Market Response to Liquidity Shocks in the Limit Order Book

Kartikey Asthana[*], Roberto A. Colón Quiñones[†], Joshua Romero[‡]
Stanford University, CS229 Machine Learning, Autumn 2015

Abstract

This report provides empirical evidence that the bid-ask spread and mid-price exhibit strong mean reversion following a liquidity shock in the limit order book. The mean behavior of buy/sell initiated shocks is to increase/decrease the macroscopic mid-price of the security. Linear as well as non-linear models are trained to incorporate pre-shock information for predicting post-shock behavior. However, the features tend to add very little predictive power beyond the statistical mean behavior. We show that this is largely on account of high volatility associated with high priced securities. Defining a measure of error which is normalized by volatility indicates that features do have some intrinsic predictive capability. In this new measure, linear regression, weighted to account for heteroskedasticity, outperforms SVR yielding an efficient prediction algorithm for high frequency data.

1 Introduction

Exchange-traded financial instruments often incorporate a transparent buyer-seller matching system known as a central limit order book. For each security, this book maintains a real-time record of bid/ask prices and volumes that market participants are willing to trade at. The difference between the highest bid and lowest ask price is referred to as the bid-ask spread and is typically narrow for actively traded securities. This spread is a quantitative measure of the security's liquidity since a narrow spread implies a smaller premium to be paid for buying/selling and vice-versa.

A *trade* is executed when a buyer/seller is willing to cross the bid-ask spread. If the buyer/seller places a large enough order that consumes all available volume at the best price, then the bid-ask spread changes, leading to a *liquidity shock*. Such a shock is followed by a period during which participants replenish the order book with new bid/ask orders. Such limit orders that update the best bid/ask price are referred to as *quote* events.

The rate of recovery or mean reversion of the spread and mid-price is indicative of the market's resiliency and is of great importance to traders, exchanges and regulators. An efficient and accurate model for market resiliency can directly benefit traders, allowing them to reduce market impact costs by splitting large orders across time. Moreover, such a model can improve existing back-testing platforms for trading strategies by including a feedback from the market.

The goal of this project is to develop an accurate model for predicting the short-term response of the market to liquidity shocks in the order book. In particular, we are interested in forecasting values of the best bid and best ask prices for fifty time steps following an order that triggers a liquidity shock. The input to our algorithm consists of information on fifty pre-shock transactions, the shock-causing transaction, as well as macroscopic indicators regarding each security. We then use variants of linear regression (LR) and support vector regression (SVR) to predict increments in mid-price and bid-ask spread for the fifty time steps after the shock.

2 Related work

Numerous stochastic models have been proposed for the state of a limit order book. These include static models [1] where the trader's choice to place limit or market orders is made exogenously, as well as dynamic models [2] that make such a choice internally by factoring waiting costs. It has been shown [3] that liquidity shocks temporarily increase the bid-ask spread which reverts to a competitive level as new orders arrive. We empirically verify this observation in Section 3.1. Dynamics of market resiliency have also been studied in [4] where the authors show that the two principal factors governing the rate of mean reversion are the proportion of patient traders and rate of order arrival.

It is important to note that a liquidity shock in the limit order book is a high frequency phenomenon wherein the change in spread is on the order of a few ticks. It must be contrasted with macroscopic shocks during financial crises wherein the bid-ask spread widens catastrophically impeding traders from closing positions. Extensive investigations have been conducted into the dynamics of macroscopic liquidity across equity and bond markets [5], premia associated with shocks in equilibrium markets [6], impact of macroscopic shocks on HFT participants [7], shocks due to federal monetary policies [8] etc.

Linear Regress.	k-Nearest Neighbors	SVR	Random Forest	k-Means
0.778	0.786	0.796	0.797	0.798

Table 1: Relative performance (RMSE) of algorithms utilized by a top-ranking kaggle.com competition participant.

This problem was part of a competition sponsored by the Capital Markets Cooperative Research Centre, hosted on-

[*]PhD candidate, kasthana@stanford.edu
[†]PhD candidate, racolon@stanford.edu
[‡]PhD candidate, jdromero@stanford.edu

line at kaggle.com [9]. The competition ended in January 2012 and a discussion ensued wherein participants shared performance of their algorithms on an undisclosed test set. Table 1 shows the root mean squared error obtained from six different algorithms. The proximity of results indicates that there is negligible predictive ability in the features! In this report, we show how this observation can be verified, explained and then circumvented.

3 Data and Benchmark

Trade and quote (TAQ) data for 110,000 examples of liquidity shocks across several securities and trading times were obtained directly from the competition website. Each example includes information on 50 transactions leading to the shock and an equal number after the shock. All prices are measured in units of pence. For pre-shock transactions, the data provides (i) nature (trade/quote), (ii) time, (iii) best bid price, and (iv) best ask price for each transaction. Regarding the shock itself, we are provided (i) volume-weighted average price, (ii) volume, and (iii) nature (buy/sell initiated) of the trade causing the liquidity shock. Additionally, we are also provided (i) the number of previous day on-market trades, and (ii) sum of previous day on-market trade values. For post-shock transactions, only data on bid and ask prices is available.

Of the total data, 80,000 examples were randomly chosen to form the training set, 20,000 were held out for optimizing parameters in SVR and the remaining 10,000 were designated as the test set.

3.1 Mean-reversion of mid and spread

Prior to exploring learning algorithms, we begin with a statistical characterization of the provided high-frequency data to verify the observations in [3]. Denoting B_t, A_t as the best bid and ask prices for a given time t, the mid price and bid-ask spread are respectively defined as

$$M_t = \frac{A_t + B_t}{2}, \quad S_t = A_t - B_t. \tag{1}$$

For liquid securities, both the mid-price and spread are often assumed to be mean reverting over short time horizons, i.e. they have a tendency to return to some non-observable mean levels. The time series for both quantities can then be represented by Ornstein-Uhlenbeck (OU) processes [10],

$$\begin{aligned} \mathrm{d}M_t &= \alpha_M(\mu_M - M_t)\mathrm{d}t + \sigma_M \mathrm{d}B_t^M, \\ \mathrm{d}S_t &= \alpha_S(\mu_S - S_t)\mathrm{d}t + \sigma_S \mathrm{d}B_t^S, \end{aligned} \tag{2}$$

where B_t^M, B_t^S are standard Brownian motions, σ_M, σ_S are the respective volatilities, μ_M, μ_S are the mean levels, and the critical quantities α_M, α_S are the rates of mean reversion. The two processes are correlated $\mathrm{Cov}(B_t^M, B_t^S) = \rho t$. Owing to linearity in the diffusion term, one can show that the expected mid and spread are given by

$$\begin{aligned} \mathrm{E}[M_t] &= (M_{t_0} - \mu_M)e^{-\alpha_M(t-t_0)} + \mu_M, \\ \mathrm{E}[S_t] &= (S_{t_0} - \mu_S)e^{-\alpha_S(t-t_0)} + \mu_S, \end{aligned} \tag{3}$$

where $t_0 = 49$ is the time of incidence of the shock. The exponential damping in time is a strong indicator of mean-reversion and can be verified using the empirical estimators

$$\begin{aligned} \hat{\mathrm{E}}[M_t - M_{t_0}] &= \frac{1}{m}\sum_{i=1}^{m}({M_t}^{(i)} - {M_{t_0}}^{(i)}), \\ \hat{\mathrm{E}}[S_t - S_{t_0}] &= \frac{1}{m}\sum_{i=1}^{m}({S_t}^{(i)} - {S_{t_0}}^{(i)}), \end{aligned} \tag{4}$$

as a function of time, where m denotes the number of examples. The post-shock values of these quantities for the training set are plotted in Fig. 1. A distinction is made between buy-initiated and sell-initiated shocks as they influence the mid-price in opposite directions. We see that both the mid-price and bid-ask spread exhibit strong mean reversion. In the case of a buy-initiated shock, the mean level is higher than than the value prior to the shock, i.e. the shock increases the long-term or macro level price of the security. The corresponding decrease happens for a sell-initiated stock. This phenomenon captures the basic process by which securities change in value over time. The mean level for the spread is slightly higher than the pre-shock level. This highlights the impact of the shock in reducing liquidity of the security.

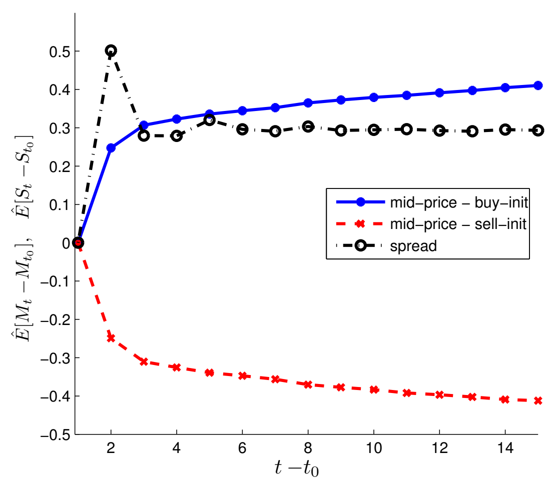

Figure 1: Mean reversion of mid-price and bid-ask spread

3.2 Benchmark

We use the statistical estimators in Eqn. (4) as our benchmark algorithm,

$$\begin{aligned} \hat{M}_t &= M_{t_0} + \hat{\mathrm{E}}[M_t - M_{t_0}], \quad \hat{S}_t = S_{t_0} + \hat{\mathrm{E}}[S_t - S_{t_0}], \\ \hat{B}_t &= \hat{M}_t - \hat{S}_t/2, \quad \hat{A}_t = \hat{M}_t + \hat{S}_t/2, \end{aligned} \tag{5}$$

where $\hat{B}_t, \hat{A}_t$ are the predicted values. This corresponds to a linear mean model consisting of only the intercept term and no features. As dictated by the competition organizers, accuracy is measured in terms of the root mean square error across the entire set,

$$\mathrm{RMSE}^{(i)} = \left[\frac{1}{100}\sum_{t=t_0+1}^{t_0+50}(B_t^{(i)} - \hat{B}_t^{(i)})^2 + (A_t^{(i)} - \hat{A}_t^{(i)})^2\right]^{1/2},$$

$$\mathrm{RMSE} = \left[\frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{RMSE}^{(i)}\right)^2\right]^{1/2}, \tag{6}$$

Figure 2 records the histogram of $\log(\mathrm{RMSE}^{(i)}), i = 1, \ldots, m$. We see that the data has a relatively fat right tail with a few isolated errors as large as 45 pence! The extreme values are outliers generated by trading events at the beginning of a day's trading [9]. However, the smooth part of the right tail is due to the fact that higher priced securities have higher volatilities (see Section 5). This was also observed in [5] where the authors suggested that past volatility is a strong indicator of liquidity. Performance of the benchmark is recorded in Table 3.

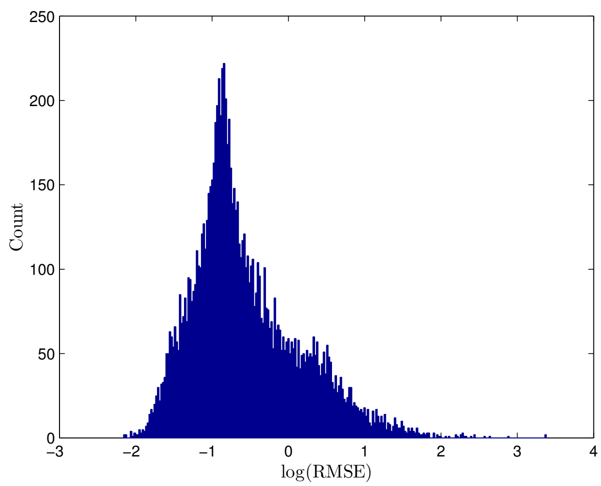

Figure 2: Histogram of $\log(\mathrm{RMSE}^{(i)})$ for the benchmark.

4 Linear Regression & Diagnostics

The simplest extension of the benchmark algorithm is to incorporate first order terms of available features in the mean model. This can be achieved by regressing linearly onto all 206 available features (pre-shock transactions, at shock transaction and security specific scalars). Three sets of models are trained to predict the fifty discrete post-shock values of (i) mid-price for a buy-initiated shock, (ii) mid-price for a sell-initiated shock, and (iii) spread. Extensive analysis is performed to determine the bias/variance trade-off of this simple approach, and to measure the predictive ability of the feature set.

4.1 Coefficient of determination

The effect of incorporating linear feature-terms can be quantified by measuring the fraction of the response's variance that is explained by the features. This is expressed by the coefficient of determination [11],

$$R^2 = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}} = \frac{SS_{\mathrm{reg}}}{SS_{\mathrm{tot}}}, \tag{7}$$

Figure 3 plots this statistic against the 50 post-shock response variables for each of the three types of responses. We see that for $t - t_0 > 2$, the linear model captures less than 10% of the variance in the response. In other words, high frequency market fluctuations (increments in mid or spread) are essentially indifferent to the type of security or pre-shock transaction history. The linear model merely captures the mean behavior expressed by the benchmark. This explains why top participants in the competition failed to improve on the performance from linear regression despite trying non-linear supervised and unsupervised learning algorithms. Note that the response right after the shock has $R^2 = 1$; this is a consequence of repetition in the data supplied by the problem organizers.

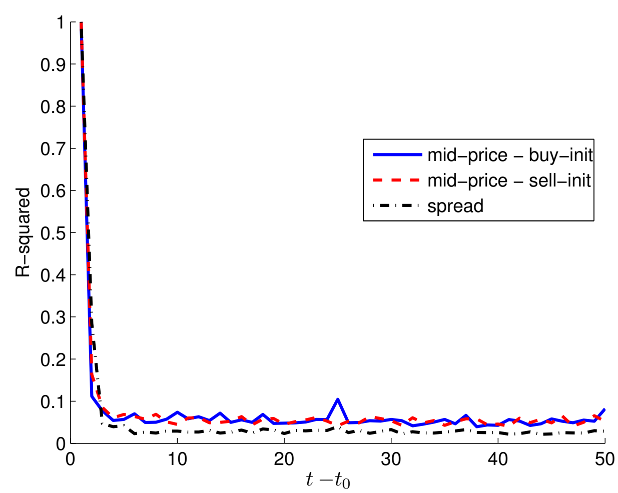

Figure 3: R^2 for the 50 post-shock variables

4.2 Non-normality of residuals

Linear regression is formulated under the assumption that the response is distributed normally about the mean model. We can test this assumption by plotting the empirical quantiles of the regression residuals against the quantiles of a normal distribution with zero mean and variance equal to the empirical variance of the residuals. Figure 4 records the QQ plot of the spread-residual at $t - t_0 = 2$. We see that the data is exceptionally fat-tailed with the largest values greater then 10 pence!. This is in line with the observations in Section 3.2.

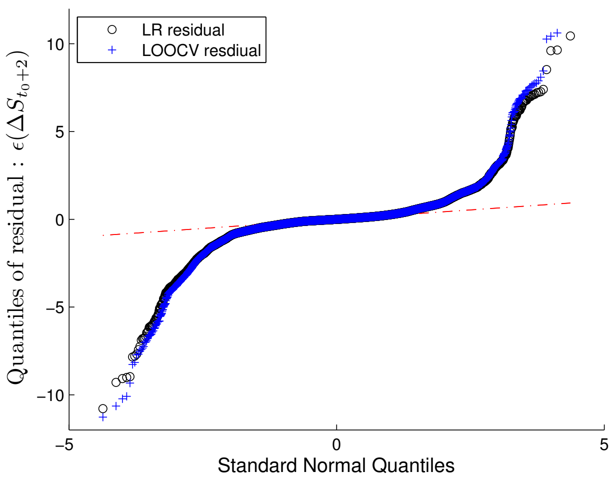

Figure 4: QQ plot for Leave-One-Out Cross Validation

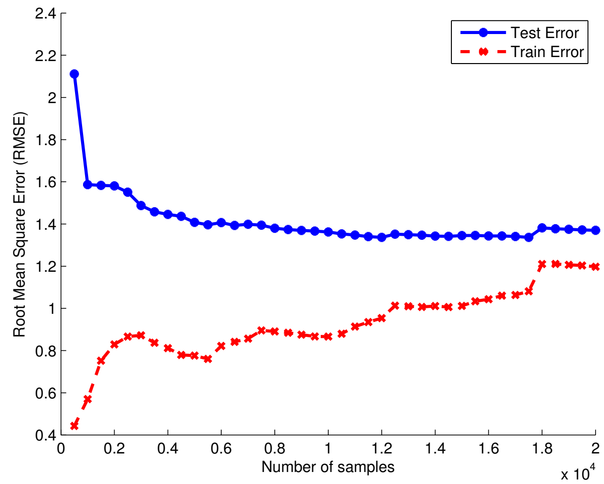

Figure 5: RMSE learning curve for linear regression

4.2.1 Leave-one-out CV

We can further verify the fat-tailed nature of the response by performing leave-one-out cross validation for the linear model and observing the change in prediction for the held-out example,

$$\epsilon_i = y_i - \hat{y}_i^{-i}, \quad i = 1, 2, \ldots, m, \tag{8}$$

where y_i is the observed response for the i^{th} example, and $\hat{y}_i^{-i}$ is the prediction for the i^{th} example using all except the i^{th} example. In the case of linear regression, a simplification can be achieved [11] using the Sherman-Morrison formula that allows for an analytical expression of the form,

$$\epsilon_i = \frac{\hat{y}_i}{1 - H_{ii}}, \quad i = 1, 2, \ldots, m, \tag{9}$$

where $\hat{y}_i$ is the prediction for the i^{th} example using all m examples and H_{ii} is the leverage of the i^{th} example. The QQ plot of leave-one-out residuals is also recorded in Figure 4. Once again. the fat tails are evident. In section 5, we show, empirically, that these fat-tails are largely on account of heteroskedasticity, since different securities have different inherent volatilities.

4.3 Learning curve

We have already seen that the entire set of 206 linear features has very little prediction power beyond the benchmark for most of the response variables. Figure 5 plots the learning curve for RMSE (Eqn. (6)) where the test error is calculated on the 20,000 examples in the held-out set mentioned in Section 3. The plot shows that the linear model suffers from high bias, with a rapid plateau of test error and increasing training error. This indicates that either the current model is not capturing nonlinear dependence on the features, or that the overall RMSE is largely controlled by the outliers mentioned in Section 4.2.1. In the upcoming sections, we provide some evidence of the latter.

5 k-means Clustering

As mentioned in Sections 3.2 and 4.2.1, we claim that the non-normality of the regression residuals is largely due to different variances for different training examples. This claim is motivated from the observation that price processes are often modeled as geometric Brownian motions wherein the volatility is proportional to the price, i.e. we expect higher priced securities to have higher volatilities. In particular, from Eqn. (6),

$$\text{RMSE}^2_{\text{bench}} \simeq \frac{1}{m} \sum_{i=1}^{m} \text{E}\left[\text{Var}(B_t^{(i)}) + \text{Var}(A_t^{(i)})\right], \tag{10}$$

which shows that each example contributes error proportional to its variance. This can be empirically verified by training separate models on clusters sorted by bid price. Towards this end, we use k-means to identify 3 centroids in the training data, train 3 sets of models, and record the corresponding RMSE values for clusters in the test set. Table 2 shows that RMSE grows with growing cluster price, confirming our claim. Hence, the RMSE for the full test set is largely dictated by the proportion of high priced securities.

	Clusters		
Centroid	£3.27	£11.75	£22.77
RMSE	0.42p	1.60p	1.90p
RMSFE	0.33	0.34	0.34

Table 2: Clustering with 3-means based on bid price

In order to circumvent this limitation, we define an alternative metric, the Root Mean Square Fractional Error,

$$\text{RMSFE} \tag{11}$$

$$= \left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{100}\sum_{t=t_0+1}^{t_0+50}\left(\frac{B_t^{(i)} - \hat{B}_t^{(i)}}{\delta}\right)^2 + \left(\frac{A_t^{(i)} - \hat{A}_t^{(i)}}{\delta}\right)^2\right]^{1/2},$$

$$\delta = \max\{\max_t B_t - \min_t B_t, \max_t A_t - \min_t A_t\}, \tag{12}$$

which measures the fraction of volatility of the response that is explained by the learning algorithm. Table 2 shows that RMSFE remains nearly constant across the clusters as desired.

6 Weighted LR: heteroskedasticity

The empirical observations from the previous section show that the normality assumption in vanilla linear regression is clearly violated. An improvement can be readily incorporated by estimating the volatility of each example from pre-shock prices,

$$\sigma_M^{(i)^2} \simeq (\hat{\sigma}_M^{(i)})^2 = \frac{1}{t_0 - 1}\sum_{t=1}^{t_0}(M_t^{(i)} - \bar{M}^{(i)})^2,$$
$$\sigma_S^{(i)^2} \simeq (\hat{\sigma}_S^{(i)})^2 = \frac{1}{t_0 - 1}\sum_{t=1}^{t_0}(S_t^{(i)} - \bar{S}^{(i)})^2, \qquad (13)$$

These differing volatilities can be accommodated in the regression by weighting the examples inversely by their variances. In particular,

$$S_t^{(i)} = \sum_j \theta_j z_j^{(i)} + \epsilon_S^{(i)}, \;\; i = 1, \dots, M, \; t > t_0$$
$$\epsilon_S^{(i)} \sim \mathcal{N}\left(0, \sigma_S^{(i)^2}\right), \qquad (14)$$

where $z_j, j = 1, \dots, 206$ denote the features. Denoting ${\Sigma_S^{-2}}_{i,j} = (1/\sigma_S^{(i)^2}) 1_{i=j}$ and using boldface for denote vectors, the normal equations for this case are given by

$$\boldsymbol{Z}^T \Sigma_S^{-2} \boldsymbol{Z} \boldsymbol{\theta} = \boldsymbol{Z}^T \Sigma_S^{-2} \boldsymbol{S}_t. \qquad (15)$$

Such a weighted linear regression results in significant improvement of the RMSFE (see Table 3) while the RMSE remains roughly the same, as expected.

7 Support Vector Regression

In this section we show that the lack of fit for linear regression is not due to a non-linear interrelation between the response and features. This further builds upon the discussion from Section 4.3. Towards this end, we use LIBSVM [12] to train a non-linear regression model based on ν-SVR,

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{\xi} \succeq 0, \boldsymbol{\xi}^* \succeq 0, \epsilon \geq 0} \; & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\nu\epsilon + \frac{C}{l} 1^T(\boldsymbol{\xi} + \boldsymbol{\xi}^*), \\ \text{s.t.} \;\; & \boldsymbol{w}^T\phi(\boldsymbol{z}_i) + b - y_i \leq \epsilon + \xi_i, \\ & -\boldsymbol{w}^T\phi(\boldsymbol{z}_i) - b + y_i \leq \epsilon + \xi_i^*, \quad i = 1, \dots, m, \end{aligned} \qquad (16)$$

where $\boldsymbol{z}$ is the vector of features, y is the response and C, ν are regularization parameters. We choose the radial basis function as our kernel for the inner product,

$$< \phi(\boldsymbol{z}_1), \phi(\boldsymbol{z}_2) >= \exp\left(-\gamma \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2^2\right). \qquad (17)$$

As directed by the authors in [13], we begin by determining reasonable choices for the parameters C and γ. This is

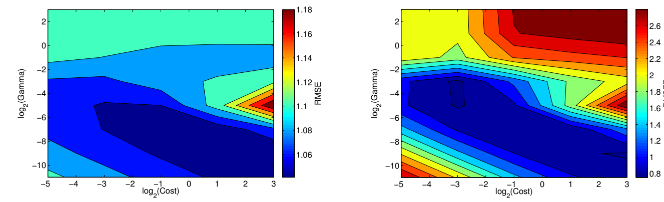

Figure 6: ν-SVR : RMSE/RMSFE on a grid of (C,γ).

achieved by calculating errors over the 50 response variables for a grid of parameter values. For computational ease, ν is fixed at 1/2. As the training set is massive and for each point 50 models must be trained, we perform this grid search using a small training set of 5,000 examples cross validated against a held-out set of 1,000 samples. The contour plot for both RMSE and RMSFE is recorded in Fig. 6. We see that the smallest generalization error is obtained for very small values of γ in the RBF kernel, i.e. the optimal model is essentially linear in the features.

Limited by computational expense, we use optimal values from Fig. 6 and train on the held-out set containing 20,000 samples, instead of the full training set.

8 Performance and Conclusion

	Train		**Test**	
Algorithm	RMSE	RMSFE	RMSE	RMSFE
Benchmark	1.41	2.23	1.23	1.26
Linear Regression	1.26	0.51	1.16	0.40
Weighted LR	1.36	0.35	1.20	0.34
SVR	—	—	1.19	0.45

Table 3: Comparative performance of the algorithms.

Table 3 records the training and testing RMSE (Eqn. (6)) and RMSFE (Eqn. (11)) for the four algorithms described in the previous sections. The results for RMSE are similar to those obtained by top kaggle-competition participants in Table 1. Note that the actual values differ since the test sets are different (the competition test set was not made public). All algorithms produce RMSE values very close to that of the benchmark showing that learning based on pre-shock data was not beneficial for RMSE. However, for the fractional error, RMSFE, learning algorithms provide significant improvement over the benchmark. Among the three supervised learning models, weighted LR performs best, yielding an efficient prediction algorithm for high frequency data.

The table strongly supports the claims made in Sections 4.1, 5 and 7. The RMSE of the test set is largely dictated by high priced securities associated with high volatility. Thus, adding linear/non-linear features adds very little predictive capability in comparison to the benchmark. However, RMSFE, which normalizes error by volatility, indeed decreases as the algorithm is refined, showing that the features do have some intrinsic predictive capability in this new measure.

References

[1] Lawrence R. Glosten, 'Is the electronic open limit order book inevitable?', The Journal of Finance 49.4 1127-1161 (1994).

[2] Christine A. Parlour, 'Price dynamics in limit order markets', Review of Financial Studies 11.4 789-816 (1998).

[3] Bruno Biais, Pierre Hillion, and Chester Spatt, 'An empirical analysis of the limit order book and the order flow in the Paris Bourse', Journal of Finance 1655-1689 (1995).

[4] Thierry Foucault, Ohad Kadan, and Eugene Kandel, 'Limit order book as a market for liquidity', Review of Financial Studies 18.4 1171-1217 (2005).

[5] Tarun Chordia, Asani Sarkar, Avanidhar Subrahmanyam, 'An Empirical Analysis of Stock and Bond Market Liquidity', The Review of Financial Studies Vol. 18, No. 1 (2005).

[6] Ming Huang, 'Liquidity shocks and equilibrium liquidity premia', Journal of Economic Theory 109 104–129 (2003).

[7] Bruno Biais, Paul Woolley, 'High Frequency Trading', Toulouse School of Economics, London School of Economics, (2011).

[8] John H. Cochrane, Monika Piazzesi, 'The fed and interest rates: a high-frequency identification', http://www.nber.org/papers/w8839, National bureau of economic research, (2002).

[9] URL: https://www.kaggle.com/c/AlgorithmicTradingChallenge/.

[10] G. E. Uhlenbeck, L. S. Ornstein, 'On the theory of Brownian Motion', Phys. Rev. 36: 823–841 (1930).

[11] Peter Dalgaard,'Introductory statistics with R. Springer Science & Business Media', (2008).

[12] Chih-Chung Chang, Chih-Jen Lin. 'LIBSVM: A library for support vector machines', ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 : 27 (2011).

[13] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, 'A practical guide to support vector classification', (2003).