Financial Market Frictions and Learning from the Stock Price

Let’s “Pay attention to the plumbing” (Levitt 2003)

Dominik Rösch

University at Buffalo

drosch@buffalo.edu

February 13, 2020

This work is supported in part by NSF ACI-1541215.
Mandelbrot Set

\[ z_{n+1} = z_n^2 + c, \quad \text{with } z_0 = 0 \]

"Can there be a more striking demonstration of the enormous complexity hidden in the simplest laws?" (Peitgen, Richter 1986)
Mandelbrot Set

\[ c \in \mathbb{C} = (\mathbb{R}, \mathbb{R}) \text{ converges to infinity in:} \]

\[ z_{n+1} = z_n^2 + c, \quad \text{with } z_0 = 0 \]

"Can there be a more striking demonstration of the enormous complexity hidden in the simplest laws?" (Peitgen, Richter 1986)
Mandelbrot Set

coloured by how fast $c \in \mathbb{C} = (\mathbb{R}, \mathbb{R})$ converges to infinity in:

$$z_{n+1} = z_n^2 + c, \text{ with } z_0 = 0$$
Mandelbrot Set

coloured by how fast \( c \in \mathbb{C} = (\mathbb{R}, \mathbb{R}) \) converges to infinity in:

\[
z_{n+1} = z_n^2 + c, \text{ with } z_0 = 0
\]

“Can there be a more striking demonstration of the enormous complexity hidden in the simplest laws?” (Peitgen, Richter 1986)
Data: Stock Price

What does it mean to learn from the stock price?
Data: Stock Price

- What does it mean to learn from the stock price?
- Berkshire trades at both over USD 300,000 for its A-share and just a bit over USD 200 for its B-share
What does it mean to learn from the stock price?

Berkshire trades at both over USD 300,000 for its A-share and just a bit over USD 200 for its B-share

Companies often have different share classes, such companies are often dropped from empirical analysis.

Prices are (almost) meaningless as they are under the control of the company.
Data: Stock Price

- What does it mean to learn from the stock price?
- Berkshire trades at both over USD 300,000 for its A-share and just a bit over USD 200 for its B-share
  - Companies often have different share classes, such companies are often dropped from empirical analysis.
- Prices are (almost) meaningless as they are under the control of the company
- Prices are often more understood as ratios, e.g., price to earnings, price to book, or just returns.
Data: Returns

- Imagine the following investment, each day you flip a coin on head you get 11% and on tail you lose 10%
- Is this a good investment?
Data: Returns

- Imagine the following investment, each day you flip a coin on head you get 11% and on tail you lose 10%
- Is this a good investment?
- The expected return is 0.5% ... annualized more than 100%
- Starting with USD 1, after less than 30 years you are a billionaire!
Imagine the following investment, each day you flip a coin on head you get 11% and on tail you lose 10%

Is this a good investment?

The expected return is 0.5% ... annualized more than 100%

Starting with USD 1, after less than 30 years you are a billionaire!

Or are you? Actually, in the long-term you lose, the expected continuously compounded return is negative!

Empirically, should you use simple or log-returns? What about dividends? (convention: assumed re-invested, except in special papers)
In our case Stock Price: Tobins Q (Corporate Finance)

- Tobins Q is the “market value of equity (price times shares outstanding from CRSP) plus book value of assets minus the book value of equity, scaled by book assets.”

\[ Q = \frac{mkt\text{cap} + ATQ - bvse}{ATQ}, \text{ where } ATQ \text{ is quarterly total assets from Compustat, and } bvse \text{ is stockholders' equity from Compustat} \]

- Actually, \( Q = \frac{mkt\text{cap}/1000 + atq - bvse}{atq} \) with \( mkt\text{cap} = |Price| \times shrout \) because of scaling and in CRSP Price can be negative.
In our case Stock Price: Tobins Q (Corporate Finance)

- Tobins Q is the “market value of equity (price times shares outstanding from CRSP) plus book value of assets minus the book value of equity, scaled by book assets.”

- “\( Q = \frac{\text{mktcap} + \text{ATQ} - \text{bvse}}{\text{ATQ}} \), where ATQ is quarterly total assets from Compustat, and bvse is stockholders’ equity from Compustat”
In our case Stock Price: Tobins Q (Corporate Finance)

- Tobins Q is the “market value of equity (price times shares outstanding from CRSP) plus book value of assets minus the book value of equity, scaled by book assets.”

- “Q = (mktcap + ATQ − bvse)/ATQ, where ATQ is quarterly total assets from Compustat, and bvse is stockholders’ equity from Compustat”

- Actually, Q = (mktcap/1000 + atq − bvse)/atq with mktcap = |Price| × shrout
In our case Stock Price: Tobins Q (Corporate Finance)

- Tobins Q is the “market value of equity (price times shares outstanding from CRSP) plus book value of assets minus the book value of equity, scaled by book assets.”

- “\( Q = \frac{mktcap + ATQ - bvse}{ATQ} \), where ATQ is quarterly total assets from Compustat, and bvse is stockholders’ equity from Compustat”

- Actually, \( Q = \frac{mktcap/1000 + atq - bvse}{atq} \) with \( mktcap = |Price| \times shrout \)

- because of scaling and in CRSP \( Price \) can be negative
Financial Market Frictions (Market Microstructure)

- Illiquidity
  - Quoted Spreads
  - Effective Spreads
- Inefficiency
  - OIB Predictability
  - Variance ratio
  - Hasbrouck
  - Put-call parity
- Not necessarily, Inefficiency $\neq$ Illiquidity.
“Financial Market Frictions and Learning from the Stock Price”

▶ “Learning”: Investment-Sensitivity to $Q$ (ISQ), $Q$-Theory

Investment is measured, e.g., as CAPX which is reported as a running sum within each fiscal year.

Do frictions increase or decrease ISQ ($\beta_1$)?

We find: $\beta_1 > 0$, ISQ increases with frictions, ISQ is lower for more liquid stocks (because of noise trading).

“[w]hat’s needed for a liquid market causes prices to be less efficient” (Black, 1986, p. 532).

Normally, now argue that this relation is causal.
“Learning” : Investment-Sensitivity to $Q$ (ISQ), $Q$-Theory

\[ \text{Inv}_{i,q} = \beta_{0,q} + \beta_{1,q} \text{Fric}_{i,q-1} \times Q_{i,q-1} + \beta_{2,q} \text{Fric}_{i,q-1} + \beta_{3,q} Q_{i,q-1} + \sum \beta_q \text{Contr}_{i,q-1} + \epsilon_{i,q} \]
“Learning”: Investment-Sensitivity to $Q$ (ISQ), $Q$-Theory

$\text{Inv}_{i,q} = \beta_{0,q} + \beta_{1,q} Fric_{i,q-1} \times Q_{i,q-1} + \beta_{2,q} Fric_{i,q-1} + \beta_{3,q} Q_{i,q-1} + \sum \beta_q \text{Contr}_{i,q-1} + \epsilon_{i,q}$

Investment is measured, e.g., as CAPX which is reported as a running sum within each fiscal year.
“Learning”: Investment-Sensitivity to Q (ISQ), Q-Theory

\[ \text{Inv}_{i,q} = \beta_{0,q} + \beta_{1,q}Fric_{i,q-1} \times Q_{i,q-1} + \beta_{2,q}Fric_{i,q-1} + \beta_{3,q}Q_{i,q-1} + \Sigma \beta_q \text{Contr}_{i,q-1} + \epsilon_{i,q} \]

- Investment is measured, e.g., as CAPX which is reported as a running sum within each fiscal year.
- Do frictions increase or decrease ISQ (\(\beta_{1,q}\))
“Financial Market Frictions and Learning from the Stock Price”

- “Learning”: Investment-Sensitivity to $Q$ (ISQ), $Q$-Theory

- $\text{Inv}_{i,q} = \beta_{0,q} + \beta_{1,q} \text{Fric}_{i,q-1} \times Q_{i,q-1} + \beta_{2,q} \text{Fric}_{i,q-1} + \beta_{3,q} Q_{i,q-1} + \sum \beta_q \text{Contr}_{i,q-1} + \epsilon_{i,q}$

- Investment is measured, e.g., as CAPX which is reported as a running sum within each fiscal year.

- Do frictions increase or decrease ISQ ($\beta_{1,q}$)

- We find: $\beta_{1,q} > 0$, ISQ increases with frictions, ISQ is lower for more liquid stocks (because of noise trading)

- “[w]hat’s needed for a liquid market causes prices to be less efficient” (Black, 1986, p. 532).

- Normally, now argue that this relation is causal.
4 Bytes of Financial Data: 2922.07
4 Bytes of Financial Data: 2922.07

- 2kb: 2922.07 is an average of across around 500 prices
- 20kb: Each stock is priced on around 10 venues.
- 400kb: At each time each stock has around 20 prices [depth].
- 4mb/second: Prices often change 10 times per second
- 93GB/day: A trading day has 6.5 hours.
4 Bytes of Financial Data: 2922.07

- 2kb: 2922.07 is an average of across around 500 prices
- 20kb: Each stock is priced on around 10 venues.
- 400kb: At each time each stock has around 20 prices [depth].
- 4mb/second: Prices often change 10 times per second
- 93GB/day: A trading day has 6.5 hours.
- if you compress the data you need around 10GB/day.
US tick-by-tick data (TAQ)

- Data contains more than just prices, such as sizes, and other identifiers. Data does not contain depth, just the best available buy and sell prices per venue.
My first attempt: Mysql framework

- I started writing my stored procedure:
  - get_ohlc_for(ticker, between, bucket)

- when I coded:

  ```
  SUBSTRING_INDEX(
    MAX(CONCAT(time, '_', price)), '_', -1
  ) AS 'close'
  ```

- I realized MySQL might not be the best DB for financial data.
My second attempt: OneTick framework
Step 0: Setting up reference data

- Exchange trading times (e.g., NYSE: 09:00-16:30)
- Primary Exchange (e.g., IBM is NYSE)
- Symbology mappings (PERMNO to TICKER, e.g. Citigroup)
  - 70519——19861029000000—19890119000000—CCC—
  - 70519——19890120000000—19931231000000—PA—
  - 70519——19940103000000—19981007000000—TRV—
  - 70519——19981008000000—19981203000000—CCI—
  - 70519——19981204000000—20161230000000—C—
- Corporate actions (stock splits, dividends)
Step 0: Setting up reference data

- Exchange trading times (e.g., NYSE: 09:00-16:30)
- Primary Exchange (e.g., IBM is NYSE)
- Symbology mappings (PERMNO to TICKER, e.g. Citigroup)
  - 70519——19861029000000—19890119000000—CCC—
  - 70519——19890120000000—19931231000000—PA—
  - 70519——19940103000000—19981007000000—TRV—
  - 70519——19981008000000—19981203000000—CCI—
  - 70519——19981204000000—20161230000000—C—
- Corporate actions (stock splits, dividends)
- After REF_DATA and importing TAQ: Lets compute effective spreads...
Effective spreads: Data Processing

- \( ESPR_{i,t} = 2 \times \text{sign}_{i,t} \times (TRD_{i,t} - MID_{i,t-1}) \)
- Need TRD prices and midpoint prices prevailing TRD, \( MID_{i,t-1} \)
Effective spreads: Data Processing

- \( ESPR_{i,t} = 2 \times \text{sign}_{i,t} \times (TRD_{i,t} - MID_{i,t-1}) \)
- Need TRD prices and midpoint prices prevailing TRD, \( MID_{i,t-1} \)
- “For data before 1996, we delay quotes by 5 seconds following Lee and Ready (1991).”
- “We use monthly TAQ data from 1993 to 2014, with trades and quotes timestamped by the second, and daily TAQ data from 2014 onwards.”
- “We apply the Holden and Jacobsen (2014) interpolated time technique for all data from 1983 to 2014”
- \( \text{sign}_{i,t} \) use Lee and Ready (1991) using NBBO (?), or take absolute values (equivalent?)
Step 1.0: QTE need National Best Bid Offer (NBBO)

- Each exchange reports prices.
- Easy for TRD: “sum” them up.
- Difficult for QTE: at each point in time look at all ASK (BID) across exchanges and take best one, i.e., lowest (highest).
- This is called National Best Bid and Offer (NBBO)
- TAQ has a file with NBBO prices, but with errors.
- Before NBBO, do not ignore 0’s, this means orders are cancelled and current exchange doesn't have valid price.
Step 1.0: Compute NBBO
Step 1.0: Compute NBBO, “Limit order book”

<table>
<thead>
<tr>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>53.01</td>
<td></td>
</tr>
<tr>
<td>53.00</td>
<td>15</td>
</tr>
<tr>
<td>52.99</td>
<td></td>
</tr>
<tr>
<td>1 @ 52.98</td>
<td>13</td>
</tr>
<tr>
<td>52.97</td>
<td></td>
</tr>
<tr>
<td>52.96</td>
<td></td>
</tr>
<tr>
<td>52.95</td>
<td></td>
</tr>
<tr>
<td>52.94</td>
<td></td>
</tr>
<tr>
<td>52.93</td>
<td></td>
</tr>
</tbody>
</table>
Step 1.0: Compute NBBO, “Limit order book”

- construct a virtual order book based on Best Bid Ask quotes from different exchanges.
Step 1.1: MTAQ vs DTAQ: Holden and Jacobsen (2014) interpolated time

- MTAQ is timestamped at second accuracy, DTAQ at least milliseconds.
- Imagine one trade $t$ with many quotes in second $s$.
- You want quote prevailing $t$, but which one?
Step 1.1: Compute interpolated time
Step 1.2: Delay quotes

- Why? TRD and QTE are reported separately. TRD not as important, therefore reported with delay.
Step 1.2: Delay quotes

- Why? TRD and QTE are reported separately. TRD not as important, therefore reported with delay.

- Timestamps of both are when published by SIP (Securities Information Processor, see Bartlett and McCrory JFM, 2019)

- DTAQ provides several timestamps like “Participant Timestamp”, “Trade Reporting Facility (TRF) Timestamp” (see NYSE)
Step 2: Compute Effective spreads per trade
Step 3: Aggregate Effective spread per day

```
BEGIN 
    build_data_for_spreads.otq::effective_spread_nbbo
    QUOTE_SIZE_AVAILABLE = $QUOTE_SIZE_AVAILABLE
    QUOTE_SOURCE_FIELDS = $QUOTE_SOURCE_FIELDS
    THRESHOLD_BAD_TICK = $THRESHOLD_BAD_TICK
    CALENDAR_NAME = $CALENDAR_NAME
    MODE_IS_CHAR = $MODE_IS_CHAR.
    MODE_IS_INT = $MODE_IS_INT.
    QUOTE_DELAY = $QUOTE_DELAY
    EFFECTIVESPREAD

END

BEGIN 
    COMPUTE
    = AVERAGE
    (PRICE_FIELD_NAME=EFFECTIVESPREAD, SIZE_FIELD_NAME=SIZE, OUTPUT_FIELD_NAME=EffectiveSpread)
    BUCKET_INTERVAL=1
    BUCKET_INTERVAL_UNITS=DAYS
```
Step 3: Aggregate Effective spread per day

- How to measure effective spreads without trades?
- How to measure quoted spreads without quotes?
- Decision to trade/quote are endogeneous.
Step 4 Run on server (using GNU Parallel)

database = ${1}
query = ${2}

for date in dates.txt ; do
    sem -k --id exp -P ot.cpus
    export.pl ${database} ${date} ${query}
done

sem --wait --id exp
Step 5: Test! TDD, write a test suite.

```perl
lives_ok {
    $one_tick->import_csv();
}
'can import data into OneTick';

my $results_summary = {
    'total_ticks' => 196232,
};

Test::More::is_deeply($one_tick->results_summary, $results_summary,
    'Got summary results as expected');
```
Step 6: Write a paper

- RStudio (easiest at Vidia CCR or RStudio Cloud) within Rstudio:
  - Version Control
    - “File” → “New Project” → “Version Control Git”
    - Use “https” address, better set up “ssh”
  - Reproducible (R Sweave)
    - “New File” → “R Sweave”
  - Access to Data (ODBC)
    - “Connections” → “New Connection” (right)
Step 6.1: How to make sure code is consistent? **Version Control!**
Step 6.1: How to make sure code is consistent? git diff
Step 6.2: Pull results into favourite statistical software: ODBC

- Regardless which DB, use stored procedures.

```r
library("RODBC")
channel <- odbcConnect("ONETICK_DEFAULT_DSN")

test <- sqlQuery(channel, "SELECT *
    FROM OTQ_FILES.examples::get_pespr QTE
    WHERE (QTE.TIMESTAMP >= '2004-01-01 00:00:00 GMT'
    AND (QTE.TIMESTAMP < '2004-01-03 00:00:00 GMT'
    AND (param_assign('query_paras','')=1)
"
)
```
Step 6.3: How to make sure paper is reproducible? RSweave!

- Embed ODBC / R-code into Latex using RSweave.
Step 6.3: How to make sure paper is reproducible? RSweave!

- Embed ODBC / R-code into Latex using RSweave.

\begin{table}[H]
\caption{\textbf{Caption ...} This table reports}

<table>
<thead>
<tr>
<th>#Stocks</th>
<th>Mean</th>
</tr>
</thead>
</table>

\texttt{\texttt{Sexpr{pretty\_numbers(format\_integer, data\_obs['trades'])}}} & \texttt{\texttt{Sexpr{pretty\_numbers(format\_integer, data\_means['trades'])}}} \\
\end{table}
What about sizes?

- Nasdaq Dealer Market: Double counting.
What about sizes?

- Nasdaq Dealer Market: Double counting.
- How are trades reported, an order of 100,000 shares will be broken into several trades.
- But will a trade against several limit orders reported for each limit order it hits?
Resources

- TAQ:
  nyse.com/publicdocs/nyse/data/Daily_TAQ_Client_Spec_v3.0d.pdf

- ODBC: en.wikipedia.org/wiki/Open_Database_Connectivity

- github.com

- rstudio.com

- Good video:
  rstudio.com/resources/webinars/working-with-big-data-in-r/