*. Int. J. Bus. Intell. Data Min. 9, 2 (October 2014), 145-160. DOI=http://dx.doi.org/10.1504/IJBIDM.2014.065091