ORF 474: High Frequency Trading
Notes 5b

Robert Almgren

March 4, 2020

As we have been discussing, there are two primary sources of trading costs:

- The bid-ask spread, which has 3 components:
  1. Fixed trading costs (Roll, 1984), including the discrete price grid,
  2. Adverse selection or information costs (Glosten and Milgrom, 1985), and
  3. Inventory risks; and
- Market impact (Kyle, 1985): large trades carry more information than small trades, and the market maker adjusts her price to compensate.

There are three reasons to be interested in this decomposition:

1. *Intellectual and scientific.* We would like to understand how markets work and why there is a spread at all.
2. *Regulatory.* If you have a mandate to reduce trade costs for the benefit of society, then you may implement different rules to address different parts of the cost.
3. *Trading.* To successfully implement a trading strategy, you need some idea what kinds of costs you will incur, and how they change in response to your trading. This may steer you to different products that have lower costs for your kind of trading, or they may limit the total amount of assets that you can take under management.

Although these models are somewhat crude, they are able to give us some insight into the structure and behavior of markets.

Today we present a simple model for inventory risk, then several statistical models that will estimate the various components that contribute to these trading costs, based on trade and quote data from TAQ.

1 Inventory Risk

The third reason for a spread is the market risk taken by the dealer or market maker) each time he takes a position. This exposition is heavily indebted to Foucault et al. (2013).

As before, let x_t be the cash held by the dealer and y_t be the share holdings. Let z_t be the number of shares that he sells at time t: $z_t > 0$ if he sells at the ask, and $z_t < 0$ if he buys at the bid. In practice, $z_t = d_t = \pm 1$, but for now let us let the size be arbitrary.

Then x_t and y_t evolve according to

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

The total wealth is

$$w_t = x_t + p_t y_t$$

which evolves according to

$$\begin{aligned} w_{t+1} &= x_{t+1} + p_{t+1} y_{t+1} \\ &= x_t + p_t z_t + p_{t+1}(y_t - z_t) \\ &= w_t + (p_{t+1} - p_t) y_{t+1}. \end{aligned}$$

We mark the position to market using the last trade price p_t, rather than say the midpoint. Implicit in this construction is the assumption that the trade price is efficient: there are no temporary transaction costs as in the Roll model. But we do have price uncertainty ϵ_t from one step to the next, which is independent of the trade direction. The trade direction carries no information as it does in the Glosten-Milgrom model.

We assume the market maker is *myopic:* she cares only about her position value at time $t+1$. Let $v = p_{t+1}$ which is random at time t, so that

$$w_{t+1} \;=\; x_t + p_t z_t \;+\; v(y_t - z_t). \tag{1}$$

We assume that x_t, p_t, z_t, and y_t are known at time t (after the completion of trade t). The only risk comes from the future price v. In fact, when the market maker sets bid and ask prices at t, there is uncertainty associated with the random submission of a buy or sell order. But we assume that this uncertainty is much smaller than the price uncertainty carried from t to $t+1$. The distribution of the final value is controlled by p_t and z_t, which the market maker influences by setting the bid and ask quotes.

Risk pricing

This is a large and complex subject and we will only say a few words. The general framework is that you have a random future wealth w, which is controlled by some parameter α, so we write $w(\alpha)$. You can choose α, which determines the distribution of w, but you do not know which part of that distribution will be realised. For example, one value of α

may give you a w with a narrow possible range; another value may give you a w with a higher mean than for the first value of α, but a larger uncertainty. In general, you want to choose α so that w has a high mean but also a low uncertainty, and the question is how you trade these off against each other.

Mean-variance optimization Define an objective function

$$U(\alpha) \;=\; \mathbb{E}\big(w(\alpha)\big) \;-\; \lambda \operatorname{Var}\big(w(\alpha)\big)$$

where $\lambda \geq 0$ is a *risk-aversion* parameter. Then solve

$$\max_{\alpha} U(\alpha).$$

This was the insight of Markowitz (1952):[1] penalizing the variance of a portfolio is a natural and effective way to model the importance of diversification.

Mean-variance optimization can have perverse consequences. As an example, suppose that you have two choices. Choice (A) pays you \$0 with certainty; choice (B) pays you \$0 with probability $1-p$, and \$1 with probability p. (B) is clearly better. But choice (B) has $\mathbb{E}(B) = p$ and $\operatorname{Var}(B) = p(1-p)$, so $U = p\big(1-\lambda(1-p)\big)$. If $\lambda > 1/(1-p)$ then you will decline choice (B), because you penalize the risk of gain just as much as the risk of loss.

Mean-variance optimization can be interpreted as an approximation of utility function maximization, going back to von Neumann and Morgenstern (1944). One maximises $U(\alpha) = \mathbb{E}\big(u(w)\big)$ where $u(w)$ is increasing (more money is always better) and concave (each extra dollar is less useful, the more you have initially). A sophisticated mathematical framework can be built around this but it does not necessarily address the messy human reality of risk aversion.

Mean-standard deviation optimization Here the objective function is

$$U(\alpha) \;=\; \mathbb{E}\big(w(\alpha)\big) \;-\; \lambda \operatorname{sd}\big(w(\alpha)\big)$$

where $\operatorname{sd}(x) = \sqrt{\operatorname{Var}(x)}$ is standard deviation. We use the same label λ for the risk aversion as for mean-variance, though the interpretation is different.

Mean-standard deviation has a slight connection to value at risk (VaR), defined as a chosen percentile of the loss distribution. If the wealth distribution is normal, then the percentile can be specified as the mean minus a coefficient times the standard deviation. Of course, VaR is mostly useful when the loss distribution is far from normal, so this analogy should not be pushed too far.

[1] 1990 Nobel Memorial Prize in Economic Sciences, jointly with Merton Miller and William Sharpe.

Comparison To contrast the two risk measures, suppose you have a risky asset whose expected return on some horizon is μ , and whose variance is σ^2. That is, for example, $P(T) = P(0) + \mu + \sigma\xi$ with $\xi \sim \mathcal{N}(0,1)$. If you hold an amount x of this asset (x is the control parameter labelled α above), then your increase in wealth w has $\mathbb{E}(w) = \mu x$, $\mathrm{Var}(w) = \sigma^2 x^2$, and $\mathrm{sd}(w) = \sigma|x|$. The expected value and standard deviation are *linear* in portfolio size x, while variance is *quadratic* in size.

- For mean-variance optimization, you solve $\max_x(\mu x - \lambda\sigma^2 x^2)$, giving $x = \mu/2\lambda\sigma^2$. There is a specific portfolio size that you should optimally hold, defined by the ratio of return to volatility and your risk aversion.
- For mean-standard deviation, you solve $\max_x(\mu x - \lambda\sigma|x|)$. In this case, if $\mu > \lambda\sigma$ then there is no finite optimal portfolio: the larger your (positive) position the better. If $|\mu| < \lambda\sigma$ then the optimum is $x = 0$. If $\mu < -\lambda\sigma$ then you go unbounded negative.

Critique Two important points to keep in mind about risk pricing:

- In practice, no one has any idea what their risk aversion parameter λ is. That is why people talk about "efficient frontiers," where you draw the entire set of optima for the range of all values of λ (that is how Markowitz presented it). Somewhat similarly, the curvature of a utility function in general depends on the total level of wealth (except for certain classes for which the sensitivity is independent of total wealth). It is not clear whether that total wealth should be that of the individual trader, of her desk, or of the firm as a whole. So it is hard to use these models to make concrete forecasts.
- In the context of high frequency trading, risky bets are often diversified across a large number of events. For example, if you make one bet every 5 minutes, then you have 78 events per day. You may not care at all about the riskiness of each one if you are averaging across so many. The importance of risk is critically dependent on the diversification and this is hard to capture in a model for single events.

Risk-averse market making

With those caveats, let us return to the market making problem. The market maker's wealth is given by (1), for which

$$\begin{aligned} \mathbb{E}(w_{t+1}) &= x_t + p_t z_t + \mu_t(y_t - z_t) \\ \mathrm{Var}(w_{t+1}) &= \sigma^2(y_t - z_t)^2 \\ \mathrm{sd}(w_{t+1}) &= \sigma|y_t - z_t| \end{aligned}$$

with $\mu_t = \mathbb{E}(v)$ and $\sigma^2 = \mathbb{E}(\epsilon_t^2)$. The control variables are the trade size z_t and the trade price p_t, which are not precisely specified by the market maker but are interrelated in a way we must discuss.

Mean-variance optimization The objective function is

$$U(z_t) \;=\; x_t \;+\; p_t z_t \;+\; \mu_t(y_t - z_t) \;-\; \lambda\,\sigma^2(y_t - z_t)^2.$$

The first derivative of this,

$$U'(z_t) \;=\; p_t \;-\; \mu_t \;+\; 2\lambda\sigma^2(y_t - z_t),$$

represents the additional utility that the market maker receives by selling an additional unit of stock at price p_t, when he has already sold z_t. The condition $U'(z_t) = 0$ determines the level at which he is indifferent, given that sale quantity and that price:

$$z_t = y_t + \frac{p_t - \mu_t}{2\lambda\sigma^2} \qquad \text{or} \qquad p_t = \mu_t - 2\lambda\sigma^2 y_t + 2\lambda\sigma^2 z_t.$$

To apply this to the bid-ask spread, we consider the particular case $z_t = d_t = \pm 1$, giving

$$\begin{aligned} a_t \;&=\; p_t(z_t = +1) \;=\; \mu_t - 2\lambda\sigma^2 y_t + 2\lambda\sigma^2 \\ b_t \;&=\; p_t(z_t = -1) \;=\; \mu_t - 2\lambda\sigma^2 y_t - 2\lambda\sigma^2 \end{aligned}$$

Thus the midpoint is

$$m_t \;=\; \mu_t - 2\lambda\sigma^2 y_t.$$

This is shifted by the market maker's current share inventory y_t. If the market maker is currently *long* the asset, so $y_t > 0$, then she desires to sell shares to drive her position back to zero and eliminate risk. Therefore she *lowers* the bid and ask prices: she must be paid an additional premium to buy at the bid, and a smaller premium to sell at the ask. Conversely, if she is *short* the asset, then she *raises* the prices. The amount of the shift is determined by the current share holdings, multiplied by the volatility and risk aversion, which represent the market maker's urgency to flatten her position.

The bid-ask spread is

$$S \;=\; 4\lambda\sigma^2,$$

again determined by the risk of taking the additional position, multiplied by risk aversion.

We have said nothing about how the external traders respond to the shift in midpoint. A natural interpretation would be that when the market maker *lowers* the prices, then external traders are more likely to *buy*, helping her to flatten her position; conversely, when she *raises* the prices they are more likely to sell. But that is not part of this model.

The above is the price demanded by the market maker to buy or sell *one* share $d_t = \pm 1$. The *second* share bought or sold demands *twice* the premium. Thus the inside quotes would be expected to be very thin. In fact, the market maker sets a continuous demand curve, where the cumulative quantity supplied or demanded at each price level increases linearly with distance away from the midpoint.

One final question is *whose* inventory sets the inside quotes. Above, we have invoked competition between multiple market makers to argue for efficient prices, and they may not all have the same inventory level. One interpretation is to say that y_t is the *total*

or *net* position of all the market makers together. This assumes that they have some relatively efficient way to transfer inventory between themselves, which is plausible in some markets but not in all.

Another interpretation is that the inside quote on each side will be set by the individual market maker who has the largest position on the appropriate side. The inside dealer is said to have the "axe" in that security. Because that individual has a position which she wants to trade for her own reasons, she is willing to quote a favorable price. But just as in the Glosten-Milgrom model, this dealer does not want to release the information that she has that position, so this inside quote may be disseminated only privately to selected potential counterparties.

Mean-standard deviation optimization The objective function is

$$\begin{aligned} U(z_t) &= x_t + p_t z_t + \mu_t(y_t - z_t) - \lambda\sigma|y_t - z_t| \\ &= \begin{cases} x_t + p_t z_t + \mu_t(y_t - z_t) + \lambda\sigma(z_t - y_t) & \text{if } z_t \le y_t \\ x_t + p_t z_t + \mu_t(y_t - z_t) - \lambda\sigma(z_t - y_t) & \text{if } z_t \ge y_t \end{cases} \end{aligned}$$

so

$$U'(z_t) = \begin{cases} p_t - \mu_t + \lambda\sigma & \text{if } z_t < y_t \\ p_t - \mu_t - \lambda\sigma & \text{if } z_t > y_t. \end{cases}$$

We claim that this determines bid and ask prices

$$\begin{aligned} a_t &= \mu_t + \lambda\sigma \\ b_t &= \mu_t - \lambda\sigma. \end{aligned}$$

To see this, suppose that at least one market maker has $|y_t| < 1$: his current inventory is less than one share positive or negative. For a buy order, $z_t \ge 1$ so $U'(z_t) = p_t - \mu_t - \lambda\sigma$. Then the trade price p_t must have $p_t = \mu_t + \lambda\sigma$, so that the market maker makes neither a positive nor a negative expected profit on an incoming order of arbitrary size. The opposite is true for a sell order. Therefore these are the market equilibrium prices. The market makers are willing to trade arbitrarily large quanitities at these prices.

If a market maker, or the market makers considered as whole, have inventory $y_t > 1$ or $y_t < 1$, then the situation is a little less satisfactory. For example, if $y_t > 1$, then for a 1-lot trade, always $z_t < y_t$ whether $z_t = d_t = +1$ or -1. Then both bid and ask prices will be $b_t = a_t = \mu_t - \lambda\sigma$. The market maker is willing to sell at least one unit at the bid because the price concession is balanced by the risk reduction of eliminating one unit of inventory. For larger trade sizes, the price will revert to the "normal" ask price of $\mu_t + \lambda\sigma$. So this is a somewhat more realistic depiction of bid and ask prices than the mean-variance interpretation but perhaps not by much.

This model can be extended to cover liquidation across multiple periods, but the details are not very satisfying and the conclusions are qualitatively the same. Of course the risk will be increased if the time taken to liquidate is increased, and that is itself an important part of such a model.

The central intuition is two points:

- The midpoint price, and the bid and ask prices, are shifted up or down depending on the inventory of the market maker or makers. They adjust the prices to compensate the added or reduced risk of taking the trade, depending on whether it adds to or subtracts from their current holdings. In practice, this shift could also be expected to induce the external traders to buy more or sell more, and that would steer the market makers' inventory back to zero, but that is not included in this model.
- The bid-ask spread is determined by the risk premium of adding the shares to the position of the market maker.

2 Empirical Tests

Our tools for doing this will be trade and quote data. Specifically, we will use

- trade prices and sizes, and
- trade signs. These latter are a bit less solid since they depend on matching trade data with quote data, except for markets that specifically report the aggressor direction.

We will try not to use the following inputs:

- the "true" value or "fair" value, since it is not observable;
- posted bid and ask prices, since the "effective" spread may be different than the posted spread; and
- the posted midpoint price, though we could use this if we needed.

The outputs of our model will be coefficients for fixed cost, information cost, inventory risk, and market impact, as they contribute to trade costs, as part of a linear regression.

For future reference,

$$
\begin{aligned}
\mu_t &= \text{market maker's best estimate of "true value" } \textit{after} \text{ trade } t \\
\epsilon_t &= \text{information revealed between } t-1 \text{ and } t, \text{ or "news"} \\
m_t &= \text{posted midpoint just before trade } t \\
S_t &= \text{bid-ask spread just before trade } t\text{:} \quad S = a_t - b_t \\
b_t, a_t &= \text{posted bid and ask just before trade } t\text{:} \; b_t = m_t - S/2, \; a_t = m_t + S/2 \\
d_t &= \text{direction of trade } t\text{:} \quad d_t = \pm 1 \\
q_t &= \text{signed trade size of trade } t\text{:} \quad q_t = d_t \cdot \texttt{size}_t, \; d_t = \operatorname{sgn}(q_t)
\end{aligned}
$$

We will denote changes in variables as

$$\Delta p_t \;=\; p_t \;-\; p_{t-1}$$

and similarly for other variables.

Roll model

We write the model of Roll (1984) in a systematic way that we can extend below. Trades do not carry information, and the market maker sets her midpoint at the expected value:

$$\begin{aligned} \mu_t &= \mu_{t-1} + \epsilon_t \\ m_t &= \mu_t \\ p_t &= \mu_t + \gamma\, d_t \end{aligned}$$

Thus,

$$\begin{aligned} \Delta p_t &= \Delta\mu_t + \gamma\, \Delta d_t \\ &= \gamma\, \Delta d_t + \epsilon_t. \end{aligned}$$

We could therefore regress Δp_t against Δd_t, with a zero intercept. The two values d_t, d_{t-1} combine into the single value $\Delta d_t = d_t - d_{t-1}$, which is used to determine the single coefficient γ. In R, this would be something like

```
lm( deltaP ~ 0 + deltaD )
```

where the 0 enforces the zero intercept. The resulting coefficient would be γ.

The difficulty with this regression is that since d_t takes only two values $\{-1, +1\}$, the difference Δd_t takes only the three values $\{-2, 0, 2\}$. Effectively, the slope will be something close to $\gamma = (\pi_+ - \pi_-)/2$ where $\pi_\pm = \langle\, \Delta p_t \,|\, \Delta d_t = \pm 2 \,\rangle$. The middle value $\langle\, \Delta p_t \,|\, \Delta d_t = 0 \,\rangle$ does not affect the regression, but should hopefully be close to zero. Also, since the price grid is discrete, Δp_t also will only take a discrete set of values.

Roll (1984) used a "trick" to avoid including trade sign: since

$$\Delta p_t \cdot \Delta p_{t-1} = (\gamma\, \Delta d_t + \epsilon_t)(\gamma\, \Delta d_{t-1} + \epsilon_{t-1})$$

then if everything is serially uncorrelated, by averaging we have

$$\operatorname{Cov}(\Delta p_t, \Delta p_{t-1}) = -\gamma^2$$

which gives us γ directly from trades, with no use of quote data.

Trade size We may suppose that the trade cost depends on trade size as well as sign:

$$p_t = \mu_t + \gamma_0\, d_t + \gamma_1\, q_t.$$

There are two possible interpretations for γ_1 depending on sign:

- If $\gamma_1 > 0$, then larger trades cost more. This would happen if a nontrivial fraction of all trades are large enough to exhaust liquidity at the inside quote, and go to worse prices deeper in the book. If $\gamma_1 = 0$, then most trades would fill at the inside quote. This intepretation is not likely in this model, since this price change is not incorporated into future prices but is experienced only on this trade; market impact usually moves the price permanently.

- If $\gamma_1 < 0$, then large trades receive a discount in terms of the price per unit traded. This cost reduction would be possible in markets where dealers quote specifically to clients. Examples of such markets would be foreign exchange, which is commonly traded over-the-counter, or corporate bonds, which trade mostly through dealers (TRACE dataset). This decreasing cost structure would not be realistic for a product traded on a public limit order book, where everyone sees the same price.

With this model,

$$\Delta p_t \;=\; \gamma_0 \, \Delta d_t \;+\; \gamma_1 \, \Delta q_t \;+\; \epsilon_t$$

which is easy to estimate: the four input values $d_t, d_{t-1}, q_t, q_{t-1}$ group into the two values $\Delta d_t = d_t - d_{t-1}$ and $\Delta q_t = q_t - q_{t-1}$, which are used to fit the two coefficients γ_0 and γ_1.

From now on, we shall not use this extension, so we will just write $p_t = \mu_t + \gamma d_t$.

Glosten-Milgrom model

The model of Glosten and Milgrom (1985) introduces the information, or adverse selection, effect of trades. Here we also include fixed transaction cost γ and news arrival ϵ_t, which we did not before. Each model we write only adds effects to the preceding ones.

As we presented it last week, the information effect depends only on the trade sign; or equivalently, all trades are of the same unit size. Thus we write

$$\begin{aligned} m_t &= \mu_{t-1} + \epsilon_t \\ \mu_t &= m_t + \lambda \, d_t \\ p_t &= \mu_t + \gamma \, d_t \end{aligned}$$

The interpretation is that μ_t is the value that the market maker is comfortable having bought or sold at, *after* the trade d_t provides its information.

To do a regression, we must eliminate the unobservable μ_t. We see

$$\begin{aligned} \Delta p_t &= \Delta \mu_t + \gamma \, \Delta d_t \\ &= \lambda \, d_t + \gamma \, \Delta d_t + \epsilon_t. \end{aligned}$$

The two values d_t, d_{t-1} (via the combination $\Delta d_t = d_t - d_{t-1}$) are used to fit the two parameters γ and λ for the fixed cost and the information cost respectively.

Regression The independent variables d_t and Δd_t take only the values $d_t \in \{-1, 1\}$ and $\Delta d_t \in \{-2, 0, 2\}$. The pairs $(d_t, \Delta d_t) = (-1, 2)$ and $(d_t, \Delta d_t) = (1, -2)$ are impossible, so the entire regression is determined by the four averages of Δp_t

$$\begin{aligned} \Pi_{-,-} &= \langle \, \Delta p_t \mid d_t = -1, \Delta d_t = -2 \, \rangle & \qquad \Pi_{+,0} &= \langle \, \Delta p_t \mid d_t = +1, \Delta d_t = 0 \, \rangle \\ \Pi_{0,+} &= \langle \, \Delta p_t \mid d_t = -1, \Delta d_t = 0 \, \rangle & \qquad \Pi_{+,+} &= \langle \, \Delta p_t \mid d_t = +1, \Delta d_t = +2 \, \rangle \end{aligned}$$

Spread estimation Since in this model

$$\begin{aligned} \Delta p_t &= (\lambda + \gamma) d_t - \gamma d_{t-1} + \epsilon_t \\ \Delta p_{t-1} &= (\lambda + \gamma) d_{t-1} - \gamma d_{t-2} + \epsilon_{t-1} \end{aligned}$$

the covariance is

$$C = -\gamma(\lambda + \gamma).$$

If the market has $\lambda > 0$, but you use the estimator $\hat{S} = 2\sqrt{-C}$ as though $\lambda = 0$, then

$$\hat{S} = 2\sqrt{\gamma(\lambda + \gamma)} = \sqrt{\frac{\gamma}{\lambda + \gamma}}\, 2(\lambda + \gamma) = \sqrt{\frac{\gamma}{\lambda + \gamma}}\, S < S.$$

Your number is too *small;* the spread is *larger* than you think. A part of the bid-ask spread comes from information cost, which is not seen in the reversion of trade prices.

Trade size It makes sense to also add trade size as a variable, writing

$$\begin{aligned} \mu_t &= m_t + \lambda_0 d_t + \lambda_1 q_t \\ p_t &= m_t + \lambda_0 d_t + \lambda_1 q_t + \gamma d_t. \end{aligned}$$

Then

$$\Delta p_t = \gamma \Delta d_t + \lambda_0 d_t + \lambda_1 q_t + \epsilon_t$$

The 3 values d_t, d_{t-1}, q_t (via $\Delta d_t = d_t - d_{t-1}$) are used to fit the 3 parameters γ, λ_0, and λ_1 for the fixed cost and the two components of the information cost.

From now on, we shall keep the component λ_1 but drop the component λ_0, so our model that combines Roll with Glosten-Milgrom is

$$\begin{aligned} p_t &= m_t + \gamma d_t + \lambda q_t \\ \Delta p_t &= \gamma \Delta d_t + \lambda q_t + \epsilon_t. \end{aligned}$$

The coefficient γ represents *temporary* trading cost: a cost which is paid by the liquidity demander on each trade, but which is not carried forward to future prices. The coefficient λ (like the parameter λ of Kyle (1985)) represents *permanent* trading cost: a cost which is paid on this trade, but which also shifts all future prices.

Inventory Cost

Now we look at the inventory model. Recall that the dealer's inventory y_t changes by $\Delta y_t = -q_t$. It is not entirely clear *whose* inventory we are talking about. In an active public market like a stock or futures market, limit orders may be placed by any participant, and it will be very difficult to distinguish the inventory of any particular liquidity provider. But we shall discuss the model as though there is a single market maker.

Now the market maker requires a trade price

$$p_t \;=\; \mu_t \;-\; \beta\, y_t \;+\; \gamma\, d_t$$

including (1) the fair value after the information transmitted by trade t (2) a risk compensation for the inventory after trade t, and (3) the fixed cost paid on trade t itself.

We readily calculate

$$\begin{aligned}\Delta p_t \;&=\; \Delta\mu_t \;-\; \beta\,\Delta y_t \;+\; \gamma\,\Delta d_t \\ &=\; \lambda\, q_t \;+\; \beta\, q_t \;+\; \gamma\,\Delta d_t \;+\; \epsilon_t \\ &=\; (\lambda+\beta)\, q_t \;+\; \gamma\,\Delta d_t \;+\; \epsilon_t.\end{aligned}$$

But now there is a problem: λ and β appear only in combination. Although we have three input variables q_t, d_t, and d_{t-1}, it is impossible to distinguish the three parameters γ, λ, and β for the fixed cost, the information cost, and the inventory cost.

We might argue that this is not a problem in practice: if they are indistinguishable then why do we not simply lump them together? That would be true if our only interest were to design trading. But from the scientific point of view we would like to understand the dynamics, and if we had the responsibility to regulate these markets then we would need to understand which is more important.

These two effects can be distinguished only if we allow information to enter the model in a richer way. More specifically, we include the serial correlation, or autocorrelation, of order signs. This may arise for two reasons:

- Positive autocorrelation may come about either because of
  - trade splitting (an interesting reason), or
  - the specifics of how the exchange reports trades (something you should fix).

  These effects are likely dominant in public liquid markets.
- Negative autocorrelation may arise if the market makers or dealers are successfully steering their inventory to keep it from getting too large positive or negative. That is, if a series of buy orders leaves them net short, they will encourage the execution of sell orders to flatten their position, and conversely. This effect may be very hard to observe, except in markets that are strongly dealer-dominated.

Now let us denote

$$\eta_t \;=\; q_t \;-\; \mathbb{E}\left(\, q_t \,\middle|\, \Omega_{t-1} \,\right)$$

the *surprise* in signed trade size. Here, Ω_{t-1} denotes the market maker's information just before trade t. The idea is that information costs will be driven by η_t, while inventory costs still respond to the actual trade q_t.

So we write

$$\mu_t \;=\; \mu_{t-1} \;+\; \lambda\,\eta_t \;+\; \epsilon_t$$

and

$$\begin{aligned}\Delta p_t &= \Delta\mu_t - \beta\,\Delta y_t + \gamma\,\Delta d_t \\ &= \lambda\,\eta_t + \beta\,q_t + \gamma\,\Delta d_t + \epsilon_t\end{aligned}$$

This is not quite ready to implement, because η_t is unobservable. To make it observable, we need to introduce a specific model for information. The simplest is the AR(1) process

$$q_t = \phi\,q_{t-1} + \eta_t$$

so that

$$\eta_t = q_t - \phi\,q_{t-1}.$$

The reversion coefficient ϕ can be estimated separately from the signed trade sizes. Then

$$\Delta p_t = (\lambda+\beta)q_t - \lambda\phi\,q_{t-1} + \gamma\,\Delta d_t + \epsilon_t.$$

This is finally the form that we want. There are 3 observable variables q_t, q_{t-1}, and d_t, to estimate 3 coefficients λ, β, and γ, and all appear independently.

More complicated models Much more complicated models appear in the literature, to take account of long-range dependencies. For example, Hasbrouck (1991) writes

$$\begin{aligned}\Delta m_t &= \sum_{j=1}^{\infty} a_j\,\Delta m_{t-j} + \sum_{j=1}^{\infty} b_j\,q_{t-j} + \epsilon_t \\ q_t &= \sum_{j=1}^{\infty} c_j\,\Delta m_{t-j} + \sum_{j=1}^{\infty} h_j\,q_{t-j} + \eta_t\end{aligned}$$

so that both midpoint changes and signed trade sizes depend on all history. This includes all of the above as special cases.

References

Foucault, T., M. Pagano, and A. Röell (2013). *Market Liquidity: Theory, Evidence, and Policy*. Oxford: Oxford University Press.

Glosten, L. R. and P. R. Milgrom (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *J. Financial Econ. 14*(1), 71–100.

Hasbrouck, J. (1991). Measuring the information content of stock trades. *J. Finance 46*(1), 179–207.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica 53*, 1315–1336.

Markowitz, H. (1952). Portfolio selection. *J. Finance 7*, 77–91.

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

von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.