A Practical Machine Learning Approach for Dynamic Stock Recommendation

Hongyang (Bruce) Yang[1], Xiao-Yang Liu[2], Qingwei Wu[2]
[1]Dept. of Statistics, Columbia University
[2]Dept. of Electrical Engineering, Columbia University
Email: {hongyang.yang, XL2427, QW2208}@columbia.edu,

***Abstract*—Stock recommendation is vital to investment companies and investors. However, no single stock selection strategy will always win while analysts may not have enough time to check all S&P 500 stocks (the Standard & Poor's 500). In this paper, we propose a practical scheme that recommends stocks from S&P 500 using machine learning. Our basic idea is to buy and hold the top 20% stocks dynamically. First, we select representative stock indicators with good explanatory power. Secondly, we take five frequently used machine learning methods, including linear regression, ridge regression, stepwise regression, random forest and generalized boosted regression, to model stock indicators and quarterly log-return in a rolling window. Thirdly, we choose the model with the lowest Mean Square Error in each period to rank stocks. Finally, we test the selected stocks by conducting portfolio allocation methods such as equally weighted, mean-variance, and minimum-variance. Our empirical results show that the proposed scheme outperforms the long-only strategy on the S&P 500 index in terms of Sharpe ratio and cumulative returns.**

***Index Terms*—Stock recommendation, fundamental value investing, machine learning, model selection, risk management**

I. Introduction

Earning reports play a key role in stock recommendation. Analysts use company earning reports to do stock buy-and-sell recommendation. Future earnings estimates are important factors to value a firm. Earnings forecasts are based on analysts' estimation of company growth and profitability. To predict earnings, most analysts build financial models that estimate prospective revenues and costs. However, it could be very difficult for analysts to accurately estimate earnings. Many researchers are trying to build a robust model to predict earnings. For example, earnings generated by the cross-sectional model are considered superior to analysts forecasts to estimate the implied costs of capital (ICC), which play a key role in firm valuation [1]. Moreover, regression-based models [2] can be used to predict scaled and un-scaled net income [3]. Numerous recent papers consider using deep learning algorithms to model stock market data [4]. Deep neural networks models also can be trained to predict future fundamentals such as book-to-market ratio, and as a result, investors can use predicted fundamentals to rank current stocks [5].

There are two traditional approaches. The first approach is selecting stocks based on a preset criteria such as price to earnings (P/E) ratio [6]. Stocks are ranked by P/E ratios using historical data. Then, a portfolio will contain stocks with lowest P/E ratios. This approach is unsatisfactory in practical situations since the selection with P/E ratio only is unstable (e.g., selecting top quantile stocks may result in a less predictive power.) The second approach jointly uses several criterion to rank stocks, such as P/E ratio, price to sales (P/S) ratio, price/earnings to growth (PEG) ratio, etc. However, this approach does not take the correlations among different predictor factors into consideration. Consequently, weights of these factors are assigned relatively subjective, which increases the risk.

Value investing has been widely used today by investors and portfolio managers. Graham first comes up with the concept of an intrinsic value for a stock that is independent of the market [7]. He emphasizes the importance of an intrinsic value that is reflected by a companys market size, assets level, dividends, financial strength, earnings stability, earnings growth. Focusing on this value, he believes would prevent an investor from misjudgment and misinterpretation during a bull or bear market. In the long run, we expect stock prices should eventually be regression towards the company's intrinsic value. Many fundamental financial ratios such as P/E ratio, earnings per share (EPS), return on equity (ROE), profit margin and quick ratio indicate overall profitability, stability, operating efficiency, capital structure, ability of generating future cash flows and other valuable information of the corresponding companies. Thus, these financial ratios could reflect a companys intrinsic value and should have predictive power on the future performance [8]. Additionally, financial ratios provide normalization so that all companies would have the same scale of data and thus, those with large capital will have equal influence.

In this paper, we propose a novel scheme that predict stock's future price return based on earnings factors by machine learning. We use five machine learning algorithms (linear regression, random forest, ridge, stepwise regression, and generalized boosting regression) to assign weights to each factor dynamically, and select top 20% stocks each quarter based on the ranking of predicted returns generated by the best performing algorithm over the past training periods before each re-balancing day on a rolling basis. The five models also have high explanatory power. By using the lowest MSE to choose the best model, we provide reliability to business decision, thus increase the security to financial

investments. After we select stocks for our portfolio on each re-balancing day, we test asset allocation methodologies: mean-variance, min-variance, and equally-weighted allocation on selected stocks using in sample data (1990-2007). Risk management is involved by the method of maximum Sharpe ratio used in portfolio allocation methodologies. We are aiming to balance the expected return and the standard deviation of the portfolio to achieve the best risk to reward. Finally, we compare the P&L of our strategy with S&P 500 index [1], all three portfolio allocation methods outperform the market, summarize the competency of our strategy.

This paper proceeds as follows. Section II describes the rolling window, trading time, the data, and also presents the methodology and implementation of our scheme. Section III contains the portfolio allocation methods, risk management and transaction cost. Section IV presents the performance and Section V concludes the paper.

II. Proposed Stock Recommendation Scheme

A. Rolling Window Based Data Separation

Rolling windows can be employed to divide data for multiple purposes (i.e., training and testing). Rolling windows for training ranges from 16-quarter (4-year) to a maximum of 40-quarter (10-year). This training rolling window is followed by an one-year window for testing and we trade according to the test results. The training-testing-trading cycle of our strategy can be summarized by Fig. 1. We also extend the trade date by two months lag beyond the standard quarter end date in case some companies have a non-standard quarter end date, e.g. Apple released its earnings report on 2010/07/20 for the second quarter of year 2010. Thus for the quarter between 04/01 and 06/30, our trade date is adjusted to 09/01 (same method for other three quarters).

B. Data Preprocessing

The data for this project is mainly taken from Compustat database accessed through Wharton Research Data Services (WRDS) [9]. The dataset used here consists of the data over the period of 27 years (from 06/01/1990 to 06/01/2017). We use all historical S&P 500 component stocks (about 1142 stocks) as the S&P 500 pool are updated quarterly. The adjusted close price goes on a daily basis (trading days) and generates 6,438,964 observations. The fundamental data goes on a quarterly basis and generates 91,216 observations. In addition, we delete outlier records that indicate a release date (rdq) after the trade date, which include about 0.84% of the dataset. We assure that on our trade date, 99% of the companies have their earnings reports ready to be used. In order to preserve an out-of-sample period sufficiently long for back-testing the relationship, the dataset has been divided into three periods in Fig. 2.

To build the dataset for training, we select top twenty most popular financial ratios in table I [2] and calculated these factors from the fundamental raw data from the WRDS. Also, in order to build a sector-neutral portfolio, we split the dataset by the Global Industry Classification Standard (GICS) sectors. We handle missing data separately by sector: if one factor has more than 5% missing data, we delete this factor; if a certain stock generates the most missing data, we delete this stock. In this way, we've removed 46 stocks and the overall missing data is reduced to less than 7% of each sector. Finally, we delete this 7% missing data.

TABLE I
20 Financial Indicators

Revenue Growth	Price to cash flow ratio
Earnings per share (EPS)	Cash ratio
Return on asset (ROA)	Enterprise multiple
Return on equity (ROE)	Enterprise value/cash flow from operations
Price to earnings (P/E) ratio	Long term debt to total assets
Price to sales (P/S) ratio	Working capital ratio
Net profit margin	Debt to equity ratio
Gross profit margin	Quick ratio
Operating margin	Days sales of inventory
Price to book (P/B) ratio	Days payable of outstanding

C. Methodology

Our goal is to predict S&P 500 forward quarter log-return r^{qtr}_{T+f} given predictors X_T constructed from historical data of the twenty financial factors over a particular quarter T and S&P 500 horizon f. At a given time T of the financial horizon, the 1-quarter forward log-returns of a certain stock price S are defined as:

$$r^{qtr}_{T+f,i} = \ln(S_{T+f,i}/S_{T,i}), \qquad i = 1, ..., n_T, \quad (1)$$

where n_T is the companies whose stock price and earnings factors are available at time T.

A general estimator is the ordinary least square:

$$r^{qtr}_{T+f,i} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{T,i,j} + \epsilon, \qquad j = 1, ..., 20, \quad (2)$$

where j is the number of the twenty financial ratios, p is the total factors we used in the model, β_0 is the intercept of the model, X_j corresponds to the jth predictor variable of the model, β_j is the coefficients of the predictor variable and ϵ is the random error with expectation 0 and variance σ^2. Moreover, regularized linear OLS estimators have a higher accuracy in many aspects [10]. We need to use multiple regression estimators to increase accuracy. [3] has summarized the prediction rule and estimator selection rule for using multiple estimators:

$$r^{qtr}_{T+f,i} | X_{T,i,j}, \theta, \qquad i = 1, ...n_T, \ j = 1, ..., 20, \quad (3)$$

$$r^{qtr}_{t+f,i} = g_\theta(X_{t,i,j}) + \epsilon, \quad t = T, ..., T-h, \ i = 1, ..., n_t, \quad (4)$$

where h is the historical estimation period, $g_\theta(X_{t,i,j})$ is used to estimate θ through historical regressions. It is noticeable to point out that (2) is the basic estimator of (4).

We pick five models for g_θ: linear regression, forward and backward stepwise regression under Akaike information criterion (AIC), regularized linear OLS estimator ridge

[1] The Standard Poor's 500 is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ.

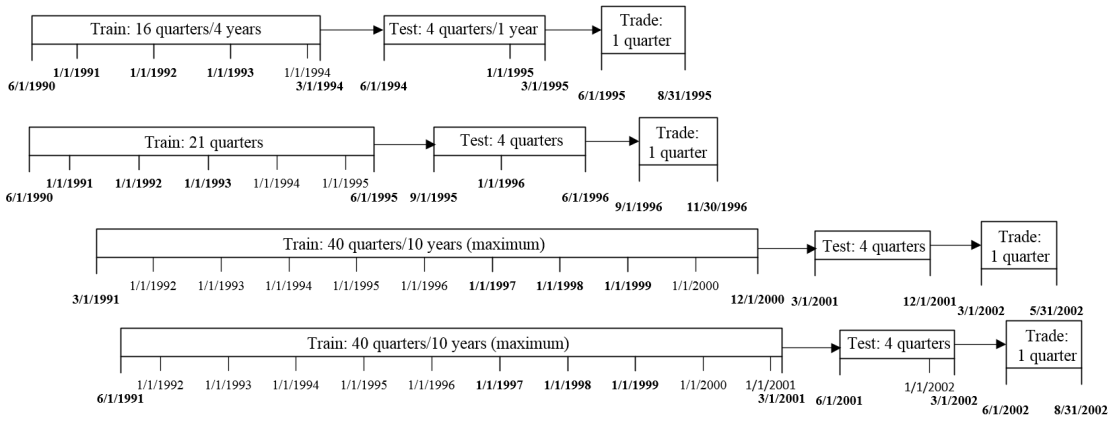

Fig. 1. Rolling Window Based Data Seperation: Training rolling window is followed by a testing rolling window. There is a two-month delay after the end of the training rolling window.

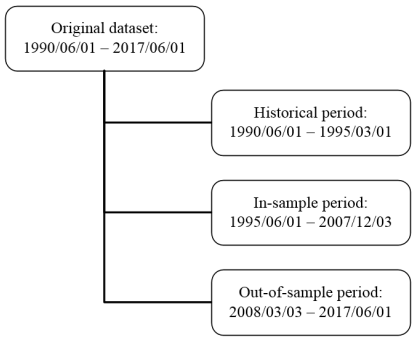

Fig. 2. Dataset Division

regression, tree based nonlinear model random forest and generalized boosted regression model (GBM) using gaussian distribution which implements AdaBoost algorithm and Friedman's gradient boosting machine. All of the algorithms are facilitated by standard R packages [11]. For linear regression and stepwise regression we use lm and $step$ and for ridge we use $glmnet$ and *MASS* [12], [13]. For random forest we use $randomForest$ [14]. For gbm we use gbm [15]. The reason of using these five models is that we need feature selection methods to remove undesirable features, thus reducing the overfitting issues, improving model accuracy and expediting the training procedure. We also have a white-box model that we can observe every single factor with its coefficients in our model.

Mean Squared Error (MSE) [3] is used as the metric for our evaluation.

D. Implementation

Our implementation can be summarized as the following four steps:

Step 1. Train and test the model to get the MSE for each of the five models. Our current methodology basically selects the minimum MSE. We assign 1 to the selected model and 0 to other models.

Step 2. Choose the model that has the lowest MSE in that certain period. For example, in Table II, we choose

TABLE II
MODEL ERROR AND SELECTED MODEL FOR SECTOR 10, ENERGY

trading date	MSE linear	MSE RF	MSE ridge	MSE step	MSE gbm
19950601	0.02238	0.02180	0.02161	0.02205	0.02443
19950901	0.01908	0.01828	0.01870	0.01841	0.02098
19951201	0.01852	0.01641	0.01820	0.01855	0.01996
19960301	0.02040	0.01822	0.01981	0.01879	0.02192
19960603	0.02442	0.01885	0.02394	0.02340	0.02210

TABLE III
PREDICTED RETURN ON TRADE DATE: 1995/06/01 SECTOR 10, ENERGY

	Linear return	RF return	Ridge return	Step return	GBM return
WMB	10.42%	5.12%	9.24%	9.01%	3.51%
OKE	7.55%	4.12%	7.42%	8.53%	2.56%
RRC	4.16%	7.01%	3.74%	4.84%	1.83%
PXD	4.63%	0.96%	3.66%	3.91%	0.23%
VLO	3.48%	2.99%	3.47%	4.04%	2.56%
EQT	2.47%	4.78%	2.34%	2.36%	1.83%
HES	1.80%	4.30%	1.61%	1.81%	0.38%
BHI	1.33%	−0.70%	1.15%	1.99%	−0.27%
MUR	1.11%	1.07%	1.01%	0.49%	0.38%
NE	1.16%	−6.33%	0.94%	0.85%	−2.21%

Ridge regression as our model to select stocks on Jun. 1^{st}, 1995. We choose Random Forest as our model to select stocks on Sept. 1^{st}, 1995.

Step 3. Use the predicted return in the selected model to pick up top 20% stocks from each sector. We predict next quarter return (predicted y) using current information (test Xs) based on the trained model.

In this example, in Table III we use ridge predicted return to pick stocks for the trade period 1995/06/01, the selected top 20% stocks are: WMB, OKE, RRC, PXD, VLO, EQT, HES, BHI, MUR, and NE. We then trade these stocks during the period between 1995/06/01 and 1995/09/01. In the second trade period of 1995/09/01, in Table IV we use random forest to pick the stocks, the top 20% stocks are: CHK, SFS.1, WMB, RRC, VLO, BJS.1, MDR, PZE.1, HP, and CVX. We then trade these stocks from 1995/09/01 to 1995/12/01. As for the stocks owned at previous quarters, such as WMB, RRC, and VLO, we just need to use the portfolio weights to adjust their shares.

Step 4. We check on the corresponding models features and its coefficients or importance level to ensure that there

TABLE IV
PREDICTED RETURN ON TRADE DATE: 1995/09/01 SECTOR 10, ENERGY

	Linear return	RF return	Ridge return	Step return	GBM return
CHK	0.97%	12.45%	2.17%	4.86%	0.03%
SFS.1	4.69%	7.35%	4.27%	4.49%	0.78%
WMB	5.87%	6.37%	5.15%	5.46%	0.77%
RRC	1.78%	6.13%	1.52%	1.76%	0.09%
VLO	2.50%	4.83%	2.65%	3.01%	0.77%
BJS.1	5.36%	4.23%	5.28%	6.38%	−1.71%
MDR	1.54%	3.96%	1.26%	1.14%	0.77%
PZE.1	0.96%	3.78%	0.55%	0.78%	−1.71%
HP	−1.63%	3.50%	−1.44%	−1.62%	0.77%
CVX	−0.73%	3.19%	−0.71%	−1.12%	0.09%

TABLE V
RIDGE COEFFICIENTS: 1995/06/01

Factor	Coefficient	Factor	Coefficient
ROA	0.205677	DPO	0.000004
GPM	0.116841	DSI	−0.000018
REVGH	0.081318	EM	−0.000405
(Intercept)	0.049237	WCR	−0.000422
NPM	0.015640	PS	−0.002042
CR	0.004123	PCFO	−0.003140
EPS	0.002562	LTDTA	−0.026227
QR	0.002322	PB	−0.032136
DE	0.000612	OM	−0.055619
EVCFO	0.000013	ROE	−0.078489
PE	0.000004		

TABLE VI
RANDOM FOREST IMPORTANCE TABLE: 1995/09/01

Factor	Importance	Factor	Importance
PB	14.2818	PE	5.6404
EPS	9.0556	CR	5.5080
ROE	8.6929	ROA	5.3972
PS	8.4976	PCFO	5.1313
WCR	8.4904	EVCFO	4.6672
EM	7.3808	LTDTA	4.6447
GPM	7.2091	OM	4.2249
QR	7.1337	DE	3.3296
DPO	6.6917	DSI	2.4621
NPM	5.9743	REVGH	0.1632

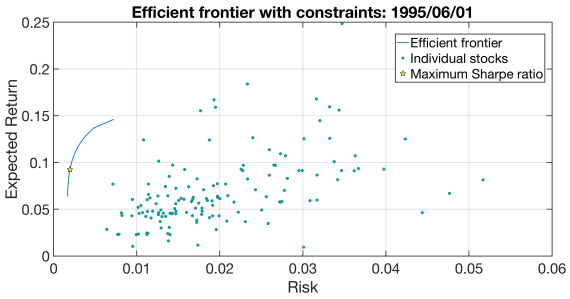

Fig. 3. This figure shows the efficient frontier.

are no abnormal results. in Table V and VI, e.g. assign 0 to all features.

We finish these steps for all eleven GICS sectors. Then we get a final table of all selected stocks with its tic name, predicted returns for next quarter, and the corresponding trading periods to conduct portfolio allocation.

III. PORTFOLIO ALLOCATION AND RISK MANAGEMENT

Portfolio allocation is crucial to an investment strategy because its balance risk and return by modeling individual asset's weights. Mean-variance and minimumu-variance are two typical methods for portfolio allocation. They perform diversification by constraining mean, volatility and correlation inputs to reduce sampling error [16]. In our portfolio we used mean-variance and min-variance to decide the weights of each stock, and then use equal-weighted portfolio as our benchmark. We perform these methods by Matlab Financial Toolbox-Portfolio Object [17].

A. Mean-Variance and Minimum-Variance Constraints

We first use mean-variance optimization to allocate the stocks we have picked. In. Fig. 3. the yellow star on the curve is the mean-variance result during our first trade time; the rest of the points are stocks plotted based on its predicted return and standard deviation. The result shows that our approach is legitimate.

We set the following constraints for mean-variance:

- Expected return: predicted return of next quarter.
- Covariance matrix: use 1 year historical daily return.
- Long only: upper bound 5% and Lower bound 0%.
- Fully invest our capital: sum of weights=100%.
- Take no leverage: LowerBudget = UpperBudget = 1.

Then, we try the min-variance optimization approach to allocate the stocks we have picked. The criteria are almost the same as the mean-variance method except that we set the expected return to be 0.

B. Transaction Costs

Generally, fees for each trade are measured based on broker fees, exchange fees and SEC fees. In the real-world scenarios, a fund or trading firm might have different execution costs for many reasons. Despite these possible variations in cost, after going through several scenarios we consider our transaction cost to be 1/1000 of the value of that trade. We believe our fee assumption to be sufficient and reasonable for the study.

We use the following formula to calculate the transaction cost:

$$\sum_{i=1}^{n} |S_{t,i} - S_{t-1,i}| \cdot P_i \times 0.1\% \tag{5}$$

where $S_{t,i}$ is the shares we need to buy or sell share based on the portfolio weights at current time t and $S_{t-1,i}$ is the shares left at previous time $t-1$. P_i is the current stock price of stock i.

C. Risk Management

After the procedure of building portfolio and structuring with appropriate functions, we equip decision rules that would be applied to risk management of each trade. Fundamentally, due to the nature of long-only strategy, the risk was controlled internally through our portfolio optimization methods. We minimize variance and maximum Sharpe ratio, have limits on position sizes (maximum of position size is 5% of portfolio value), and don't take any leverage.

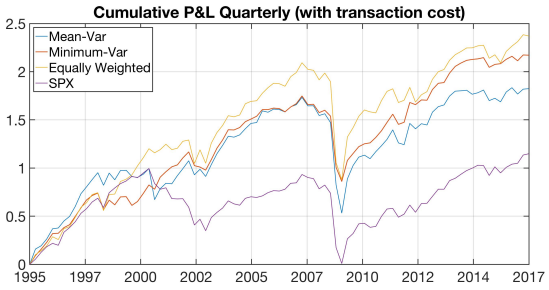

Fig. 4. This figure shows the P&L.

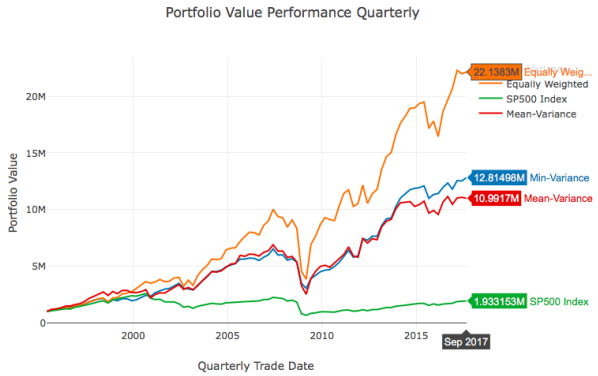

Fig. 5. This figure shows Portfolio Value starts with 1 million.

IV. Performance Evaluation

In Fig. 4, the results show that our strategy outperforms the market. The following back-test figures including equally weighted portfolio, mean-variance portfolio and minimum-variance portfolio indicate better performance than that of the benchmark S&P 500 index. More importantly, all the statistics can show that the portfolio outperforms the S&P 500 index not only in the in-sample training period (before 2007) but also in the overall trade period in Table VIII. Our codes are available online [18].

We conclude that our scheme for stock selections does generate a better result than the market portfolio does. If we only analyze the portfolio value performance from the figure, it is noticeable that the equally weighted portfolio, the benchmark, has higher value than min-variance and mean-variance portfolio. However, the portfolio value is not the only consideration when selecting the optimal portfolio. We have two reasons to conclude that the min-variance portfolio is a better method in the real trade period. First, the equally-weighted portfolio is not robust enough. We notice that the performance of this method fully depends on the predicted returns that we calculate from our model. The predicted returns will vary every time we run our model. Secondly, min-variance portfolio allocation takes the risk factor into consideration, making it more reliable in real trade. From the Table VII, we find that the min-variance allocation has a higher Sharpe ratio than that of the equally-weighted allocation during the in-sample period. We thus choose the min-variance as our portfolio allocation method.

TABLE VII
In sample data result: 1995-2007

	Mean-Var	Equally	Min-Var	S&P 500
Annualized Return	13.17%	16.12%	13.29%	7.12%
Annualized Std	17.0%	16.4%	12.9%	13.8%
Sharpe Ratio	0.687	0.887	0.917	0.406

TABLE VIII
Overall Performance

(Risk-Free: 1.5%)	Mean-Var	Equally	Min-Var	S&P 500
Start Value in million	1	1	1	1
End Value in million	10.9917	22.1383	12.81498	1.933153
Total Return	999.17%	2113.83%	1181.50%	93.32%
Maxiumum Drawdown	−56.89%	−57.63%	−46.30%	−66.73%
Annualized Return	8.29%	10.77%	9.87%	5.22%
Annualized Std	23.6%	26.4%	18.1%	19.1%
Sharpe Ratio	0.287	0.351	0.462	0.195

V. Conclusion

Applying machine learning algorithms to the fundamental financial data can filter out stocks with a relative bad earnings, thus providing a better way to select stocks. Minimum-variance method, the 5 percent holding rule, no short and leverage rule provide risk management and diversification, reduce the portfolio risk and thus yield a higher Sharpe ratio. Compared to the benchmark, our trading strategy outperforms the S&P 500 index. More importantly, combined with our trading strategy, the portfolio allocation method is proven to improve the overall performance. Finally, the Sharpe ratios of the three portfolio methods indicate that our strategy also outperforms the market. Future work would be dealing with anomaly data [19] in the data preprocessing stage, and applying accurate prediction schemes by modeling stock indicators as tensor time series [20] with sparsity in transform domains.

References

[1] Yinglei Zhang Kewei Hou, Mathijs A. van Dijk, "The implied cost of capital: A new approach," *Journal of Accounting and Economics 53 (3), 504526.*, 2011.

[2] Kewei Hou, Chen Xue, and Lu Zhang, "Digesting anomalies: An investment approach," *Fisher College of Business Working Paper No. WP 2012-03-021*, 2014.

[3] Joseph J. Gerakos and Robert Gramacy, "Regression-based earnings forecasts," *Chicago Booth Research Paper No. 12-26.*, 2013.

[4] Gilberto Batres-Estrada, "Deep learning for multivariate financial time series," 2015.

[5] John Alberg and Zachary C. Lipton, "Improving factor-based quantitative investing by forecasting company fundamentals," *arXiv:1711.04837*, 2017.

[6] Radim Gottwald, "The use of the p/e ratio to stock valuation," vol. 415, 2012.

[7] Benjamin Graham, Sidney Cottle, Roger F. Murray, and Frank E. Block, *Security Analysis, Fifth Edition*, McGraw-Hill, 1988.

[8] Maria Crawford Scott, "Value investing: A look at the benjain graham approach," *AAII*, 1996.

[9] Compustat Industrial [daily and quarterly Data]. (2017). Available: Standard Poor's/Compustat [2017]. Retrieved from Wharton Research Data Service., ," .

[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning, 2nd Edition*, Springer, New York, NY, 2009.

[11] 2011. R Development Core Team, "R: A language and environment for statistical computing.," *R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL http://www.R-project.org/.*

[12] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, New York, NY, 2013.
[13] Iain Johnstone Bradley Efron, Trevor Hastie and Robert Tibshirani, "Least angle regression," *The Annals of Statistics 2004, Vol. 32, No. 2, 407499*.
[14] Andy Liaw and Matthew Wiener, "Classification and regression by randomforest," *R News 2 (3), 1822.*, 2002.
[15] Greg Ridgeway, "Generalized boosted models: A guide to the gbm package," 2007, URL https://CRAN.R-project.org/package=gbm.
[16] Andrew Ang, "Mean-variance investing," *Columbia Business School Research Paper No. 12/49.*, August 10, 2012.
[17] MathWorks, "Matlab financial toolbox: Portfolio object," 2017, https://www.mathworks.com/help/finance/portfolio-object-mv.html.
[18] "Our codes," *http://www.tensorlet.com/*.
[19] Xiao-Yang Liu and Xiaodong Wang, "Ls-decomposition for robust recovery of sensory big data," *IEEE Transactions on Big Data*, 2017.
[20] Xiao-Yang Liu and Xiaodong Wang, "Fourth-order tensors with multidimensional discrete transforms," *arXiv preprint arXiv:1705.01576*, 2017.