8-2017

The Efficiency of Liquidity Resiliency

Nathan Burton

THE EFFICIENCY OF LIQUIDITY RESILIENCY

by

Nathan R Burton

A report submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Economics

Approved:

Tyler Brough
Major Professor

Jared DeLisle
Committee Member

Christopher Fawson
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah

2017

Abstract

Efficiency of Liquidity Resiliency

by

Nathan Burton, Master of Science

Utah State University, 2017

Major Professor: Tyler Brough

Department: Economics and Finance

Using a VECM to estimate the dynamics of liquidity, in this case bid-ask spread, I run simulations for stocks of varying market capitalizations and find that lower market cap stocks require more orders to return to equilibrium spread following a shock, suggesting less efficiency of price discovery in lower cap stocks. Despite the greater number of order necessary for lower cap stocks, the return to equilibrium spread is still very fast, suggesting a relatively efficient market for NYSE and NASDAQ stocks in the upper three market cap quartiles.

Contents

List of Tables

List of Figures

Introduction

Liquidity is an essential component of asset pricing and market behavior (Amihud & Mendelson 1986). Some asset pricing methods require an appropriate estimation of liquidity resilience, i.e. the return to equilibrium of liquidity in the case of a perturbations from 'normal'. This measure of liquidity resilience is indeed difficult to define because of the multi-variable dynamics involved.

Lo & Hall (2015) provides one of the more complex treatments of short-term liquidity resilience by using a VECM on Australian Securities Exchange (ASX) data. I analyze US data from the NYSE and NASDAQ using a VECM (cointegrated VAR) similar to L&H to compare liquidity resiliency for differing market cap levels. Using event-driven analysis (as opposed to time-driven) I find that lower cap stocks require more order events before returning to liquidity equilibrium than higher cap stocks. This suggests less efficiency in the price discovery process for low cap stocks, not only in terms of time required but trades/order required.

Liquidity Resiliency: What, How, Why...

Liquidity, as a measure of the ease of selling an asset, is an important component of an investor's view of the worth of an asset. Periods of higher than normal *illiquidity* can have a drastic effect on the transaction cost, uncertainty, etc. and thus influential on the price investors are willing to buy (or sell) at. These periods of high illiquidity are the topic of interest in this study. It is highly observable that illiquidity bubbles return to a more typical equilibrium type state after a period of time, but the speed (or even the change in speed) of this resilience is an important consideration in the pricing of the asset.

L&H introduce a method where an impulse response VECM is estimated, then the estimated values are used to simulate the response to a 'shock' to equilibrium. L&H investigate a number of scenarios for the theoretical cause of the shock using data from the ASX,

VAR, VECM, and Beyond

VAR

In brief review, the autoregressive (AR) model is a typical tool in time-series analysis in which any given value in the series is dependent on one or more of the preceding values. A generalized example, given a variable y, the AR(1) process could be represented in the form $y_t = \beta y_{t-1} + \varepsilon_t$ where y_{t-1} is the observed value of y at the time $t-1$, β is the coefficient, and ε_t is an iid error term at time t.

A similar VAR of two variables may take the form

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \beta_1 & \beta_2 \\ \delta_1 & \delta_2 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

or more concisely

$$\mathbf{y}_t = \boldsymbol{\beta} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$$

where $\mathbf{y}_t$ is the vector of all y variables at time t, $\boldsymbol{\beta}$ is the matrix of coefficients for the vector $\mathbf{y}_{t-1}$ containing the singe-period lagged variables, and $\boldsymbol{\varepsilon}_t$ is the vector of concurrent error terms.

More generally, we may specify a VAR of p number of lags

$$\mathbf{y}_t = \sum_{n=1}^{p} \boldsymbol{\beta}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t$$

or

$$\mathbf{y}_t = \mathbf{B} \cdot \mathbf{Y}_L + \boldsymbol{\varepsilon}_t$$

where $\mathbf{B}$ is a vector of coefficient matrices $\boldsymbol{\beta}$ and $\mathbf{Y}_L$ is a vector of corresponding lag vectors $\mathbf{y}$.

VECM

The simple AR process $y = \beta y_{t-1} + \varepsilon_t$ may not be stationary, but if the process is integrated of order 1 i.e. $I(1)$, then we would find that $\Delta y_t = \beta \Delta y_{t-1} + \varepsilon_t$ is stationary. In the case that multiple processes are cointegrated we can account for this relationship by using an Error Correction Model (ECM). An ECM (in terms of the Engel-Granger two-step method) uses the residuals of the variable values differenced on one another which residuals are included in a differenced model to account for the cointegration. Defining our referenced residuals as $u_{t-1} = y_{t-1} - \zeta x_{t-1} - \alpha$, our simple example becomes

$$\Delta y_t = \delta \Delta x_{t-1} + \gamma u + \varepsilon_t$$

or

$$\Delta y_t = \delta \Delta x_{t-1} + \gamma (y_{t-1} - \zeta x_{t-1} - \alpha) + \varepsilon_t$$

We can then finally fulfill the next step of defining a cointegrated VAR, i.e. Vector Error Correction Model (VECM) in the same process with which went from an AR model to a VAR:

$$\Delta \mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\alpha} \boldsymbol{\beta} \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Machine Learning for Feature Selection

In every study there is the troubling question of what should be in a model and how the variables should interact with each other. Traditionally, one simply chooses variables by theory and building upon previous studies. With increased computing power industry has often resorted to machine learning for selection of variables. Many academics are apprehensive of the machine learning approach which often conjures the (antiquated) buzzword *data mining.* Methodically regressing every combination of variables to simply find the combination that gives the lowest RSS on a single dataset is theoretically unsound, likely unhelpful (even in industry), and computationally costly to the point of impractical. Modern machine learning techniques are much more refined and can be far more useful in both academia and industry when employed properly.

Regression Coefficient Penalties

Both when employing regressions of a polynomial type ($y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$) or simply an unclear set of variables ($y = \beta_0 + \beta_1 x_1 + \beta_2 x_x + \cdots + \beta_k x_k$) the problem of overspecification (especially spurious correlation) is of constant concern in the attempt to determine causality and real-world associations. Overspecification has a tendency to increase the absolute value of regression coefficients and so an early attempt to bridle overspecification was to run a version of the regression for which the problem changed from $minRSS$ to $min\ RSS + \lambda\ell_2$ where ℓ_2 is the L_2 distance/$\ell_2 norm$, and λ is an arbitrary coefficient chosen by the statistician (or, by proxy, the computer) which reduces the coefficients in the optimization problem. A problem in research with this approach is that the absolute value of regression coefficients are simply reduced for predictive purposes, not rendering any help in exactly which variables should or should not be included.

Lasso

Tibshirani (1996) presented a method referred to as *lasso* in which ℓ_1 is used instead, providing the optimization problem $min\ RSS + \lambda\ell_1$. A subtle and not immediately apparent advantage to the lasso method is that as λ is increased, the coefficients decrease linearly resulting in some being driven to exactly $\beta_k = 0$ while the other, more influential variables remain at $|\beta_k| > 0$. Lasso provides a method to actually decide which variables meet some demand of specification, given a certain λ. Important to note, is that in order to give each variable an equal penalty weight, the data should be normalized before utilizing lasso.

Initially the concern of spurious correlation is still apparent using lasso and there is the problem of selecting an appropriate λ. This is where the 'learning' of *machine learning* comes in. Given a dataset, the data is randomly split into training, validation, and test sets. A variety of λ values can be selected to provide a series of model estimations on the training data (presumably with some coefficients optimizing to 0 using lasso). Each model is then run on the validation set, and the model with the lowest sum of residuals on the validation set is presumed most appropriate. In the case of smaller datasets, training and validation sets can be recombined and new test and validation sets formed to run the same process of λ selection. After making a final decision on the value of λ the model can be run on the test set to assess the finalized model.

Still problematic is that in lasso the non-zero coefficients have still been reduced and possibly resulting in bias, in addition to 'standard errors' having essentially no meaning in the context of a lasso penalty regression. To circumvent this, academic researchers can simply treat the lasso method as a tool for variable selection, after which selection the appropriate non-zero variables can then be used in a traditional style (no λ) regression. This method of specification is that which I use for this study.

Simulating the Impulse Response

Once the VECM has been estimated, simulation of the data is very straightforward where Δy_t for each variable is calculated for each iteration using the VECM estimates. Using $y_{t+1} = y_t + \Delta y_t$ gives a final $y_T = y_1 + \sum_{t=0}^{T-1} \Delta y_t$. L&H employ this method as described in Hautsch and Huang (2012). Once initialized with equilibrium state values, a shock observation is added to the system, and then the simulation run from this equilibrium-plus-shock initialization. The simulation can then provide how many iterations (i.e. order events) are necessary to return to a predetermined level of recovery. L&H use a 90% recovery, but the simulations I employ result in such a fast recovery that I am able to use 99%.

Data

The main data used in my study is TAQ tick-frequency data for 79 NYSE and NASDAQ securities. Note that some of the securities are funds, ect., but owing to their being traded on the NYSE or NASDAQ I refer to all securities in this study as 'stocks'. I order all NYSE and NASDAQ stocks by market cap, and categorized them into quartiles. I then randomly selected 20 stocks from each quartile (one stock in quartile 2 was unusable due to lack of data, leaving only 19 for the 2nd quartile). The data is for the one month period from July 1, 2013 to July 31, 2013. Observations are limited to the trading day. See Appendix A for the full list of individual stocks and statistical summaries for each. The original data includes date, time, bid, offer, bid size, and offer size, from which I calculate bid-ask spread as my liquidity measure. Figures *** provide summary statistics for each quartile of stocks.

Initial Characteristic Conclusions

Running Lasso technique for feature selection with a varying λ strongly suggests that spread is affected heavily by the other four variables available for this analysis. Analysis on NASDAQ-only ITCH data, for which numerous other variables of interest are available on a high-frequency level, could more fully use Lasso to select important variables.

A method of my own design which I will refer to as *regression series analysis* (RSA) suggests that the stocks do not exhibit what might be termed structural multi-breaks, i.e. there are not extended periods of time (say a couple hours) which can consistently be reasonably distinguished from one another via the behavior during that period. RSA splits the data into numerous subsets of a few thousand tick observations (which generally equates to a few minutes for high cap stocks), runs a VECM for each subset, then groups the data subsets according to the statistics estimated by the VECM. Running RSA on the stocks at hand essentially produces a single cluster with over 99% of the subsets while all other clusters (up to 39) contain a very small assortment of extreme outlier subsets.

L&H run a VECM for each week of their period of interest. I conclude via RSA that it is unnecessary to treat the data as having weekly structural breaks. Additionally, the assumption that statistics and parameters from one week do not apply to another week really makes the entire analysis rather pointless. I analyze the data for each stock over the entirety of the period of interest.

Model

The VECM model I use is

$$\Delta \mathbf{y}_t = \boldsymbol{\alpha}_t \boldsymbol{\beta}_t \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_{i,t} \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t$$

where $\mathbf{y}$ is the vector of both endogenous and exogenous variables chosen in the Lasso step (it does not matter whether exogenous variables are simply included in the $\mathbf{y}$ vector) and $\boldsymbol{\varepsilon}$ is the vector of associated errors for each variable in $\mathbf{y}$.

Impulse Response Results

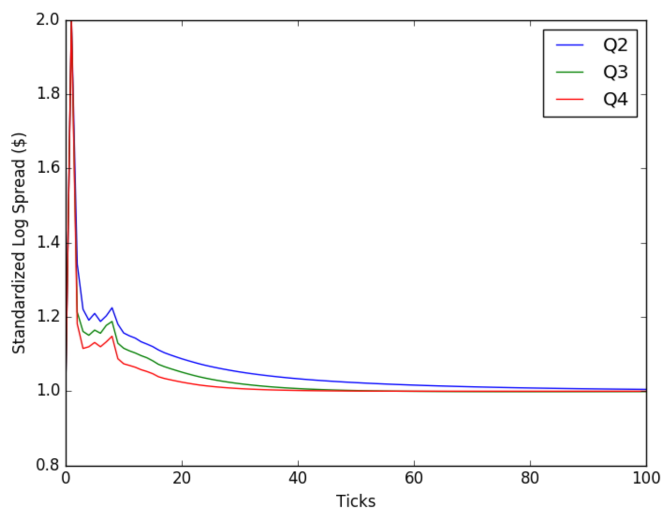

Figure 1: Impulse Response for Decreased Bid

I run a VECM for each stock individually over the period described. In order to compare the resiliency characteristics for each quartile I average each coefficient in the VECM for each quartile. Using a randomly selected interval as an initial set of observations, I then simulate forward observations to an equilibrium via the averaged coefficients for the corresponding quartile. Using the equilibrium values as a new set of initial observations, I add a ‘shock’ in which the spread is doubled, and then simulate forward observations in a return to equilibrium. I perform this post-shock simulation for three circumstances of shock creation: a decrease in bid, and increase in offer, and an equal decrease in bid and an increase in offer. All circumstances are characterized by a shock of doubled spread.

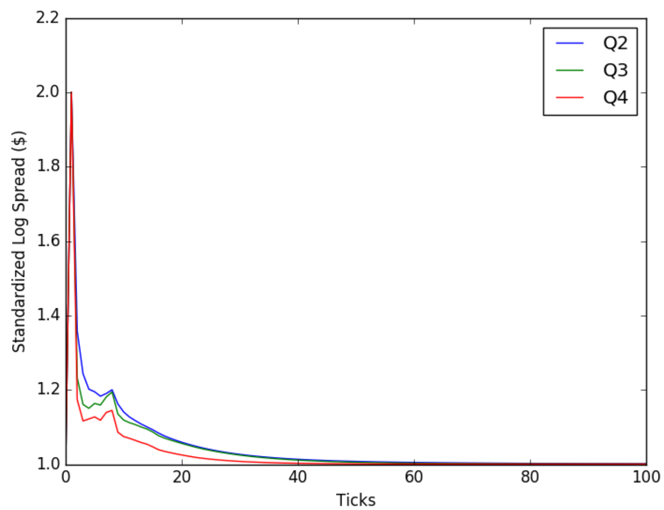

Figure 2: Impulse Response for Decreased Bid and Increased Offer

The dynamics of the individual first quartile stocks are so diverse and inconsistent that the method described above using the averaged first quartile coefficients results in the simulated spread blowing up, although the method results in a converging equilibrium if using the coefficients for an individual first quartile stock. I will not attempt to draw conclusion on these lowest cap stocks.

The three upper quartiles are much more consistent in nature and produce respective averaged coefficients which provide us with useful information on the dynamics of stocks at various market cap levels. In order to visualize the resiliency, I have standardized by dividing the log spread by its equilibrium value after completing the simulation, i.e. each log spread begins at 1.0, is shocked to 2.0, and eventually returns to 1.0. Graphs of the impulse response are shown in Figures 1, 2, and 3. Table 1 lists the number of events before the log spread has made a 99% recovery. The results of the

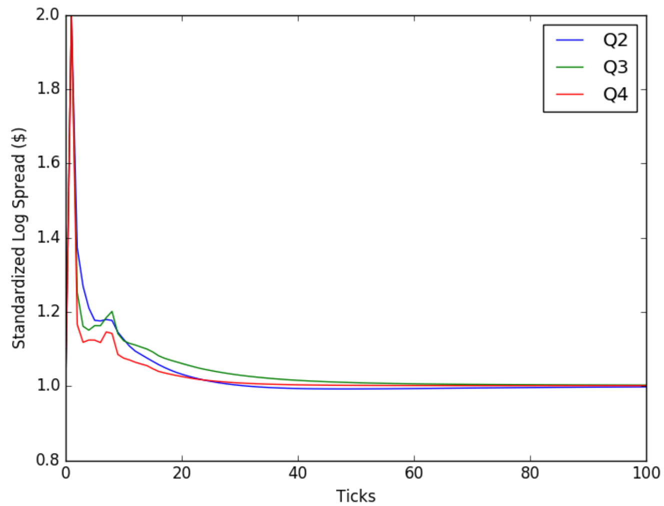

Figure 3: Impulse Response for Increased Offer

simulations are that for the number of order events needed for a return to equilibrium Quartile 2 > Quartile 3 > Quartile 4. Although in the case of increased offer the Quartile 2 simulation reaches the equilibrium value more quickly, it should be noted that it has not actually returned to equilibrium as can ben seen in its crossing the equilibrium value before converging, the intuition of which still places Quartile 2 at a slower return to equilibrium than Quartile 3.

Table 1: Number of Events to 99% Recovery

Shock Type	Quartile 4	Quartile 3	Quartile 2
Decr Bid	18	28	65
Incr Offer	20	39	48[1]
Both	19	33	36

The implications of these results is that NASDAQ and NYSE stocks return to equilibrium with relatively few events - especially at the event frequency of high cap stocks, the recovery time is likely less than a second. Also of great importance is that lower cap stocks are less efficient in reaching equilibrium i.e. a component of price discovery in terms of number of orders, not just in terms of time.

not provide evidence for structural breaks via clustering intervals. In comparison, clustering the observations individually produces a reasonable separation, for example MSFT very consistently separates into five clusters of 39%, 16%, 35%, .001%, and 9%. The RSA and individual clustering outcomes, in tandem, suggest that switching, if it occurs, happens on a frequency less than every few minutes. This leads me to conclude that a probabilistic Markov switching model is far more appropriate than a simple multiplicity of structural breaks.

A problem in the analysis of high frequency data is the computationally heavy load. Both this and the L&H publication analyze a regrettably small number of stocks for the sake of plausibility. GPU computing has been gaining momentum, and could provide an important way to perform the many calculations necessary for a large number of stocks much more quickly by distributing the load among hundreds or thousands of cores.

Conclusion

A major conclusion of this study include that on average higher cap stocks exhibit a more efficient liquidity resiliency than lower cap stocks. We can see that the efficiency of price discovery is thus dependent on the market cap of the specific stock, with microeconomic implications that the absolute size of a market affects the overall efficiency of that market. Additionally, notable is that though lower cap Quartile 2 stocks have a less efficient price discovery, it is not exceptionally less efficient, requiring two to four times the number of order events as higher cap Quartile 4 stocks. This still suggests a relatively efficient price discovery for Quartile 2 stocks.

Appendix A: Summary of Stocks

Tables 2, 3, and 4 contain summaries of the securities used in this study, including the security name, market cap, and the following averages for the period of the study (July 1 to July 31, 2013): average bid size, average offer size, average log(bid), average log(offer), and average log spread defined as $log(offer) - log(bid)$.

Table 2: Quartile 2 Securities

Stock Symbol	Stock Name	Market Cap	Avg. Bid Size	Avg. log(bid)	Avg. log(offer)	Avg. log Spread [log(offer)-log(bid)]	Avg. Offer Size
JRO	Nuveen FRIO	451M	7	2.53044594499	2.55825757273	0.0278116277376	4
SLMAP	SLM Corp	165M	3	3.05657424991	4.71490657485	1.65833232494	2
HTF	Horizon Technology	298M	2	3.12411970932	3.32388752606	0.199767816732	2
HPS	John Hancock Preferred Income III	602M	3	2.8777865324	2.90024892405	0.0224623916512	3
MCRI	Monarch Casino and Resort	518M	2	2.88156521865	2.91840229557	0.0368370769276	3
EXA	Exa Corp.	201M	1	2.01414559995	2.78674351886	0.772597918913	1
IBCP	Independent Bank Corp.	448M	2	1.99623069131	2.06790943199	0.071678740682	2
PEIX	Pacific Ethanol	260M	6	1.39798030905	1.42841831581	0.0304380067551	5
UIHC	United Insurance Holdings	677M	3	1.94744070365	2.02519190967	0.0777512060206	2
GAIN	Gladstone Investment	310M	4	1.98825182398	2.01463431062	0.0263824866345	7
SMM	Salient Midstream and MLP Fund	211M	2	3.12771452168	3.21319125374	0.0854767320544	2
BGY	Blackrock Enhanced International	711M	5	2.01890424428	2.03087160344	0.0119673591636	10
XBKS	Xenith Bank Shares	635M	4	1.5832191861	1.86153829743	0.278319111338	2
NSSC	Napco Security	194M	4	1.49923714968	1.69393867707	0.1947015274	5
HDSN	Hudson Technologies	333M	10	0.881300484963	0.923168068865	0.0418675839027	5
NNA	Navios Maritime Acquisition	218M	4	1.28502550822	1.31373312589	0.0287076176765	4
I	Intelsat	358M	1	3.05734759762	3.07677853763	0.0194309400042	2
NTZ	Natuzzi	152M	6	0.702067703614	0.822032359882	0.119964656268	2
USLV	Credit Suisse Velocity 3x Long Silver	302M	41	1.7318764731	1.73802635908	0.00614988598636	42

Table 3: Quartile 3 Securities

Stock Symbol	Stock Name	Market Cap	Avg. Bid Size	Avg. log(bid)	Avg. log(offer)	Avg. log Spread [log(offer)-log(bid)]	Avg. Offer Size
TYG	Tortoise Energy	1.5B	2	3.80015616339	3.86112969012	0.0609735267326	2
IMKTA	Ingles Markets	906M	2	3.23042202405	3.38244240831	0.152020384255	2
GAM	General American Investors	1.15B	2	3.45532590509	3.50972831167	0.0544024065824	2
ASTE	Astec	1.27B	2	3.50485837277	3.66815382841	0.163295455638	2
TPH	TRI Pointe	2.23B	2	2.77028109403	2.79520394468	0.0249228506596	2
SHLM	A Schulman	779M	1	3.3197251587	3.39116761151	0.0714424528023	1
NHI	National Health	3.09B	1	4.13573896175	4.14607438688	0.010335425131	1
TAC	TransAlta	1.89B	4	2.63690558614	2.64079729985	0.00389171372337	4
HLX	Helix Energy	830M	2	3.21238720779	3.21770897085	0.00532176306261	2
SANM	Sanmina	3.20B	3	2.73651211443	2.74391052604	0.00739841161014	4
FOLD	Amicus Therapeutics	2.20B	5	0.8258995565	0.85880811999	0.0329085634899	6
WWE	World Wrestling Entertainment	1.66B	2	2.37093768598	2.38450456088	0.0135668748935	2
MDRX	Allscripts Heathcare Solutions	2.23B	9	2.71217754246	2.71455746925	0.0023799267932	10
UHT	Universal Health Realty Income Trust	1.13B	1	3.7902426762	3.81529533151	0.0250526553065	1
MATX	Matson	1.28B	2	3.2977025955	3.32159006031	0.0238874648106	2
HTH	Hilltop Holdings	2.52B	2	2.83643090926	2.84473526219	0.00830435292933	3
CEM	ClearBridge Energy MLP	1.09B	3	3.31952001666	3.3404634733	0.0209434566323	3
MTDR	Matador Resources	2.32B	2	2.49626647693	2.50980545535	0.0135389784214	2
CDE	Coeur Mining	1.57B	4	2.55300629512	2.55588563309	0.00287933796785	4
RP	RealPage	3.21B	2	2.97117842385	2.99833939311	0.0271609692612	2

Table 4: Quartile 4 Securities

Stock Symbol	Stock Name	Market Cap	Avg. Bid Size	Avg. log(bid)	Avg. log(offer)	Avg. log Spread [log(offer)-log(bid)]	Avg. Offer Size
LVS	Las Vegas Sands	49.84B	4	3.99139619572	3.99232697985	0.000930784133815	4
MCHP	Microchip Technology	18.26B	4	3.66103430149	3.66279787305	0.00176357155732	4
TLLP	Tesoro Logistics LP	5.62B	1	4.03186100229	4.04911106948	0.0172500671921	2
CTXS	Citrix Systems	12.22B	2	4.19136339175	4.19402326761	0.00265987586076	2
OAK	Oaktree Capital	7.36B	2	3.96638031993	3.97973397643	0.0133536564988	2
JPM	JpMorgan Chase and Co.	323.20B	11	4.00835378145	4.00876138972	0.000407608270764	11
CFX	Colfax	5.16B	2	3.95157121822	3.95980721332	0.0082359950969	2
BAK	Braskem SA	9.06B	5	2.69932520902	2.70561301226	0.00628780324573	5
AMZN	Amazon.com	495.13B	2	5.69519071705	5.69734836666	0.00215764960684	2
TCBI	Texas Capital Bancshares	3.96B	2	3.85189380558	3.86216403571	0.0102702301347	2
O	Realty Income	15.61B	3	3.78254345317	3.78524341267	0.00269995950595	3
PFE	Pfizer	198.07B	56	3.35808598219	3.35865709334	0.000571111155052	60
ENB	Enbridge USA	68.22B	4	3.77942557357	3.78083272555	0.0014071519793	4
ESRX	Express Scripts Holding Co	37.10B	3	4.17358786716	4.17485795099	0.0012700838287	4
SCHW	Charles Schwab	56.12B	29	3.07951552946	3.08032655442	0.00081102496062	27
LPL	LG Display Co Ltd. (ADR)	10.80B	10	2.49891049158	2.50334278576	0.00443229418073	10
WLK	Westlake Chemical Corp.	9.06B	2	4.60046350651	4.61426860991	0.0138051033909	2
AVT	Avnet	4.83B	2	3.58305673223	3.58681044279	0.00375371055452	2
WRB	W. R. Berkley Corp	8.69B	2	3.7534337467	3.7595938775	0.00616013079633	2
VTR	Ventas	24.24B	2	4.23975873951	4.24202427471	0.00226553520717	2

References

Amihud, Yakov and Haim Mendelson. "Asset pricing and the bid-ask spread." *Journal of Financial Economics* 17, no. 2, (1986): 223-249.

Lo, Danny and Anthony Hall. "Resiliency of the limit order book." *Journal of Economic Dynamics and Control* 61, (2015): 222–244.

Hautsch, Nikolaus and Ruihong Huang. "The market impact of a limit order." *Journal of Economic Dynamics and Control* 36, no. 4, (2012): 501-522.

Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58, no. 1 (1996): 267-288.