LECTURE NOTES 1

Transaction Costs and Market Impact[1]

Petter Kolm
NYU Courant

In addition to these notes, the required readings are:

- Familiarize yourself with the most recent transaction cost reports by Virtu Financial. They produce good U.S. equity reports.[2]

Optional reading:

- Peruse some of the research reports by Pragma such as *"On the Limits of Markouts and Venue Curation"* (October 2021) and *"Measuring Execution Quality – Finding the Signal in the Noise"* (November 2020).

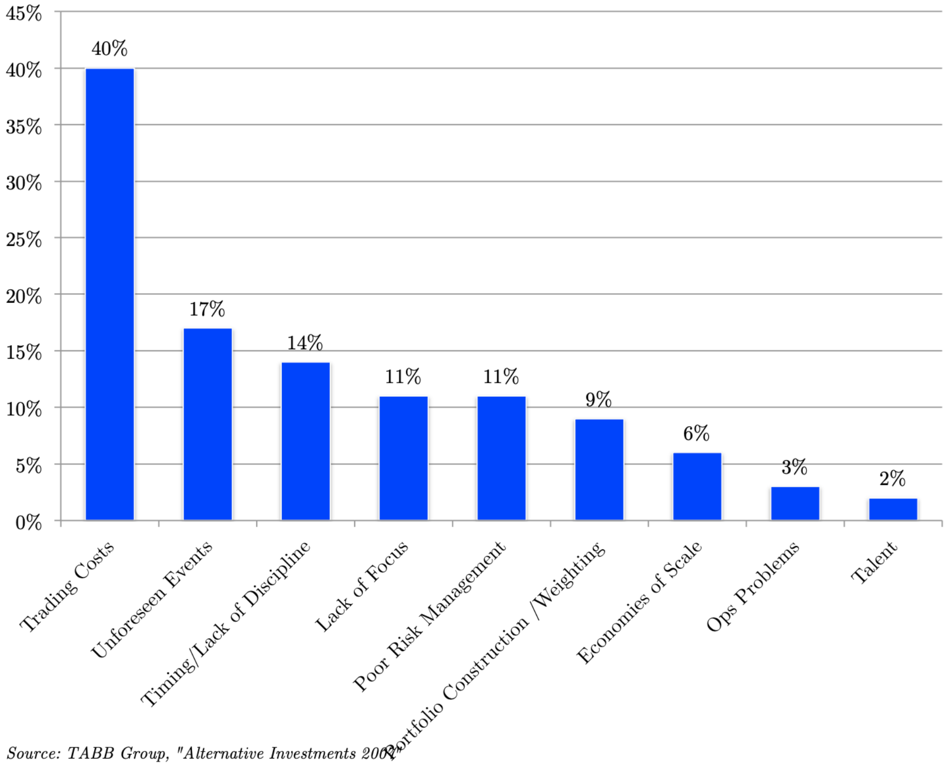
1. Transaction Costs

In a survey conducted by the TABB Group, U.S. domestic equity managers where asked where they think most of their alpha is lost. The results from the survey are shown in Figure 1. We see that trading costs by far dominated.

Some transaction costs (t-costs) are not known before we trade. We refer to t-costs that are known up front before trading as *explicit costs*, and those that are not as *implicit costs*:

[1] Version: January 31, 2023. These notes are copyrighted and intended only for students registered for the course *"Algorithmic Trading & Quantitative Strategies"* at NYU Courant. Please do not copy or post the notes on the web or pass them on to others by any means, whether digitally or otherwise. If you find any typos, I would much appreciate you email them to me.

[2] Some key questions to consider: How are costs summarized and reported? What are the costs, and what do they depend on?



Source: TABB Group, "Alternative Investments 2007"

FIGURE 1. Where U.S. domestic equity managers think most of their alpha is lost. (Source: TABB Group.)

- Explicit costs include commissions, bid-ask spreads, taxes, and foreign exchange costs.
- Implicit costs include market impact, and opportunity costs.

While not known before trading, implicit costs can be estimated using t-cost models. We will cover models for this purpose later.

There are many ways in which t-costs can be measured. One common measure is implementation shortfall. We define the *implementation shortfall* (also referred to as *slippage*) of an order, $\mathcal{C}$, as the difference between the arrival price and average execution price of the order (Treynor, 1981; Perold, 1988)

$$\mathcal{C} = \text{arrival price} - \text{average execution price} \tag{1}$$

The *arrival price* is the mid price of the security prevailing at the time the order is submitted, typically truncated to the next open price if submitted outside market hours.

As we can see in Figure 2, trading costs are much different for large (lower) and small cap (higher) stocks as given by implementation shortfall (dark blue and green color) plus commissions (light blue and green). Notice that t-costs vary with time and that the largest part are the implicit costs. This implies that it might be useful for traders to be able to forecast the implicit costs of trades. Continuing this line of thinking, we may also ask if it is possible to decrease t-costs in meaningful ways. We introduce the main financial ideas related to this topic in Section 3. In future lecture notes we will cover the resulting mathematical models and optimization problem.

How does market impact arise? To understand this, we take a look at basic limit order book mechanics next.

2. Limit Order Book Mechanics

This section follows Kolm and Maclin (2010) and Kolm and Maclin (2012). The *limit order book* contains resting limit orders. These orders rest in the book and provide liquidity as they wait to be matched with non-resting orders, which represent a demand for liquidity. The three most common types of non-resting orders are *marketable limit orders*, *market orders*, and *fill-or-kill orders*.

The *bid side* of the limit book contains resting bids to buy a certain number of shares of stock at a certain price. The *offer side* contains resting offers to sell a certain number of shares of stock at a certain price.

A market order is a demand for an immediate execution of a certain number of shares at the best possible price. To get the best possible price, a market order sweeps through one side of the limit order book – starting with the best price – matching against resting orders until the full quantity of the market order is filled or the book is completely depleted.

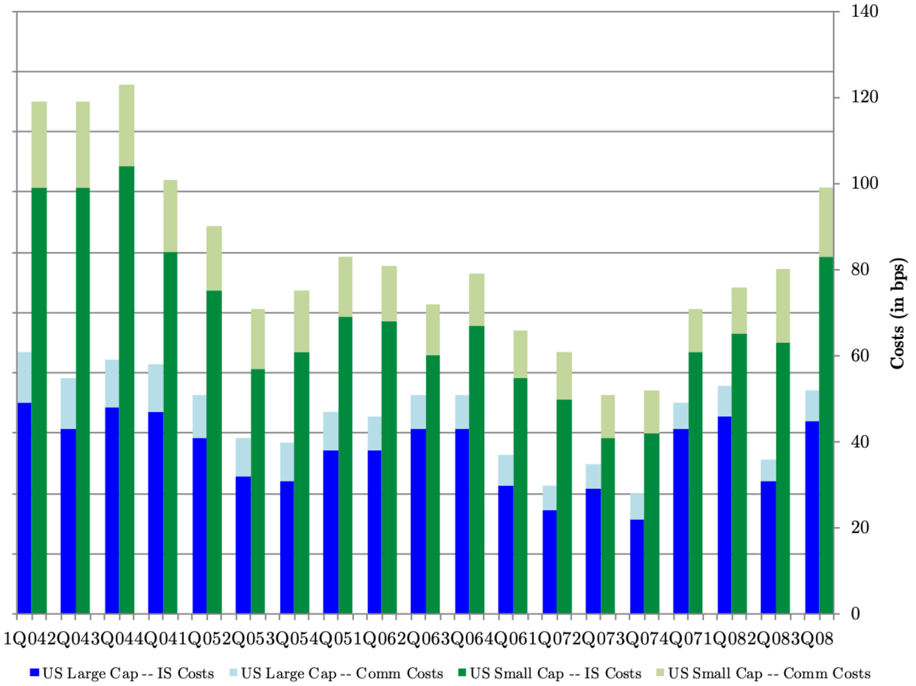


FIGURE 2. Trading costs for large (blue bars) and small cap (green) stocks as given by implementation shortfall (dark blue and green) plus commissions (light blue and green), where implementation shortfall is defined as the sum of timing delay costs and market impact costs. The time period covers 2004Q1–2008Q3. (Source: ITG.)

Unlike a market order, a marketable limit order can be executed only at a specified price or better. For example, a marketable limit order to buy 100 shares at $90.01 can match with a resting limit order to sell 200 shares at $90.00. The trade print – the price at which the trade would take place – would be $90.00.

The following examples illustrate how market orders to sell interact with resting limit orders to buy.

Figure 3 shows the idealized market impact of a two hundred share market order to sell. The x- and y-axes display the time and price, respectively.

The bid side of the limit order book contains bids to buy a certain number of shares of stock at a certain price. Resting limit orders – orders that sit in the order book – are said to *provide liquidity* by

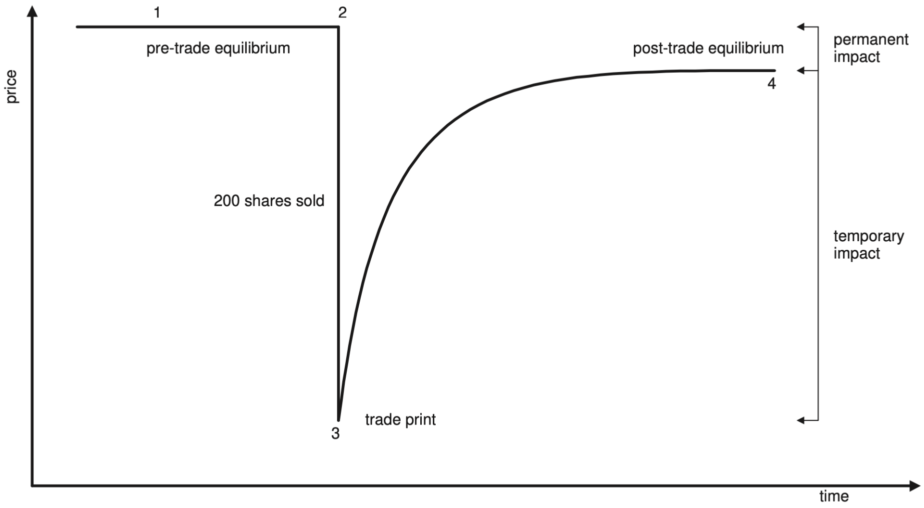


FIGURE 3. The idealized market impact from selling 200 shares in one trade.

mitigating the market impact of orders that must be filled immediately. The state of the book establishes a *pre-trade equilibrium* (1), which is disturbed by a market order to sell 200 shares (2). Market orders must be filled immediately, and therefore represent a demand for liquidity.

As the sell order depletes the bid book by matching with limit orders to buy, it obtains an increasingly less favorable (lower) trade price, resulting in the *trade print* (3). Assuming no other trading activity, over time liquidity providers replenish the bid book to (4), which is the *post-trade equilibrium.*

The difference between (4) and (1) is an information-based effect called *permanent market impact.* It is the market's response to information that a market participant has decided not to own 200 shares of this stock. This effect is typically modeled as immediate and linear in total number of shares executed. Huberman and Stanzl (2004) show that, if the effect were not linear and immediate, buying and selling at two different rates could produce an arbitrage profit.

The difference between (4) and (3) is called *temporary market impact.* The trader who initiated the trade is willing to obtain a less favorable *fill price* (3) to get their trade done immediately. This cost

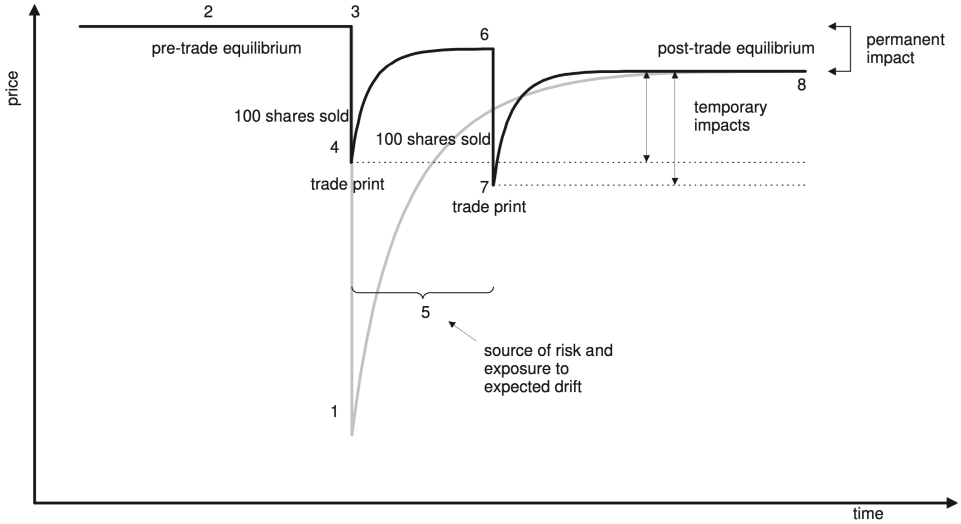


FIGURE 4. The idealized market impact from selling 200 shares in two trades.

of immediacy is typically modeled as a nonlinear function. For example, under the assumption of square root impact, with all other factors held constant, a trade of 200 shares executed over the same period of time as a trade of 100 shares would have square root of two times more temporary impact per share.

Figure 4 shows what would happen if the same trader were willing to wait some time between trades. The trade print from the previous figure is shown as a reference point (1). As in Figure 1, a pre-trade equilibrium (2) is disturbed by a 100 share market order to sell (3). As the market order depletes the bid book by matching with limit orders to buy, it obtains a fill price (4). Over time (5), liquidity providers refill the bid book with limit orders to buy. But the new post-trade equilibrium (6) is lower than the pre-trade equilibrium because it incorporates the information of the executed market order.

Our trader then places another market sell order for 100 shares (6) and obtains a trade print (7). Over time the temporary impact – (8) minus (7) – decays and results in a new post trade equilibrium (8). As the permanent impact is assumed to be linear and immediate, the post-trade equilibrium is shown to be the same for one order of 200 shares as it is for two orders of 100 shares each.

3. Introduction To Optimal Execution

While our trader waits between trades (5), they incur *price risk* – the risk that their execution will be less favorable due to the random movement of prices. In this context, a *shortfall* is the difference between the *effective execution price* and the *arrival price* – the prevailing price at the start of the execution period. If we use the variance of shortfalls as a proxy for risk, a trader's aversion to risk establishes a risk/cost trade-off. In the first scenario, they pay a higher cost – the difference between (8) and (1) – to eliminate risk. In the second scenario, they pay a lower cost – the average of the differences between (8) and (4), and (8) and (7) – but takes on a greater dispersion of shortfalls associated with the waiting time between trades (5). This is the trade-off considered in the seminal paper of Almgren and Chriss (2001), the subject of the next lecture notes.

Risk aversion increases a trader's sense of urgency and makes it attractive to pay some premium to reduce risk. The premium the trader pays is in the form of higher temporary market impact. All other factors held constant, a higher expected temporary market impact encourages slower trading, while a higher expected risk or risk aversion encourages faster trading.

Risk aversion embodies the notion that people dislike risk. For a risk-averse agent, the utility of a fair game, $u(G)$, is less than the utility of having the expected value of the game, $\mathbb{E}[u(G)]$, with certainty, where $\mathbb{E}$ denotes the expectation. The degree of risk aversion may be captured by the risk aversion parameter $\lambda \geq 0$, which is used to translate risk into a *certain dollar cost equivalent* – the smallest certain dollar amount that would be accepted instead of the uncertain payoff from the fair game. For an agent with quadratic utility the certain dollar cost equivalent is given by $\mathbb{E}[G] - \lambda \text{var}[G]$, where var denotes the variance. Hence, their degree of risk aversion is characterized by the family of risk/return pairs with the same constant trade-off between

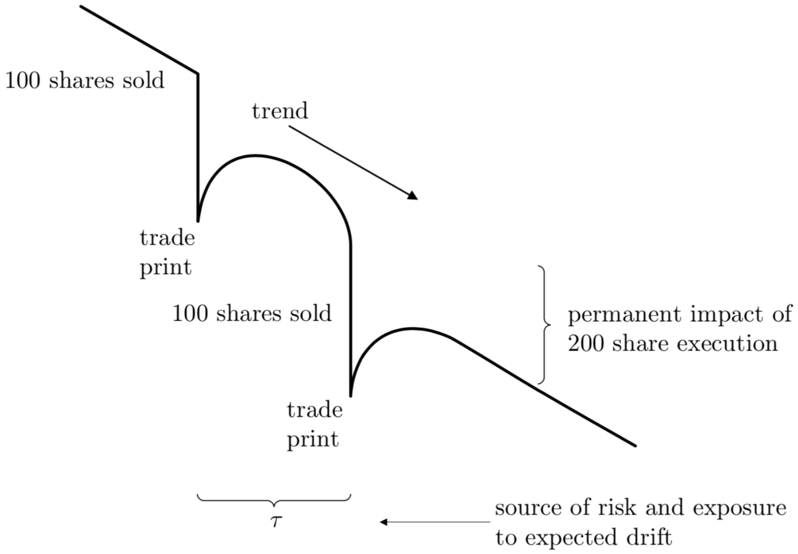


FIGURE 5. The idealized market impact from selling 200 shares in two trades with alpha.

expected return and risk. An annualized target return and standard deviation imply a risk aversion, and may be translated to a risk aversion parameter of the type used in some optimal execution algorithms.

Another factor that influences the decision to trade more quickly or more slowly is the expectation of price change. For the purpose of execution, a positive alpha is an expectation of profits per share per unit time for unexecuted shares. A faster execution captures more of the profits associated with this expectation of price change. A negative alpha is the expectation of losses per share per unit time for unexecuted shares. A slower execution incurs less of the losses associated with this expectation of price change.

For example, a trader has positive alpha if they expect prices to move lower while they are executing their sell orders. Figure 5 depicts the idealized market impact of a two hundred share market order to sell subject to alpha.

They may choose to front-weight their trade schedule – execute more rapidly at the beginning of the execution period – to obtain better execution prices. Similarly, a seller who believes that prices are moving higher may back-weight their trade schedule or delay the execution. The general form of the optimal execution problem is finding the best trade-off between the effects of risk, market impact, and alpha by minimizing risk-adjusted costs relative to a pre-specified benchmark. Common benchmarks are VWAP (volume weighted average price) and arrival price (the price prevailing at the beginning of the execution period).

The first formulations of this problem go back to the seminal papers of Bertsimas and Lo (1998), and Almgren and Chriss (2001). We will discuss the latter article in great detail in the next lecture notes. Assuming a quadratic utility function, a general formulation of this problem takes the form

$$\min_x \mathbb{E}[\mathcal{C}(x)] + \lambda \text{var}[\mathcal{C}(x)] \tag{2}$$

where $x := \{x_0, \dots, x_T\}$ is the trade schedule and $\mathcal{C}(x)$ is the cost of deviating from the benchmark. The trade schedule represents the number of shares that remains to buy/sell at each point in time. Naturally, the trader's optimal trade schedule is a function of their level of risk aversion (λ) which determines their urgency to trade, and dictates the preferred trade-off between execution cost and risk. When we want to make this dependency explicit, we write $x_\lambda := \{x_{\lambda,0}, \dots, x_{\lambda,T}\}$ instead of just x.

In the following two subsections we discuss the sell- and buy-side perspectives of the typical arrival price optimal execution models.

3.1. The Sell-Side Perspective. The typical optimal execution model uses arrival price as a benchmark and balances the trade-off between market impact, price risk, and opportunity cost. Alpha is assumed to be greater than or equal to zero, which means that delaying execution may carry an associated opportunity cost, but does not carry an expectation of profit. The optimal strategy lies somewhere between two extremes: (1) trade everything immediately at a known cost, or (2) reduce market impact by spreading the order into smaller trades over a longer horizon at the expense of increased price risk and opportunity cost.

Bertsimas and Lo (1998) proposed an algorithm for the optimal execution problem that finds the minimum expected cost of trading over a fixed period of time for a risk neutral trader, $\lambda = 0$, facing an environment where price movements are assumed to be serially uncorrelated.

Almgren and Chriss (2001) extended this concept using quadratic utility to embody the trade-off between expected cost and price risk. The more aggressive (passive) trade schedules incur higher (lower) market impact costs and lower (higher) price risk. Similar to classical portfolio theory, as λ varies the resulting set of points $(\mathrm{var}(x_\lambda), \mathbb{E}(x_\lambda))$ traces out the *efficient frontier of optimal trading strategies*. The two extreme cases $\lambda = 0$ and $\lambda \to \infty$ correspond to the *minimum impact strategy* – trading at a constant rate throughout the execution period – and the *minimum variance strategy* – a single execution of the entire target quantity at the start of the execution period.

Let us consider selling X shares, that is we want $x_0 = X$ and $x_T = 0$. Under the assumptions that asset prices follow an arithmetic Brownian motion, permanent impact is immediate and linear in total shares executed, and temporary impact is linear in the rate of trading, the solution of the Almgren and Chriss model is

$$x_t = X \frac{\sinh(\kappa(T-t))}{\sinh(\kappa T)} \tag{3}$$

where $\kappa = \sqrt{\frac{\lambda\sigma^2}{\eta}}$. Here σ and η represent stock volatility and linear temporary market impact cost.

Loosely speaking, the solution is effectively a decaying exponential $X\exp(-\kappa t)$ "adjusted" such that $x_T = 0$. It does not depend on the permanent market impact, consistent with the discussion in the previous section. The urgency of trading is embodied in κ. This parameter determines the speed of liquidation *independent* of the order size X.

For a higher risk aversion parameter or volatility – for example, representing increased perceived risk – the speed of trading increases as well. We also see that for a higher expected temporary market impact cost, the speed of trading decreases.

3.2. The Buy-Side Perspective. Optimal execution algorithms have less value to a typical portfolio manager if analyzed separately from the corresponding returns earned by their trading strategy. In fact, high transaction costs are not bad per se – they could simply prove to be necessary for generating superior returns. At present, the typical sell side perspective of algorithmic trading does not take expectation of profits or the client's portfolio objectives into account. Needless to say, this is an important component of execution.

The decisions of the trader and the portfolio manager are based on different objectives. The trader decides on the timing of the execution, breaking large parent orders into a series of child orders that, when executed over time, represent the correct trade-off between opportunity cost, market impact, and risk. The trader sees only the trading assets, whereas the portfolio manager sees the entire portfolio, which includes both, the trading assets and the static – non-trading – positions.

The portfolio manager's task is to construct a portfolio by optimizing the trade-off between opportunity cost, market impact, and risk for the full set of trading and non-trading assets. In general, the optimal execution framework described by Almgren and Chriss is not appropriate for the portfolio manager.

Engle and Ferstenberg (2007) proposed a framework that unites these objectives by combining optimal execution and classical mean-variance optimization models. They demonstrate that to correctly measure risk we must take both existing positions and unexecuted shares into account. This idea unites execution risk with portfolio risk. Portfolio construction and optimal execution are similarly united by incorporating market impact costs directly into the portfolio construction process.

Ideally, the portfolio manager would like to solve a problem similar in nature to the multi-period consumption-investment problem (see, Merton (1969)), that in addition takes market impact costs and changing probability distributions for a large universe of securities into account. This dynamic portfolio or small delta continuous trading problem represents the next step in the evolution of institutional money management. However, it presents some mathematical and computational challenges. As has been pointed out by Sneddon (2005), dynamic portfolio problem differs in several important ways from the classical multi-period consumption-investment problem. First, the return probability distributions change throughout time. Second, the objective functions for active portfolio management do not depend on predicted alpha/risk, but rather on realized return/risk. Finally, the dynamics of the model may be far more complex. Grinold (2007) and Gârleanu and Pedersen (2013) provide elegant and analytically tractable models. However, these models are hard to extend to real-world situations where impact costs are nonlinear and the portfolio might be subject to various constraints. To address these shortcomings, Kolm and Ritter (2015) propose a general computational framework based on nonlinear filtering techniques that has been adopted by a number of market practitioners. We will cover models of this kind in future lectures.

References

Almgren, Robert and Neil Chriss (2001). "Optimal Execution Of Portfolio Rransactions". In: *Journal of Risk* 3, pp. 5–40.

Bertsimas, Dimitris and Andrew W Lo (1998). "Optimal Control Of Execution Costs". In: *Journal of Financial Markets* 1.1, pp. 1–50.

Engle, R. and R. Ferstenberg (2007). "Execution Risk: Its The Same As Investment Risk". In: *Journal of Portfolio Management* 33.2, pp. 34–44.

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.

Grinold, Richard (2007). "Dynamic Portfolio Analysis". In: *Journal of Portfolio Management* 34.1, pp. 12–26.

Huberman, Gur and Werner Stanzl (2004). "Price Manipulation And Quasi-Arbitrage". In: *Econometrica* 72.4, pp. 1247–1275.

Kolm, Petter N. and Lee Maclin (2010). "Algorithmic Trading". In: *Encyclopedia of Quantitative Finance.*

— (2012). "Algorithmic Trading, Optimal Execution, and Dyna Mic Port Folios". In: *The Oxford Handbook of Quantitative Asset Management.*

Kolm, Petter N. and Gordon Ritter (Mar. 2015). "Multiperiod Portfolio Selection And Bayesian Dynamic Models". In: *Risk* 28.3, pp. 50–54.

Merton, Robert C. (1969). "Lifetime Portfolio Selection Under Uncertainty: The Continuous-Time Case". In: *The Review of Economics and Statistics*, pp. 247–257.

Perold, Andre F. (1988). "The Implementation Shortfall: Paper Versus Reality". In: *Journal of Portfolio Management* 14.3, pp. 4–9.

Sneddon, Leigh (2005). “The Dynamics Of Active Portfolios”. In: *Proceedings of Northfield Research Conference*. Westpeak Global Advisors.

Treynor, Jack L. (1981). “What Does It Take To Win The Trading Game?” In: *Financial Analysts Journal* 37.1, pp. 55–60.