Section 8. Transaction Costs

Definition 8.1. An *order* $\mathfrak{o} = (q, \tau, \tau')$ will be an instruction to buy or sell a fixed quantity q of a certain asset over the time window $[\tau, \tau']$. Per convention, an order to *sell* has $q < 0$ while an order to buy has $q > 0$. A *fill* is a statement that part of the buying or selling in a particular order has been completed at a certain time and a certain price.

Thus the order $\mathfrak{o}$, if it were completed, would lead to a sequence of fills $\{f_i : i = 1 \dots n_f\}$ where each fill $f_i = (\mathfrak{o}, t_i, n_i, p_i)$ contains the parent order, the time $t_i \in [\tau, \tau']$, the number n_i of shares filled, and the price p_i at which they were filled. Let $\mathrm{par}(f)$ denote the parent order that generated the fill.

In other words a fill $f_i = (\mathfrak{o}, t_i, n_i, p_i)$ is a statement that n_i shares have been exchanged for cash in the amount of $n_i p_i$ dollars (or other numeraire currency) at time t_i, as part of the parent order $\mathfrak{o}$. Also, we assume all of the fills associated to a fixed parent order have $\mathrm{sgn}(n_i) = \mathrm{sgn}(q)$, as is logical. Thus necessarily $\sum_i |n_i| = |\sum_i n_i| \leq |q|$.

The total amount filled is $\sum_{i=1}^{n_f} n_i$ which could be smaller in magnitude than $|q|$; in other words, not every parent order is completely filled. For example, a parent order could be an unrealistically large number of shares in a very illiquid stock. The execution algo could either decide not to fully fill based on certain agreed-upon limits, or it could simply fail to locate the shares. (If the order involves taking a short position in a stock that is hard to borrow, failing to locate is quite common.)

Definition 8.2. An *execution algorithm* or, more simply and in common usage, an *algo*, is a means of creating a sequence of fills for any given order, ie. a mapping from $\mathfrak{o} \to \{f_i\}$. Equivalently it is a means of choosing n_i and t_i in the sequence $f_i = (\mathfrak{o}, t_i, n_i, p_i)$.

For any list of pairs $L = \{(n_i, p_i) : i = 1, \dots, N_L\}$ where $p_i > 0$ are prices and $n_i \in \mathbb{N}$, we define

$$\mathrm{vwap}(L) = \frac{\sum_i n_i p_i}{\sum_i n_i}.$$

where vwap stands for "volume-weighted average price." In one common example, L is the list of all trade prices, with the volume transacted at each price, for a particular (stock, day). For analysis of intraday patterns, one could take L to be a subset of this data over a finer-grained time period, such as a minute.

Hence any order that generates at least one fill has an associated vwap:

$$\mathrm{vwap}(\mathfrak{o}) = \mathrm{vwap}\{f : \mathrm{par}(f) = \mathfrak{o}\}$$

To the portfolio manager, the end result of executing the order is essentially the same as if the entire quantity q had been filled in one shot, at price $\mathrm{vwap}(\mathfrak{o})$.

Definition 8.3. A *benchmark pricing method* is a way of assigning a theoretical price $p_0(\mathfrak{o}) \in \mathbb{R}$ to any order, ie a mapping $p_0 : \mathfrak{O} \to \mathbb{R}$ where $\mathfrak{O}$ denotes the space of all possible orders. This price need not represent the prices of any actual trades.

One popular benchmark is arrival (mid) price, or simply "arrival price." This is the last midpoint price available before the order begins being executed. This seems to be the benchmark used in Almgren et al. (2005), for example. We now come to the most important definition of this section:

Definition 8.4. Given a benchmark pricing method p_0, the *slippage* of an order relative to this benchmark pricing method is defined as

$$\mathrm{slip}(\mathfrak{o}) = \left(\sum_{i=1}^{n_f} n_i\right)[\mathrm{vwap}(\mathfrak{o}) - p_0(\mathfrak{o})] \tag{8.1}$$

Note that $\mathrm{slip}(\mathfrak{o})$ has units of whatever currency the prices are denominated in, and the sign is such that *positive* slippage denotes a *worse* result for the trader than transacting at the benchmark price. If $\mathfrak{O}$ is an entire set of orders, then

$$\mathrm{slip}(\mathfrak{O}) = \sum_{\mathfrak{o} \in \mathfrak{O}} \mathrm{slip}(\mathfrak{o})$$

As stated, the benchmark pricing method is a mathematical construct and need not correspond to tradable prices. For example, if we take the zero mapping $p_0(\mathfrak{o}) = 0 \ \forall \ \mathfrak{o} \in \mathfrak{O}$ then the slippage is simply the total traded notional value. The arrival mid price is also not achievable – how exactly are we supposed to consistently transact between the bid and the offer? Given a definition of benchmark pricing method p_0, one can then begin to devise algos which are designed to minimize slippage relative to that benchmark.

For orders that are executed incrementally over the course of an entire day, the full-day vwap is a popular benchmark price. Like the other benchmark prices discussed above, this one is not implementable either: no algo can guarantee that the vwap of your order will equal the aggregate market vwap of the day.

A very simplistic implementation of an algo benchmarked to vwap might be as follows. Divide the trading day up into 5-minute bins and predict the fraction of the day's volume that will occur within each bin. Plan to execute the same fraction of your order within that bin. When there is no "interesting" activity going on in a given stock, then algos of this sort can achieve the vwap plus noise, where the noise mostly comes from the difference between the predicted intraday volume pattern and the realized one.

In the presence of interesting activity, there is an interesting bias in the performance of algos designed with a benchmark of vwap. For example, suppose at close minus 10 minutes, a news item comes out, and it's *good news*, and this causes the price and volume both to rise continuously for the next 10 minutes. Since the typical vwap algo does not have a crystal ball and hence did not foresee this sequence of events, the vwap algo implementing a buy order will fall short of the day vwap. In the opposite case of bad news and decreasing price (but still increasing volume), vwap algos implementing sell orders once again achieve worse than day-long vwap. This proves that:

> *No execution algo always achieves VWAP-or-better prices unless the algo has reasonably accurate short-term forecasts of volume and/or price.*

Let's now consider a detailed example in which we buy a stock that is going up and sell it at a higher price and our benchmark is arrival mid price. Immediately before the buy order begins execution, the bid and the ask are $b_0 < a_0$ and immediately after the sell order begins execution, the bid and the ask are $b_1 < a_1$. We also assume that both transactions are for 100 shares which are always easily available on the inside market, and also that $a_0 < b_1$ so the transaction will be profitable.

Hence the benchmark prices are the mids $m_0 = (a_0 + b_0)/2$ and $m_1 = (a_1 + b_1)/2$. We buy at the ask and sell at the bid, so

$$\text{vwap}(\mathfrak{o}_0) = a_0, \quad \text{vwap}(\mathfrak{o}_1) = b_1.$$

Let π denote our P&L, then

$$\pi = 100(b_1 - a_0) = 100(m_1 - m_0) - \text{slip}(\mathfrak{o}_0) - \text{slip}(\mathfrak{o}_1) \tag{8.2}$$

where $\text{slip}(\mathfrak{o}_0) = 100(a_0 - m_0)$ and $\text{slip}(\mathfrak{o}_1) = 100(m_1 - b_1)$. Note that as per our convention $\text{slip}(\mathfrak{o}_0)$ and $\text{slip}(\mathfrak{o}_1)$ are both positive. Eq (8.2) is easily verified by simple arithmetic. Furthermore,

$$\pi = 100(b_1 - a_0) = hR - \mathrm{slip}(\mathfrak{o}_0) - \mathrm{slip}(\mathfrak{o}_1), \quad R := \frac{m_1 - m_0}{m_0}, \quad h := 100m_0. \tag{8.3}$$

We can interpret (8.3) as stating that if we price our intended holding of 100 shares at the arrival mid, so in dollars our intended holding is worth $h = 100m_0$, then the P&L π can be represented as the holding value times the return R (which must be price-return using the benchmark price!) minus the total slippage from both orders.

Eq.(8.3) generalizes to portfolios in the following way. Suppose that we hold portfolio $h_0 \in \mathbb{R}^n$ now and intend to trade into portfolio $h \in \mathbb{R}^n$. Suppose there are two times t_0 and t_1, and at t_0 we will begin trading from h_0 to h at t_0, we will reach h before t_1, and then at t_1 we will begin liquidating h. Let $\mathfrak{O}_i$ denote all orders started at t_i. Then the P&L can be written

$$\pi = h'R - \mathrm{slip}(\mathfrak{O}_1) - \mathrm{slip}(\mathfrak{O}_2) \tag{8.4}$$

where $R \in \mathbb{R}^n$ is the vector of returns over the interval $[t_0, t_1]$ computed with respect to benchmark price. We could equivalently write

$$\pi = h'R - \mathrm{slip}(h_0, h) - \mathrm{slip}(h, 0) \tag{8.5}$$

where $\mathrm{slip}(x, y)$ denotes the slippage of the orders needed to trade from portfolio x into portfolio y. The second term $\mathrm{slip}(h, 0)$ is the slippage incurred from liquidating the final portfolio h. As in the simple example above, it is necessary to liquidate the final portfolio to actually realize all profits in dollars; otherwise some portion of the profits will be left as "unrealized" and any unrealized profits will be subject to slippage before they are "realized" or translated to dollars.

Definition 8.5. *Liquidation slippage* of a portfolio h is defined as $\mathrm{slip}(h, 0)$ in the notation above, i.e. the slippage incurred on the full set of orders necessary to convert the holdings entirely to cash. The liquidation slippage of h will be denoted by

$$\mathrm{liqslip}(h) := \mathrm{slip}(h, 0).$$

For a perspective on optimal execution algos with fairly similar notation to ours, see Almgren and Chriss (1999) and Almgren and Chriss (2001). For a more advanced multiperiod treatment in the spirit of optimizing (8.4) see Kolm and Ritter (2015).

Note that $\mathrm{slip}(\mathfrak{o})$ is not knowable at the order creation time τ (as it involves future prices).

Definition 8.6. A *predictive slippage model* is a model for the conditional density $p(\mathrm{slip}(\mathfrak{o}) \,|\, I_\tau)$ where I_τ denotes the information set available at time τ. Many researchers simply model $\mathbb{E}[\mathrm{slip}(\mathfrak{o}) \,|\, I_\tau]$ directly without modeling the full distribution.

A number of prominent academics have studied the problem of predicting $\mathrm{slip}(\mathfrak{o})$ as function of the order quantity q and attributes of the asset being traded. Some attributes that have been found to be predictive include that asset's volatility, volume, and the window $T = [\tau, \tau']$ over which the orders are filled. One of the most oft-cited such studies is Almgren et al. (2005).

Let's now suppose that our prior holding in some asset is h_0 dollars and we are considering a trade of

$$\delta := h - h_0$$

so that our new holding will be h. Suppose we translate δ into a quantity of shares q using the arrival price, so that up to roundoff errors, $q = \delta/p_0$. If we assume that the order *will be fully executed* then we can algebraically manipulate the definition (8.1) to express it in terms of price return and order value:

$$\mathrm{slip}(\mathfrak{o}) \quad = \quad \delta \cdot R_S(\mathfrak{o}) \quad \text{where} \quad R_S(\mathfrak{o}) := \frac{\mathrm{vwap}(\mathfrak{o}) - p_0}{p_0} \tag{8.6}$$

and the quantity $R_S(\mathfrak{o})$, referred to as "slippage price return" in these notes, is such that the slippage equals return on (signed) dollars traded, using this number as the return.

The slippage price return is called J in the paper of Almgren et al. (2005), and is referred to there as "realized impact" although it need not literally be due to impact if we are talking about tiny trades in liquid names. Furthermore the model of Almgren et al. (2005) for slippage price return is

$$J = R_S(\mathfrak{o}) = \frac{1}{2}\gamma\sigma\frac{q}{V}\left(\frac{\Theta}{V}\right)^{1/4} + \mathrm{sgn}(q)\eta\sigma\left|\frac{q}{VT}\right|^{3/5}$$

where V is average daily share volume, Θ is shares outstanding, and γ, η are constants that must be fit to market data. Almgren et al. (2005) denote the quantity traded, q, instead by X. Note that the predicted slippage price return always has the same sign as the order, hence $\mathbb{E}[\mathrm{slip}(\mathfrak{o})]$ is always positive in this model.

Noting that $\delta = p_0 q$ we can take the expression for J above and re-write it in terms of dollar-denominated quantities, by multiplying both terms by $1 = p_0/p_0$, which gives

$$
\begin{aligned}
(8.7)\ \ R_S(\mathfrak{o}) \quad &= \quad \frac{\gamma}{2}\sigma\frac{\delta}{\text{Advp}}\left(\frac{\Theta}{V}\right)^{1/4} + \operatorname{sgn}(\delta)\eta\sigma\left|\frac{\delta}{\text{Advp}\cdot T}\right|^{3/5} \\
(8.8)\ \ \text{slip}(\delta) \quad &= \quad \delta\cdot R_S(\mathfrak{o}) = \sigma\left[\frac{\gamma}{2}\frac{\delta^2}{\text{Advp}}\left(\frac{\Theta}{V}\right)^{1/4} + \eta|\delta|\left|\frac{\delta}{\text{Advp}\cdot T}\right|^{3/5}\right]
\end{aligned}
$$

where $\text{Advp} := p_0 V$ denotes average daily volume times price.

Someone obsessed with mathematical simplicity over empirical goodness-of-fit might take the approach of ignoring the more complicated second term in (8.8) and re-fitting the coefficient γ in the first term to account for the second term's absence. In this modified model, the impact would be a purely quadratic function of δ, as seems to be assumed by Gârleanu and Pedersen (2013).

Mean-variance optimization is concerned with the problem

$$\max_h \left\{\mathbb{E}[\pi(h)] - \frac{\kappa}{2}\mathbb{V}[\pi(h)]\right\}$$

where

$$\pi(h) = h'R - [\text{slip}(h_0, h) + \text{liqslip}(h)]$$

and the latter expression for π is of course copied from (8.5). Suppose we assume that $\mathbb{V}[\pi(h)]$ is well approximated by the variance of the term $h'R$ in $\pi(h)$, in other words $\mathbb{V}[\pi(h)] \approx h'\Sigma h$ where $\Sigma := \text{cov}(R)$. Then the mean-variance problem becomes

$$(8.9) \qquad \max_h \left\{h'\mathbb{E}[R] - \frac{\kappa}{2}h'\Sigma h - \mathbb{E}[\text{slip}(h_0, h) + \text{liqslip}(h)]\right\}$$

Note that there are (at least) two functional forms for the latter two terms in (8.9) which allow for easy solution of the mean-variance maximization problem: (1) purely quadratic, and (2) quadratic plus absolute-value type penalty terms. In the first case, the entire problem remains quadratic, while in the second case, the problem becomes mathematically equivalent to a LASSO regression. The Almgren et al. (2005) form (8.8) does not lead to such a well-known procedure as LASSO, but at least the associated problem is convex and differentiable, hence standard optimization routines can be expected to perform well.

References

Almgren, Robert and Neil Chriss (1999). “Value under liquidation”. In: *Risk* 12.12, pp. 61–63.

– (2001). “Optimal execution of portfolio transactions”. In: *Journal of Risk* 3, pp. 5–40.

Almgren, Robert et al. (2005). “Direct estimation of equity market impact”. In: *Risk* 18.7, pp. 58–62.

Gârleanu, Nicolae and Lasse Heje Pedersen (2013). “Dynamic trading with predictable returns and transaction costs”. In: *The Journal of Finance* 68.6, pp. 2309–2340.

Kolm, Petter N and Gordon Ritter (2015). “Multiperiod Portfolio Selection and Bayesian Dynamic Models”. In: *Risk magazine* March issue.