ORF 474: High Frequency Trading
Notes 3b

Robert Almgren

Feb. 19, 2020

The question for the next few weeks is, why is there a bid-ask spread? (We neglect tick size, and assume that prices may have a continuum of values.) The reason this needs explanation is that if market makers can buy the asset at price b, and sell at price a, with spread $S = a - b$, then they make a profit of S on each round-trip trade. Of course the price moves randomly in between their trades, but this adds only randomness to their profits; they still consistently make money. Then you would expect competition between market makers to drive the spread $S \to 0$. But in real markets, $S > 0$.

We will invoke three separate mechanisms:

1. Fixed costs (Roll model),
2. Information flow (Glosten-Milgrom model), and
3. Inventory risk.

The models we present here are "toy models," and should not be taken too literally. But they are useful to get some insight into what the driving effects are.

1 Roll Model

Original model by Roll (1984).

Time is discrete and counted in trades, so $t = 1, 2, \ldots$. At time t, m_t is the midpoint price, satisfying the random walk

$$m_t = m_{t-1} + \epsilon_t.$$

with $\mathbb{E}(\epsilon_t) = 0$. We ignore the discrete price grid, so m_t and ϵ_t may take any real values.

We do not talk about why or how the price changes, just that it does. The interpretation is that ϵ_t represents information that is revealed between time $t-1$ and t, and is somehow incorporated into the price.

A "market maker" or "dealer" sets bid and ask prices

$$a_t = m_t + \frac{S}{2}$$
$$b_t = m_t - \frac{S}{2}.$$

The spread S is assumed to be set exogeneously and is fixed.

Outside particpants, who we call "noise traders," "liquidity traders," or just "traders," randomly send market orders, one order per time step. The trade sign d_t ($d_t = +1$ for a buy, $d_t = -1$ for a sell) is chosen randomly, with $+1$ and -1 each having probability $\frac{1}{2}$. So the trade price is

$$p_t = m_t + \frac{S}{2} d_t.$$

The external trader either buys from the dealer at the ask, so $p_t = a_t$ if $d_t = +1$, or sells to the dealer at the bid, so $p_t = b_t$ if $d_t = -1$. All trades are of unit size, and the quotes are large enough to absorb these trades without changing.

Assumptions:

- $\{\epsilon_t\}$ is serially uncorrelated: $\mathbb{E}(\epsilon_t \epsilon_s) = 0$ for $t \neq s$. That is, each ϵ_t represents new information, which could not be anticipated given information available to time $t-1$. In fact, we know that midpoint quote changes $\{\epsilon_t\}$ have substantial negative serial correlation, especially for large-tick assets. The correlation is less if you sample at times of trades rather than every quote change, but it is still not zero.

- $\{d_t\}$ is serially uncorrelated: $\mathbb{E}(d_t d_s) = 0$ for $t \neq s$. This is because the outside traders have no information and no longer-term trade goals. In fact, we know empirically that $\{d_t\}$ has high positive serial correlation, in part because outside traders often are executing large positions across many successive individual trades.

- $\{\epsilon_t\}$ and $\{d_t\}$ are uncorrelated: $\mathbb{E}(\epsilon_t d_s) = 0$ for all t, s. In fact we know that

  - d_t is positively correlated with *subsequent* price changes $\epsilon_{t+1}, \epsilon_{t+2}, \ldots$ for two reasons: market impact and alpha. Market impact means that submission of a buy order at time t will cause the price to rise in future and conversely for a sell. Alpha means that that the buy order was submitted because the trader correctly anticipated that the price would rise, for exogeneous reasons. These two effects are very hard to distinguish empirically.
  - d_t is correlated in various ways with *previous* price changes $\epsilon_t, \epsilon_{t-1}, \ldots$. The order was perhaps submitted as a response to what had been happening in the market immediately before.

Then the change in successive trade prices is

$$\begin{aligned} p_t - p_{t-1} &= m_t - m_{t-1} + \frac{S}{2}(d_t - d_{t-1}) \\ &= \epsilon_t + \frac{S}{2}(d_t - d_{t-1}). \end{aligned} \tag{1}$$

To talk about probabilities, let us denote

$$\Omega_t \;=\; \{\, \text{information available to the market maker just } \textit{after} \text{ trade } t \,\}.$$

(This is a bit heavy for the Roll model, but will be useful in the models below.) Then

$$\{\, \text{Information available to the market maker just } \textit{before} \text{ trade } t \,\} \;=\; \Omega_{t-1} \cup \{\epsilon_t\}$$

(we do not give a name to this information set) and

$$\Omega_t \;=\; \Omega_{t-1} \cup \{\epsilon_t, d_t\}.$$

This is not quite the right notation for information flow, which should be represented by a nested family of filtrations, but it captures the essence: between times $t-1$ and t the information ϵ_t is revealed, and then in the trade at time t the trade sign d_t is revealed. We specify that this is the information available to the market maker, because we have no idea when the other participants receive information or make their trade decisions.

Then we can say that

$$\mathbb{E}(\, p_t - p_{t-1} \,|\, \Omega_s \,) \;=\; 0 \quad \text{for } s < t-1.$$

That is, before the market maker observes the trade at $t-1$ (or the price change ϵ_{t-1}), all of d_{t-1}, d_t, and ϵ_t are random with expectation zero, so the expected price change is zero. But once the market maker observes the trade d_{t-1}, then

$$\mathbb{E}(\, p_t - p_{t-1} \,|\, \Omega_{t-1} \,) \;=\; -\frac{S}{2}\, d_{t-1}.$$

In the calculation below, we take expectations in an information set Ω_s for $s < t-1$, so the expectation of the price change $p_t - p_{t-1}$ is zero.

Changing t to $t+1$ in (1) we have

$$p_{t+1} - p_t = \epsilon_{t+1} + \frac{S}{2}(d_{t+1} - d_t).$$

Then the *unconditional* (taken at any time s for $s < t-1$) covariance of successive price changes is

$$\begin{aligned} C \equiv \operatorname{Cov}(\, p_t - p_{t-1},\, p_{t+1} - p_t \,) \;&=\; \mathbb{E}((p_t - p_{t-1})(p_{t+1} - p_t)) \\ &=\; \frac{S^2}{4}\mathbb{E}((d_t - d_{t-1})(d_{t+1} - d_t)) \;=\; -\frac{S^2}{4}. \end{aligned}$$

$C < 0$ because of bid-ask bounce. The Roll estimator for the effective spread is

$$S = 2\sqrt{-C}.$$

We may relax two assumptions of the model:

Correlated trade signs Suppose that

$$\mathbb{E}(d_t d_{t+1}) = \delta \qquad \text{and also} \qquad \mathbb{E}(d_{t-1} d_{t+1}) = \delta^2.$$

The second does not necessarily follow from the first, but would be the case, for example, if autocorrelation decays exponentially. Typically, $\delta \geq 0$. Then

$$\begin{aligned} C &= \frac{S^2}{4}\mathbb{E}((d_t - d_{t-1})(d_{t+1} - d_t)) = \frac{S^2}{4}\mathbb{E}(d_t d_{t+1} - d_{t-1}d_{t+1} - d_t^2 + d_{t-1}d_t) \\ &= \frac{S^2}{4}(\delta - \delta^2 - 1 + \delta) = -\frac{S^2}{4}(1-\delta)^2 \end{aligned}$$

and the Roll spread estimator is

$$S = \frac{2}{1-\delta}\sqrt{-C} = \frac{1}{1-\delta}\, S(\delta = 0) > S(\delta = 0) \quad \text{if } \delta > 0..$$

If the market has $\delta > 0$, but you use the Roll estimator $S = 2\sqrt{-C}$ as though $\delta = 0$, then the spread will be *larger* than you think. The serial covariance of price changes will be reduced by the prevalance of successive trades in the same direction, and you will think this is because the spread is smaller.

Correlated midpoint changes Suppose that

$$\mathbb{E}(\epsilon_t \epsilon_{t+1}) = \rho\sigma^2 \qquad \text{with} \qquad \mathbb{E}(\epsilon_t^2) = \sigma^2.$$

That is, σ is a typical size of each ϵ_t, and ρ is the serial correlation. Typically $\rho \leq 0$. Then

$$C = \rho\sigma^2 - \frac{S^2}{4} \qquad \text{or} \qquad S = 2\sqrt{\rho\sigma^2 - C} < S(\rho = 0) \quad \text{if } \rho < 0.$$

If the market has $\rho < 0$, but you use the Roll estimator $S = 2\sqrt{-C}$ as though $\rho = 0$, then the spread will be *smaller* than you think. Some of the mean reversion of the trade prices will be due to midpoint reversion, but you will attribute that motion to bid-ask bounce, which will require you to estimate a larger spread. This may not be entirely wrong.

Market Maker's P&L

Let us look at the problem from the point of view of the market maker. The market maker fixes S at the start of trading. But she[1] knows m_t just before the tth trade, and uses that to set the bid and ask prices. Let x_t be her cash position after the tth trade, and y_t be the share holdings. Then

$$\begin{aligned} x_t &= x_{t-1} + p_t d_t \\ y_t &= y_{t-1} - d_t. \end{aligned}$$

[1] I use a package that chooses "he" or "she" randomly each time I process the text file.

We assume that she marks her position to market after each trade using the midpoint price m_t. Then her wealth is

$$\begin{aligned} V_t &= x_t + m_t y_t \\ &= x_{t-1} + p_t d_t + m_t (y_{t-1} - d_t) \\ &= x_{t-1} + (p_t - m_t) d_t + (m_{t-1} + \epsilon_t) y_{t-1} \\ &= V_{t-1} + \frac{S}{2} + \epsilon_t y_{t-1}. \end{aligned}$$

So the market maker's expected wealth is

$$\mathbb{E}(V_t) \;=\; \mathbb{E}(V_{t-1}) \;+\; \frac{S}{2}.$$

She makes an expected profit of the half-spread on each trade.

Obviously, the market maker would like to increase S as much as possible, to make maximum possible profit, assuming (as we do) that the larger spread would not reduce the frequency of trading. In fact, the discrete price grid to some extent plays the role of forcing a minimum value of S.

But in a competitive market, any finite value of $S > 0$ set by one market maker will be undercut by a smaller value of $S > 0$ set by another market maker, to capture all the flow while still making a positive profit. The only stable value will be $S \to 0$. The market makers cannot make a steady positive profit in a competitive environment. Therefore this model does not answer the question of why the spread exists.

There are three possible ways to modify the model to get a positive spread:

- Transaction costs. Suppose that each trade costs $\gamma > 0$ to the market maker, for exchange fees, order processing, personnel costs, *etc.* Then the change in cash is

  $$x_t = x_{t-1} + p_t d_t - \gamma$$

  and the change in wealth is

  $$V_t = V_{t-1} + \left(\frac{S}{2} - \gamma \right) + \epsilon y_{t-1}.$$

  The overall money flow per trade is

  $$\text{Traders} \quad \underset{S/2}{\longrightarrow} \quad \text{Market Makers} \quad \underset{\gamma}{\longrightarrow} \quad \text{External}$$

  where "External" is outside participants such as technology providers, landlords, etc.

  If $S/2 < \gamma$ then the market makers are losing money and will exit the market; if $S/2 > \gamma$ then competition will drive $S/2$ down to γ but no lower. Presumably the noise traders are making enough money from their long-term trades to support the cost $S/2$ per trade, and presumably the "external" participants have their own costs which means they do not make excess profits.

- Inventory risk. The change in wealth has a risky term $\epsilon_t y_{t-1}$, representing the effect of the price change at time t on the position already held. Above, we have assumed a risk-neutral market maker who is indifferent to this risk. In fact, the market maker likely has some degree of risk aversion, and this riskiness must be priced just like a fixed trade cost.
- Adverse selection. The market maker may lose money despite the spread, if $\mathbb{E}(\epsilon_t y_{t-1}) < 0$, that is,

$$\mathbb{E}\left(\epsilon_t \left(y_0 - \sum_{s=0}^{t-1} d_s \right) \right) \; < \; 0.$$

  But as noted above, there very likely is a positive correlation between ϵ_t and d_s for $s < t$, representing either the market impact of the trade at time s, or the correct anticipation by the trader at time s of the future price change at time t. That is, the market maker buys the stock from the trader when the trader wants to sell, which may be because the trader has information that the price will go down. That is the topic of the next model.

Volatility

Finally, let us calculate the volatility that would be determined from trade prices. For any lag ℓ we compute

$$\begin{aligned} p_{t+\ell} - p_t \; &= \; m_{t+\ell} - m_t \; + \; \frac{1}{2} S (d_{t+\ell} - d_t) \\ &= \; \epsilon_{t+1} + \cdots + \epsilon_{t+\ell} \; + \; \frac{1}{2} S (d_{t+\ell} - d_t) \end{aligned}$$

Since each of these terms is independent,

$$\begin{aligned} \mathbb{E}\Big((p_{t+\ell} - p_t)^2 \Big) \; &= \; \mathbb{E}(\epsilon_{t+1}^2) + \cdots + \mathbb{E}(\epsilon_{t+\ell}^2) \; + \; \frac{1}{4} S^2 \Big(\mathbb{E}(d_{t+\ell}^2) + \mathbb{E}(d_t^2) \Big) \\ &= \; \ell \, \sigma^2 \; + \; \frac{1}{2} S^2 \end{aligned}$$

since $d_t = \pm 1$, and with $\sigma^2 = \mathbb{E}(\epsilon_t^2)$. Then the *volatility at lag* ℓ is

$$\sigma(\ell)^2 \; \equiv \; \frac{1}{\ell} \, \mathbb{E}\Big((p_{t+\ell} - p_t)^2 \Big) \; = \; \sigma^2 \; + \; \frac{S^2}{2\ell}.$$

This is equal to the volatility per trade, plus an additional term that grows large as ℓ decreases to 1, because of the negative serial correlation.

References

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *J. Finance 39*, 1127–1139.