THE UNIVERSITY OF CHICAGO

MACHINE LEARNING MODELS APPLIED TO HIGH FREQUENCY FINANCIAL DATA

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE UNIVERSITY OF CHICAGO

BOOTH SCHOOL OF BUSINESS

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

AN QI

CHICAGO, ILLINOIS

AUGUST 2019

TABLE OF CONTENTS

Chapter 2: How useful are machine learning tools in predicting high frequency returns?

LIST OF FIGURES

Chapter 1: A dynamic network model for large dimension order flows in financial markets

Chapter 2: How useful are machine learning tools in predicting high frequency returns?

LIST OF TABLES

Chapter 1: A dynamic network model for large dimension order flows in financial markets

Chapter 2: How useful are machine learning tools in predicting high frequency returns?

ACKNOWLEDGMENTS

Foremost, I would like to thank my advisor and committee chair Professor Jeffrey Russell. This thesis became a reality because of his knowledge, encouragement, support, and motivation.

I would like to thank the rest of my committee, Professor Tengyuan Liang, Professor Dacheng Xiu, and Professor Ruey Tsay for their insights and guidance.

I would like to thank Professor Max Farrell, Professor Christian Hansen, and Professor Mladen Kolar for their helpful comments.

Finally, I would like to acknowledge with eternal gratitude the love and support of my parents.

CHAPTER 1:

A dynamic network model for high frequency order flows in financial markets

An Qi
University of Chicago
Booth School of Business

June 12, 2019

Abstract

This work constructs a network model for high frequency trading volume data using a regularized vector autoregression moving average (VARMA) method. The network models how trading activity in one economic sector or asset group impacts trading in other asset groups. I explore the extent to which current trading volume is predictable from past trading volume history and how bursts of trading activity in one asset group is transmitted to other asset groups. I construct network connectedness measures which quantify the impact of these volume shocks. The results reveal that trading volume has a good deal of predictability: for the 51 asset groups considered, the model generates an average R^2 of .16. About 45% of the network impacts as measured by impulse response functions are due to cross asset shocks. The results also reveal clear clustering of assets into groups that match economic intuition.

1 Introduction

In this work, I construct a network model for high frequency trading volume data using a regularized vector autoregression moving average (VARMA) method. The network models how trading activity in one economic sector or asset group impacts trading in other asset groups. To measure trading activity in an economic sector, I use the minute by minute trading volume of exchange traded funds with holdings in that economic sector. I explore the extent to which current trading volume is predictable from past trading volume history and how bursts of trading activity in one asset group is transmitted to other asset groups. I construct network connectedness measures which quantify the impact of these volume shocks.

Methodologically, the work builds on the network literature in empirical economics and finance by introducing sparse vector autoregressive moving average (VARMA) estimation to the problem of network connectedness. VARMA models are capable of representing the dynamics of persistent series in a much more parsimonious way compared with VAR models, which would require a high number of lags and estimated parameters, resulting in loss of precision. But while both VAR and the one dimensional autoregressive moving average are highly popular in empirical economics and finance, their multidimensional fusion VARMA is rarely used due to severe problems with estimation. The efficient maximum likelihood methods used to estimate single dimensional ARMA models become much more computationally intensive with higher dimension datasets, leading to optimization problems such as slow convergence and non-robustness with respect to initial conditions. An alternative method to MLE is the iterative ordinary least squares (IOLS) estimator (Hannen and Risssanen (1982), Dias and Kapetanios (2017)), which is feasible for large datasets. IOLS is used to recover the latent residuals, which are then used to construct the vector moving average terms to be shrinked and regularized along with the vector autoregressive terms. Note that the estimator is not truly "high dimensional", as it requires the sample size to be much higher than the number of regressors to obtain reasonable estimates. However, this method is perfect for analysis of certain high frequency financial time series, which has high sample size as well as persistency.

Empirically, the work's analysis of high frequency volume fills an important gap in the empirical finance literature. As in all markets, the intersection of asset supply and asset demand determines an equilibrium price, returns, and an equilibrium quantity, volume. But compared with the massive literature on returns, academic work on trading volume has been fairly sparse. An early survey of the price-volume relationship is presented in Karpoff (1987). A more recent theoretical and empirical treatment of volume is given in Lo and Wang (2009). This sparsity of literature extends into the high dimensional and high frequency areas as

well. Works in high dimensional, high frequency returns have been emerging. Xiu et al. (2017) apply high dimensional inference techniques to evaluating the contribution of pricing factors to asset returns. Chinco et al. (2018) apply sparse regression methods to the prediction of minute by minute returns. However, I am not aware of any such works for high dimensional, high frequency volume. Moreover, an analysis of high frequency volume does not need to be a purely academic exercise. As the old Wall Street adage goes, "it takes volume to make prices move." A deeper understanding of the dynamics of high frequency volume may shed light on the dynamics of high frequency returns.

This work follows a large body of previous work in network analysis, broadly characterized as studying the interdependence in large multivariate systems. Diebold and Yilmaz (2009) study the spillover of return and volatility in global equity markets. Diebold and Yilmaz (2014) measure the connectedness of US financial firms, and Demirer et al (2018) measure the connectedness of global financial firms. Billio et al. (2012), Allen et al. (2012), and Hautsch et al (2015) examine connectedness as it relates to systemic risk in the financial sector. Acemoglu et. at (2012) applies network analysis to study cascade effects where idiosyncratic shocks can lead to aggregate fluctuations. Theoretical work has focused on learning the underlying network structure of high dimensional systems under the assumption of sparsity (Meinshausen and Buhlman (2006), Friedman et al, (2008), Peng et al. (2009)). Theoretical work on sparse estimation of vector autoregression include Basu and Michailidis (2015), Nicholson et al (2017), and Barigozzi and Brownlees (2018).

Section 2 describes the data. Section 3 describes the methodology. Section 4 gives model results and network connectedness measures. In section 5, I apply a community detection algorithm that groups together asset categories whose trading volume tends to move together and provide a graphical visualization the network. In section 6, I disaggreggate total volume into buy volume and sell volume and model them separately. Section 7 concludes.

2 Data

2.1 Exchange Traded Funds

I use exchange traded fund (ETF) trading data to measure trading activity in an economic sector. ETFs are marketable securities that track an index, a commodity, bonds, or a basket of assets like an index fund. Unlike mutual funds, an ETF trades like a common stock on a stock exchange. The fund owns the underlying assets (shares of stock, bonds, oil futures, gold bars, foreign currency, etc.) and divides ownership of those

assets into shares. I choose to use ETF trading data rather than the trading data of their underlying assets in true big data fashion for two reasons. The first reason is scope. To measure trading activity in all cap equities, I can use the data from a few of the most highly traded equity ETFs (SPY, QQQ, DIA, VTI) rather than using the entire universe of stocks. Similarly for corporate bonds, treasuries, sector equities, and other broad asset classes. The second reason is liquidity. ETFs tend to be more liquid than their underlying assets. Attempting to model the volume of such illiquid assets at the minute frequency would not be possible, given the percentage of the minute bins in the sample that would be 0.

The ETF trading data is taken from the Trade and Quote (TAQ) database of the Wharton Research Data Service. I use data from 225 ETFs from January 2015 - December 2016. For each transaction, the TAQ trade data logs the time stamp, the transaction price, and the transaction volume. I create minute by minute dollar volume for the 225 ETFs by assigning each transaction to a minute bin and summing the dollar volume within the minute. Then the 225 ETFs are grouped into 51 asset categories. The minute trading volume for an asset category is given by the sum of the minute trading volumes of the constituent funds. The following table gives the 51 asset categories.

Table 2.1: Asset Categories

Fixed Income	**Broad Based Equity**	**Sector Equity**
Corporate Bonds	All Cap Equities	Building and Construction Equities
Emerging Market Bonds	Large Cap Blend Equities	Consumer Discretionary Equities
Government Bonds	Large Cap Growth Equities	Consumer Staples Equities
High Yield Bonds	Large Cap Value Equities	Energy Equities
Inflation Protected Bonds	Mid Cap Blend Equities	Financial Equities
International Government Bonds	Mid Cap Growth Equities	Health and Biotech Equities
Money Market	Mid Cap Value Equities	Industrials Equities
Mortgage Backed Securities	Small Cap Blend Equities	Materials Equities
National Munis	Small Cap Growth Equities	Master Limited Partnership Equities
Preferred Stock/Convertible Bonds	Small Cap Value Equities	Technology Equities
Total Bond Market		Utilities Equities
International Equity	**Commodities**	**Inverse/Leveraged/Other**
Small Cap Blend Equities	Commodities	Inverse Treasuries
Small Cap Growth Equities	Gold And Silver	Inverse Commodities
Small Cap Value Equities	Oil and Gas	Inverse Equities
Asia Pacific Equities		Leveraged Commodities
China Equities		Leveraged Equities
Emerging Market Equities		Real Estate
Europe Equities		Currency
Foreign Large Cap Equities		
Foreign Small and Mid Cap Equities		
India Equities		
Japan Equities		
Latin America Equities		

Fixed income ETFs hold different types of bonds, such as corporate bonds or treasuries. Broad based equity ETFs track equity indices that include stocks of a certain market cap or growth/value characteristics. Sector and international equity ETFs are self explanatory. Commodities ETFs either focus on a broad range of commodities, typically using futures contracts. Or they focus on one commodity, such as oil, gas, gold, or silver, sometimes by holding it in storage and sometimes using contracts. Inverse and leveraged ETFs track indices for broad based categories of assets, and seek to deliver either 2 or 3 three times the return of the index in the case of leveraged ETFs, or -1, -2, or -3 times the return in the case of inverse ETFs. These funds act as instruments to lever up bets. Currency ETFs consist of baskets of exchange rate futures contracts, from euros, yen, pounds, etc... Real estate ETFs primarily invest in real estate investment trusts, which are companies that own and in most cases operate income producing real estate.

The ETFs included in the corporate bonds category is given by the following table. A full list of the ETFs included in these asset categories is included in the appendix.

Table 2.2: ETFs included in the corporate bonds asset category

Category	Ticker	ETF Name
Corporate Bonds	LQD	iShares iBoxx $ Investment Grade Corporate Bond ETF
	VCSH	Vanguard Short-Term Corporate Bond ETF
	VCIT	Vanguard Intermediate-Term Corporate Bond ETF
	VCLT	Vanguard Long-Term Corporate Bond ETF

I drop the first ten minutes and last ten minutes of each trading day, as I am not interested in modeling the fulfillment of overnight orders in the next morning or the flurry of end of day activity from traders closing out their positions. I drop the Wednesday and Friday between Thanksgiving, Christmas Eve, and the days between Christmas and New Year's due to depressed amounts of trading compared to the typical trading day. I also drop the data from August 24, 2015, when a flash crash occurred which caused many ETFs to trade at a market value far below that of its underlying assets. This leaves 370 minutes of data for 51 assets categories over 487 trading days.

2.2 Data Transformations

Let $x_{i,t,d}$ be the dollar volume of transactions for asset category i in minute t of day d. There is typically a U shaped diurnal trend in trading data, as trading during the morning and afternoon is higher than trading during midday. This trend is removed by first regressing $x_{i,t,d}$ on minute of day t for each asset category using local linear regression with span .25. Call the asset diurnal trend $\tilde{x}_{i,t}$. After removing the diurnal trend, the data is then log transformed, which is necessary to account for the heavy right skew of the data.

Let $n_{i,t,d} = \log\left(\frac{1+x_{i,t,d}}{1+\tilde{x}_{i,t}}\right)$ be the normalized logged dollar volume.

It is possible that there are common shocks that could increase/decrease trading in all assets. I am interested in how activity in one economic sector affects activity in other economic sectors, not in how common shocks affect trading activity economy wide. The presence of a common shock could lead to spurious linkages between asset categories being detected. I remove this common trend using principal component analysis.

Stack the data day by day to obtain N, a $(370)(487) \times 51$, or (minutes)(days)×assets, data matrix.

$$N = \begin{bmatrix} n_{1,1,1} & n_{2,1,1} & \dots & n_{51,1,1} \\ \dots & \dots & \dots & \dots \\ n_{1,370,1} & n_{2,370,1} & \dots & n_{51,370,1} \\ n_{1,1,2} & n_{2,1,2} & \dots & n_{51,1,2} \\ \dots & \dots & \dots & \dots \\ n_{1,370,2} & n_{2,370,2} & \dots & n_{51,370,2} \\ \dots & \dots & \dots & \dots \\ n_{1,1,487} & n_{2,1,487} & \dots & n_{51,1,487} \\ \dots & \dots & \dots & \dots \\ n_{1,370,487} & n_{2,370,487} & \dots & n_{51,370,487} \end{bmatrix}$$

Let C be N's correlation matrix. Looking at the scree plot of the correlation matrix, I extract one factor.

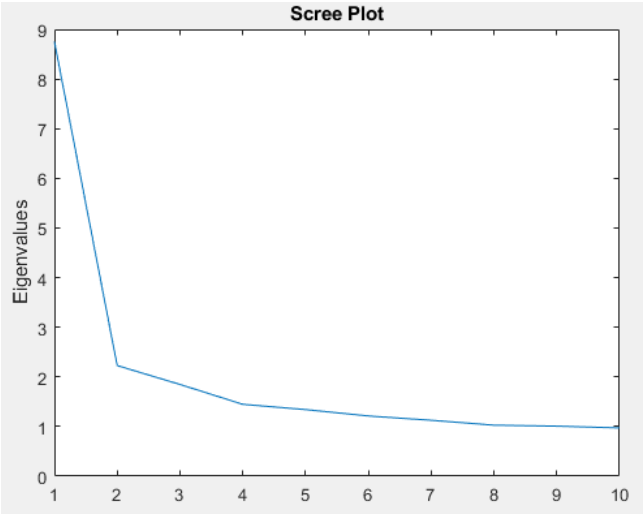


Figure 2.1: Scree plot, transactions

Let w be the normalized eigenvector associated with the largest eigenvalue of C. Then construct factor $F = w^T N$. Let N_i be the stacked data for asset i, the ith column of N. Then set $Y_i = N_i - D(D^T D)^{-1} D^T N_i$, where $D = \begin{bmatrix} 1 & F \end{bmatrix}^T$. Unstacking Y_i, $y_{i,t,d}$ gives the transformed volume of asset i in minute t and day d. This procedure removes contemporaneous shocks that affects all assets.

3 Methodology

3.1 Iterative Least Squares Estimation of VARMA Models

The standard VAR, as used by Diebold and Yilmaz (2014) and Demirer et al (2018) for network construction, is ill suited to modeling volume data; VAR estimates of

$$y_t = \alpha + \sum_{k=1}^{p} \Pi_k y_{t-k} + \epsilon_t$$

with successively increasing numbers of lags p revealed slowly declining AR coefficients. The following plot gives the absolute sum of the off diagonal coefficients of Π_k for lag k for multiple VAR models.

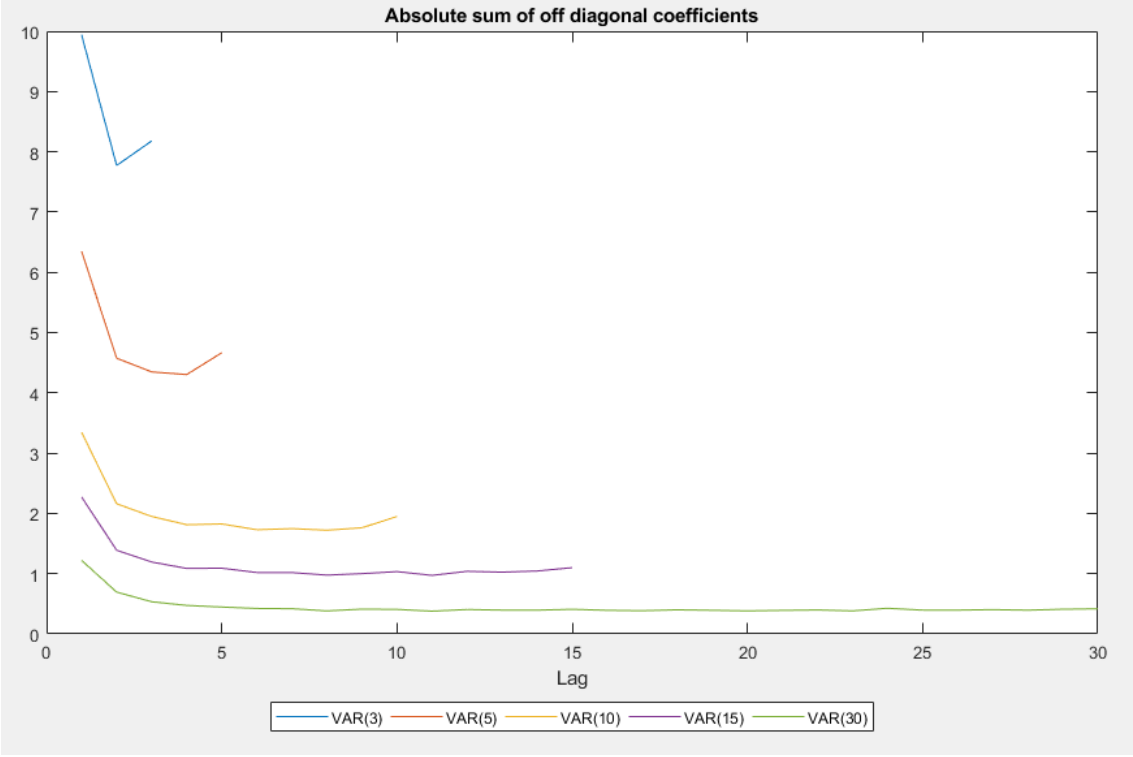


Figure 3.1: Absolute sum of off diagonal coefficients for VAR models

These diagnostics suggest the need for moving average terms. I use iterative least squares to estimate the VARMA model, which avoids the pitfalls of traditional MLE methods.

A stationary invertible VARMA model admits a VAR(∞) representation

$$y_t = \alpha + \sum_{k=1}^{\infty} \Pi_k y_{t-k} + \epsilon_t$$

Hannen and Rissanen (1982) give a two step least squares procedure to estimate a VARMA(p,q) model. In the first stage, the VAR(∞) representation is truncated at lag $\tilde{p}$. OLS is used to obtain coefficient estimates $\hat{\alpha}$, $\hat{\Pi}_k$, and residual estimates $\hat{\epsilon}_t^0 = y_t - \hat{\alpha} - \sum_{k=1}^{\tilde{p}} \hat{\Pi}_k y_{t-k}$. In the second stage, the residuals estimates from the first stage is plugged into the equation

$$y_t = \alpha + \sum_{k}^{p} \beta_k y_{t-k} + \sum_{k}^{q} \gamma_k \hat{\epsilon}_{t-k}^0 + \epsilon_t$$

which is estimated with OLS. Dufour and Jouini (2005) derive consistency of the parameter estimators for $\tilde{p}$ increasing at a rate below $T^{1/2}$ and asymptotic normality for $\tilde{p}$ increasing at a rate below $T^{1/4}$.

Dias and Kapetanios (2017) suggest using this OLS procedure iteratively. The initial VAR($\tilde{p}$) is used as a warm start to get initial residuals estimates. The parameter estimates from the previous OLS estimation are used to construct residuals which are used to perform another OLS estimation. The algorithm is given as follows:

1. For long lag length $\tilde{p}$, fit the VAR $y_t = \alpha + \sum_k^{\tilde{p}} \Pi_k y_{t-k} + \epsilon_t$, obtain estimates $\hat{\alpha}^0$, $\hat{\Pi}_k$ and $\hat{\epsilon}_t^0 = y_t - \hat{\alpha}^0 - \sum_k^{\tilde{p}} \hat{\Pi}_k y_{t-k}$
2. For AR lag length p and MA lag length q, fit the model $y_t = \alpha + \sum_k^p \beta_k y_{t-k} + \sum_k^q \gamma_k \hat{\epsilon}_{t-k}^0 + \epsilon_t$, obtain estimates $\hat{\alpha}^1$, $\hat{\beta}_k^1$, $\hat{\gamma}_k^1$ and $\hat{\epsilon}_t^1 = y_t - \hat{\alpha}^1 - \sum_k^p \hat{\beta}_k^1 y_{t-k} - \sum_k^q \hat{\gamma}_k^1 \hat{\epsilon}_{t-k}^0$
3. For AR lag length p and MA lag length q, fit the model $y_t = \alpha + \sum_k^p \beta_k y_{t-k} + \sum_k^q \gamma_k \hat{\epsilon}_{t-k}^1 + \epsilon_t$, obtain estimates $\hat{\alpha}^2$, $\hat{\beta}_k^2$, $\hat{\gamma}_k^2$ and $\hat{\epsilon}_t^2 = y_t - \hat{\alpha}^2 - \sum_k^p \hat{\beta}_k^2 y_{t-k} - \sum_k^q \hat{\gamma}_k^2 \hat{\epsilon}_{t-k}^1$
4. Iterate until convergence: $||\epsilon^s - \epsilon^{s-1}|| < \delta$

The algorithm is consistent and has the same asymptotic properties as the Hannen-Rissanen method. The authors conduct Monte Carlo studies that show that the iterative OLS procedure has good finite sample performance compared with MLE in a variety of settings, including in high dimensions.

3.2 L1 Regularization and the Lasso

Given consistent estimates of the residuals, I perform a final L1 regularization step. This is done for two reasons. The first reason is model selection: if the past of asset j is not predictive of current asset i volume, L1 regularization will set the regression coefficients associated with asset j to 0. The second reason is shrinkage. Shrinkage makes the parameter estimates less noisy, trading off variance for an increase in the bias of the parameter estimates. For moderate levels of shrinkage, this decreases the mean squared error of predictions. This also makes the resulting network constructed from the coefficients less noisy and more interpretable.

L1 regularization was first proposed by Tibshirani (1996). The least absolute shrinkage and selection operator (lasso) minimizes

$$\arg\min_{\alpha,\beta} ||y - \alpha - X\beta||_2^2 + \lambda||\beta||_1$$

The additional penalty term applies an L1 penalty to the regression coefficients, which sets certain coefficients to 0, encouraging sparsity, and shrinks the other coefficients towards 0, as in ridge regression. The parameter λ controls the degree of regularization, with higher values leading to sparser models. This parameter can be chosen using cross validation.

If the regressors possess a known group structure, a variant of the lasso, the group lasso of Yuan and Lin (2007), can be fit:

$$\arg\min_{\alpha,\beta} ||y - \alpha - \sum_{l=1}^{L} X_l\beta_l||_2^2 + \lambda\sum_{l=1}^{L}\sqrt{p_l}||\beta_l||_2$$

The regressors are grouped into L groups, with p_l the size of group l. As before, λ controls the degree of regularization, with higher values leading to sparser models. The group lasso penalty encourages sparsity at the group level, causing entire groups of covariates to drop out. There is no sparsity within selected groups.

3.3 The Model

I fit the group lasso for each asset, equation by equation, with residuals given by the last iteration of the iterative least squares procedure:

$$y_{i,t} = \alpha_i + \sum_{j=1}^{51}[\sum_{k=1}^{p}\beta_{i,j,k}y_{j,t-k} + \sum_{k=1}^{p}\gamma_{i,j,k}\hat{\epsilon}^s_{j,t-k}]$$

For convenience I set the AR and MA lag length to be equal. The groups are given by

$G_{i,j} = \{\beta_{i,j,1}, \beta_{i,j,2}, ..., \beta_{i,j,p}, \gamma_{i,j,1}, \gamma_{i,j,2}, ..., \gamma_{i,j,p}\}$, for a total of 51 groups for each asset equation. This has

the effect of turning entire assets on or off. The penalty parameter is chosen by 3-fold cross validation, where an entire day of data is assigned to a fold rather than individual observations. Stacking the parameter estimates equation by equation gives the following regularized VARMA model

$$y_t = \alpha + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{p} \gamma_k \epsilon_{t-k} + \epsilon_t$$

To choose a lag length, I look at in-sample MSE for different lag choices:

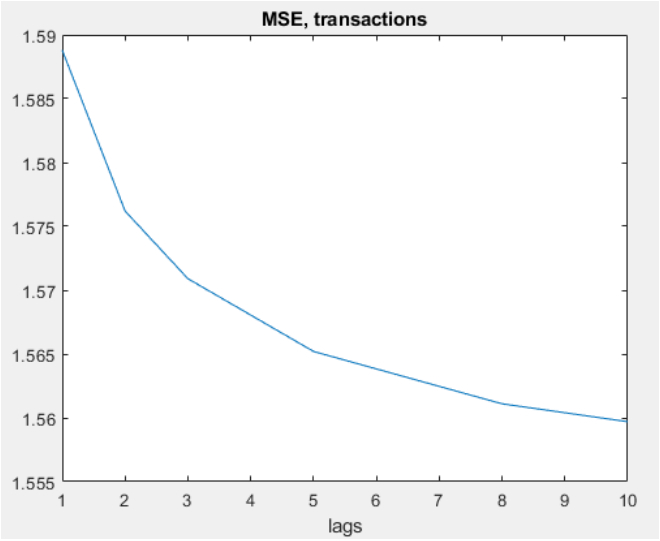


Figure 3.2: MSE for VARMA models, transactions

As there isn't a clear "elbow" where additional lags yield little further predictability, I pick, somewhat arbitrarily, a lag length of $p = 3$. The VARMA iterative least squares procedure reached sufficient convergence after 20 iterations.

4 Results

4.1 Model Estimates

For each asset equation, I obtain the number of active cross asset groups $\{G_{i,j} | i \neq j\}$ and the R^2, plotted in the histograms below. The estimation procedure selects fairly populated models, with a mean number of cross assets selected of 36. The model also shows that volume has a good amount of predictability. The mean R^2 is .16, and Gold and Silver and Energy Equities both have R^2 above .3.

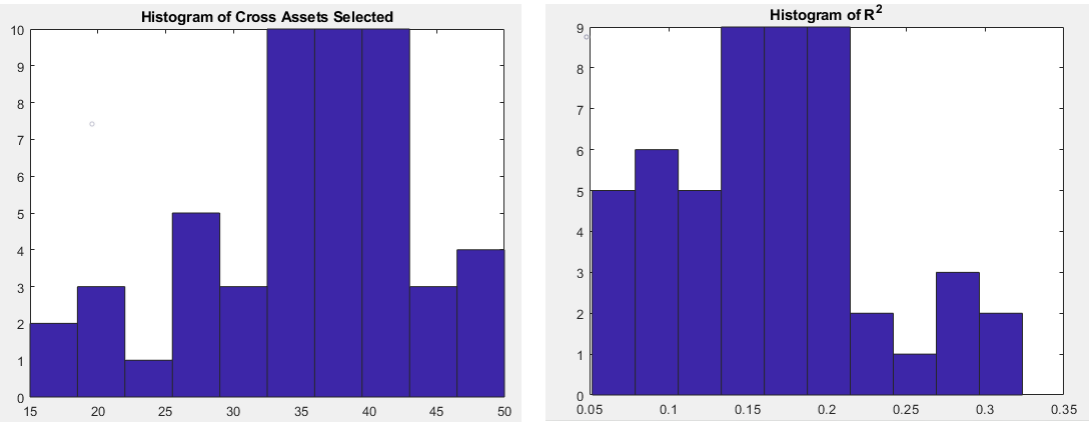


Figure 4.1: Histograms of cross assets selected and R^2, transactions

4.2 Impulse Response Network

I invert the VARMA model to obtain the VMA(∞) representation

$$y_t = \gamma_0 + \sum_{k=1}^{\infty} \gamma_k^{VMA} \epsilon_{t-k} + \epsilon_t$$

where γ_k^{VMA} is the lag k coefficient matrix. The (i,j) entry of γ_k gives the effect on asset i to a unit shock to asset j, holding the shocks to all assets $j \neq i$ constant. As the variables are in logs, this represents the percent above daily trend for asset i if asset j is shocked to be 1% above daily trend. This is the nonorthogonalized impulse response function; I ignore contemporaneous correlation of the shocks. Summing the impulse response over lags give the cumulative impulse response function

$$G_k = \sum_{k'=1}^{k} \gamma_{k'}^{VMA}$$

This object can be interpreted as a network, with the off-diagonal elements giving the strength of the connection between nodes. This is a bi-directional network: the (i, j) term gives the connection strength of asset j's past on asset i's present, whereas the (j,i) term gives the connection strength of asset i's past on asset j's present. Note that although the estimating equation induces sparsity, the resulting network is a nonlinear transformation of the VARMA coefficients and is therefore generally not sparse.

The following table gives the 10 largest edges $|G_{i.j,15}|$, $i \neq j$ at a horizon of $k = 15$.

Table 4.1: 10 largest edges, transactions network

Leveraged Commodities	activates	Inverse Commodities	0.739
All Cap Equities	activates	Leveraged Equities	0.582
All Cap Equities	activates	Inverse Equities	0.532
Energy Equities	activates	Oil and Gas	0.499
Leveraged Commodities	activates	Oil and Gas	0.390
Oil and Gas	activates	Inverse Commodities	0.315
Inverse Commodities	activates	Leveraged Commodities	0.259
Leveraged Equities	activates	Inverse Equities	0.237
Inverse Commodities	activates	Oil and Gas	0.233
Energy Equities	activates	Inverse Commodities	0.221

This table shows that the most active effects are levered bets for or against the stock market, levered bets for or against commodities, and trading in oil and energy.

Now I construct network connectivity measures. For each asset i, I consider two measures: an intensity measure $\sum_{j \neq i} |G_{j,i,15}|$, and a connectedness measure $\frac{\sum_{j \neq i} |G_{j,i,15}|}{\sum_{j=1}^{51} |G_{j,i,15}|}$. The intensity measure sums up the outward edges emanating from node i. The connectedness measure gives the percentage of the total cumulative impulse response to a shock to asset i that is due to cross asset responses. The intensity measure is analogous to a node's degree, from elementary graph theory, defined as the number of edges connected to a node. A node is said to be central if it has a high degree. Interpreting the volume impulse response function as a weighted, bi-directional network, the intensity of a node is a degree centrality measure. It measures how much a shock to asset i impacts the entire network. The connectedness measure accounts for the presence of own node effects, which are absent in a traditional graph setting. It measures the extent to which a shock to asset i spills over to other nodes as opposed to being confined to the own node.

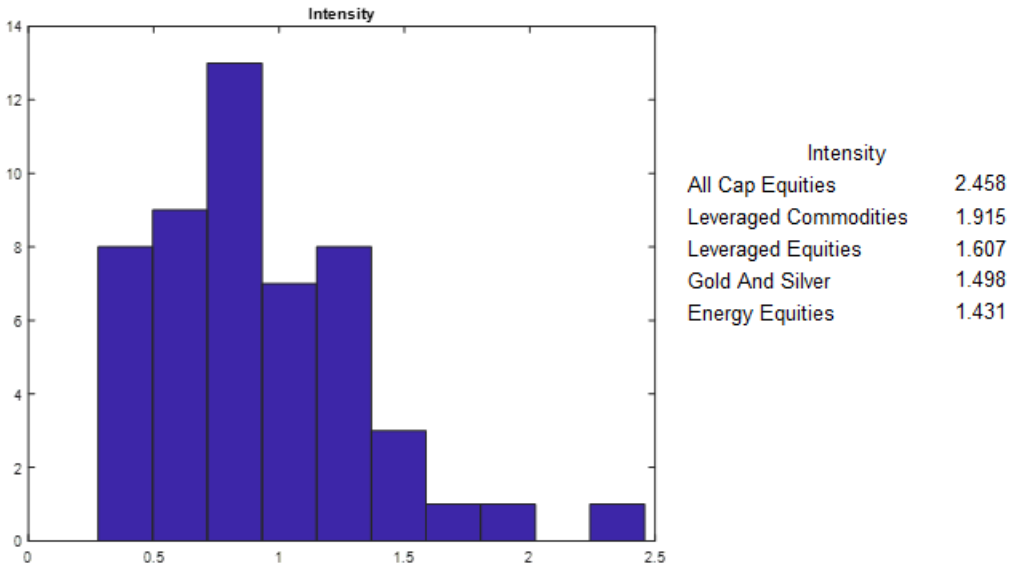


Figure 4.2: Histogram of Intensity

Not surprisingly, the asset category with the broadest network impact is all cap equities, consisting of ETFs that hold the entire stock market. Other asset categories with high intensities are leveraged commodities, leverage equities, gold and silver, and energy equities. At the other end, foreign small and mid cap equities, real estate, and technology have the lowest network impacts.

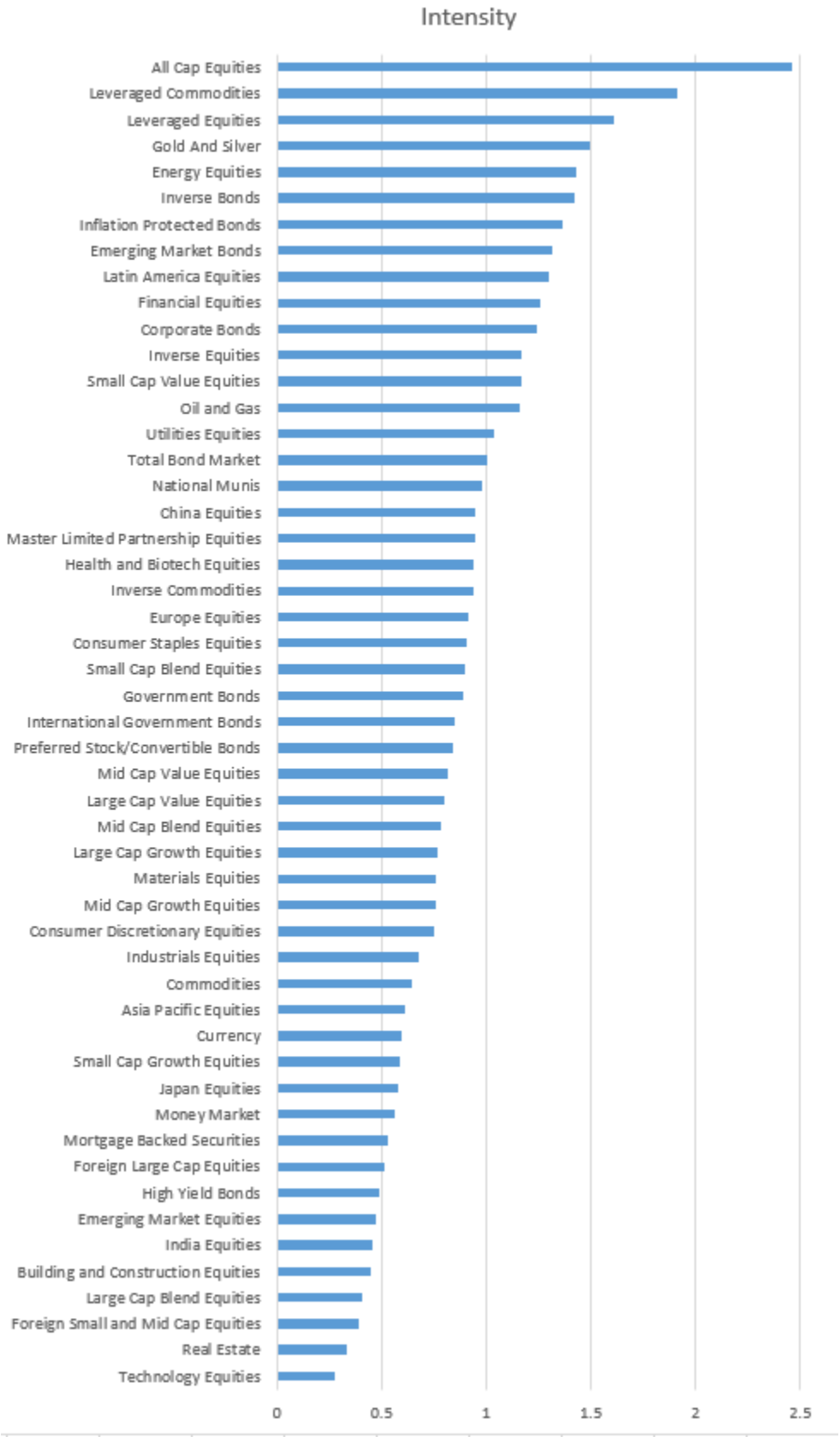


Figure 4.3: Intensity

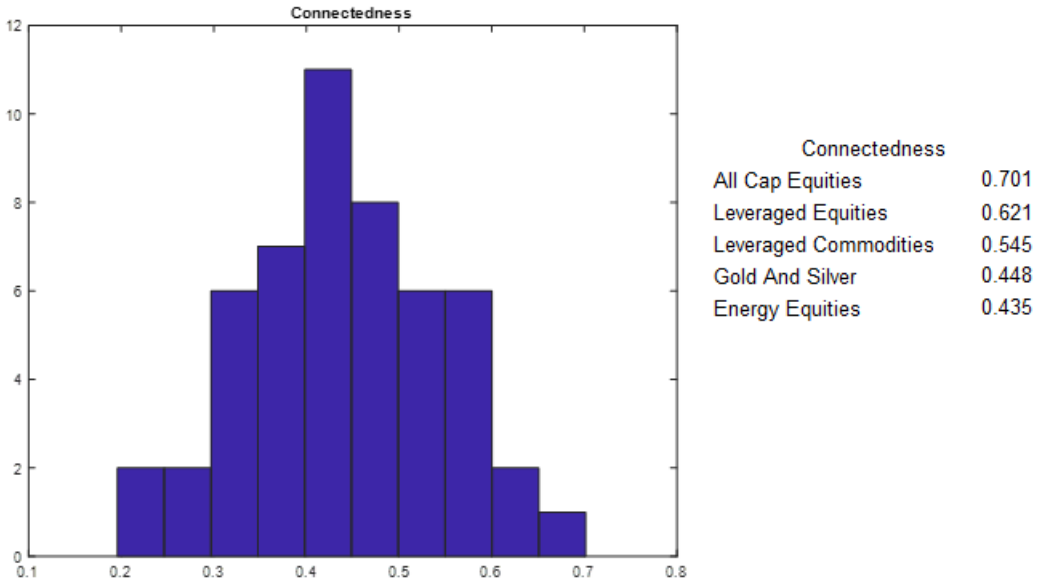


Figure 4.4: Histogram of Connectedness

Also not surprisingly, the asset category with the highest spillover is all cap equities. Other asset categories with high connectedness are leveraged equities, leverage commodities, gold and silver, and energy equities. At the other end, large cap blend equities, real estate, and technology have the lowest spillovers. 15 of the 51 asset categories have connectedness above .5, indicating that the majority of the network effects from shocks to these assets spill over to cross assets.

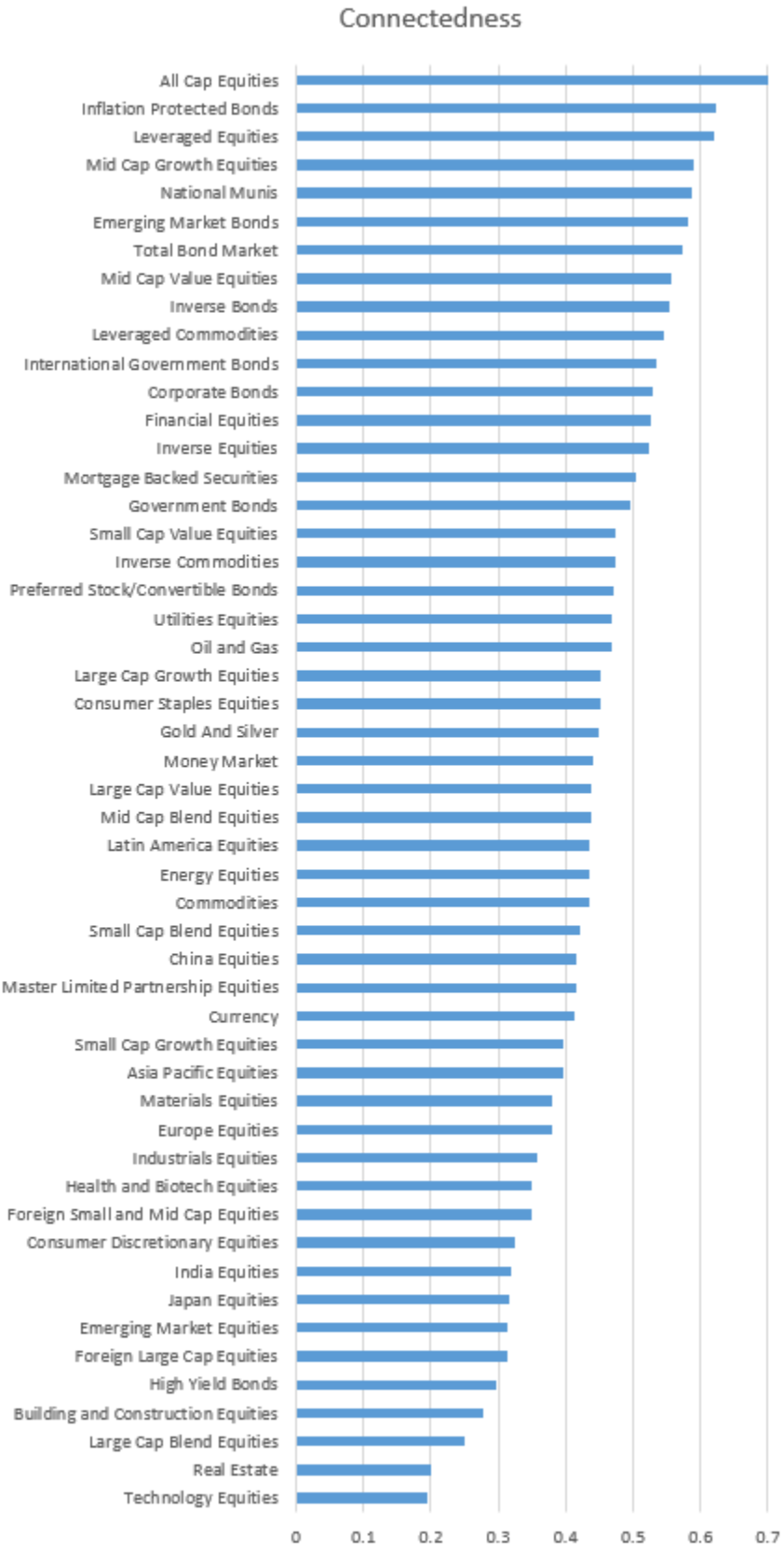


Figure 4.5: Connectedness

I now construct a total connectedness measure

$$\frac{\sum_i \sum_{j \neq i} |G_{i,j,15}|}{\sum_i \sum_j |G_{i,j,15}|} = 0.4518$$

This is the ratio of the absolute sum of cross asset effects to the absolute sum of total effects. 45% of the network impacts are due to cross asset effects.

4.3 Impulse Response: All Cap Equities, Total Bond Market, Oil and Gas

I examine the cumulative impulse response for a major equity, fixed income, and commodity node.

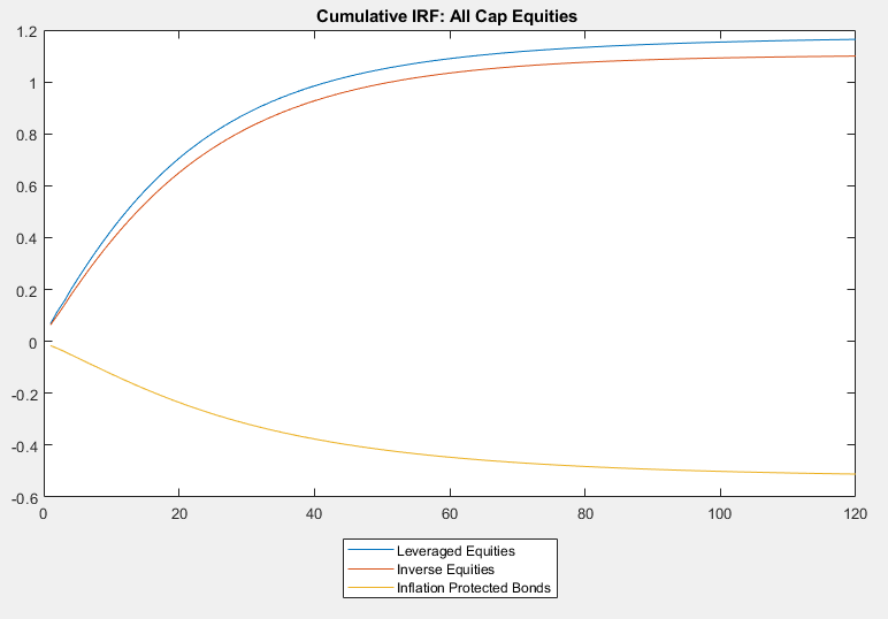


Figure 4.6: Cumulative IRF, three highest responses to a shock to all cap equities

The above gives the cumulative impulse response functions of the three asset categories with the highest responses to a 1% shock to all cap equities. Increased trading in equities predicts increased trading in bets for or against equities, with trading in leveraged and inverse equities increasing by .3% at a 15 minute horizon and asymptoting to a 1.1% increase. It also predicts less trading in inflation protected bonds, with a -.2% effect at 15 minutes and asymptoting to -.5%. This suggests that news in the equities markets causes traders to focus on equity trading at the expense of trading riskless assets such as Treasury Inflation Protected Securities (TIPS).

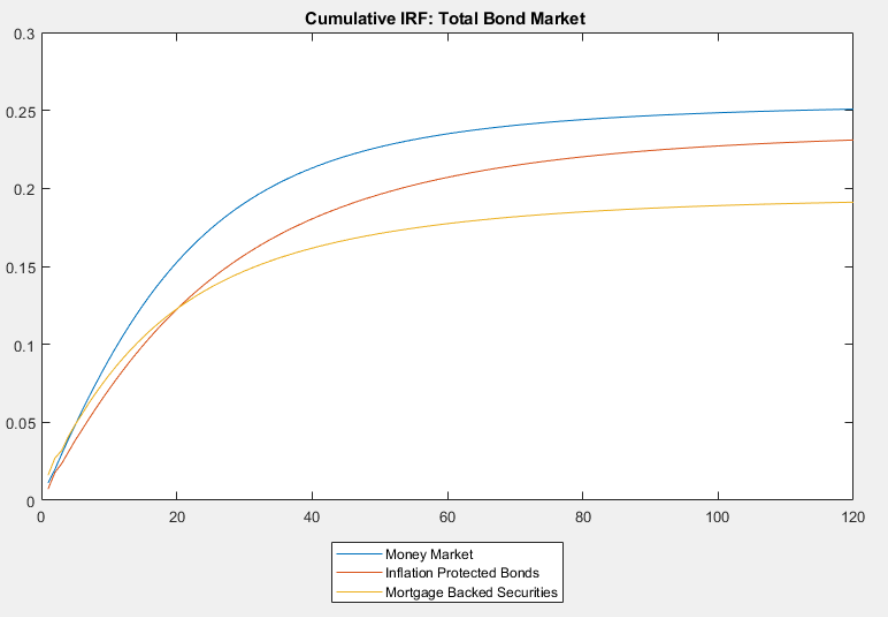


Figure 4.7: Cumulative IRF, three highest responses to a shock to total bond market

The above gives the cumulative impulse response functions of the three asset categories with the highest responses to a 1% shock to the total bond market. As expected, the largest effects are felt in segments of the bond market. Trading in the money market increases by .125% at 15 minutes, asymptoting to .25%. Trading in TIPS increases by .1% at 15 minutes, asymptoting to .225%. Trading in MBS increases by .1% at 15 minutes, asymptoting to .2%.

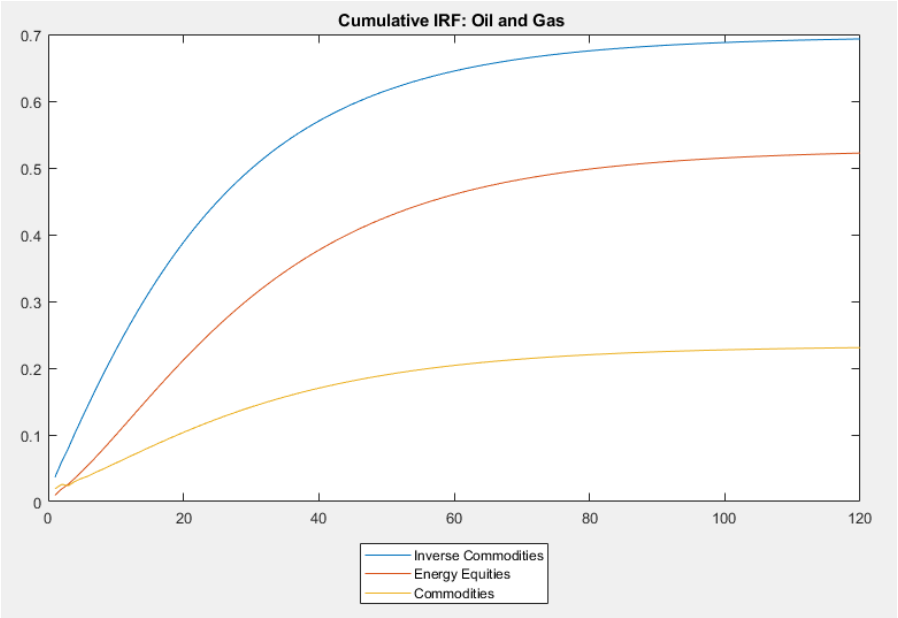


Figure 4.8: Cumulative IRF, three highest responses to a shock to oil and gas

The above gives the cumulative impulse response functions of the three asset categories with the highest responses to a 1% shock to oil and gas. The largest effects are felt in the commodities market, as well as in the stock of oil and gas companies. Trading in inverse commodities increases by .3% at 15 minutes, asymptoting to .7%. Trading in commodities increases by .1% at 15 minutes, asymptoting to .225%. Trading in energy equities increases by .2% at 15 minutes, asymptoting to .5%.

5 Community Detection

5.1 Clustering Algorithm

The nodes of a network are often found to form natural groups or communities where the edges between nodes in the same community are stronger than edges between nodes in different communities. In practice, this community structure is typically unknown, and methods for its recovery has been studied extensively (Fortunato 2010). Gudmundsson and Brownlees (2018) study this problem in the context of vector autoregression. It is easy to adapt their VAR network detection algorithm to the VARMA setting. Recall that a VARMA model admits a VMA(∞) representation $y_t = \gamma_0 + \sum_{k=1}^{\infty} \gamma_k^{VMA} \epsilon_{t-k} + \epsilon_t$. For a truncation lag p and number of clusters k, apply the following:

1. Construct the summed symmetrised MA matrix $\Phi = \sum_{k=1}^{p} (\gamma_k^{VMA} + (\gamma_k^{VMA})')$
2. Construct the $51 \times k$ eigenvector matrix U such that its columns are the eigenvectors corresponding to the k largest absolute eigenvalues of Φ
3. Standardize each row of U by its norm
4. Apply k-means clustering to the rows of U

I use a truncation lag $p = 60$. The first step constructs a similarity matrix by taking the summed symmetrized coefficients from the VMA representation. The next steps apply a spectral clustering algorithm. The usual method to pick the number of clusters in a spectral clustering algorithm is to examine the plot of eigenvalues of the similarity matrix:

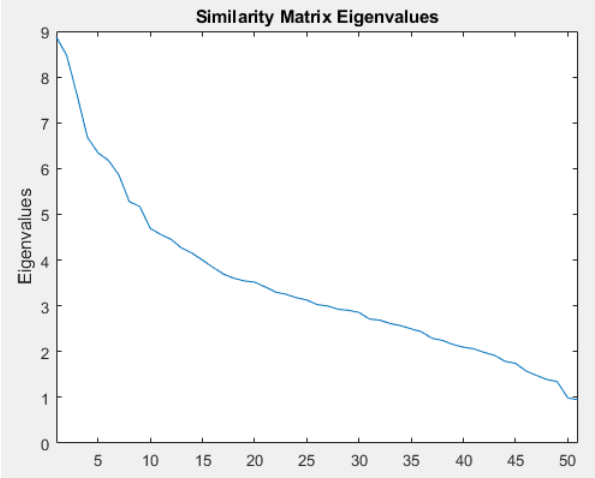


Figure 5.1: Similarity matrix eigenvalues

As there is no clearly defined elbow, I choose, somewhat arbitrarily, 10 clusters, as this choice returned economically meaningful clusters of assets which one would expect to move together. The cluster assignments are given by the following table:

Table 5.1: Cluster Assignment

Cluster 1 (Energy and Energy Speculation)
14: Oil and Gas
38: Energy Equities
48: Inverse Commodities
50: Leveraged Commodities

Cluster 2 (Fixed Income)
1: Corporate Bonds
2: Emerging Market Bonds
3: Government Bonds
5: Inflation Protected Bonds
6: International Government Bonds
7: Money Market
8: Mortgage Backed Securities
9: National Munis
10: Preferred Stock/Convertible Bonds
11: Total Bond Market
22: Mid Cap Value Equities
25: Small Cap Value Equities
39: Financial Equities

Cluster 3 (Broad Equity 1)
16: All Cap Equities
17: Large Cap Blend Equities
35: Building and Construction Equities
49: Inverse Equities
51: Leveraged Equities

Cluster 4 (Developed Market Equity)
29: Europe Equities
30: Foreign Large Cap Equities
33: Japan Equities

Cluster 5 (High Growth Equity)
23: Small Cap Blend Equities
40: Health and Biotech Equities
44: Technology Equities

Cluster 6 (Misc)
4: High Yield Bonds
13: Gold And Silver
42: Materials Equities
43: Master Limited Partnership Equities

Cluster 7 (Safe Equity)
36: Consumer Discretionary Equities
37: Consumer Staples Equities
41: Industrials Equities
45: Utilities Equities

Cluster 8 (Developing Market Equity)
26: Asia Pacific Equities
27: China Equities
28: Emerging Market Equities
32: India Equities
34: Latin America Equities

Cluster 9 (Misc)
12: Commodities
15: Currency
46: Real Estate
47: Inverse Bonds

Cluster 10 (Broad Equity 2)
18: Large Cap Growth Equities
19: Large Cap Value Equities
20: Mid Cap Blend Equities
21: Mid Cap Growth Equities
24: Small Cap Growth Equities
31: Foreign Small and Mid Cap Equities

Cluster 3 contains the very active all cap equity, leveraged and inverse equity nodes. Cluster 10 is a broad equity cluster. Cluster 4 and 8 are clusters of developed and developing country equity, respectively. Cluster 5 contains equities that are usually considered to be high growth, such as biotech and tech. Cluster 8 contains equities that are usually considered to be safe, like utilities and industrials. Cluster 6 is an energy and energy speculation cluster, containing oil and gas, energy equities, and leveraged and inverse commodities. Cluster 2 contains all of the fixed income asset categories.

To measure the effect of the clustering, I aggregate the network G_{15} by community. For communities C_i, C_j, define

$$\tilde{G}_{i,j,15} = \sum_{i' \in C_i} \sum_{j' \in C_j} |G_{i',j',15}|$$

Define a between cluster connectedness measure

$$\frac{\sum_{i=1}^{10} \sum_{j \neq i} |\tilde{G}_{i,j,15}|}{\sum_{i=1}^{51} \sum_{j=1}^{51} |G_{i,j,15}|} = 0.2902$$

and a within cluster connectedess measure

$$\frac{\sum_{i=1}^{10} |\tilde{G}_{i,i,15}| - \sum_{i=1}^{51} |G_{i,i,15}|}{\sum_{i=1}^{51} \sum_{j=1}^{51} |G_{i,j,15}|} = 0.1616$$

The between cluster connectedness measure gives the percent of the network impact due to cross community effects, whereas the within cluster connectedness measure (excluding own asset effects) gives the percent of network impact due to within community effects. The measures state that of the 45% of between asset connectedness derived earlier, 16% is due to within community effects and 29% is due to cross community effects. Clustering therefore can account for around a third of between asset effects.

5.2 Network Visualization

I now present the network object G_{15} estimated in Section 4 graphically, with the assets grouped together according to the communities identified in Section 5.1. Also in the cluster assignment table 5.1 are asset-number pairs which identify the asset's node on the graph. The colors of the node give community membership. An edge is included if the absolute value of the cumulative IRF is above .05, corresponding to a .05% response to a 1% shock. An arrow from node 1 to node 2 indicates that a past shock to asset 1 affects current asset 2 volume. A green edge is a positive response or activation, whereas a red edge is a negative response or sup-

pression. The width of the edge reflects the strength of the effect. Solid edges denote connections between nodes in the same community, whereas dotted faint edges denote connections between nodes in different communities.

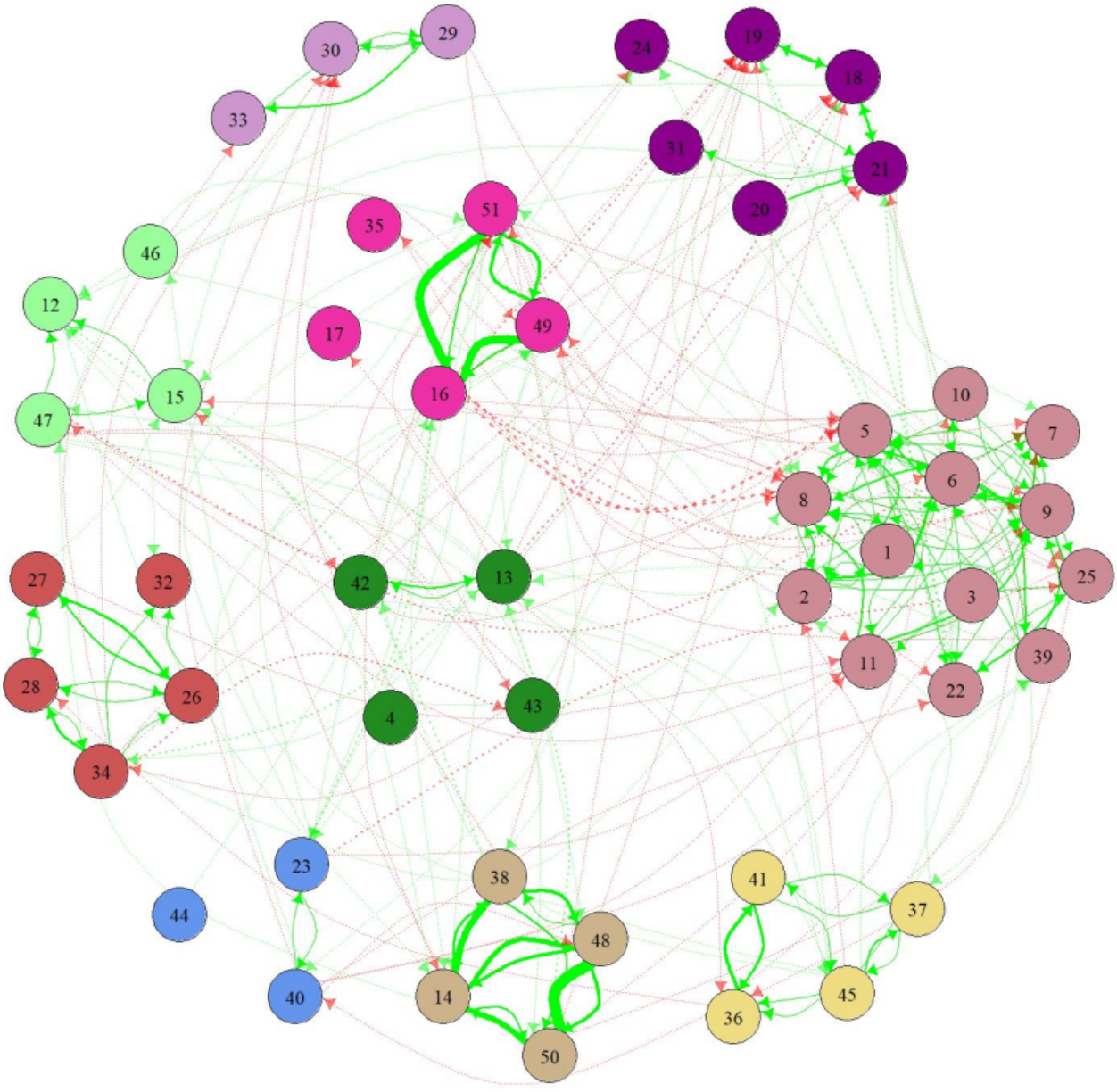


Figure 5.2: Transactions network

The graph visualization shows that for the most part, the largest effects are within community effects. The largest edges are in the broad equity 1 (pink) and energy (beige) communities.

The 10 largest within community edges are

Table 5.2: 10 largest within community edges

Leveraged Commodities	activates	Inverse Commodities	Energy	0.739
All Cap Equities	activates	Leveraged Equities	Broad Equity 1	0.582
All Cap Equities	activates	Inverse Equities	Broad Equity 1	0.532
Energy Equities	activates	Oil and Gas	Energy	0.499
Leveraged Commodities	activates	Oil and Gas	Energy	0.390
Oil and Gas	activates	Inverse Commodities	Energy	0.315
Inverse Commodities	activates	Leveraged Commodities	Energy	0.259
Leveraged Equities	activates	Inverse Equities	Broad Equity 1	0.237
Inverse Commodities	activates	Oil and Gas	Energy	0.233
Energy Equities	activates	Inverse Commodities	Energy	0.221

and the 5 largest across community edges are

Table 5.3: 5 largest across community edges

All Cap Equities	suppresses	Inflation Protected Bonds	Broad Equity 1 to Fixed Income	-0.184
All Cap Equities	suppresses	Mortgage Backed Securities	Broad Equity 1 to Fixed Income	-0.156
All Cap Equities	activates	Small Cap Blend Equities	Broad Equity 1 to High Growth Equity	0.151
Mid Cap Blend Equities	activates	Mid Cap Value Equities	Broad Equity 2 to Fixed Income	0.146
Large Cap Value Equities	activates	Mid Cap Value Equities	Broad Equity 2 to Fixed Income	0.136

6 Buy and Sell

6.1 Model

I now disaggregate volume into buy volume and sell volume and model them separately. This is done because buy volume and sell volume contain different informational content. Increasing aggregate volume above trend signals that traders have received news about an asset category, and this information is pushing up trading activity. Whether this news is good news or bad news can be inferred from disaggregated volume. If buy volume is above trend, it may be an indication that traders have received positive news and are rushing to buy, whereas the opposite is true when sell volume is above trend. Modeling buys and sells separately can give an idea of how good news in the form of a buy volume shock propagates through the network, and how this differs from how bad news in the form of a sell volume shock propagates.

To infer the trade direction, I use the TAQ quote data and the algorithm of Lee and Ready (1991). The TAQ quote data logs the time of a quote, the bid price, the bid size, the ask price, and ask size. For each transaction in the trade data, a corresponding quote is matched to it using the time stamp. Then apply the quote test. If the price of a trade is higher (lower) than the midpoint of the matched bid and ask, classify trade as a buy (sell). If the price of the trade is at the midpoint, apply the tick test. If the price is higher (lower) than the previous price, classify trade as a buy (sell). Then, separately, the buy and sell volume for ETFs is aggregated into minute bins and asset categories, the diurnal trend is removed using local linear regression, and the detrended series is logged, as with the transactions data in 2.2.

Let $n^b_{i,t,d}$ be the normalized logged dollar buy volume of asset i at minute t day d. Define $n^s_{i,t,d}$ analogously for sells. As with transactions, I remove the common contemporaneous component. Stack the buy volume data into a (minutes)(days)×assets matrix

$$N^b = \begin{bmatrix} n^b_{1,1,1} & n^b_{2,1,1} & \cdots & n^b_{51,1,1} \\ \cdots & \cdots & \cdots & \cdots \\ n^b_{1,370,1} & n^b_{2,370,1} & \cdots & n^b_{51,370,1} \\ n^b_{1,1,2} & n^b_{2,1,2} & \cdots & n^b_{51,1,2} \\ \cdots & \cdots & \cdots & \cdots \\ n^b_{1,370,2} & n^b_{2,370,2} & \cdots & n^b_{51,370,2} \\ \cdots & \cdots & \cdots & \cdots \\ n^b_{1,1,487} & n^b_{2,1,487} & \cdots & n^b_{51,1,487} \\ \cdots & \cdots & \cdots & \cdots \\ n^b_{1,370,487} & n^b_{2,370,487} & \cdots & n^b_{51,370,487} \end{bmatrix}$$

Let N^b_i be the stacked buy data for asset i, the ith column of N^b. Recall F, the stacked factor constructed before using the aggregate volume data. Then set $Y^b_i = N^b_i - D(D^TD)^{-1}D^TN^b_i$, where $D = \begin{bmatrix} 1 & F \end{bmatrix}^T$. Unstacking Y^b_i, $y^b_{i,t,d}$ gives the transformed buy volume of asset i in minute t and day d. Define $y^s_{i,t,d}$ analogously for sells.

The model for buys:

$$y^b_{i,t} = \alpha^b_i + \sum_{j=1}^{51}\sum_{k=1}^{p}[\beta^{b,b}_{i,j,k}y^b_{j,t-k} + \gamma^{b,b}_{i,j,k}\epsilon^b_{j,t-k}] + \sum_{j=1}^{51}\sum_{k=1}^{p}[\beta^{b,s}_{i,j,k}y^s_{j,t-k} + \gamma^{b,s}_{i,j,k}\epsilon^s_{j,t-k}] + \epsilon^b_{i,t}$$

The model for sells:

$$y^s_{i,t} = \alpha^s_i + \sum_{j=1}^{51}\sum_{k=1}^{p}[\beta^{s,b}_{i,j,k}y^b_{j,t-k} + \gamma^{s,b}_{i,j,k}\epsilon^b_{j,t-k}] + \sum_{j=1}^{51}\sum_{k=1}^{p}[\beta^{s,s}_{i,j,k}y^s_{j,t-k} + \gamma^{s,s}_{i,j,k}\epsilon^s_{j,t-k}] + \epsilon^b_{i,t}$$

I stack the buy and sell data, and fit the model using the iterative OLS and group lasso procedure. For the asset i buy model, define the 102 groups as 51 buy groups $G^{b,b}_{i,j} = \{\beta^{b,b}_{i,j,1}, \beta^{b,b}_{i,j,2}, ..., \beta^{b,b}_{i,j,p}, \gamma^{b,b}_{i,j,1}, \gamma^{b,b}_{i,j,2}, ..., \gamma^{b,b}_{i,j,p}\}$ and 51 sell groups $G^{b,s}_{i,j} = \{\beta^{b,s}_{i,j,1}, \beta^{b,s}_{i,j,2}, ..., \beta^{b,s}_{i,j,p}, \gamma^{b,s}_{i,j,1}, \gamma^{b,s}_{i,j,2}, ..., \gamma^{b,s}_{i,j,p}\}$. Define the 102 groups for the asset i sell model $G^{s,b}_{i,j}$, $G^{s,s}_{i,j}$ analogously.

As before, I fix the AR and MA lag length to be equal and choose a lag length by looking at in-sample MSE for different lag choices.

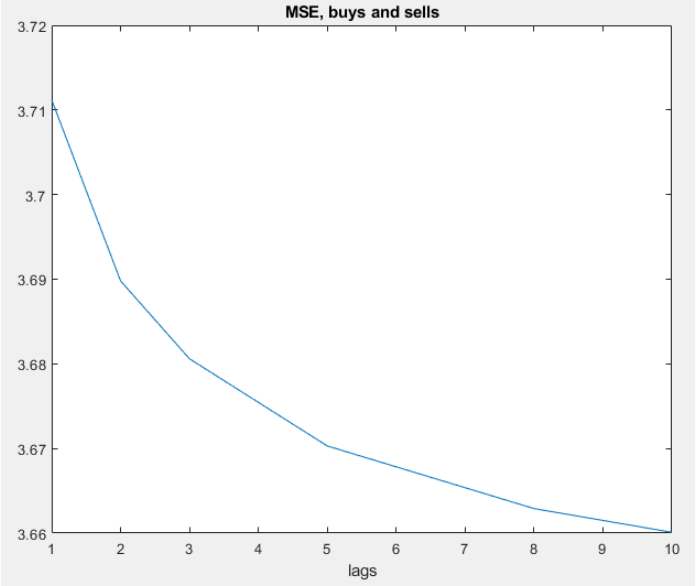


Figure 6.1: MSE for VARMA models, buys and sells

Again, there is no clear elbow, so for symmetry I choose lag $p = 3$ as before.

Stacking the data $y_t = \begin{pmatrix} y^b_t \\ y^s_t \end{pmatrix}$ and the estimated coefficients $\beta_k = \begin{pmatrix} \beta^{b,b}_k & \beta^{b,s}_k \\ \beta^{s,b}_k & \beta^{s,s}_k \end{pmatrix}$, $\gamma_k = \begin{pmatrix} \gamma^{b,b}_k & \gamma^{b,s}_k \\ \gamma^{s,b}_k & \gamma^{s,s}_k \end{pmatrix}$, I obtain the VARMA dynamics

$$y_t = \alpha + \sum_{k=1}^{3}\beta_k y_{t-k} + \sum_{k=1}^{3}\gamma_k\epsilon_{t-k} + \epsilon_t$$

The VARMA admits the VAR(∞) representation

$$y_t = \beta_0 + \sum_{k=1}^{\infty} \beta_k^{VAR} y_{t-k} + \epsilon_t, \beta_k^{VAR} = \begin{pmatrix} \beta_k^{VAR,b,b} & \beta_k^{VAR,b,s} \\ \beta_k^{VAR,s,b} & \beta_k^{VAR,s,s} \end{pmatrix}$$

the VMA(∞) representation

$$y_t = \gamma_0 + \sum_{k=1}^{\infty} \gamma_k^{VMA} \epsilon_{t-k} + \epsilon_t, \gamma_k^{VMA} = \begin{pmatrix} \gamma_k^{VMA,b,b} & \gamma_k^{VMA,b,s} \\ \gamma_k^{VMA,s,b} & \gamma_k^{VMA,s,s} \end{pmatrix}$$

and the cumulative impulse response function

$$G_k = \sum_{k'=1}^{k} \gamma_{k'}^{VMA}$$

.

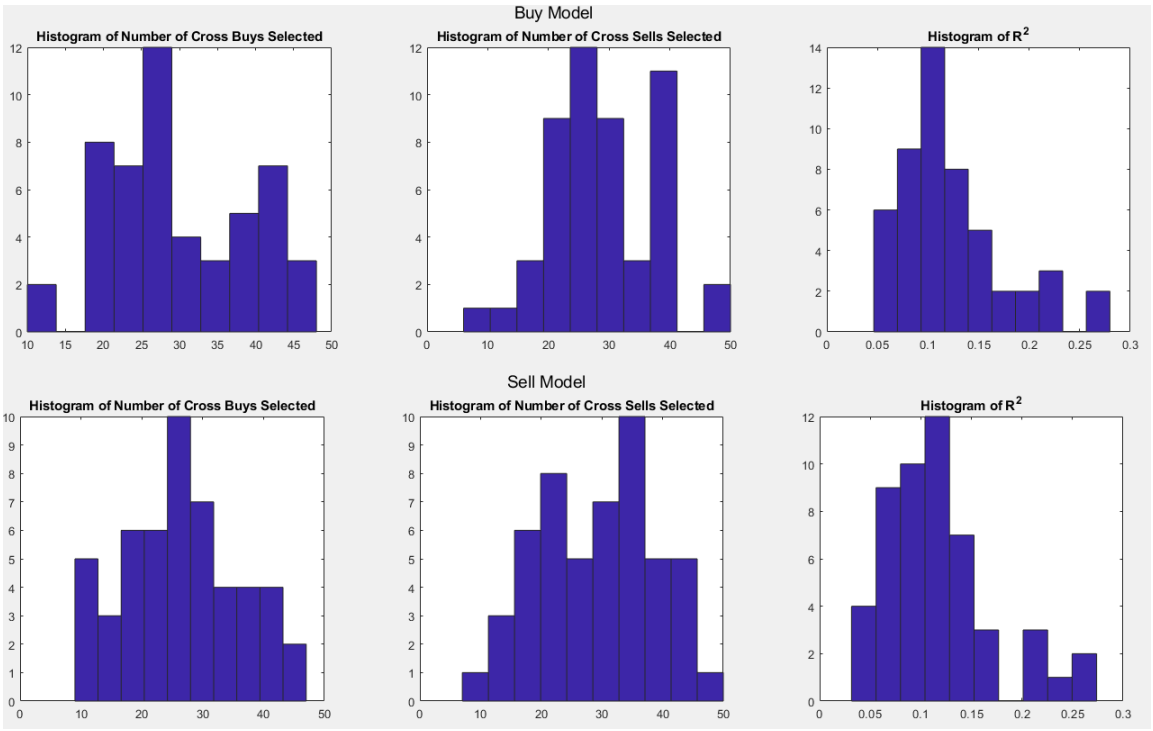


Figure 6.2: Histograms of cross assets selected and R^2, buys and sells

For the buy equations, the procedure selects on average 30 cross buys, 29 cross sells, with an average R^2 of .12. For the sell equations, the procedure selects on average 27 cross buys, 30 cross sells, with an average R^2 of .12. Compared with the aggregate volume model, which selected 36 cross assets with an average R^2 of .16, the disaggregated model tends to be slightly less populated and have slightly less predictive power.

6.2 Network Visualization

Now I present a graphic visualization of the buy and sell impulse response network at a horizon of 15 minutes G_{15}. The assets are grouped together according to the clusters defined in section 5.2, but with two nodes for each asset, a buy node labeled "b" in the shape of a circle and a sell node labeled "s" in the shape of a square. As before, an edge is included if the absolute value of the cumulative IRF is above .05. The edge between buy and sell nodes of the same asset are suppressed.

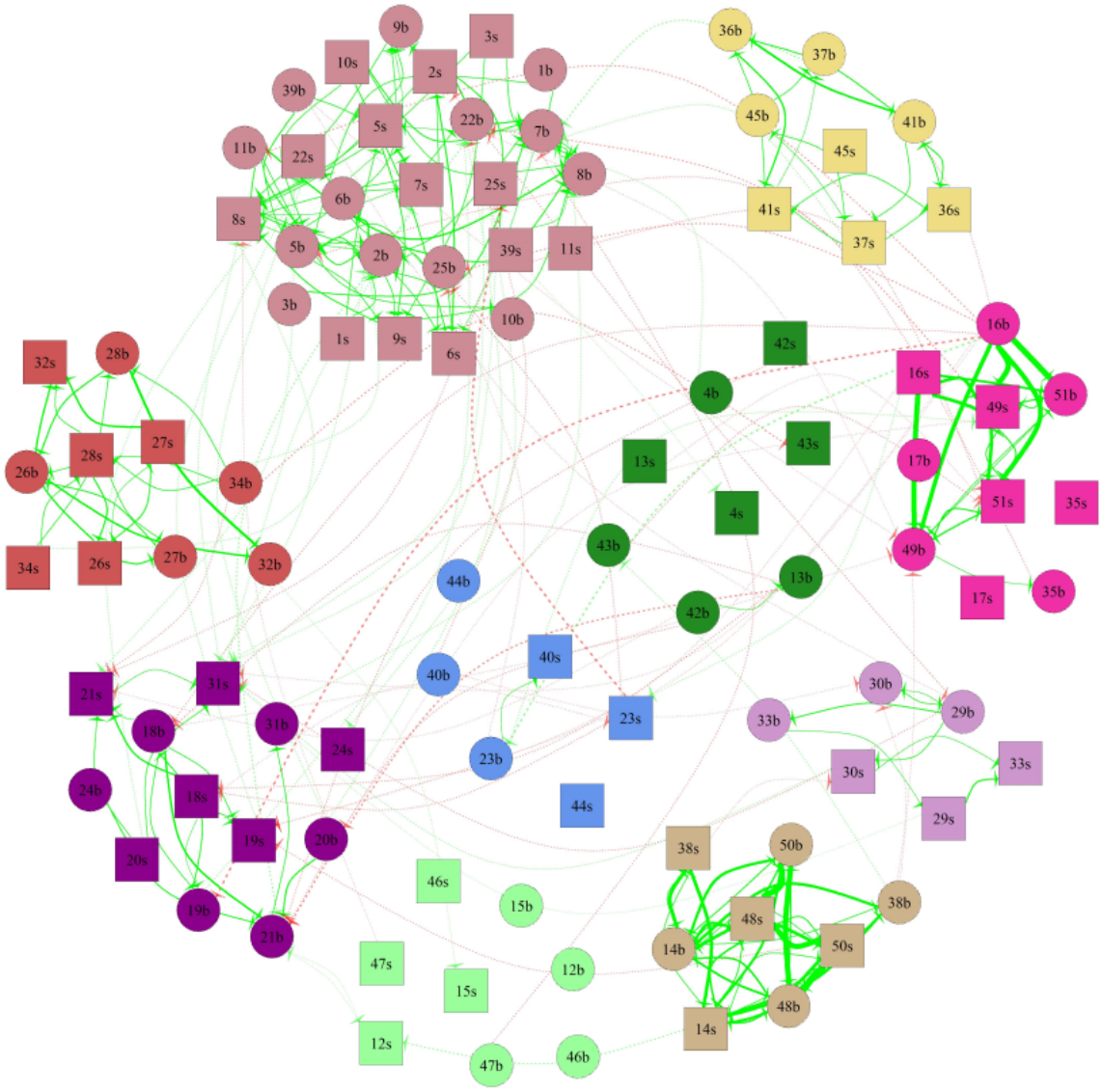
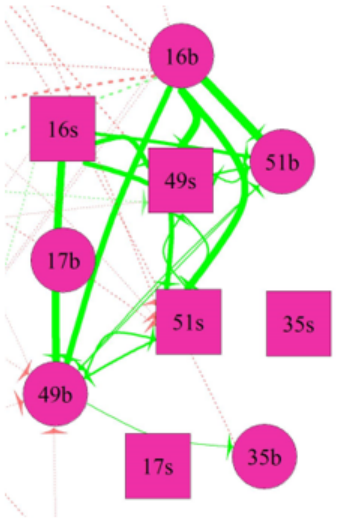


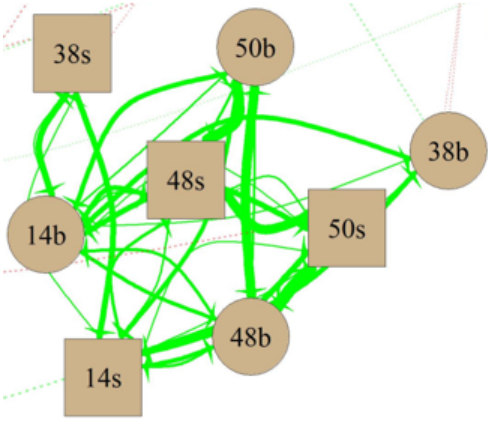
Figure 6.3: Buy and sell network



All Cap Equities Buy	activates	Leveraged Equities Buy	0.435
All Cap Equities Buy	activates	Inverse Equities Sell	0.387
All Cap Equities Sell	activates	Inverse Equities Buy	0.387
All Cap Equities Buy	activates	Leveraged Equities Sell	0.354
All Cap Equities Buy	activates	Inverse Equities Buy	0.308
All Cap Equities Sell	activates	Leveraged Equities Sell	0.245
All Cap Equities Sell	activates	Inverse Equities Sell	0.234
All Cap Equities Sell	activates	Leveraged Equities Buy	0.172
Leveraged Equities Sell	activates	Inverse Equities Buy	0.121
Inverse Equities Buy	activates	Leveraged Equities Sell	0.107

Figure 6.4: 10 largest edges, broad equity 1 cluster

Zooming in on the 10 largest edges in the active broad equity 1 cluster, the three largest edges can be interpreted as directional bets. Buys of all cap equities signal that equity is becoming more attractive, which spur future buys of leveraged equity and sells of inverse equity. Sells of all cap equity signal that equity is becoming less attractive, which spur buys of inverse equity.



Leveraged Commodities Sell	activates	Inverse Commodities Buy	0.518
Leveraged Commodities Buy	activates	Inverse Commodities Sell	0.452
Leveraged Commodities Buy	activates	Inverse Commodities Buy	0.407
Leveraged Commodities Sell	activates	Inverse Commodities Sell	0.403
Energy Equities Sell	activates	Oil and Gas Buy	0.279
Energy Equities Sell	activates	Oil and Gas Sell	0.260
Leveraged Commodities Sell	activates	Oil and Gas Buy	0.234
Leveraged Commodities Sell	activates	Oil and Gas Sell	0.225
Energy Equities Buy	activates	Oil and Gas Buy	0.225
Leveraged Commodities Buy	activates	Oil and Gas Buy	0.223

Figure 6.5: 10 largest edges, energy and energy speculation cluster

Similarly for the 10 largest edges in the energy and energy speculation cluster, the two largest edges are directional bets. Sells of leveraged commodities spurring buys of inverse commodities represent a bet against commodities. Buys of leveraged commodities spurring sells of inverse commodities represent a bet for commodities.

7 Conclusion

I estimated and examined the network graph for trading volume in broad economic sectors. The results reveal that trading volume has a good deal of predictability, with an average R^2 of .16, that about 45% of the network impacts as measured by impulse response functions are due to cross asset shocks, and that the most active trades are in levered bets on equity and commodities. The results also reveal clear clustering of assets into groups that match economic intuition.

Despite introducing the work as high frequency and high dimensional, it is really neither. High frequency traders operate at the second, not minute, horizon. And using only 51 asset categories doesn't fully utilize the power of sparse estimation techniques. This is due to liquidity issues with the data. The ETFs simply aren't liquid enough to model at a second by second frequency. This is also the rationale behind grouping the ETFs into 51 asset categories rather than attempting to model the 225 ETFs separately: the ETFs by themselves aren't liquid enough to support modeling at even the minute frequency. However, the methods and measures introduced here are well suited to studying connectedness and the propagation of shocks for large networks at a high frequency. Interest will surely grow as trading becomes more computerized and the quality and quantity of financial data increases.

8 References

Acemoglu, D., Carvalho, V. M., Ozdaglar, A., & Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. Econometrica (80), 1977–2016.

Allen, L., Bali, T., & Tang, Y. (2012). Does systemic risk in the financial sector predict future economic downturns? Review of Financial Studies (25), 3000–3036.

Basu, S., & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics (43), 1535-1567.

Barigozzi, M., Brownlees, C. (2018). Network estimation for time series. Journal of Applied Econometrics, Forthcoming.

Billio, M., Getmansky, M., Lo, A., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics (104), 535–559.

Chinco, A., Clark-Joseph, A., & Ye, M. (2018). Sparse Signals in the Cross-section of Returns. Journal of Finance, Forthcoming.

Demirer, M., Diebold, F., Liu, L., Yilmaz, K. (2018), Estimating global bank network connectedness. Journal of Applied Econonometrics (33), 1–15

Dias, G. F. & Kapetanios, G. (2017), Estimation and forecasting in vector autoregressive moving aver-

age models for rich datasets, Journal of Econometrics (202). 75-91

Diebold, F., & Yilmaz, K. (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. Economic Journal (119), 158–171.

Diebold, F., & Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of Econometrics (182), 119–134.

Dufour, J. M. & Jouini, T. (2005), Asymptotic distribution of a simple linear estimator for varma models in echelon form, in P. Duchesne & B. Remillard, eds., Statistical Modeling and Analysis for Complex Data Problems, Kluwer/Springer-Verlag, New York, chapter 11.

Fortunato, S. (2010). Community detection in graphs. Physics Reports (486) 75–174.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. Biostatistics (9), 432–441.

Gudmundsson, G. & Brownlees, C.T., Community Detection in Large Vector Autoregressions (2018). Working Paper, Available at SSRN: https://ssrn.com/abstract=3072985

Hannan, E. J. & Rissanen, J. (1982), Recursive estimation of mixed autoregressive-moving average order, Biometrika (69), 81–94.

Hautsch, N., Schaumburg, J., & Schienle, M. (2015). Financial network systemic risk contributions. Review of Finance (19), 685-738.

Karpoff, J. M. (1987), The Relation Between Price Changes and Trading Volume: A Survey. The Journal of Financial and Quantitative Analysis (22), 109-126.

Lee, C.M. & Ready, J.M., (1991). Inferring Trade Direction from Intraday Data, Journal of Finance (46), 733-746.

Lo, A.W. & Wang, J. (2009), Stock Market Trading Volume, in Y. Aıt-Sahalia & L. Hansen, eds., The Handbook of Financial Econometrics, New York: North-Holland, chapter 9.

Meinshausen, N. & Buhlmann, P. (2006). High dimensional graphs and variable selection with the LASSO. The Annals of Statistics (34), 1436-1462.

Nicholson, W.B., Matteson, D.S, & Bien, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. International Journal of Forecasting (33), 627-651.

Peng, J., Wang, P., Zhou, N., & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association (104) 735-746.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (58), 267–288.

Xiu, D., Giglio, S. & Feng, G. (2017), Taming the Factor Zoo. Chicago Booth Research Paper No. 17-04.

Yuan, M. & Lin, Y. (2007) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society (68), 49–67.

9 Appendix - Asset categories

Category	Symbol	ETF Name
Corporate Bonds	LQD	iShares iBoxx $ Investment Grade Corporate Bond ETF
	VCSH	Vanguard Short-Term Corporate Bond ETF
	VCIT	Vanguard Intermediate-Term Corporate Bond ETF
	VCLT	Vanguard Long-Term Corporate Bond ETF
Emerging Market Bonds	EMB	iShares J.P. Morgan USD Emerging Markets Bond ETF
	EMLC	VanEck Vectors J.P. Morgan EM Local Currency Bond ETF
	EBND	SPDR Barclays Capital Emerging Markets Local Bond ETF
Government Bonds	SHY	iShares 1-3 Year Treasury Bond ETF
	IEI	iShares 3-7 Year Treasury Bond ETF
	IEF	iShares 7-10 Year Treasury Bond ETF
	GOVT	iShares U.S. Treasury Bond ETF
High Yield Bonds	JNK	SPDR Barclays Capital High Yield Bond ETF
	HYG	iShares iBoxx $ High Yield Corporate Bond ETF
Inflation Protected Bonds	TIP	iShares TIPS Bond ETF
	VTIP	Vanguard Short-Term Inflation-Protected Securities ETF
	SCHP	Schwab U.S. TIPS ETF
International Government Bonds	BWX	SPDR Barclays Intl Treasury Bond ETF
	PCY	PowerShares Emerging Markets Sovereign Debt
	VWOB	Vanguard Emerging Markets Government Bond ETF
	IGOV	iShares International Treasury Bond ETF
	BWZ	SPDR Barclays Capital Short Term International Treasury Bond ETF
	ISHG	iShares 1-3 Year International Treasury Bond ETF
Money Market	SHV	iShares Short Treasury Bond ETF
	SCHO	Schwab Short-Term U.S. Treasury ETF
	BIL	SPDR Barclays 1-3 Month T-Bill ETF
	VGSH	Vanguard Short-Term Treasury ETF
	NEAR	iShares Short Maturity Bond ETF
	GSY	Guggenheim Ultra Short Duration ETF
Mortgage Backed Securities	MBB	iShares MBS ETF
	MBG	SPDR Barclays Capital Mortgage Backed Bond ETF
	VMBS	Vanguard Mortgage-Backed Securities ETF
	CMBS	iShares CMBS ETF
National Munis	MUB	iShares National Muni Bond ETF
	CMF	iShares California Muni Bond ETF
	NYF	iShares New York AMT-Free Muni Bond ETF
	SHM	SPDR Barclays Short Term Municipal Bond
	TFI	SPDR Barclays Capital Municipal Bond ETF
	HYD	VanEck Vectors High-Yield Municipal Index ETF
	ITM	VanEck Vectors AMT-Free Intermediate Municipal Index ETF
	PZA	PowerShares National AMT-Free Municipal Bond Portfolio
Preferred Stock/Convertible Bonds	PFF	iShares U.S. Preferred Stock ETF
	PGX	PowerShares Preferred Portfolio
	CWB	SPDR Barclays Capital Convertible Bond ETF
Total Bond Market	AGG	iShares Core U.S. Aggregate Bond ETF
	BND	Vanguard Total Bond Market ETF
	SCHZ	Schwab U.S. Aggregate Bond ETF

Category	Symbol	ETF Name
Commodities	DBC	PowerShares DB Commodity Index Tracking Fund
	GSG	iShares S&P GSCI Commodity-Indexed Trust
	DJP	iPath Dow Jones-UBS Commodity ETN
	GCC	WisdomTree Continuous Commodity Index Fund
	USCI	United States Commodity Index Fund
Gold And Silver	GLD	SPDR Gold Shares ETF
	IAU	iShares Gold Trust
	SLV	iShares Silver Trust
Oil and Gas	USO	United States Oil Fund
	UNG	United States Natural Gas Fund
	UGA	United States Gasoline Fund
Currency	UUP	PowerShares DB US Dollar Index Bullish Fund
	USDU	WisdomTree Bloomberg U.S. Dollar Bullish Fund
	FXE	CurrencyShares Euro Trust
	FXB	CurrencyShares British Pound Sterling Trust
	FXC	CurrencyShares Canadian Dollar Trust ETF
	FXF	CurrencyShares Swiss Franc
	FXY	CurrencyShares Japanese Yen Trust
All Cap Equities	SPY	SPDR S&P 500 ETF
	QQQ	PowerShares QQQ ETF
	DIA	SPDR Dow Jones Industrial Average ETF
	VTI	Vanguard Total Stock Market ETF
Large Cap Blend Equities	IWB	iShares Russell 1000 ETF
	VV	Vanguard Large Cap ETF
	IVV	iShares Core S&P 500 ETF
Large Cap Growth Equities	IWF	iShares Russell 1000 Growth ETF
	VUG	Vanguard Growth ETF
	IVW	iShares S&P 500 Growth ETF
Large Cap Value Equities	IWD	iShares Russell 1000 Value ETF
	VTV	Vanguard Value ETF
	IVE	iShares S&P 500 Value ETF
Mid Cap Blend Equities	IWR	iShares Russell Midcap ETF
	VO	Vanguard Mid-Cap Index ETF
	IJH	iShares Core S&P Mid-Cap ETF
Mid Cap Growth Equities	IWP	iShares Russell Mid-Cap Growth ETF
	VOT	Vanguard Mid-Cap Growth ETF
	IJK	iShares S&P MidCap 400 Growth ETF
Mid Cap Value Equities	IWS	iShares Russell Mid-Cap Value ETF
	VOE	Vanguard Mid-Cap Value ETF
	IJJ	iShares S&P MidCap 400 Value ETF
Small Cap Blend Equities	IWM	iShares Russell 2000 ETF
	VB	Vanguard Small Cap ETF
	IJR	iShares Core S&P Small-Cap ETF
Small Cap Growth Equities	IWO	iShares Russell 2000 Growth ETF
	VBK	Vanguard Small Cap Growth ETF
	IJT	iShares S&P SmallCap 600 Growth ETF
Small Cap Value Equities	IWN	iShares Russell 2000 Value ETF
	VBR	Vanguard Small Cap Value ETF
	IJS	iShares S&P SmallCap 600 Value ETF

Category	Symbol	ETF Name
Asia Pacific Equities	VPL	Vanguard FTSE Pacific ETF
	EPP	iShares MSCI Pacific ex Japan ETF
	EWY	iShares MSCI South Korea ETF
	EWT	iShares MSCI Taiwan ETF
	EWH	iShares MSCI Hong Kong ETF
	EWS	iShares MSCI Singapore ETF
	EWM	iShares MSCI Malaysia ETF
	EIDO	iShares MSCI Indonesia ETF
China Equities	FXI	iShares China Large-Cap ETF
	GXC	SPDR S&P China ETF
	MCHI	iShares MSCI China ETF
Emerging Market Equities	EEM	iShares MSCI Emerging Markets ETF
	IEMG	iShares Core MSCI Emerging Markets ETF
	VWO	Vanguard FTSE Emerging Markets ETF
	SCHE	Schwab Emerging Markets Equity ETF
Europe Equities	VGK	Vanguard FTSE Europe ETF
	IEV	iShares Europe ETF
	EZU	iShares MSCI EMU ETF
	EWG	iShares MSCI Germany ETF
	EWU	iShares MSCI United Kingdom ETF
	EWL	iShares MSCI Switzerland ETF
	EWP	iShares MSCI Spain ETF
	EWQ	iShares MSCI France ETF
	EWI	iShares MSCI Italy ETF
	EWN	iShares MSCI Netherlands ETF
	EWK	iShares MSCI Belgium ETF
Foreign Large Cap Equities	EFA	iShares MSCI EAFE ETF
	IEFA	iShares Core MSCI EAFE ETF
	VEU	Vanguard FTSE All-World ex-US ETF
	SCHF	Schwab International Equity ETF
	FNDF	Schwab Fundamental International Large Company Index ETF
Foreign Small and Mid Cap Equities	SCZ	iShares MSCI EAFE Small-Cap ETF
	VSS	Vanguard FTSE All-World ex-US Small-Cap ETF
	FNDC	Schwab Fundamental International Small Company Index ETF
	SCHC	Schwab International Small-Cap Equity ETF
	GWX	SPDR S&P International Small Cap ETF
India Equities	INDA	iShares MSCI India ETF
	PIN	PowerShares India Portfolio
	SCIF	VanEck Vectors India Small-Cap Index ETF
	SMIN	iShares MSCI India Small-Cap ETF
Japan Equities	EWJ	iShares MSCI Japan ETF
	SCJ	iShares MSCI Japan Small Cap ETF
	DXJ	WisdomTree Japan Hedged Equity Fund
Latin America Equities	EWZ	iShares MSCI Brazil ETF
	EWW	iShares MSCI Mexico ETF
	ILF	iShares Latin American 40 ETF

Category	Symbol	ETF Name
Building and Construction Equities	XHB	SPDR S&P Homebuilders ETF
	ITB	iShares U.S. Home Construction ETF
	PKB	PowerShares Dynamic Building & Construction
Consumer Discretionary Equities	XLY	Consumer Discretionary Select Sector SPDR Fund
	XRT	SPDR S&P Retail ETF
	VCR	Vanguard Consumer Discretionary ETF
	IYC	iShares US Consumer Services ETF
	FDIS	Fidelity MSCI Consumer Discretionary Index ETF
Consumer Staples Equities	XLP	Consumer Staples Select Sector SPDR Fund
	VDC	Vanguard Consumer Staples ETF
	IYK	iShares U.S. Consumer Goods ETF
	FSTA	Fidelity MSCI Consumer Staples Index ETF
Energy Equities	XLE	Energy Select Sector SPDR Fund
	XOP	SPDR S&P Oil & Gas Exploration & Production ETF
	XES	SPDR S&P Oil & Gas Equipment & Services ETF
	VDE	Vanguard Energy ETF
	OIH	VanEck Vectors Oil Services ETF
	KOL	VanEck Vectors Coal ETF
	IYE	iShares U.S. Energy ETF
	IEO	iShares U.S. Oil & Gas Exploration & Production ETF
	FENY	Fidelity MSCI Energy Index ETF
Financial Equities	XLF	Financial Select Sector SPDR Fund
	KBE	SPDR S&P Bank ETF
	KRE	SPDR S&P Regional Banking ETF
	IYF	iShares U.S. Financials ETF
	IYG	iShares U.S. Financial Services ETF
	FNCL	Fidelity MSCI Financials Index ETF
Health and Biotech Equities	XLV	Health Care Select Sector SPDR Fund
	XBI	SPDR S&P Biotech ETF
	XPH	SPDR S&P Pharmaceuticals ETF
	IBB	iShares Nasdaq Biotechnology ETF
	IYH	iShares U.S. Healthcare ETF
	IHI	iShares U.S. Medical Devices ETF
	VHT	Vanguard Healthcare ETF
	FHLC	Fidelity MSCI Health Care Index ETF
Industrials Equities	XLI	Industrial Select Sector SPDR Fund
	VIS	Vanguard Industrial ETF
	IYJ	iShares U.S. Industrials ETF
	FIDU	Fidelity MSCI Industrials Index ETF
Materials Equities	XLB	Materials Select Sector SPDR Fund
	XME	SPDR S&P Metals & Mining ETF
	VAW	Vanguard Materials ETF
	IYM	iShares U.S. Basic Materials ETF
	FMAT	Fidelity MSCI Materials Index ETF

Category	Symbol	ETF Name
Master Limited Partnership Equities	AMLP	Alerian MLP ETF
	AMJ	JPMorgan Alerian MLP Index ETN
Technology Equities	XLK	Technology Select Sector SPDR Fund
	XSD	SPDR S&P Semiconductor ETF
	VGT	Vanguard Information Technology ETF
	IYW	iShares U.S. Technology ETF
	SOXX	iShares PHLX Semiconductor ETF
	SMH	VanEck Vectors Semiconductor ETF
	FTEC	Fidelity MSCI Information Technology Index ETF
Utilities Equities	XLU	Utilities Select Sector SPDR Fund
	VPU	Vanguard Utilities ETF
	IDU	iShares U.S. Utilities ETF
	FUTY	Fidelity MSCI Utilities Index ETF
Real Estate	IYR	iShares U.S. Real Estate ETF
	VNQ	Vanguard Real Estate Index Fund
	SCHH	Schwab US REIT ETF
Inverse Bonds	TBF	ProShares Short 20+ Year Treasury
	TBX	ProShares Short 7-10 Year Treasury
	TBT	UltraShort Barclays 20+ Year Treasury
	PST	UltraShort Barclays 7-10 Year Treasury
	TTT	UltraPro Short 20+ Year Treasury
	TMV	Direxion Daily 20-Year Treasury Bear 3X
	TYO	Direxion Daily 7-10 Year Treasury Bear 3X
Inverse Commodities	DGLD	VelocityShares 3x Inverse Gold ETN
	DSLV	VelocityShares 3x Inverse Silver ETN
	ZSL	ProShares UltraShort Silver
	GLL	ProShares UltraShort Gold
	SCO	ProShares UltraShort Bloomberg Crude Oil
	DGAZ	VelocityShares 3x Inverse Natural Gas
Inverse Equities	SH	ProShares Short S&P 500
	SDS	ProShares UltraShort S&P 500
	SPXU	ProShares UltraPro Short S&P 500
Leveraged Commodities	UGLD	VelocityShares 3x Long Gold ETN
	USLV	VelocityShares 3x Long Silver ETN
	AGQ	ProShares Ultra Silver
	UGL	ProShares Ultra Gold
	UCO	ProShares Ultra Bloomberg Crude Oil
	UGAZ	VelocityShares 3x Long Natural Gas
Leveraged Equities	SSO	ProShares Ultra S&P 500
	UPRO	ProShares UltraPro S&P 500

CHAPTER 2:

How useful are machine learning tools in predicting high frequency returns?

An Qi
University of Chicago
Booth School of Business

June 12, 2019

Abstract

This work explores the predictability of high frequency returns for 63 exchange traded funds by applying recently popularized machine learning techniques. I consider factor models, LASSO, and random forests. From a baseline of predicting an asset's return using its own return history, I build a sequence of predictive models of increasing complexity, and evaluate them using out of sample predictability. The results show that using the return history of all 63 assets improves out of sample predictability compared with the baseline model that uses only own return history. Incorporating market microstructure information in the form of volume, depth, and spread data into the model improves out of sample predictability further. However, introducing interactions, whether included in a linear model or created using a random forest, does not improve out of sample predictability. This predictability can be turned into substantial trading revenue, but transaction costs would mitigate these profits.

1 Introduction

This work explores the predictability of high frequency returns for 63 exchange traded funds, which are chosen to represent broad sectors of the economy. I apply recently popularized machine learning techniques for regression and prediction: factor models, LASSO, and random forests. Candidate predictor variables include past return, volume, depth, and spread information. I build a sequence of predictive models of increasing complexity, and evaluate them using out of sample predictability. I find that LASSO models work much better compared with factor models in incorporating the information from the same predictor set. The baseline OLS model using only past own asset return history yields no out of sample predictability, with an average out of sample R^2 of 0.0029 over the 63 ETFs. Using the past return history of all 63 assets in a LASSO model improves this number to 0.0192, and using microstructure information in the form of volume, depth, and spread data in addition to returns further improves this to 0.0347. Incorporating nonlinearities through two way interaction effects or through the random forest fails to improve out of sample predictability, and so are omitted for reasons of parsimony. I then create a trading strategy with the best performing model, and demonstrate that a hypothetical trader can earn substantial daily returns (6.17%), although transaction costs would mitigate these profits.

The data for the analysis comes from the Trade and Quote (TAQ) database of the Wharton Research Data Service (WRDS). TAQ first began tracking intraday second by second trading data in 1993 and millisecond data a decade later in 2003. Factor models for economic forecasting were proposed by Geweke (1977). The least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996). The random forest was proposed by Breiman (2001). This paper seeks to tie these strands together: do recently popularized machine learning techniques, applied to now readily available intraday trading data, have value in predicting high frequency returns?

There is a large, burgeoning literature exploring return predictability. Early work demonstrated that stock returns were close to unpredictable and therefore that market prices already contain most information about fundamental value. This view, now known as the efficient markets hypothesis, is epitomized by Fama (1970) and Malkiel's (1973) book, A Random Walk Down Wall Street. These findings were extensively revised by later work, which showed that certain stock level characteristics predict returns. These include dividend-price ratio (Rozeff, 1984), earnings-price ratio (Campbell and Shiller, 1988), and book-to-market ratio (Kothari and Shanken, 1997). Recently, works in high dimensional, high frequency returns have been emerging. Xiu et al. (2018) apply machine learning techniques to predicting asset returns using stock level characteristics.

Chinco et al. (2018) apply LASSO to the prediction of minute by minute returns using past return history. This work extends Xiu (2018) by applying these machine learning techniques to the intraday high frequency trading setting. It extends Chinco (2018) by incorporating economic factors and potential nonlinearities in the predictive models. The use of high frequency microstructure data also represents a novel contribution.

Section 2 describes the data and the predictive variables. Section 3 builds predictive models for returns. Section 4 presents trading strategies created from the best performing predictive model. Section 5 concludes.

2 Data

The returns I examine are returns for 63 exchange traded funds (ETFs). ETFs are marketable securities that tracks an index, a commodity, bonds, or a basket of assets like an index fund. Unlike mutual funds, an ETF trades like a common stock on a stock exchange. The fund owns the underlying assets (shares of stock, bonds, oil futures, gold bars, foreign currency, etc.) and divides ownership of those assets into shares.

The funds are selected for scope and liquidity. To measure trading activity in all cap equities, I can use the data from a few of the most highly traded equity ETFs (SPY, QQQ, DIA) rather than using the entire universe of stocks. Similarly for corporate bonds, treasuries, sector equities, and other broad asset classes. The ETFs include fixed income ETFs, which hold different types of bonds such as corporate bonds or treasuries; an aggregate commodity ETF focusing on a range of commodity futures contracts as well as ETFs that track a single commodity; ETFs that track the yen and euro; broad based equity ETFs tracking indices that include stocks of a certain market cap or growth/value characteristic; sector and international equity ETFs; and ETFs which place a -1x or a 2x bet on the S&P 500. The relatively small number of funds selected give a diversity of assets that reflect activity in broad sectors of the economy. These funds will respond to movements in a wide variety of economic variables, such as interest rate changes, exchange rate changes, industry specific shocks, etc... In addition, ETFs tend to be very liquid assets. Liquidity mitigates the impact of stale information and makes ETFs easier to model compared with the underlying assets.

The following table lists the 63 ETFs:

Table 2.1: Exchange traded funds included in analysis

LQD	iShares iBoxx $ Investment Grade Corporate Bond ETF	EWY	iShares MSCI South Korea ETF
EMB	iShares J.P. Morgan USD Emerging Markets Bond ETF	MCHI	iShares MSCI China ETF
SHY	iShares 1-3 Year Treasury Bond ETF	EEM	iShares MSCI Emerging Markets ETF
IEI	iShares 3-7 Year Treasury Bond ETF	EZU	iShares MSCI EMU ETF
IEF	iShares 7-10 Year Treasury Bond ETF	EWG	iShares MSCI Germany ETF
JNK	SPDR Barclays Capital High Yield Bond ETF	EWU	iShares MSCI United Kingdom ETF
TIP	iShares TIPS Bond ETF	EFA	iShares MSCI EAFE ETF
BWX	SPDR Barclays Intl Treasury Bond ETF	INDA	iShares MSCI India ETF
SHV	iShares Short Treasury Bond ETF	EWJ	iShares MSCI Japan ETF
MBB	iShares MBS ETF	EWZ	iShares MSCI Brazil ETF
MUB	iShares National Muni Bond ETF	EWW	iShares MSCI Mexico ETF
PFF	iShares U.S. Preferred Stock ETF	XHB	SPDR S&P Homebuilders ETF
AGG	iShares Core U.S. Aggregate Bond ETF	XLY	Consumer Discretionary Select Sector SPDR Fund
BND	Vanguard Total Bond Market ETF	XRT	SPDR S&P Retail ETF
DBC	PowerShares DB Commodity Index Tracking Fund	XLP	Consumer Staples Select Sector SPDR Fund
IAU	iShares Gold Trust	XLE	Energy Select Sector SPDR Fund
SLV	iShares Silver Trust	XOP	SPDR S&P Oil & Gas Exploration & Production ETF
USO	United States Oil Fund	XLF	Financial Select Sector SPDR Fund
UNG	United States Natural Gas Fund	KBE	SPDR S&P Bank ETF
FXE	CurrencyShares Euro Trust	KRE	SPDR S&P Regional Banking ETF
FXY	CurrencyShares Japanese Yen Trust	XLV	Health Care Select Sector SPDR Fund
SPY	SPDR S&P 500 ETF	XBI	SPDR S&P Biotech ETF
QQQ	PowerShares QQQ ETF	XLI	Industrial Select Sector SPDR Fund
DIA	SPDR Dow Jones Industrial Average ETF	XLB	Materials Select Sector SPDR Fund
IWB	iShares Russell 1000 ETF	XME	SPDR S&P Metals & Mining ETF
IWF	iShares Russell 1000 Growth ETF	XLK	Technology Select Sector SPDR Fund
IWD	iShares Russell 1000 Value ETF	XLU	Utilities Select Sector SPDR Fund
IWR	iShares Russell Midcap ETF	IYR	iShares U.S. Real Estate ETF
IWP	iShares Russell Mid-Cap Growth ETF	SH	ProShares Short S&P 500
IWS	iShares Russell Mid-Cap Value ETF	SSO	ProShares Ultra S&P 500
IWM	iShares Russell 2000 ETF		
IWO	iShares Russell 2000 Growth ETF		
IWN	iShares Russell 2000 Value ETF		

The TAQ trade data logs the time a trade occurred, the transaction price, and the transaction volume. The quote data logs the time of a quote, the bid price, the bid size, the ask price, and ask size. For each transaction in the trade data, a corresponding quote is matched to it using the time stamp. Using the quote, the algorithm of Lee and Ready (1991) is used to infer the trade direction of a transaction. First, apply the quote test. If the price of a trade is higher (lower) than the midpoint of the matched bid and ask, classify trade as a buy (sell). If the price of the trade is at the midpoint, apply the tick test. If the price is higher (lower) than the previous price, classify trade as a buy (sell).

The resolution of the models is at the minute frequency. To create minute by minute returns for an asset, I first create a minute by minute price series by taking the minute price to be the midpoint of the bid and ask of the last quote within the minute bin, and then taking the log first difference. If there is no quote in the minute bin, the price is set to the midquote of the last available quote from a previous minute bin. I use the midquote because it is a commonly used measure of the fair market value of an asset. A common assumption is that the quotes are set symmetrically about the expected value of the asset conditional on all public information. See, for instance, Hasbrouck (1991). Modeling returns with transaction prices would have resulted in microstructure noise from the bid ask bounce being fit. Using midquotes therefore gives a more accurate measure of the true return.

To motivate the construction of predictor variables, I look to works in market microstructure. In perfectly competitive markets, equilibrium prices reflect the demand curves of investors with perfect information. However, short term deviations between transaction prices and fundamental values arise because of frictions reflecting asymmetric information and strategic behavior. Market microstructure is the process by which investors' asymmetric information and latent demand is translated into prices. Models of this type were developed by Copeland and Galai (1983), Glosten and Milgrom (1985), Easley and O'Hara (1987), and Lee et. al (1993). Assume there are noise traders and informed traders. Informed traders generally trade on one side of the market. Trade direction and volume provide signals to market makers who update their price expectations accordingly. Compared with noise traders, informed traders prefer to trade in larger quantities, and the larger the trade size, the more likely it is that the market maker is trading with an informed trader. This suggests that the number of trades, and not just the volume of trades, provide information. The bid-ask spread and market depth also reflect market information. The market maker expects to lose in trades with informed traders, but gains in trades with noise traders from the bid-ask spread. The optimal bid-ask spread reflects this tradeoff: the higher the bid-ask spread, the lower the loss to informed traders, at the cost of the profit from noise traders with reservation prices inside the spread. If the market maker sees an increase in trading volume, potentially from informed traders, the market maker increases spread and decreases depth to protect itself from information asymmetry.

Motivated by the discussion on market microstructure, I construct 18 variables reflecting volume, depth, and spread from the trade and quote data. In addition to returns, this gives a total of 19 predictive variables.

Let $r_{i,t,d}$ be the returns for asset i, minute t, and day d.

From the trade data, I include the number of buy transactions within the minute ($bt_{i,t,d}$), the dollar volume of buys within the minute ($bv_{i,t,d}$), the number of sells ($st_{i,t,d}$), and the dollar volume of sells ($sv_{i,t,d}$). From the quote data, I include the average bid size within the minute bin ($bsa_{i,t,d}$), the bid size from the last quote within the minute bin ($bsl_{i,t,d}$), the average ask size ($asa_{i,t,d}$), and last ask size ($asl_{i,t,d}$). If there is no quote within the minute, the average bid size and last bid size is set to that from the last available quote, and similarly with the average ask size and last ask size.

Because there is typically a U-shaped diurnal trend in trading activity, as trading during the morning and afternoon is higher than trading during midday, the 8 variables defined above are detrended. First, for each asset, define a transactions trend $tt_{i,t} = \frac{1}{D}\sum_{d=1}^{D}(bv_{i,t,d} + sv_{i,t,d})$, the mean over trading days of the sum of buy volume and sell volume. Let $\tilde{tt}_{i,t}$ be the fit of $tt_{i,t}$ to a quadratic B-spline with 3 knots. Then set

$$bt_{i,t,d} \to \log\left(\frac{1+bt_{i,t,d}}{1+\tilde{tt}_{i,t}}\right)$$
$$bv_{i,t,d} \to \log\left(\frac{1+bv_{i,t,d}}{1+\tilde{tt}_{i,t}}\right)$$
$$st_{i,t,d} \to \log\left(\frac{1+st_{i,t,d}}{1+\tilde{tt}_{i,t}}\right)$$
$$sv_{i,t,d} \to \log\left(\frac{1+sv_{i,t,d}}{1+\tilde{tt}_{i,t}}\right)$$

Rather than fitting a separate trend for each of the four variables, the above procedure forces a common trend, as it makes little sense for the diurnal trend for buys and sells of an asset to look different. Similarly, there is no reason for number of transactions and volume to have different daily trends. The log transformation is needed to account for the heavy right tail of the data, with 1 added to account for the 0's in the data. This transformation changes the variables' interpretation to the effect on returns of a percentage above or below daily trend of the predictor.

Define a depth trend $dt_{i,t} = \frac{1}{D}\sum_{d=1}^{D}(bsa_{i,t,d} + asa_{i,t,d})$, the mean over trading days of the sum of average bid size and average ask size. Let $\tilde{dt}_{i,t}$ be the fit of $dt_{i,t}$ to a quadratic B-spline with 3 knots. Then apply the same transformation as before.

$$bsa_{i,t,d} \to \log\left(\frac{1+bsa_{i,t,d}}{1+\tilde{dt}_{i,t}}\right)$$
$$bsl_{i,t,d} \to \log\left(\frac{1+bsl_{i,t,d}}{1+\tilde{dt}_{i,t}}\right)$$
$$asa_{i,t,d} \to \log\left(\frac{1+asa_{i,t,d}}{1+\tilde{dt}_{i,t}}\right)$$

$asl_{i,t,d} \rightarrow \log\left(\frac{1+asl_{i,t,d}}{1+\tilde{dt}_{i,t}}\right)$

Again, the four variables are forced to have a common trend.

The following plots show the estimated diurnal transactions trend and depth trend for the most liquid ETF, the SPY (SPDR S&P 500 ETF). The U-shape is clearly visible.

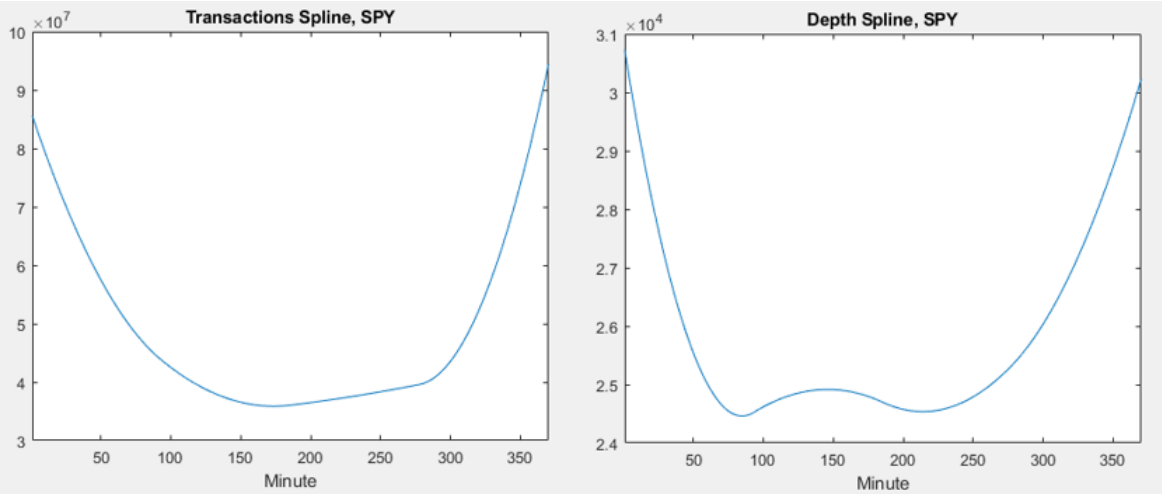


Figure 2.1: Diurnal trends for SPY (SPDR S&P 500 ETF)

Define two bid-ask spread variables by computing $\frac{\text{Ask Price}-\text{Bid Price}}{\text{Midquote}}$ for each quote within the minute. Let average bid-ask spread ($basa_{i,t,d}$) be the average of these values within the minute bin, and last bid-ask spread ($basl_{i,t,d}$) be the value from the last quote in the minute bin. If there is no quote within the minute, set the variables to the value from the last available quote

Define two variables that give relative market depth by computing $\frac{\text{Bid Size}}{\text{Bid Size}+\text{Ask Size}}$ for each quote within the minute. Let average bid tilt ($bta_{i,t,d}$) be the average of these values within the minute bin, and last bid tilt ($btl_{i,t,d}$) be the value from the last quote in the minute bin. If there is no quote within the minute, set the variables to the value from the last available quote.

Define two variables that give relative trading activity by taking the sum of the raw dollar buy volume and the sum of the raw dollar sell volume within the minute. Define buy tilt ($bti_{i,t,d}$) to be $\frac{\text{Dollar Buy Volume}}{\text{Dollar Buy Volume}+\text{Dollar Sell Volume}}$ if there are transactions within the minute bin and 0 otherwise. Define sell tilt ($sti_{i,t,d}$) to be $\frac{\text{Dollar Sell Volume}}{\text{Dollar Buy Volume}+\text{Dollar Sell Volume}}$ if there are transactions within the minute bin and

0 otherwise. Unlike with the bid tilt variables, where it wasn't necessary to construct an ask tilt variable because the sum of the bid and ask variables is always 1, both variables are needed here because of the presence of minute bins where no transactions occur, in which case both variables are 0.

Finally, define four additional spread variables for when transactions occur. For each buy transaction within the minute bin, calculate $\frac{\text{Transaction Price-Midquote}}{\text{Midquote}}$. Define average trans-mid spread for buys ($tmsba_{i,t,d}$) by taking the weighted average of these values, with weights given by $\frac{\text{Current Transaction Dollar Buy Volume}}{\text{Total Minute Dollar Buy Volume}}$. Define last trans-mid spread for buys ($tmsbl_{i,t,d}$) by taking the value from the last buy transaction in the minute bin. If there are no buys in the minute bin, set both variables to 0. Then for each sell transaction within the minute bin, calculate $\frac{\text{Transaction Price-Midquote}}{\text{Midquote}}$. Define average trans-mid spread for sells ($tmssa_{i,t,d}$) by taking the weighted average of these values, with weights given by $\frac{\text{Current Transaction Dollar Sell Volume}}{\text{Total Minute Dollar Sell Volume}}$. Define last trans-mid spread for sells ($tmssl_{i,t,d}$) by taking the value from the last sell transaction in the minute bin. If there are no sells in the minute bin, set both variables to 0.

This gives the 19 predictive variables, listed below for easy reference.

- Returns ($r_{i,t,d}$)
- Buy Transactions ($bt_{i,t,d}$)
- Buy Dollar Volume ($bv_{i,t,d}$)
- Sell Transactions ($st_{i,t,d}$)
- Sell Dollar Volume ($sv_{i,t,d}$)
- Bid Size (Average) ($bsa_{i,t,d}$)
- Bid Size (Last) ($bsl_{i,t,d}$)
- Ask Size (Average) ($asa_{i,t,d}$)
- Ask Size (Last) ($asl_{i,t,d}$)
- Bid-Ask Spread (Average) ($basa_{i,t,d}$)
- Bid-Ask Spread (Last) ($basl_{i,t,d}$)
- Bid Tilt (Average) ($bta_{i,t,d}$)
- Bid Tilt (Last) ($btl_{i,t,d}$)
- Buy Tilt ($bti_{i,t,d}$)
- Sell Tilt ($sti_{i,t,d}$)
- Trans-Mid Spread for Buys (Average) ($tmsba_{i,t,d}$)
- Trans-Mid Spread for Buys (Last) ($tmsbl_{i,t,d}$)

- Trans-Mid Spread for Sells (Average) ($tmssa_{i,t,d}$)
- Trans-Mid Spread for Sells (Last) ($tmssl_{i,t,d}$)

I drop the first ten minutes and last ten minutes of each trading day, as I am not interested in modeling the fulfillment of orders placed overnight in the next morning or the flurry of end of day activity from traders closing out their positions. This leaves a total of 370 minutes of data per trading day. The models are trained using data from January - March 2015 (60 trading days). April 2015 (21 trading days) is a hold out validation sample used to compute the out-of-sample R^2. May 2015 (20 trading days) is used to test trading strategies derived from the best performing model.

3 Predicting Returns

This section presents a sequence of predictive models of increasing complexity. The baseline model is an OLS model which predicts an ETF's return using only its past return history. Then, for two information sets, both a factor model and a LASSO model are fit: one information set includes the past returns of the 63 ETFs, and the other includes the 18 microstructure variables in addition to returns.

Factor analysis using principal components analysis is a tool that is over a century old, first appearing in Pearson (1901). Its use has exploded with the advent of computers, and entire books have been written on the subject (Flury, 1998), (Jackson, 1991). In economics, factor models have been widely used in forecasting since dynamic factor models were first introduced by Geweke in 1977. Sargent and Sims (1977) showed that two factors can forecast important quarterly macroeconomic variables, such as output, employment and prices. Surveys of the factor literature in economics include Bai and Ng (2008) and Stock and Watson (2011). These types of models will work well if there are a few latent variables driving variation in the return and microstructure data that can be used to price assets economywide.

In contrast, the LASSO model will work well if an asset's return is predicted by a sparse set of cross-asset predictors rather than economywide factors. Mathematically speaking, sparse means roughly that the value $\frac{s \log p}{n}$ is not too large, where n is the sample size, p the number of predictors, and s the number of active coefficients. Zhao and Yu (2006) give conditions under which the LASSO selects the true model consistently. Meinshausen and Yu (2009) give conditions for l_2 consistency.

Models with nonlinearities are then introduced. First, I fit a LASSO model with two way predictor interactions. This is motivated by the fact that traders don't trade one asset only, but a portfolio of assets. In the classic portfolio theory paper of Markowitz (1952), the investor forms beliefs on expected returns and variance and sets holdings in all assets simultaneously. A natural assumption then is that predictive signals from one asset affects the strength of signals from other assets. For instance, it is possible that if volume for asset i is high, then the effect of high volume for asset j is amplified compared to when the volume for asset i is low. Then I add interactions of the predictors with return factors to the LASSO model. This allows economywide shocks to affect the strength of asset level signals. If returns economywide are elevated or depressed, asset level signals are also elevated or depressed.

The final model to be fit is the random forest. This nonparametric, highly flexible learner is an ensemble of decision trees and is one of the most successful machine learning algorithms. Random forests are used in a wide variety of fields, including genomics, e-commerce, remote sensing, and text mining. It is capable of both regression and classification tasks. It can handle categorical, ordinal, and numeric data with very little preprocessing and is robust to outliers and missing data. The random forest is preferred if there are complex, nonlinear, multiway interactions that are deeper than the two way interaction models can capture. However, the ability to fit deep interactions could potentially cause the random forest to overfit. Empirical studies from Diaz-Uriarte and de Andres (2006) and Geneur et al. (2008) show that the random forest provides accurate predictions in a variety of settings and compares favorably to other machine learning algorithms such as boosting and support vector machines.

For any variable x, define

$$\tilde{x}_{i,t,d} = \begin{pmatrix} x_{i,t-1,d} \\ x_{i,t-2,d} \\ x_{i,t-3,d} \\ \sum_{k=1}^{15} \exp(-(k-1)(.025))x_{i,t-k,d} \\ \sum_{k=1}^{15} \exp(-(k-1)(.1))x_{i,t-k,d} \\ \sum_{k=1}^{15} \exp(-(k-1)(.2))x_{i,t-k,d} \\ \sum_{k=1}^{15} \exp(-(k-1)(.5))x_{i,t-k,d} \end{pmatrix}^T$$

-

Predictors take the above form. The past 3 lagged minutes are explicitly included as predictors. The other

4 terms are exponential decays. These terms allow for signals past 3 minutes in a parsimonious manner: rather than include a lagged variable for every minute past 3, impose that signals decay exponentially and capture this using just 4 terms. This forces signals at lag t to be always stronger than signals at lag t+1. The signal dies off completely after 15 minutes. 4 exponential terms are included to give the model some flexibility. The following gives a plot of the decay rates.

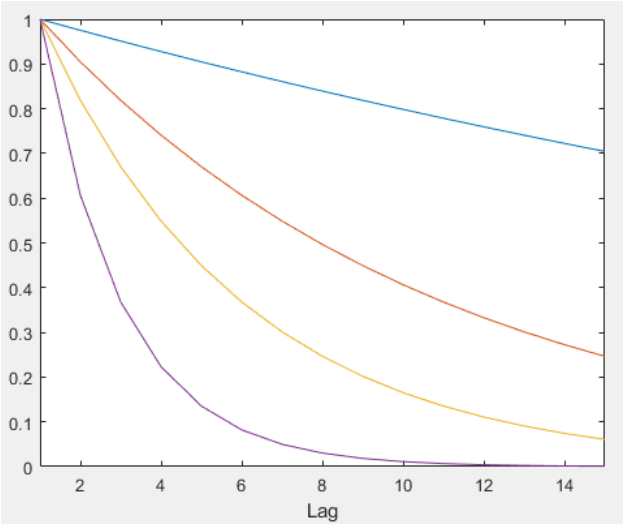


Figure 3.1: Decay rates

As the predictors use 15 minutes of the past, the sample size for the training sample is (370-15=355 minutes)×(60 days) = 21300 and for the validation sample (370-15=355 minutes)×(21 days) = 7455.

3.1 Baseline - Own Asset Returns

Consider the baseline model, where each asset's return is predicted using only the history of its past returns:

$$r_{i,t,d} = \alpha_i + \tilde{r}_{i,t,d}\beta_{i,i} + \epsilon_{i,t,d}$$

Using the 60 days of training data to fit, the OLS estimate is

$$\begin{pmatrix} \hat{\alpha}_i & \hat{\beta}_{i,i} \end{pmatrix} = \arg\min_{\alpha,\beta} \sum_{d=1}^{60} \sum_{t=16}^{370} (r_{i,t,d} - \alpha_i - \tilde{r}_{i,t,d}\beta_{i,i})^2$$

This model is fit for each of the 63 ETFs, and an in sample R^2 and out of sample R^2 are computed. The following histograms report these R^2 values.

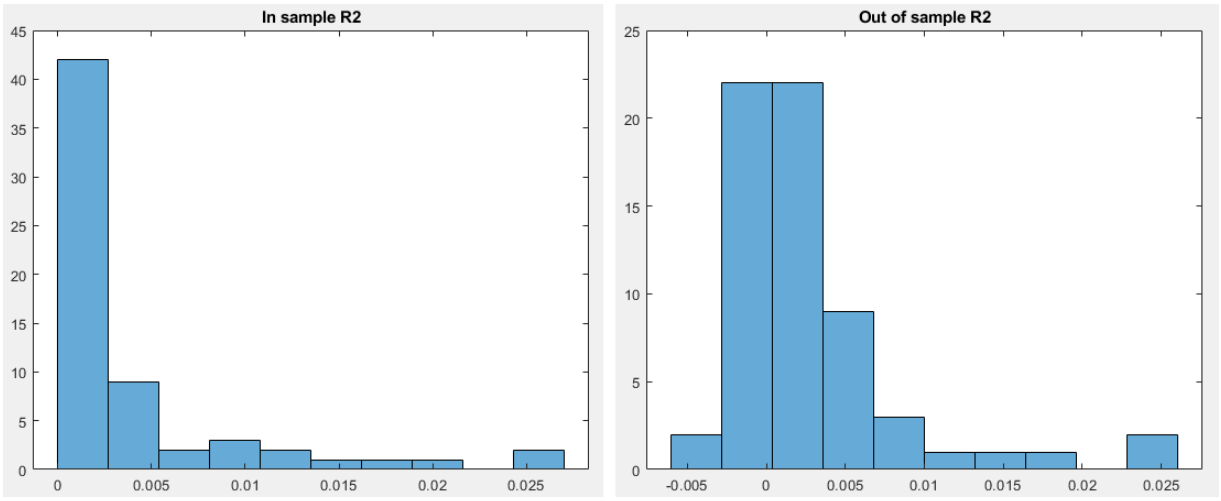


Figure 3.2: Histogram of R^2: baseline OLS

This shows that outside of a very small handful of assets, there is no out of sample predictive power. The average in sample R^2 is 0.0043 and the average out of sample R^2 is 0.0029.

3.2 Linear Models - Factor Models

3.2.1 Returns

I construct a factor model for the returns of the 63 ETFs. A survey of factor models for asset returns by Connor and Korajczyk (2010) shows that these types of models work well in predicting returns in the low frequency setting.

I use principal components analysis to construct factors for returns. Stack the return data day by day:

$$X = \begin{pmatrix} r_{1,1,1} & r_{2,1,1} & \dots & r_{63,1,1} \\ \dots & \dots & \dots & \dots \\ r_{1,370,1} & r_{2,370,1} & \dots & r_{63,370,1} \\ r_{1,1,2} & r_{2,1,2} & \dots & r_{63,1,2} \\ \dots & \dots & \dots & \dots \\ r_{1,370,2} & r_{2,370,2} & \dots & r_{63,370,2} \\ \dots & \dots & \dots & \dots \\ r_{1,1,60} & r_{2,1,60} & \dots & r_{63,1,60} \\ \dots & \dots & \dots & \dots \\ r_{1,370,60} & r_{2,370,60} & \dots & r_{63,370,60} \end{pmatrix}$$

X is a $(370)(60) \times 63$, or (minutes)(days)×assets, data matrix. Let C be X's correlation matrix. Looking at the scree plot of the correlation matrix, I decide to extract one factor. The factor explains 42.15% of the contemporaneous variance in returns.

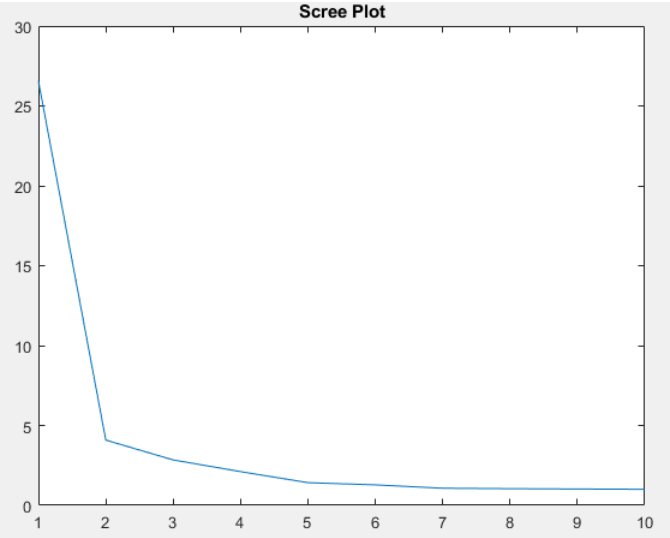


Figure 3.3: Scree plot: returns

Let w be the normalized eigenvector associated with the largest eigenvalue of C. Then for minute t and day d, define the factor for returns as

$$fr_{t,d} = \begin{pmatrix} r_{1,t,d} & \dots & r_{63,t,d} \end{pmatrix} w^T$$

Define $\tilde{fr}_{t,d}$ in the same manner as before. Then for each asset, fit the factor model

$$r_{i,t,d} = \alpha_i + \tilde{fr}_{t,d}\gamma_i + \epsilon_{i,t,d}$$

where the OLS estimate is

$$\begin{pmatrix} \hat{\alpha}_i & \hat{\gamma}_i \end{pmatrix} = \arg\min_{\alpha,\gamma} \sum_{d=1}^{60} \sum_{t=16}^{370} (r_{i,t,d} - \alpha_i - \tilde{fr}_{t,d}\gamma_i)^2$$

To generate the factor out of sample, define $fr_{t,d} = \begin{pmatrix} r_{1,t,d} & \dots & r_{63,t,d} \end{pmatrix} w^T$, where the returns are from the validation sample while w is constructed using the training sample.

The following histograms report in sample and out of sample R^2 values.

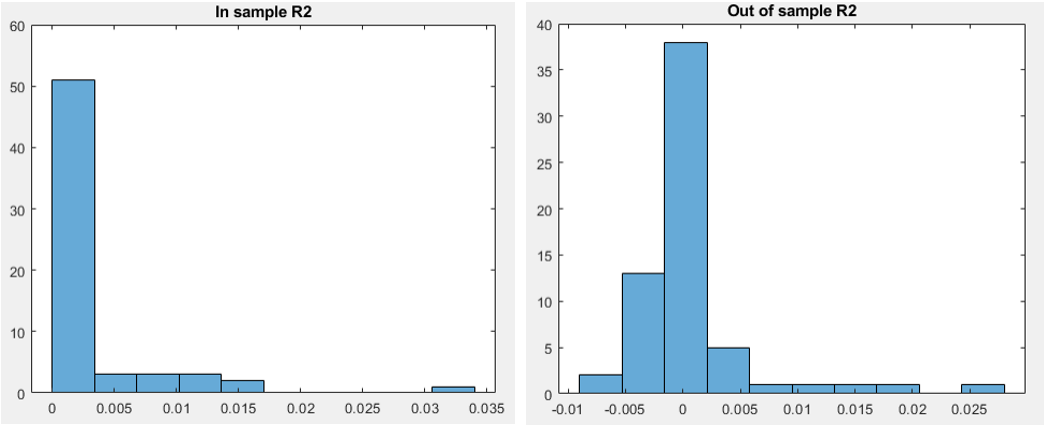


Figure 3.4: Histogram of R^2: factor model for returns

The factor model performs poorly. The average in sample R^2 is 0.0029 and the average out of sample R^2 is 0.0008.

3.2.2 Returns and Microstructure

I now construct factors using both the return and the microstructure data. This is motivated by the fact that asymmetric information risk arising from the market microstructure is unlikely to be confined to a single asset. Hasbrouck and Seppi (2001) and Russell (2017) show that there are common factors in the microstructure of equity markets.

As before, I use principal components analysis to construct factors. Stack the 19 variables day by day:

$$X = \begin{pmatrix} r_{1,1,1} & \dots & r_{63,1,1} & bt_{1,1,1} & \dots & bt_{63,1,1} & \dots & tmssl_{1,1,1} & \dots & tmssl_{63,1,1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{1,370,1} & \dots & r_{63,370,1} & bt_{1,370,1} & \dots & bt_{63,370,1} & \dots & tmssl_{1,370,1} & \dots & tmssl_{63,370,1} \\ r_{1,1,2} & \dots & r_{63,1,2} & bt_{1,1,2} & \dots & bt_{63,1,2} & \dots & tmssl_{1,1,2} & \dots & tmssl_{63,1,2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{1,370,2} & \dots & r_{63,370,2} & bt_{1,370,2} & \dots & bt_{63,370,2} & \dots & tmssl_{1,370,2} & \dots & tmssl_{63,370,2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{1,1,60} & \dots & r_{63,1,60} & bt_{1,1,60} & \dots & bt_{63,1,60} & \dots & tmssl_{1,1,60} & \dots & tmssl_{63,1,60} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{1,370,60} & \dots & r_{63,370,60} & bt_{1,370,60} & \dots & bt_{63,370,60} & \dots & tmssl_{1,370,60} & \dots & tmssl_{63,370,60} \end{pmatrix}$$

X is a (370)(60) × (63)(19), or (minutes)(days)×(assets)(variables), data matrix. Let C be X's correlation matrix. Looking at the scree plot of the correlation matrix, I decide to extract 15 factors. The factors explain 18.62% of the contemporaneous variance in returns and microstructure data.

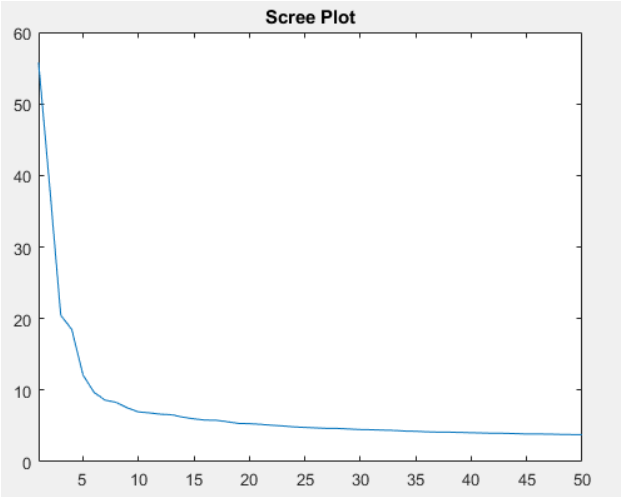


Figure 3.5: Scree plot: returns and microstructure

Let w_k be the normalized eigenvector associated with the kth largest eigenvalue of C, Then for minute t and day d, define the kth factor for all variables as

$$fv^k_{t,d} = \begin{pmatrix} r_{1,t,d} & \dots & r_{63,t,d} & \dots & tmssl_{1,t,d} & \dots & tmssl_{63,t,d} \end{pmatrix} w_k^T$$

Define $\tilde{fv}^k_{t,d}$ in the same manner as before. Then for each asset, fit the factor model

$$r_{i,t,d} = \alpha_i + \sum_{k=1}^{15} \tilde{fv}^k_{t,d} \gamma_i^k + \epsilon_{i,t,d}$$

where the OLS estimate is

$$\begin{pmatrix} \hat{\alpha}_i & \hat{\gamma}_i^1 & \dots & \hat{\gamma}_i^{15} \end{pmatrix} = \arg\min_{\alpha,\gamma} \sum_{d=1}^{60} \sum_{t=16}^{370} (r_{i,t,d} - \alpha_i - \sum_{k=1}^{15} \tilde{fv}^k_{t,d} \gamma_i^k)^2$$

The following histograms report in sample and out of sample R^2 values.

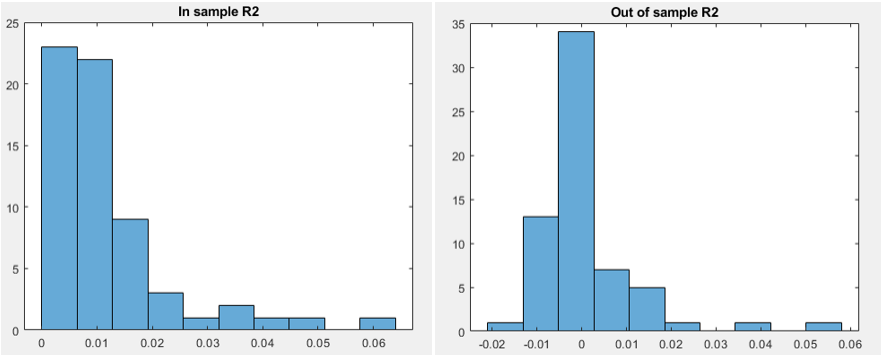


Figure 3.6: Histogram of R^2: factor model for returns and microstructure

This factor model also performs poorly. The average in sample R^2 is 0.0123 and the average out of sample R^2 is 0.0008.

The poor performance of both factor models suggests that the success of factor models in the low frequency setting does not translate to the intraday high frequency setting. It also highlights a drawback of the factor model: the factors are constructed based on the covariation among the predictors, without any consideration of how the predictors covary with future returns.

3.3 Linear Models - LASSO

The LASSO minimizes

$$\arg\min_{\alpha,\beta} ||y - \alpha - X\beta||_2^2 + \lambda||\beta||_1$$

The additional penalty term applies an L1 penalty to the regression coefficients. It performs model selection by setting certain coefficients to 0. It also shrinks the other coefficients towards 0, as in ridge regression. This makes parameter estimates less noisy, trading off variance for an increase in the bias of the parameter estimates. For moderate levels of shrinkage, this decreases the mean squared error of predictions. The parameter λ controls the degree of regularization, with higher values leading to sparser models. The penalty parameter λ is chosen by 3-fold cross validation, where an entire day of data is assigned to a fold rather than individual observations.

3.3.1 Returns

For each asset, I fit a LASSO model using the history of the returns of all 63 assets. This model is similar to that in Chinco (2018).

$$r_{i,t,d} = \alpha_i + \sum_{j=1}^{63} \tilde{r}_{j,t,d}\beta_{i,j} + \epsilon_{i,t,d}$$

The LASSO estimates are

$$\begin{pmatrix} \hat{\alpha}_i & \hat{\beta}_{i,1} & \dots & \hat{\beta}_{i,63} \end{pmatrix} = \arg\min_{\alpha,\beta} \sum_{d=1}^{60} \sum_{t=16}^{370} (r_{i,t,d} - \sum_{j=1}^{63} \tilde{r}_{j,t,d}\beta_{i,j})^2 + \lambda \sum_{j=1}^{63} ||\beta_{i,j}||_1$$

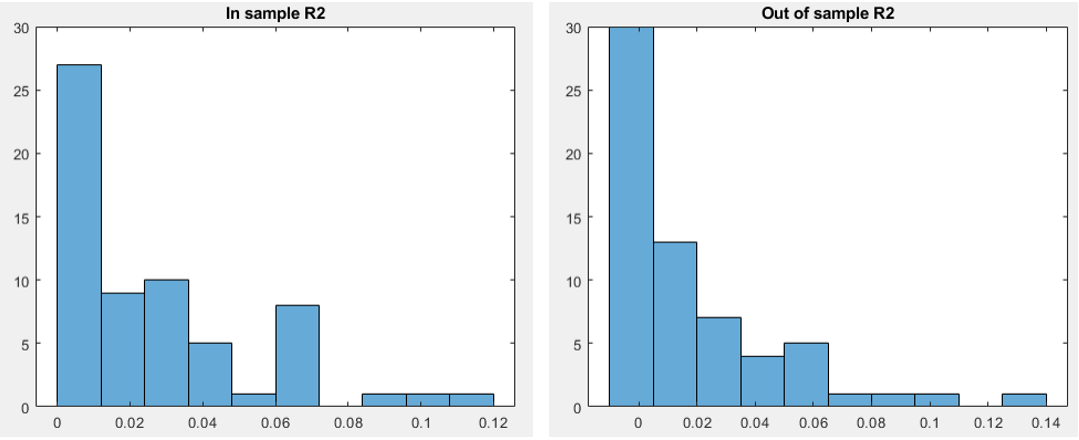


Figure 3.7: Histogram of R^2: LASSO model for returns

The average in sample R^2 is 0.0258 and the average out of sample R^2 is 0.0192. Many of the assets have a good amount of out of sample predictability.

As a measure of how important each asset is, for asset i, I count the number of times the LASSO models for assets $j \neq i$ select terms associated with asset i. The following table reports these counts (out of 62).

Table 3.1: Number of times assets are selected by non-own asset models: returns

Asset	Times Selected	Asset	Times Selected	Asset	Times Selected	Asset	Times Selected
LQD	33	SLV	30	IWN	6	XLE	23
EMB	40	USO	31	EWY	23	XOP	17
SHY	39	UNG	31	MCHI	26	XLF	11
IEI	30	FXE	34	EEM	29	KBE	30
IEF	27	FXY	33	EZU	17	KRE	17
JNK	21	SPY	12	EWG	22	XLV	24
TIP	24	QQQ	26	EWU	23	XBI	17
BWX	29	DIA	19	EFA	21	XLI	19
SHV	16	IWB	0	INDA	35	XLB	17
MBB	25	IWF	6	EWJ	23	XME	21
MUB	34	IWD	6	EWZ	24	XLK	10
PFF	30	IWR	7	EWW	26	XLU	20
AGG	30	IWP	6	XHB	19	IYR	35
BND	30	IWS	4	XLY	12	SH	13
DBC	21	IWM	27	XRT	15	SSO	13
IAU	35	IWO	13	XLP	17		

The average of these counts is 21.8, indicating that cross asset effects are important for prediction.

3.3.2 Returns and Microstructure

I now construct LASSO models that predict using the 18 microstructure variables and returns. Define

$$\tilde{M}_{t,d} = \begin{pmatrix} \tilde{r}_{1,t,d} & \dots & \tilde{r}_{63,t,d} & \tilde{bt}_{1,t,d} & \dots & \tilde{bt}_{63,t,d} & \dots & \tilde{tmssl}_{1,t,d} & \dots & \tilde{tmssl}_{63,t,d} \end{pmatrix}$$

Each asset's return is predicted using all 19 variables for all 63 assets:

$$r_{i,t,d} = \alpha_i + \tilde{M}_{t,d}\beta_i + \epsilon_{i,t,d}$$

The LASSO estimates are

$$\begin{pmatrix} \hat{\alpha}_i & \hat{\beta}_i \end{pmatrix} = \arg\min_{\beta} \sum_{d=1}^{60} \sum_{t=16}^{370} (r_{i,t,d} - \alpha_i - \tilde{M}_{t,d}\beta_i)^2 + \lambda ||\beta_i||_1$$

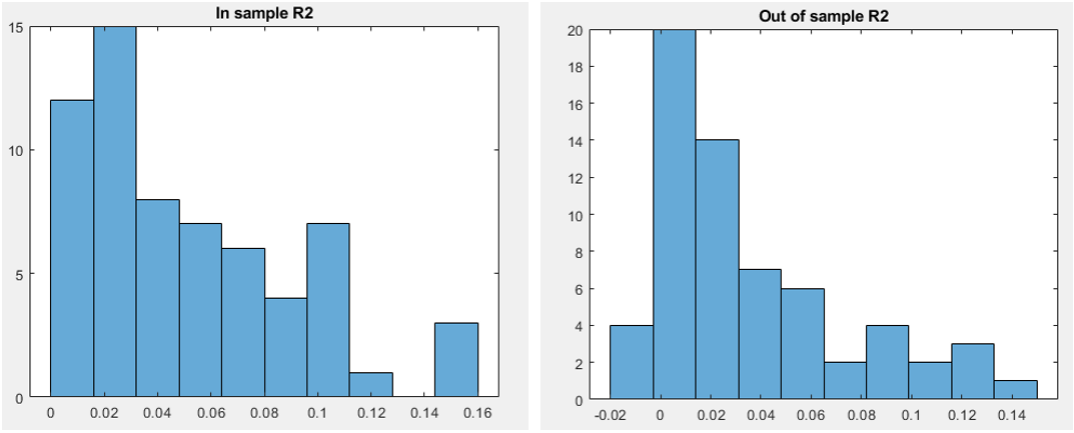


Figure 3.8: Histogram of R^2: LASSO model for returns and microstructure

The average in sample R^2 is 0.0517 and the average out of sample R^2 is 0.0347. Compared with the LASSO model using only returns, which had an out of sample R^2 of 0.0192, using volume, depth, and spread data in addition to returns increased out of sample predictability by 81%.

As a measure of how important each asset is, for asset i, I count the number of times the LASSO models for assets $j \neq i$ select terms associated with asset i. The following table reports these counts (out of 62).

Table 3.2: Number of times assets are selected by non-own asset models: returns and microstructure

Asset	Times Selected	Asset	Times Selected	Asset	Times Selected	Asset	Times Selected
LQD	17	SLV	23	IWN	21	XLE	36
EMB	17	USO	22	EWY	29	XOP	14
SHY	8	UNG	10	MCHI	10	XLF	16
IEI	40	FXE	54	EEM	17	KBE	33
IEF	19	FXY	14	EZU	51	KRE	31
JNK	15	SPY	26	EWG	22	XLV	43
TIP	22	QQQ	24	EWU	35	XBI	17
BWX	15	DIA	35	EFA	49	XLI	24
SHV	8	IWB	29	INDA	18	XLB	15
MBB	4	IWF	20	EWJ	6	XME	22
MUB	32	IWD	9	EWZ	12	XLK	6
PFF	27	IWR	28	EWW	26	XLU	15
AGG	22	IWP	27	XHB	23	IYR	22
BND	44	IWS	13	XLY	27	SH	10
DBC	24	IWM	39	XRT	3	SSO	13
IAU	23	IWO	14	XLP	30		

The average of these counts is 22.5. As with the return only model, cross asset information helps greatly with prediction.

As a measure of how important each variable is, I count the number of LASSO models (out of 63) that select a particular variable as a predictor.

Table 3.3: Number of times variables are selected: returns and microstructure

Variable	Times Selected	Variable	Times Selected
Return	56	Bid-Ask Spread (Average)	51
Buy Transactions	34	Bid-Ask Spread (Last)	60
Buy Dollar Volume	30	Bid Tilt (Average)	36
Sell Transactions	37	Bid Tilt (Last)	59
Sell Dollar Volume	17	Buy Tilt	29
Bid Size (Average)	23	Sell Tilt	40
Bid Size (Last)	39	Trans-Mid Spread For Buy (Average)	39
Ask Size (Average)	33	Trans-Mid Spread For Buy (Last)	48
Ask Size (Last)	43	Trans-Mid Spread For Sells (Average)	46
		Trans-Mid Spread For Sells (Last)	32

The average count for the non-return variables is 38.7, confirming that the volume, depth, and spread variables encode information not included in the return data.

The LASSO models do a much better job of incorporating the predictive information in the returns and microstructure data compared with the factor models.

3.4 Nonlinear Models - Two Way Interactions

3.4.1 Variable Interactions

To allow for the possibility of nonlinear effects, I introduce a model with two way interactions. This allows predictive signals from one asset to affect the strength of signals from other assets. This is in contrast to the main effects only model of 3.3.2 that includes just levels. In the main effects only model, the strength of a signal for an asset's returns do not depend on the levels of other predictors.

Define

$$K_{t,d} = \begin{pmatrix} r_{1,t,d} & \dots & r_{63,t,d} & bt_{1,t,d} & \dots & bt_{63,t,d} & \dots & tmssl_{1,t,d} & \dots & tmssl_{63,t,d} \end{pmatrix}$$

and

$$I_{i,t,d} = \begin{pmatrix} r_{i,t-1,d}K_{t-1,d} & \dots & tmssl_{i,t-1,d}K_{t-1,d} & \dots & r_{i,t-3,d}K_{t-3,d} & \dots & tmssl_{i,t-3,d}K_{t-3,d} \end{pmatrix}$$

The interaction vector $I_{i,t,d}$ contains 3 lags of own asset-own asset and own asset-cross asset interactions. For computational reasons, cross asset-cross asset interactions are omitted. $I_{i,t,d}$ is a vector of size (19 own asset variables)(63 assets)(19 variables)(3 lags) = 68229; including cross asset-cross asset interactions would have added another (62 cross assets)(19 cross asset variables)(63 assets)(19 variables)(3 lags) = 4230198 terms. This omission is necessary to keep memory use and computation time down. Interaction effects die off after 3 minutes; unlike with main effects, I do not include any decay terms. Let $\tilde{M}_{t,d}$ be defined as in 3.3.2. For each asset, fit the model

$$r_{i,t,d} = \alpha_i + \tilde{M}_{t,d}\beta_i + I_{i,t,d}\delta_i + \epsilon_{i,t,d}$$

The LASSO estimates are

$$\begin{pmatrix} \hat{\alpha}_i & \hat{\beta}_i & \hat{\delta}_i \end{pmatrix} = \arg\min_{\beta} \sum_{d=1}^{60} \sum_{t=16}^{370} (r_{i,t,d} - \alpha_i - \tilde{M}_{t,d}\beta_i - I_{i,t,d}\delta_i)^2 + \lambda(||\beta_i||_1 + ||\delta_i||_1)$$

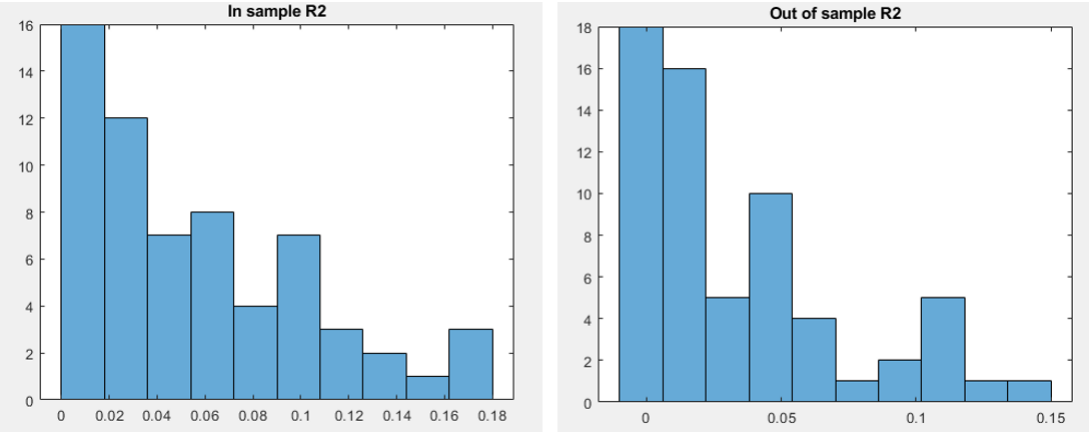


Figure 3.9: Histogram of R^2: variable interactions

The average in sample R^2 is 0.0576 and the average out of sample R^2 is 0.0342. Compared with the LASSO model with main effects only, which had an average out of sample R^2 0.0347, interactions do not add out of sample predictability.

For each asset i, I count the number of times the LASSO models for assets $j \neq i$ select main effect and interaction terms associated with asset i.

Table 3.4: Number of times assets are selected by non-own asset models: variable interactions

	Times Selected			Times Selected			Times Selected			Times Selected	
Asset	Main	Interactions	Asset	Main	Interactions	Asset	Main	Interactions	Asset	Main	Interactions
LQD	3	24	SLV	2	18	IWN	1	31	XLE	8	25
EMB	0	21	USO	2	22	EWY	6	26	XOP	3	19
SHY	0	19	UNG	0	16	MCHI	0	27	XLF	1	23
IEI	10	20	FXE	4	54	EEM	4	18	KBE	5	27
IEF	8	33	FXY	4	36	EZU	17	44	KRE	1	21
JNK	1	22	SPY	0	26	EWG	0	21	XLV	7	41
TIP	2	41	QQQ	3	16	EWU	2	23	XBI	0	19
BWX	0	30	DIA	3	21	EFA	13	37	XLI	3	18
SHV	0	28	IWB	1	31	INDA	1	19	XLB	3	19
MBB	0	18	IWF	0	24	EWJ	0	14	XME	3	27
MUB	7	20	IWD	0	23	EWZ	1	15	XLK	1	18
PFF	3	22	IWR	2	21	EWW	6	30	XLU	3	20
AGG	7	28	IWP	0	26	XHB	0	26	IYR	7	24
BND	2	37	IWS	1	29	XLY	1	22	SH	0	16
DBC	0	24	IWM	3	37	XRT	0	17	SSO	3	24
IAU	3	25	IWO	1	21	XLP	2	18			

The average count for main effects is 2.8, and the average count for interactions is 24.8.

Now I count the number of LASSO models that select a particular variable's main effects and interactions as a predictor.

Table 3.5: Number of times variables are selected: variable interactions

	Times Selected			Times Selected	
Variable	Main	Interactions	Variable	Main	Interactions
Return	27	56	Bid-Ask Spread (Average)	8	46
Buy Transactions	1	33	Bid-Ask Spread (Last)	27	62
Buy Dollar Volume	5	34	Bid Tilt (Average)	14	38
Sell Transactions	5	30	Bid Tilt (Last)	1	60
Sell Dollar Volume	1	36	Buy Tilt	6	44
Bid Size (Average)	0	43	Sell Tilt	8	41
Bid Size (Last)	0	54	Trans-Mid Spread For Buy (Average)	2	43
Ask Size (Average)	2	40	Trans-Mid Spread For Buy (Last)	0	56
Ask Size (Last)	8	56	Trans-Mid Spread For Sells (Average)	5	55
			Trans-Mid Spread For Sells (Last)	0	47

For the main effects, the average count for the non-return variables is 5.2, and for interactions 45.4.

The frequency with which interaction terms are selected indicate that while the average out of sample predictability for the main effects only model and the interaction model are the same, the method of action is very different.

3.4.2 Factor Interactions

I construct a model which incorporates interactions with the return factor constructed earlier. This allows economywide shocks to affect the strength of asset level signals.

Recall the definition of $fr_{t,d}$ and $\tilde{fr}_{t,d}$ from 3.2.1. Define

$$F_{t,d} = \begin{pmatrix} \tilde{fr}_{t,d} & (fr_{t-1,d})^2 & (fr_{t-2,d})^2 & (fr_{t-3,d})^2 & fr_{t-1,d}K_{t-1,d} & fr_{t-2,d}K_{t-2,d} & fr_{t-3,d}K_{t-3,d} \end{pmatrix}$$

For each asset, fit the model

$$r_{i,t,d} = \alpha_i + \tilde{M}_{t,d}\beta_i + I_{i,t,d}\delta_i + F_{t,d}\phi_i + \epsilon_{i,t,d}$$

The LASSO estimates are

$$\begin{pmatrix} \hat{\alpha}_i & \hat{\beta}_i & \hat{\delta}_i & \hat{\phi}_i \end{pmatrix} = \arg\min_{\beta} \sum_{d=1}^{60} \sum_{t=16}^{370} (r_{i,t,d} - \alpha_i - \tilde{M}_{t,d}\beta_i - I_{i,t,d}\delta_i - F_{t,d}\phi_i)^2 + \lambda(||\beta_i||_1 + ||\delta_i||_1 + ||\phi_i||_1)$$

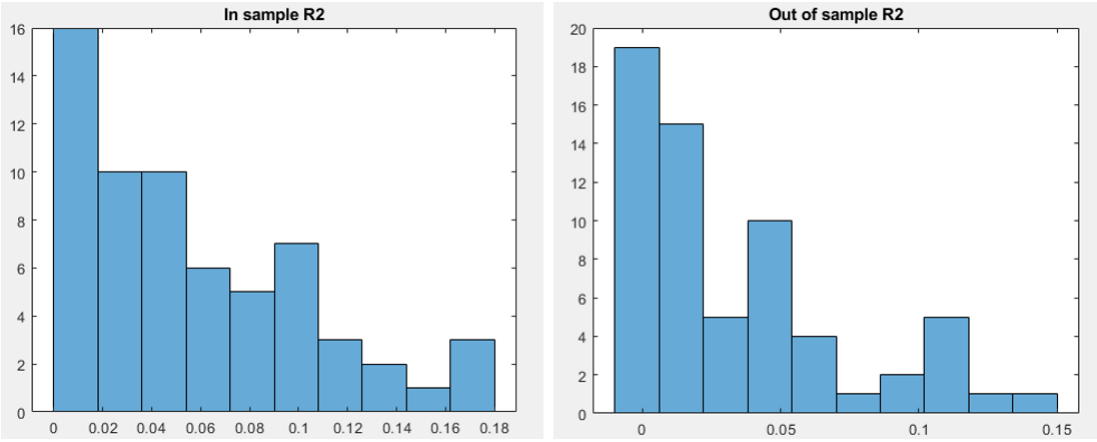


Figure 3.10: Histogram of R^2: factor interactions

The average in sample R^2 is 0.0578 and the average out of sample R^2 is 0.0341. Compared with the LASSO model with main effects only, which had an average out of sample R^2 of 0.0347, return factor interactions also do not add out of sample predictability.

For each asset i, I count the number of times the LASSO models for assets $j \neq i$ select main effect and nonfactor interaction terms associated with asset i.

Table 3.6: Number of times assets are selected by non-own asset models: factor interactions

	Times Selected			Times Selected			Times Selected			Times Selected	
Asset	Main	Interactions	Asset	Main	Interactions	Asset	Main	Interactions	Asset	Main	Interactions
LQD	3	24	SLV	1	17	IWN	0	28	XLE	8	20
EMB	0	20	USO	4	21	EWY	4	25	XOP	1	22
SHY	0	20	UNG	0	17	MCHI	0	27	XLF	0	24
IEI	10	19	FXE	4	54	EEM	3	17	KBE	4	26
IEF	8	34	FXY	4	36	EZU	18	43	KRE	1	19
JNK	0	19	SPY	1	26	EWG	0	18	XLV	6	39
TIP	2	39	QQQ	3	15	EWU	2	27	XBI	0	18
BWX	0	28	DIA	1	20	EFA	14	36	XLI	1	19
SHV	0	24	IWB	1	30	INDA	0	17	XLB	1	21
MBB	0	16	IWF	0	22	EWJ	0	13	XME	1	26
MUB	6	21	IWD	0	21	EWZ	2	14	XLK	2	18
PFF	2	24	IWR	1	22	EWW	5	28	XLU	4	18
AGG	7	28	IWP	0	24	XHB	0	24	IYR	8	23
BND	3	37	IWS	1	27	XLY	1	24	SH	0	15
DBC	0	24	IWM	1	36	XRT	0	16	SSO	2	22
IAU	2	24	IWO	1	21	XLP	3	19			

The average count for main effects is 2.5, and the average count for interactions is 24.1.

Now I count the number of LASSO models that select a particular variable's main effects and nonfactor interactions as a predictor.

Table 3.7: Number of times variables are selected: factor interactions

	Times Selected			Times Selected	
Variable	Main	Interactions	Variable	Main	Interactions
Return	28	55	Bid-Ask Spread (Average)	7	46
Buy Transactions	1	31	Bid-Ask Spread (Last)	26	62
Buy Dollar Volume	5	30	Bid Tilt (Average)	14	36
Sell Transactions	5	31	Bid Tilt (Last)	0	60
Sell Dollar Volume	0	32	Buy Tilt	7	39
Bid Size (Average)	0	41	Sell Tilt	8	39
Bid Size (Last)	0	54	Trans-Mid Spread For Buy (Average)	3	43
Ask Size (Average)	2	40	Trans-Mid Spread For Buy (Last)	1	56
Ask Size (Last)	6	55	Trans-Mid Spread For Sells (Average)	4	54
			Trans-Mid Spread For Sells (Last)	0	46

For the main effects, the average count for the non-return variables is 4.9, and for interactions 44.2.

44 of the 63 LASSO models select at least one interaction of the return factor with one of the 19 predictors, indicating that economywide shocks affect the strength of asset level signals. Again, the main effects only model and the factor interaction model have the same out of sample predictability despite having different methods of action. For reasons of parsimony, I prefer the main effects only LASSO model to both of the interaction models introduced in this section.

3.5 Nonlinear Models - Random Forest

I implement a random forest, which allows for much more flexible nonlinearities and interactions compared with the two way interaction models in the previous section.

3.5.1 Model Description

The random forest builds a collection of decorrelated regression trees and then averages their outputs. A regression tree is grown in a sequence of steps. At each step, a branch forms which partitions the data into two nodes based on one of the predictor variables. For data (x_i, y_i), $x_i = (x_{i1,\ldots,}x_{ip})$, and K partitions R_k, the predicted value is

$$f(x) = \sum_{k=1}^{K} c_k I(x \in R_k)$$

For the sum of squares loss, the best prediction is $\hat{c}_k = \text{ave}(y_i|x_i \in R_k)$, the average of y_i in partition k. Starting with the data in a particular node, consider a splitting variable j and split point s. Define the half planes $R_1(j,s) = \{X|X_j \leq s\}$ and $R_2(j,s) = \{X|X_j > s\}$. Then search for the splitting variable j and split point s which minimizes

$$\min_{j,s}\left[\sum_{x_i \in R_1(j,s)} (y_i - \text{ave}(y_i|x_i \in R_1(j,s))^2 + \sum_{x_i \in R_2(j,s)} (y_i - \text{ave}(y_i|x_i \in R_2(j,s))^2\right]$$

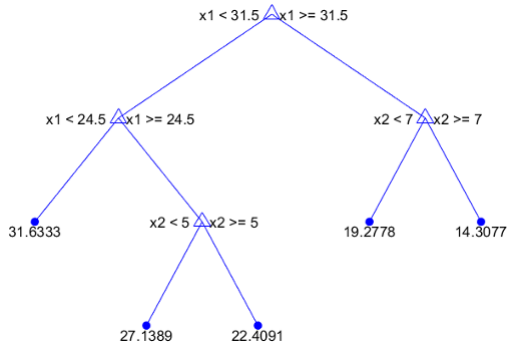


Figure 3.11: Regression tree

To control the depth of the tree, set a minimum terminal node size. If a split results in a child node that contains less data points than the terminal node size, stop. Otherwise, accept the split and create two child nodes. Repeat this process for all the nodes. The minimum terminal node size is a tuning parameter for the tree that controls overfitting.

The random forest averages the output of multiple trees. While trees can capture complex nonlinearities and interactions, they are very noisy, and therefore reap great benefits from averaging. The following random forest algorithm is taken from the textbook Elements of Statistical Learning of Hastie, Tibshirani, and Friedman (2009).

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

There are three tuning parameters: the size of the bootstrap sample $Z^\star$, the number of variables randomly sampled as candidates at each split m , and the minimum node size n_{min}. The tuning criteria used is out-of-bag prediction error. To generate an out-of-bag prediction for data point (x_i, y_i), take the average of the predictions from only those trees corresponding to bootstrap samples in which (x_i, y_i) did not appear. The out-of-bag error estimate is almost identical to that obtained by N-fold cross validation, and incurs less computational costs.

I use the R package tuneRanger from Probst, Wright, and Boulesteix (2019) to fit the random forest. The package uses sequential model-based optimization as a tuning strategy, which gives a "smart" way to search over parameter values. Details are given in their paper. I set the number of trees in the forest to 500.

3.5.2 Model Results

Recall the definition of $\tilde{M}_{t,d} = \begin{pmatrix} \tilde{r}_{1,t,d} & ... & \tilde{r}_{63,t,d} & \tilde{bt}_{1,t,d} & \dots & \tilde{bt}_{63,t,d} & \dots & \tilde{tmssl}_{1,t,d} & ... & \tilde{tmssl}_{63,t,d} \end{pmatrix}$. Each of the tilde terms includes 3 lags terms and 4 decay terms. $\tilde{M}_{t,d}$ is therefore a vector of (19 variables)(63 assets)(7) = 8379 terms. Define

$$R_i = \begin{pmatrix} r_{i,1,1} \\ \vdots \\ r_{i,370,1} \\ \vdots \\ r_{i,1,60} \\ \vdots \\ r_{i,370,60} \end{pmatrix}$$

and

$$M = \begin{pmatrix} \tilde{M}_{1,1} \\ \vdots \\ \tilde{M}_{370,1} \\ \vdots \\ \tilde{M}_{1,60} \\ \vdots \\ \tilde{M}_{370,60} \end{pmatrix}$$

Recall the main effects only LASSO estimate for asset i $\hat{\beta}_i = (\hat{\beta}_{i,1}, ..., \hat{\beta}_{i,8379})$ from 3.3.2. Let P_i be the active set $P_i = \{p|\hat{\beta}_{i,p} \neq 0\}$ and M_{P_i} be the columns of M corresponding to the active set of the LASSO estimate. M_{P_i} serves as the matrix of predictors for R_i. The choice to use M_{P_i} rather than M is made due to two reasons. In the setting with low signal and a large number of predictors, the ability of the random forest to fit complex nonlinear interactions causes the predictor to fit a lot of noise. The LASSO prunes the set of predictors, shutting off several potential branches of noise, and helps with the overfitting problem. Also, the random forest's runtime is roughly linear in the number of predictors. If the set of predictors it searches over is smaller, the algorithm will run faster, giving another reason to use M_{P_i} rather than M. For each asset, I fit a random forest, with all three tuning parameters n_{min}, $Z^\star$, and m tuned using the tuneRanger package.

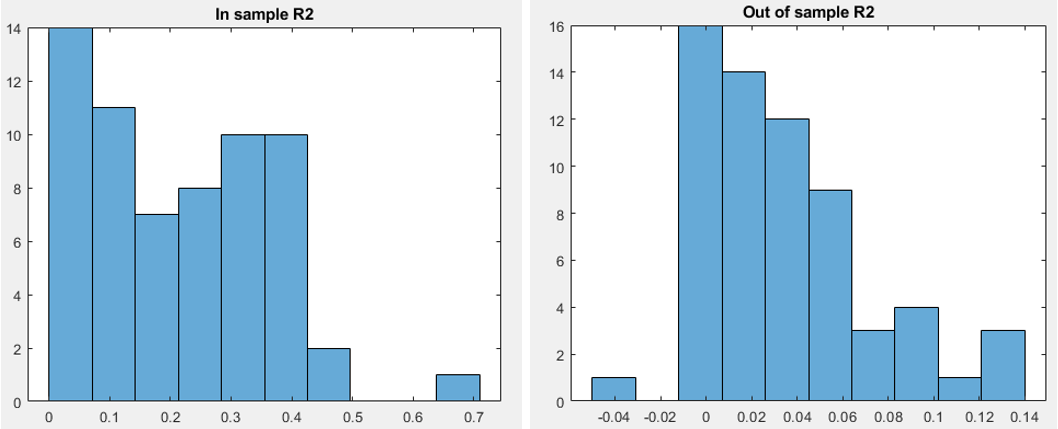


Figure 3.12: Histogram of R^2: random forest

The average in sample R^2 is 0.2144 and the average out of sample R^2 is 0.0334. Comparing the in sample and out of sample R^2, it is clear that there is still substantial overfitting. As n_{min} is the tuning parameter which controls the depth of the trees, it is responsible for the amount of overfitting. In the following experiment, I try setting this value manually. For each asset i, I fix a value for n_{min}, set $Z^\star$ to be one-third of the sample, and let the tuning algorithm choose m. The following table reports the average in sample and out of sample R^2 for values of n_{min}.

Table 3.8: R^2 for values of n_{min}

n_{min}	R^2 in sample	R^2 out of sample sample
100	0.1922	0.0323
500	0.1094	0.0354
1000	0.0862	0.0356
1500	0.0742	0.0349
2000	0.065	0.0343

Increasing n_{min} decreases in sample R^2. While this may indicate that it helps with overfitting, out of sample predictability stays roughly the same. Compared with the LASSO model with main effects only, which had an average out of sample R^2 of 0.0347, adding a random forest layer on top of the main effects only model produces only a very modest improvement in out of sample R^2 for certain specifications and performs worse in others. The difficulties in tuning and overall poor performance are a surprise given the random forest literature and its wide use. In Xiu's (2018) high dimensional analysis of low frequency returns, random forests had been one of the better predictors and outperformed linear models. For reasons of parsimony, I prefer the main effects only LASSO model and dispense with adding an additional random forest layer.

4 Trading Strategy

Predicting high frequency returns is not a purely academic exercise. The question I examine now is whether this return predictability can be turned into trading profits. The model I use to construct the trading strategy is the main effects only LASSO model from 3.3.2, since it gives the highest out of sample predictability with the least model complexity. Recall the LASSO estimates $\begin{pmatrix} \hat{\alpha}_i & \hat{\beta}_i \end{pmatrix}$, constructed using data from January-March 2015. Let $\tilde{M}_{t,d}$ be defined as before, except now d will be taken from May 2015, the sample used to test trading strategies. Let the realized return be $r_{i,t,d}$ and the predicted return be $\hat{r}_{i,t,d} = \hat{\alpha}_i + \tilde{M}_{t,d}\hat{\beta}_i$.

Assume that a hypothetical trader transacts at the midquote. Consider a trader trading fund i on day d. She ignores the first 10 minutes of the trading day. Then she collects the data from the next 15 minutes, and uses it to construct $\tilde{M}_{16,d}$ and $\hat{r}_{i,16,d}$. Then set

$$D_{i,16,d} = \begin{cases} 1 & \hat{r}_{i,16,d} \geq 0 \\ -1 & \hat{r}_{i,16,d} < 0 \end{cases}$$

The trader will enter into a long position if the predicted return is positive, and a short position if the predicted return is negative. At the end of the minute, she liquidates her position, realizing the return $\pi_{i,16,d} = r_{i,16,d}D_{i,16,d}$. She then repeats this for $t = 17...370$. Finally, she ignores the last 10 minutes of the trading day. Her trading return for the day is

$$\Pi_{i,d} = \sum_{t=16}^{370} \pi_{i,t,d}$$

Compare this with a buy and hold strategy which results in return

$$R_{i,d} = \sum_{t=16}^{370} r_{i,t,d}$$

For each day in May 2015, generate the trading strategy return and the buy and hold return. Let $\bar{\Pi}_i = \frac{1}{20}\sum_{d=1}^{20} \Pi_{i,d}$ and $\bar{R}_i = \frac{1}{20}\sum_{d=1}^{20} R_{i,d}$ be the average daily trading strategy return and buy and hold return. The following table reports these two values for all 63 funds.

Table 4.1: Average daily returns for trading strategy and buy and hold

Asset		Strategy	Buy and Hold	Asset		Strategy	Buy and Hold
LQD	iShares iBoxx $ Investment Grade Corporate Bond ETF	0.84%	-0.03%	EWY	iShares MSCI South Korea ETF	1.00%	-0.01%
EMB	iShares J.P. Morgan USD Emerging Markets Bond ETF	0.46%	0.01%	MCHI	iShares MSCI China ETF	1.40%	0.03%
SHY	iShares 1-3 Year Treasury Bond ETF	0.10%	0.00%	EEM	iShares MSCI Emerging Markets ETF	1.57%	-0.03%
IEI	iShares 3-7 Year Treasury Bond ETF	0.63%	0.00%	EZU	iShares MSCI EMU ETF	2.14%	0.01%
IEF	iShares 7-10 Year Treasury Bond ETF	0.72%	-0.02%	EWG	iShares MSCI Germany ETF	2.93%	0.02%
JNK	SPDR Barclays Capital High Yield Bond ETF	0.60%	0.01%	EWU	iShares MSCI United Kingdom ETF	2.74%	0.01%
TIP	iShares TIPS Bond ETF	0.98%	-0.04%	EFA	iShares MSCI EAFE ETF	1.20%	0.02%
BWX	SPDR Barclays Intl Treasury Bond ETF	1.21%	-0.01%	INDA	iShares MSCI India ETF	1.29%	0.07%
SHV	iShares Short Treasury Bond ETF	0.00%	0.00%	EWJ	iShares MSCI Japan ETF	2.21%	0.05%
MBB	iShares MBS ETF	0.78%	-0.01%	EWZ	iShares MSCI Brazil ETF	2.08%	-0.27%
MUB	iShares National Muni Bond ETF	0.30%	-0.02%	EWW	iShares MSCI Mexico ETF	0.77%	-0.06%
PFF	iShares U.S. Preferred Stock ETF	0.91%	0.00%	XHB	SPDR S&P Homebuilders ETF	2.54%	0.22%
AGG	iShares Core U.S. Aggregate Bond ETF	0.55%	-0.03%	XLY	Consumer Discretionary Select Sector SPDR Fund	1.18%	0.00%
BND	Vanguard Total Bond Market ETF	0.84%	-0.03%	XRT	SPDR S&P Retail ETF	0.71%	0.08%
DBC	PowerShares DB Commodity Index Tracking Fund	3.35%	-0.03%	XLP	Consumer Staples Select Sector SPDR Fund	1.64%	0.02%
IAU	iShares Gold Trust	2.17%	-0.02%	XLE	Energy Select Sector SPDR Fund	0.59%	-0.12%
SLV	iShares Silver Trust	4.28%	-0.27%	XOP	SPDR S&P Oil & Gas Exploration & Production ETF	0.99%	-0.18%
USO	United States Oil Fund	3.57%	-0.05%	XLF	Financial Select Sector SPDR Fund	2.80%	0.06%
UNG	United States Natural Gas Fund	4.71%	-0.31%	KBE	SPDR S&P Bank ETF	2.07%	0.17%
FXE	CurrencyShares Euro Trust	0.84%	0.01%	KRE	SPDR S&P Regional Banking ETF	1.67%	0.22%
FXY	CurrencyShares Japanese Yen Trust	0.58%	-0.02%	XLV	Health Care Select Sector SPDR Fund	1.09%	0.09%
SPY	SPDR S&P 500 ETF	0.03%	0.03%	XBI	SPDR S&P Biotech ETF	1.90%	0.54%
QQQ	PowerShares QQQ ETF	0.89%	0.08%	XLI	Industrial Select Sector SPDR Fund	1.56%	-0.01%
DIA	SPDR Dow Jones Industrial Average ETF	0.03%	0.03%	XLB	Materials Select Sector SPDR Fund	1.78%	0.06%
IWB	iShares Russell 1000 ETF	0.41%	0.03%	XME	SPDR S&P Metals & Mining ETF	2.34%	-0.19%
IWF	iShares Russell 1000 Growth ETF	0.55%	0.05%	XLK	Technology Select Sector SPDR Fund	2.03%	0.07%
IWD	iShares Russell 1000 Value ETF	0.60%	0.03%	XLU	Utilities Select Sector SPDR Fund	1.81%	-0.01%
IWR	iShares Russell Midcap ETF	0.75%	0.06%	IYR	iShares U.S. Real Estate ETF	1.38%	-0.09%
IWP	iShares Russell Mid-Cap Growth ETF	0.72%	0.06%	SH	ProShares Short S&P 500	2.89%	-0.04%
IWS	iShares Russell Mid-Cap Value ETF	0.80%	0.06%	SSO	ProShares Ultra S&P 500	0.18%	0.06%
IWM	iShares Russell 2000 ETF	0.51%	0.21%				
IWO	iShares Russell 2000 Growth ETF	0.57%	0.25%				
IWN	iShares Russell 2000 Value ETF	0.70%	0.15%				

The trading strategy outperforms buy and hold in all cases. Some of the ETFs generate fairly substantial average daily returns; for example see DBC, IAU, SLV, USO, and UNG, the commodity ETFs. The average over assets of strategy returns is 1.36%, compared with 0.01% for buy and hold.

Recall that out of sample R^2 has been constructed for these funds from April 2015 data. The correlation between out of sample R^2 and strategy returns is .456. This makes sense; one cannot make money trading an asset if an asset's return is not predictable. This suggests a refinement to the strategy of trading just one fund. The trader wants to trade assets that show return predictability and for which she just received a strong signal. Suppose that on the night of April 30, 2015, she has a list of ETFs sorted by out of sample R^2. She wishes to trade assets that show return predictability, and therefore decides to trade only assets that have an R^2 above .075. There are 10 of these, shown in the following table.

Table 4.2: Assets with out of sample R^2 above .075

Asset		Out of Sample R^2
EWU	iShares MSCI United Kingdom ETF	0.145
XLF	Financial Select Sector SPDR Fund	0.130
SH	ProShares Short S&P 500	0.128
DBC	PowerShares DB Commodity Index Tracking Fund	0.118
IAU	iShares Gold Trust	0.107
BWX	SPDR Barclays Intl Treasury Bond ETF	0.106
BND	Vanguard Total Bond Market ETF	0.099
EWJ	iShares MSCI Japan ETF	0.098
JNK	SPDR Barclays Capital High Yield Bond ETF	0.086
MBB	iShares MBS ETF	0.082

Let $I = \{EWU, ..., MBB\}$. Suppose now that it is May 2015, and the trader has received predictive signals $\{\hat{r}_{i,t,d}\}_{i \in I}$. The trader wants to trade the asset for which the signal is strongest, so she trades asset $i^\star_{t,d}$, where

$$i^\star_{t,d} = \{i \in I || \hat{r}_{i,t,d}| \geq |\hat{r}_{j,t,d}|, \forall j \in I, j \neq i\}$$

Define

$$D^\star_{t,d} = \begin{cases} 1 & \hat{r}_{i^\star_{t,d},t,d} \geq 0 \\ -1 & \hat{r}_{i^\star_{t,d},t,d} < 0 \end{cases}$$

The trader realizes return $\pi^\star_{t,d} = \hat{r}_{i^\star_{t,d},t,d} D^\star_{t,d}$. For May 2015, this strategy realizes average daily returns of

$$\Pi^\star = \frac{1}{20} \sum_{d=1}^{D} \sum_{t=16}^{370} \pi^\star_{t,d} = 6.17\%$$

This is an impressive return, but a few caveats are necessary. ETF fees, fees paid to the market maker, and other transaction costs will eat into this return. The 6.17% return figure results from a $6.17/355 = 1.7$ bp return per transaction, as we are transacting every minute. Transaction costs need to be less than this 1.7 bp figure to produce positive returns.

Suppose due to transaction costs the trader decides to transact only when the signal is sufficiently strong, when the highest absolute predicted return exceeds some threshold c. The trader potentially trades asset $i^\star_{t,d}$, where

$$i^\star_{t,d} = \{i \in I || \hat{r}_{i,t,d}| \geq |\hat{r}_{j,t,d}|, \forall j \in I, j \neq i\}$$

Define

$$D^{\star}_{t,d} = \begin{cases} 1 & \hat{r}_{i^{\star}_{t,d},t,d} \geq c \\ -1 & \hat{r}_{i^{\star}_{t,d},t,d} \leq -c \\ 0 & -c < \hat{r}_{i^{\star}_{t,d},t,d} < c \end{cases}$$

If the trader chooses to transact, she holds the asset for one minute and liquidates her position at the end of the minute. The trader then realizes return $\pi^{\star}_{t,d} = \hat{r}_{i^{\star}_{t,d},t,d} D^{\star}_{t,d}$.

For different signal thresholds, I report the average daily raw return from the strategy $\frac{1}{20}\sum_{d=1}^{D}\sum_{t=16}^{370}\pi^{\star}_{t,d}$,

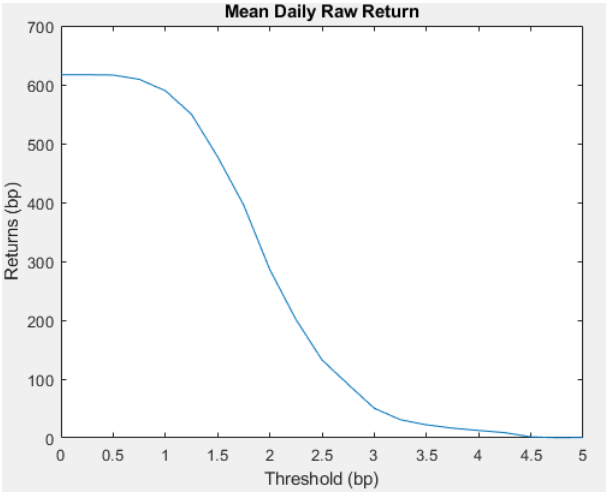


Figure 4.1: Mean daily raw return

the average number of transactions per day $\frac{1}{20}\sum_{d=1}^{D}\sum_{t=16}^{370}|D^{\star}_{t,d}|$,

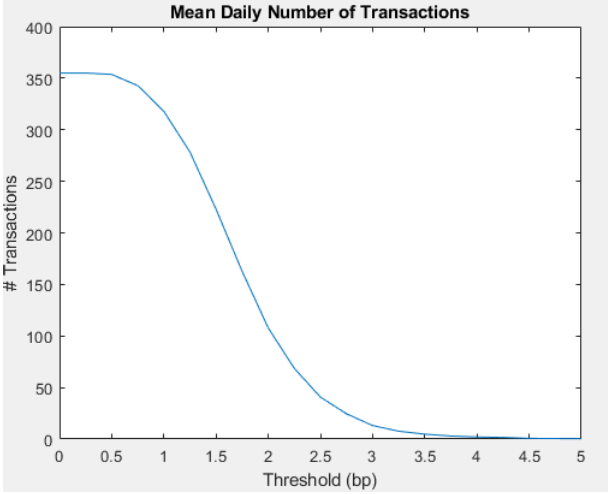


Figure 4.2: Mean daily number of transactions

and the average daily return per transaction $\frac{1}{20}\sum_{d=1}^{D}\frac{\sum_{t=16}^{370}\pi_{t,d}^{\star}}{\sum_{t=16}^{370}|D_{t,d}^{\star}|}$.

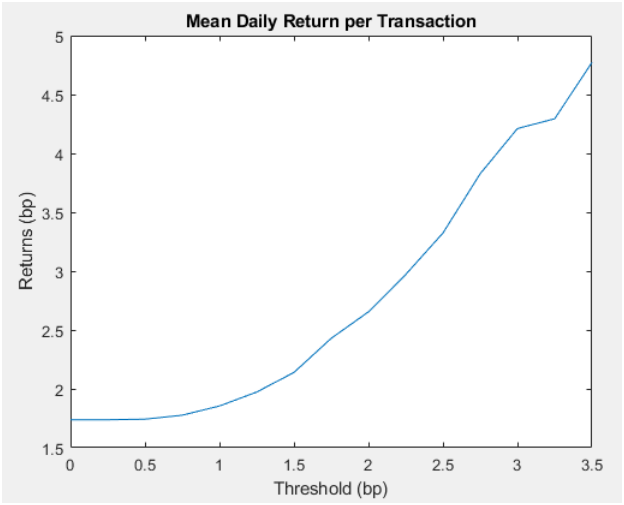


Figure 4.3: Mean daily return per transaction

The above plot gives the transaction cost hurdle: the lower right section of the plot is the region where transaction costs must be to generate positive profits. As the signal threshold increases, the transaction cost hurdle becomes easier to clear and the return per transaction increases. However, this is offset by the decreasing number of transactions. The following plot illustrates this tradeoff. Suppose that the transaction cost is 1 bp per trade. The trader incurs a total transaction cost of 2 bp over the course of the minutes in which she transacts (buy/sell at start of minute, sell/buy at end of minute). The average daily profit is given by $\frac{1}{20}\sum_{d=1}^{D}\left(\sum_{t=16}^{370}\pi_{t,d}^{\star}-(.0002)|D_{t,d}^{\star}|\right)$.

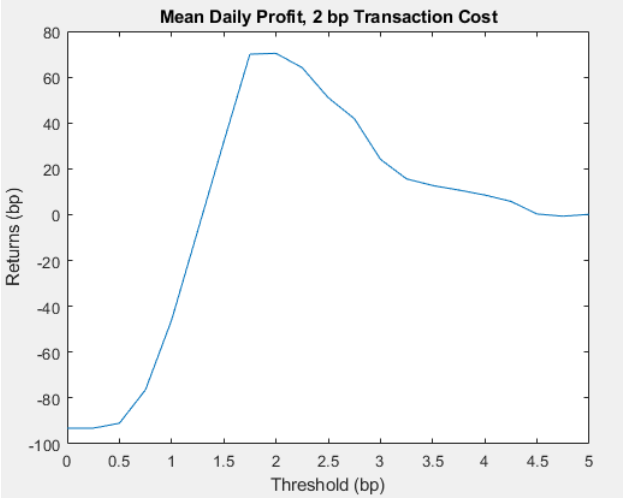


Figure 4.4: Mean daily profit, 2 bp transaction cost

The maximum 0.7% return after transaction costs is much less than the 6.17% return figure given earlier.

A final question to be asked is what do the transaction costs of ETFs look like. The following histogram gives the average bid-ask spread over the data sample for the 63 ETFs.

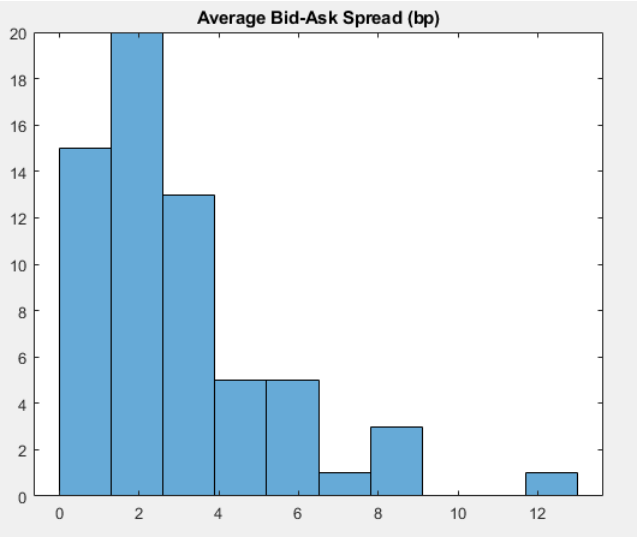


Figure 4.5: Histogram of average bid-ask spreads

The spreads range from a low of 0.48 bp for SPY (SPDR S&P 500 ETF) to a high of 12.18 bp for BWX (SPDR Barclays Intl Treasury Bond ETF). Only 6 of the 63 ETFs, and none of the 10 ETFs used in the trading exercise, have an average bid-ask spread that is lower than the 1 bp per trade figure used for the profit calculation. The ability to profit from the return predictability therefore depends on the ability of the trader to trade within the spread. This is outside the realm of possibility for the average retail investor, but not for specialized institutional investors, with access to dark pools, direct market access, and other sophisticated tools.

5 Conclusion

This work demonstrates that an asset's past returns, volume, depth, and spread contain predictive information not only for predicting its own returns, but also the returns of assets across the economy. The machine learning model which performs the best is the main effects only LASSO, which has an average out of sample R^2 of 0.0347. Additional complexity does not improve this number. The model's predictability can be turned into substantial trading revenue, although transaction costs could substantially reduce these profits.

A natural question is whether the results of this analysis generalizes to other assets. The choice to use just the 63 ETFs is because it keeps the problem small enough to estimate in a reasonable amount of time. But if enough computational resources were available, one could, for instance, perform the same analysis using the 500 stocks listed in the S&P 500 or an even larger universe of assets. Another question is whether these results generalize to other time horizons. High frequency traders now operate at the second, not minute, horizon. Modeling at the higher frequency would require much more computational resources, and even if it were available, most of the ETFs are not liquid enough to model at the second frequency. These questions will become easier to answer as trading becomes more computerized and the quality and quantity of financial data increases.

6 References

Bai, J., & Ng, S. (2008) Large Dimensional Factor Analysis. Foundations and Trends in Econometrics (3), 89-163.

Breiman, L. (2001) Random forests. Machine Learning (45), 5–32.

Campbell, J.Y., Shiller, R.J. (1988) Stock prices, earnings, and expected dividends. Journal of Finance (43), 661–676.

Chinco, A., Clark-Joseph, A., & Ye, M. (2018) Sparse Signals in the Cross-section of Returns. Journal of Finance, Forthcoming.

Connor, G. & Korajczyk, R.A. (2007) Factor Models of Asset Returns, in Encyclopedia of Quantitative Finance, ed. Cont, R., Chicester: Wiley.

Copeland, T.E. & Gabai, D. (1983) Information Effects on the Bid-Ask Spread. Journal of Finance (38), 1457-1469.

Diaz-Uriarte, R. & de Andres, S.A. (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics (7), 1471–2105.

Easley, D. & O'Hara, M. (1987) Price, Trade Size, and Information in Securities Markets. Journal of Financial Economics (19), 69-90.

Fama, E. (1970) Efficient capital markets: a review of theory and empirical work. Journal of Finance (25), 383–417.

Flury, B. (1988) Common principal components and related models. New York: Wiley.

Genuer, R., Poggi, J.M., & Tuleau, C. (2008) Random Forests: Some Methodological Insights. arXiv:0811.3619.

Geweke, J. (1977) The Dynamic Factor Analysis of Economic Time Series, in Latent Variables in Socio-Economic Models, eds. Aigner D.J. & Goldberger, A.S., Amsterdam: North-Holland.

Glosten, L.R. & Milgrom, P.R. (1985) Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Trades. Journal of Financial Economics (14), 71-100.

Hasbrouck, J. (1991) Measuring the Information Content of Stock Trades. Journal of Finance (46), 179-207.

Hasbrouck, J. & Seppi, D. (2001) Common factors in Prices, Orders, and Liquidity. Journal of Financial Economics (3), 383-411.

Hastie, T., Tibshirani, R., & Friedman, J. (2009) The Elements of Statistical Learning. New York: Springer.

Jackson, J.E. (1991) A user's guide to principal components. New York: Wiley.

Kothari, S.P. & Shanken, J. (1997) Book-to-market, dividend yield, and expected market returns: a time-series analysis. Journal of Financial Economics (44), 169–203.

Lee, C.M., Mucklow, B., & Ready, J.M. (1993) Spreads, Depths, and the Impact of Earnings Information: An Intraday Analysis. Review of Financial Studies (6), 345–374.

Lee, C.M. & Ready, J.M. (1991) Inferring Trade Direction from Intraday Data. Journal of Finance (46), 733-746.

Malkiel, B.G. (1973) A Random Walk Down Wall Street. New York: Norton.

Markowitz, H.M. (1952) Portfolio Selection. The Journal of Finance (7), 77–91.

Meinshausen, N. & Yu, B. (2009) Lasso-Type Recovery of Sparse Representations for High-Dimensional Data. The Annals of Statistics (37), 246–270.

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine (2), 559–572.

Probst, P., Wright, M.N., & Boulesteix, A.L. (2019) Hyperparameters and tuning strategies for random forest, arXiv preprint: 1804.03515.

Rozeff, M.S. (1984) Dividend yields are equity risk premiums. Journal of Portfolio Management (11), 68–75.

Russell, J.R. (2017) Time Series and Cross-Sectional Properties of Equity Market Liquidity with Applications to the Financial Crisis. Working Paper.

Sargent, T.J., & Sims, C.A. (1977), Business Cycle Modeling Without Pretending to Have Too Much A-Priori Economic Theory, in New Methods in Business Cycle Research, eds. Sims, C.A & Granger, C.W., Minneapolis: Federal Reserve Bank of Minneapolis.

Stock, J.H. & Watson, M.W. (2011) Dynamic Factor Models, in The Oxford Handbook of Economic Forecasting, eds. Clement M.P. & Hendry, D.F., Oxford: Oxford University Press.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (58), 267–288.

Xiu, D., Kelly, B. & Gu, S. (2018) Empirical Asset Pricing via Machine Learning. Chicago Booth Research Paper No. 18-04.

Zhao, P. & Yu, B. (2006) On Model Selection Consistency of Lasso. Journal of Machine Learning Research (7), 2541–2563.