# Predicting Short-term Market Response to Liquidity Shocks

Kartikey Asthana, Roberto A. Colón Quiñones, Joshua Romero

Stanford University, CS229 Machine Learning, Autumn 2015

## Motivation and Objective

A liquidity shock is a trade that increases the bid-ask spread by consuming all available volume at the best price. The rate of recovery of spread and mid-price, following a shock, is of great importance to traders, exchanges and regulators. The goal of this project is to develop a model for forecasting values of the mid-price and bid/ask spread for a period following a liquidity shock.

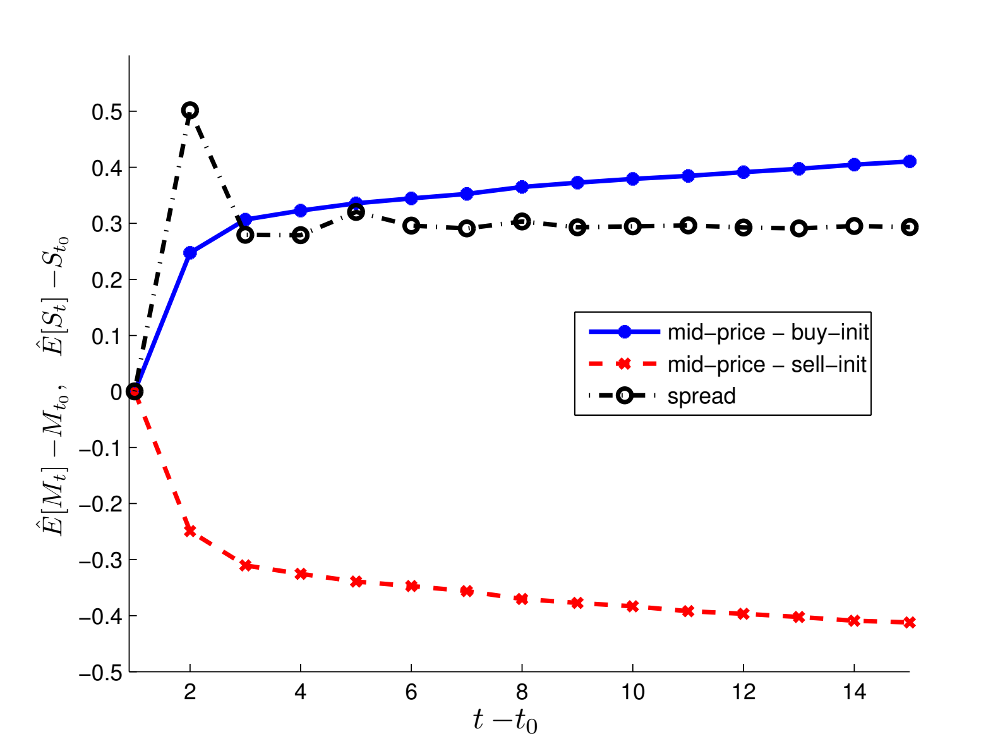
## Data and Benchmark

This problem was part of a competition sponsored by the Capital Markets Cooperative Research Centre, hosted online at kaggle.com, which provides trade and quote (TAQ) data for $> 100,000$ examples.

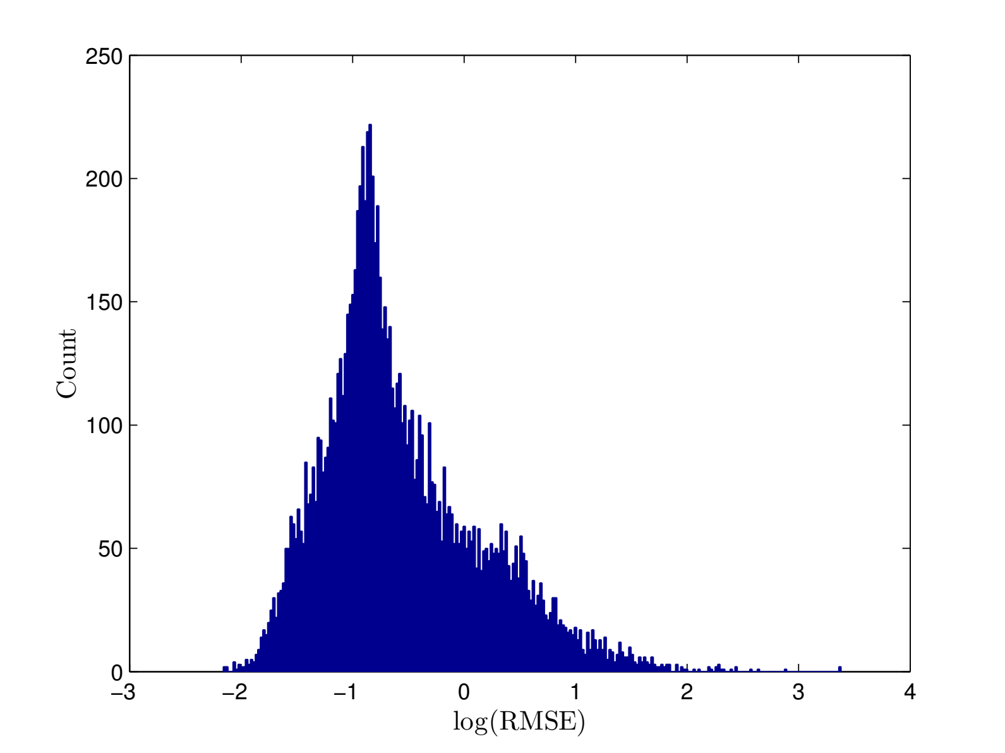
For liquid securities, both the mid and spread are often assumed to be mean reverting over short time horizons,

$$\mathrm{d}S_t = \alpha(\mu - S_t)\mathrm{d}t + \sigma \mathrm{d}B_t,$$
$$E[\mathrm{d}S_t] = \alpha(\mu - S_{t_0})e^{-\alpha(t-t_0)}\mathrm{d}t$$

where the expected increment decays exponentially in time. Mean reversion can be verified by plotting empirical estimators (Fig. a).



(a) Mean reversion for mean, spread

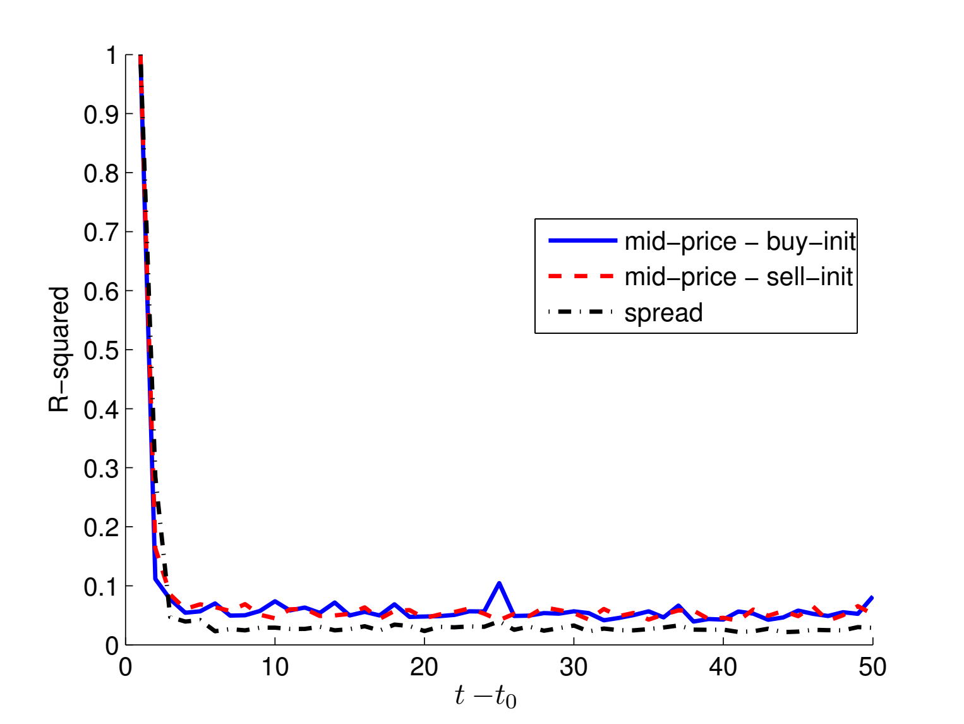


(b) RMSE for benchmark

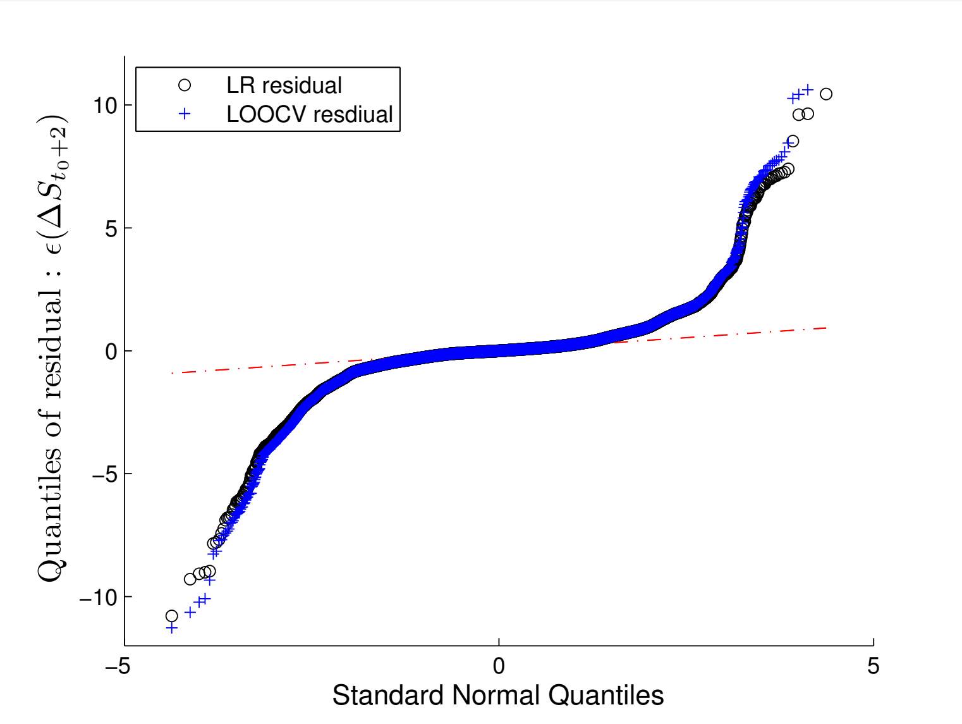
We use these statistical estimators as our benchmark. The histogram for RMSE (Fig. b) shows that the data has several outliers which we show, later, are largely on account of heteroskedasticity.

## Linear Regression

- The entire set of 206 features explains less than 10% of the variance in the post shock mid and spread. The residuals from regression / leave-one-out CV show exceptionally fat tails.

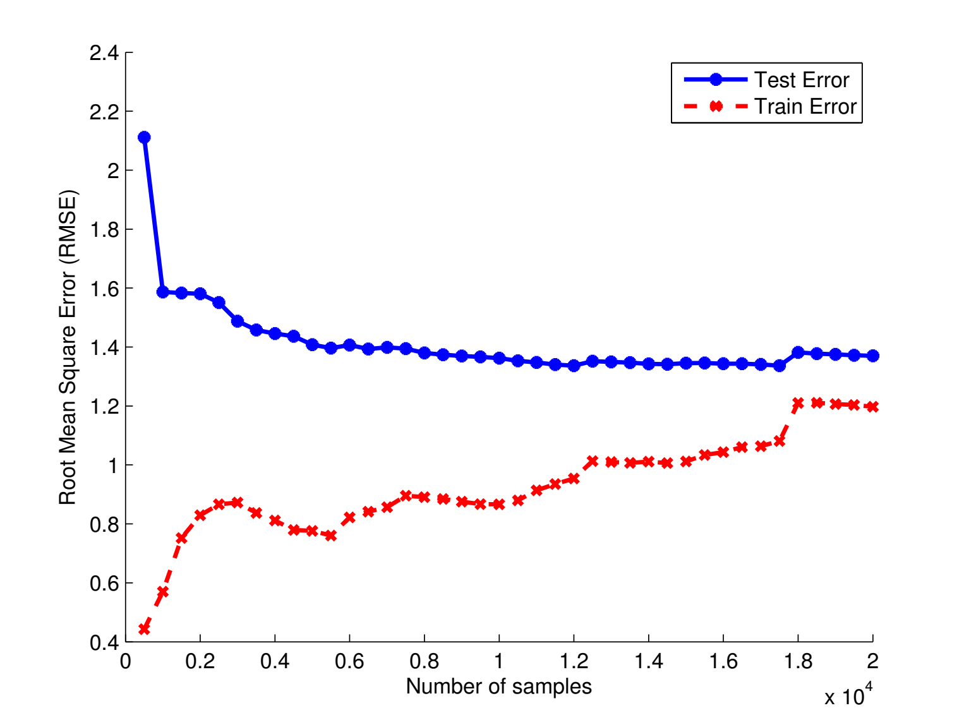


Coefficient of determination

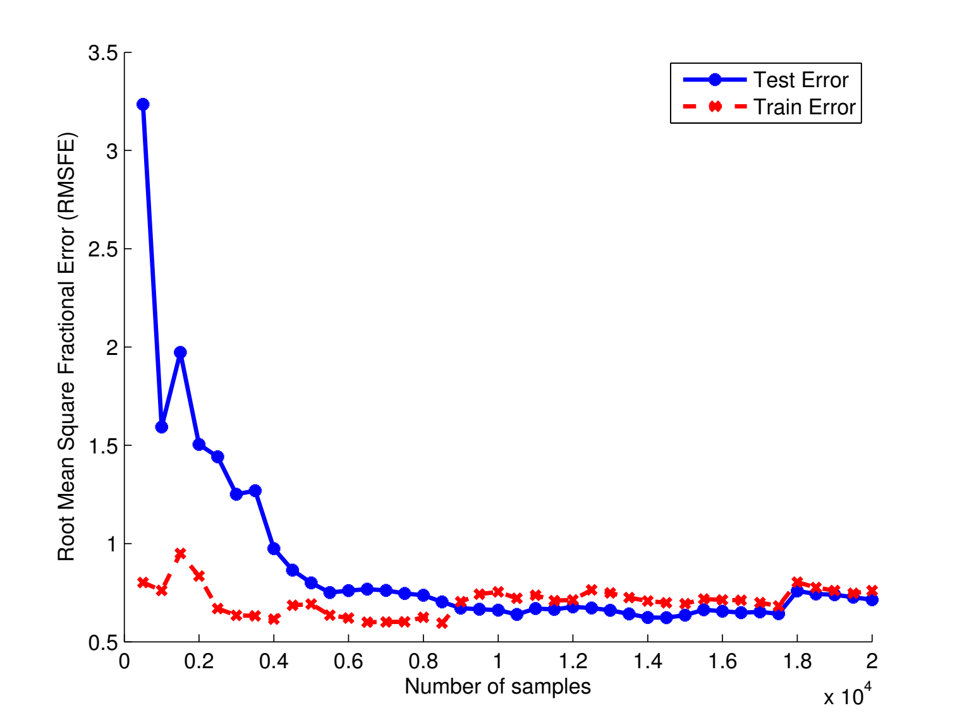


QQ plot of residuals

- The fat tails are largely contributed by non-uniform volatility of prices. While RMSE can hardly be reduced by learning, RMSFE, which normalizes by volatility, can be improved significantly.



Learning curve for RMSE



Learning curve for RMSFE

## K-means Clustering

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

Clustering based on bid price confirms heteroskedasticity

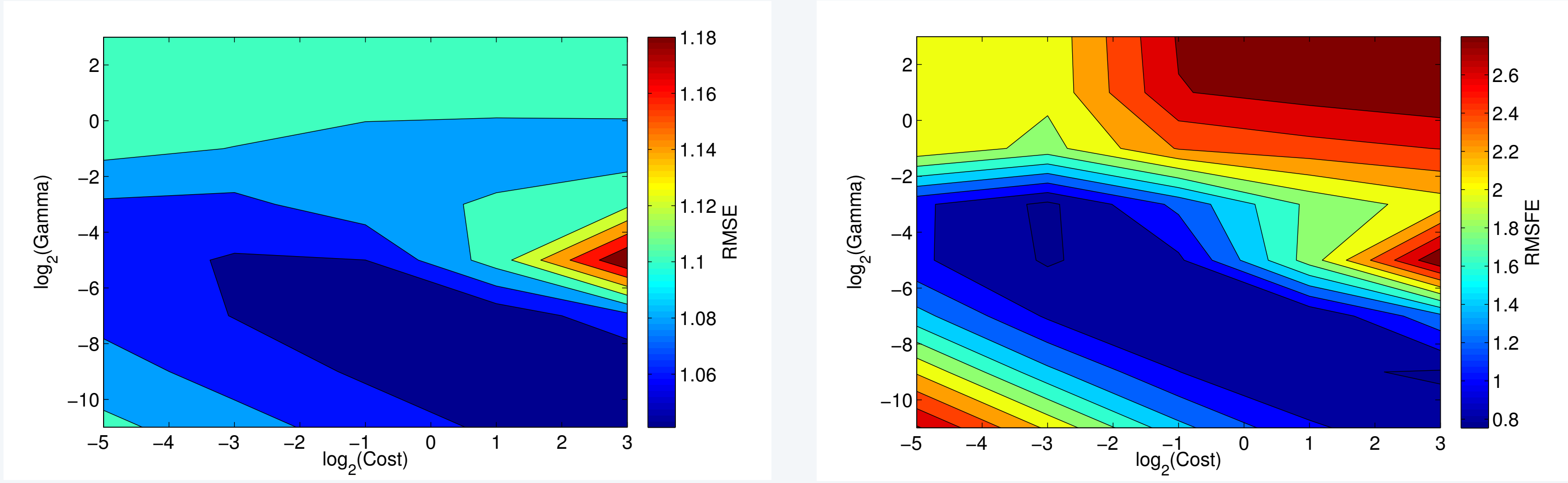
## Weighted LR : Heteroskedasticity fix

The volatility of each example can be estimated empirically from the features,

$$S_{ti} = \sum_j \theta_j z_{j_i} + \epsilon_i, \quad i = 1, \ldots, M, \; t > t_0$$
$$\epsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$$
$$\sigma_i^2 \simeq \hat{\sigma}_i^2 := \frac{1}{t_0 - 1} \sum_{\tau=1}^{t_0} (S_{\tau i} - \bar{S})^2$$

## Support Vector Regression

The lack of fit in LR is not due to large bias, but rather on account of the Markovian nature of the OU process. This can be verified by querying for optimal parameters in $\nu$-SVR,



$\nu$-SVR : RMSE/RMSFE($C$,$\gamma$). 5K Training Set, cross-validated against 1K Dev Set

which shows that smallest generalization error is obtained for very small values of $\gamma$ in the RBF kernel.

## Performance

| | 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 |