Implementation Shortfall with Transitory Price Effects

Terrence Hendershott, Charles M. Jones, and Albert J. Menkveld

March 18, 2013

Regulators and some large investors have recently raised concerns about temporary or transitory volatility in highly automated financial markets.[1] It is far from clear that high-frequency trading, fragmentation, and automation are contributing to transitory volatility, but some institutions complain that their execution costs are increasing. In this chapter, we introduce a methodology for decomposing the price process of a financial instrument into its permanent and transitory components, and we explore the insights from applying this methodology to execution cost measurement. Among other things, our methodology allows an institutional investor to accurately measure the contributions of transitory price movements to its overall trading costs. The methodology is particularly applicable to an investor that splits a large order into small pieces and executes it gradually over time.

The importance of transitory price impact has been well-known in the academic literature since early work on block trading (e.g., Kraus and Stoll (1972)).[2] While it is fairly straightforward to measure the transitory price impact of a block trade, it is a much greater challenge to measure the transitory price impact when a large institutional parent order is executed in perhaps hundreds of smaller child order executions. The key innovation of our approach is that we estimate the temporary component at each point in time, and in particular whenever a child order executes. By summing over all child orders, we can thus measure the effect of the temporary component on overall trading costs.

To be more precise, we extend the classic Perold (1988) "implementation

[1] See for example the US Securities and Exchange Commission's 2010 concept release on equity market structure (Release No. 34-61358).

[2] See Duffie (2010) for an extensive discussion of temporary price impacts from large informationless demands for liquidity.

shortfall" approach to decompose ex-post transaction costs into various components, one of which accounts for the trading costs associated with transitory pricing errors. Because trading cost analysis is often performed on an institution's daily trading, we first illustrate our transaction cost measurement approach at a daily frequency. However, our methods are much more precise when more disaggregated trading data are available. Using detailed information on the intraday child order executions from a larger institutional parent order, we show how the transitory price component evolves with trading on a minute-by-minute basis, and we show how this transitory price component contributes to overall implementation shortfall.

In some ways, our work is most closely related to Almgren et al. (2005), who assume a particular functional form for both permanent and transitory price impacts, with limited persistence in the latter. They then apply their model to a large set of institutional orders to characterize permanent and transitory components of transaction costs as a function of various stock and order characteristics.[3] In contrast, we allow the data to determine the persistence of the temporary component.

1 Implementation Shortfall

Even for those who are intimately familiar with trading cost analysis, Perold (1988) is worth a re-read. For example, he frames the discussion on p.4:

> After selecting which stocks to buy and which to sell, "all" you have to do is implement your decisions. If you had the luxury of transacting on paper, your job would already be done. On paper, transactions occur by mere stroke of the pen. You can transact at all times in unlimited quantities with no price impact and free of all commissions. There are no doubts as to whether and at what price your order will be filled. If you could transact on paper, you would always be invested in your ideal portfolio.
>
> There are crucial differences between transacting on paper and transacting in real markets. You do not know the prices at which you will be able to execute, when you will be able to execute, or even whether you will ever be able to execute. You do not know whether you will

[3] Engle and Ferstenberg (2007) also estimate implementation shortfall costs on a sample of institutional orders, focusing on the variance of the execution costs as well as their mean.

> be front-run by others. And you do not know whether having your limit order filled is a blessing or a curse - a blessing if you have just extracted a premium for supplying liquidity, a curse if you have just been bagged by someone who knows more than you do. Because you are so much in the dark, you proceed carefully, and strategically.

These comments are just as apt in 2013 as they were in 1988, except that in 2013 the concern about front-running is mainly a worry about being sniffed out by algorithmic traders. Some algorithms use sophisticated forecasting and pattern recognition techniques to predict future order flow and thus future price changes. To the extent that the slicing and dicing of large institutional orders into many smaller trades leaves a footprint in the data, algorithms may attempt to identify and trade ahead of these large institutional orders. Any such order anticipation could increase the transitory impact of a large order and thereby increase its overall cost.

With a few notational changes we follow Perold (1988, Appendix B)'s methodology for measurement and analysis of implementation shortfall. At the beginning of a measurement period, the paper portfolio is assumed to be worth the same amount as the real portfolio. At the end of the period, any differences in value capture the implementation shortfall. In general the length of the measurement period is not important. For many institutions, one day is the preferred period length, but it can be longer or shorter. The key constraint is that if implementation shortfall is to be measured for an order that is executed gradually over time, the measurement period must span the time over which the order is executed.

Assume there are N securities with one being cash. Let n_i denote the number of shares of security i in the paper portfolio, ω_i^b be the number of shares of security i in the real portfolio at the beginning of the period, and ω_i^e be the number of shares held at the end of the period. ω_i^e differs from ω_i^b by the shares traded in security i.

Denote the time of trades by $j = 1, \ldots, K$. Denote the number of shares traded in security i at time j by t_{ij}; t_{ij} is positive for buys, negative for sales, and zero when there is no trade. Therefore, the ending shareholding in security i is

$$\omega_i^e = \omega_i^b + \sum_{j=1}^{K} t_{ij}. \tag{1}$$

Denote the prices at which transactions take place by p_{ij}; p_{ij} are net of incremental costs such as commissions and transfer taxes. Let the price of security i be p_i^b at the beginning of the period and p_i^e at the end. While the p_{ij} must be transaction prices, the two benchmark prices can be either trade prices or quote midpoints.

Assuming there are no net cash flows into or out of the real portfolio, all transactions are financed with proceeds of other transactions. That is, at each time j, $\sum t_{ij} p_{ij}$ is zero when summed over $i = 1$ to N.

Let the value of the paper and real portfolios at the beginning of the period be V_b:

$$V_b = \sum n_i p_i^b. \tag{2}$$

Let the end-of-period values of the real and paper portfolios be V_p and V_r, respectively:

$$V_p = \sum n_i p_i^e \text{ and } V_r = \sum \omega_i^e p_i^e. \tag{3}$$

The performance of the paper portfolio is $V_p - V_b$, and the performance of the real portfolio is $V_r - V_b$. The implementation shortfall is the difference between the two.

The performance of the real portfolio can be expanded as

$$\begin{aligned}\sum(\omega_i^e p_i^e - \omega_i^b p_i^b) &= \sum \omega_i^e (p_i^e - p_i^b) - \sum p_i^b (\omega_i^e - \omega_i^b) \\ &= \sum \omega_i^e (p_i^e - p_i^b) - \sum\sum (p_{ij} - p_i^b) t_{ij}. \qquad (4)\end{aligned}$$

The performance of the paper portfolio can be expanded as

$$\sum n_i (p_i^e - p_i^b). \tag{5}$$

Subtracting the real portfolio performance from paper portfolio performance completes the calculation:

$$\text{Impl. Shortfall} = \underbrace{\sum\sum (p_{ij} - p_i^b) t_{ij}}_{\text{Execution Cost}} + \underbrace{\sum\sum (p_i^e - p_i^b)(n_i - \omega_i^e)}_{\text{Opportunity Cost}}. \tag{6}$$

The term $(p_{ij} - p_i^b)$ is the per-share cost of transacting at p_{ij} instead of at p_i^b, and this cost is applied to t_{ij} traded shares. The weighted sum is the total execution cost relative to the pre-trade benchmark. The term $(p_i^e - p_i^b)$ is the paper return on security i over the period. The opportunity cost is the sum of these returns weighted by the size of the unexecuted orders. While opportunity costs are a real concern for institutional investors, our methodology does not offer much insight into them, and in the rest of the chapter we focus only on the execution cost component.

2 Observed prices, efficient prices, and pricing errors

The implementation shortfall incorporates the total price impact of a large order. However, to better understand the sources of the shortfall, it may be useful to decompose the price impact into its permanent and transitory components. To do this one must define and measure the efficient price and any deviations from it at each moment in time. We take the standard approach of assuming the efficient price is unpredictable, i.e., it follows a random walk.

Absent trading frictions, the efficient price at the daily or intraday frequency can be characterized as a martingale process. Let m_j be this latent price:

$$m_j = m_{j-1} + w_t. \tag{7}$$

Sometimes the quote midpoint is assumed to represent this latent price. However, quote midpoints are not generally martingales with respect to all available order flow, in which case Hasbrouck (1995, p.1179) proposes to view the random-walk component of a Stock and Watson (1988) decomposition as the "implicit efficient price." Hasbrouck (2007, Ch.4 and Ch.8) constructs an efficient price more generally as the projection of m_t onto all available conditioning variables, i.e., the so-called filtered state estimate:

$$\tilde{m}_{ij} = E^* \left[m_j | p_{ij}, p_{i,j-1}, \ldots \right], \tag{8}$$

where E^* [.] is the linear projection of m_{ij} on a set of lagged prices.[4] A standard approach to implementing such a projection is through ARIMA time series econometrics (Hasbrouck (2007, Ch.4)). The filtered estimate can be enriched by expanding the set of conditioning variables with trade-based variables (e.g., signed order flow), news-based variables (e.g., the Reuters sentiment score of

[4] The observed p_{ij} in this section can be either trade prices or quote midpoints. In this chapter we always use midquotes.

press releases), etc.[5]

A more general approach constructs the "efficient price" based on a state-space model. This nests the ARIMA approach but has the following advantages. First, it allows for not only using past information to estimate the efficient state, but also future information. This is particularly relevant in decomposing a price change into a permanent price change (i.e., the efficient price change) and a (transitory) pricing error. For a particular in-sample price change, one does in fact want to 'peek into the future' to establish whether it was largely permanent or transitory. A state-space model produces, in addition to a filtered price estimate, a so-called smoothed price estimate that also takes future price information into account, i.e.,

$$\hat{m}_{ij} = E^* \left[m_j | \ldots, p_{i,j+1}, p_{ij}, p_{i,j-1}, \ldots \right]. \tag{9}$$

Second, the state-space approach extends naturally to multi-market trading where there are potentially multiple price quotes for the same security at any instant of time. It also accounts optimally for missing observations that arise, for example, when various markets do not perfectly overlap. Third, structural models often generate a system of equations in state-space form. This system can then be taken to the data without further (potentially imperfect) transformations. Further discussion of the approach and implementation details are in Menkveld, Koopman, and Lucas (2007).

The efficient price estimate enables one to decompose an observed price into a (smoothed) efficient price and a pricing error:

$$p_{ij} = \hat{m}_{ij} + s_{ij}. \tag{10}$$

Hereafter, the focus is mainly on the smoothed price estimate (as opposed to the filtered estimate), as implementation shortfall is about ex-post evaluation and therefore 'future' price information is available and relevant.[6]

Let us reconsider part of the quote of Perold (1988):

> And you do not know whether having your limit order filled is a

[5] Non-public information can also be incorporated into the estimation. See Hendershott and Menkveld (2011) for an application using NYSE market-maker inventory data, Menkveld (2011) using data from a high-frequency trading firm's inventory positions, and Brogaard, Hendershott, and Riordan (2012) using data on the aggregate trading of 26 high-frequency trading firms.

[6] Filtered price estimates are more natural in case of real-time trade decisions that necessarily only have historical information available.

> blessing or a curse - a blessing if you have just extracted a premium for supplying liquidity, a curse if you have just been bagged by someone who knows more than you do.

The efficient price estimate enables one to further refine the standard implementation shortfall calculation of equation (6) by recognizing the size of these two components. The execution cost component of the implementation shortfall can be rewritten as:

$$\begin{aligned}\text{Execution Cost} = &\underbrace{\sum\sum(p_{ij} - \hat{m}_{ij})t_{ij}}_{\text{Liquidity Cost}} + \underbrace{\sum\sum(\hat{m}_{ij} - \hat{m}_i^b)t_{ij}}_{\text{Informational Cost}} + \\ &\underbrace{\sum\sum(\hat{m}_i^b - p_i^b)t_{ij}}_{\text{Timing Cost}}\end{aligned} \tag{11}$$

The first component captures liquidity cost relative to the efficient price. If one buys at a price above the efficient price, one effectively pays a liquidity premium, and if one buys at a lower price one earns the premium. The liquidity costs incorporate both the bid-ask spread and any transitory price effects. For example, if a sequence of trades causes the current quoted price to differ from the efficient price, this temporary price impact is captured in the liquidity cost component.

This differs from the standard approach to measuring temporary price impact, which compares the price immediately after execution to a price some time later. In the standard approach, the temporary impact reflects the correlation between the direction of the order and subsequent price movements. For example, there is temporary impact if prices fall after the completion of a large buy order. The state-space approach captures this general idea, as it incorporates future price movements to estimate the permanent and temporary price decomposition. However, the main advantage of the state-space approach is that it calculates efficient prices throughout the execution period. The temporary component can be measured and incorporated into the liquidity cost component for each of the N executions. In contrast, the standard approach can only measure the temporary price impact at the end of the execution period based on its dissipation thereafter.

The second component of the implementation shortfall captures the informational cost, as it measures the covariation between executed signed order flow and the efficient price change. This is sometimes referred to as the permanent price impact of the trades. If for some reason signed flow does not correlate with efficient price changes, then the informational cost is zero. In most financial markets, however, the order flow is potentially informationally motivated, so this component is positive on average. For example, in a classic market-making model a liquidity supplier cannot distinguish informed from uninformed flow and therefore charges all incoming flow the same price impact (see, e.g., Glosten and Milgrom (1985)). In reality, a small informational cost component could reflect the skill of a trader or algorithm in camouflaging the order and having it perceived as uninformed. This component can also reflect variation in the information environment over time. For example, informational costs may be greater just before scheduled earnings announcements.

The third component measures whether the timing of the trade is correlated with the temporary component. If the parent order is a buy, for example, then starting it when the quote midpoint is above the efficient price increases the overall cost of the trade, all else equal. Conversely, starting a buy order when the price is below the efficient price should improve its overall execution. We capture this by assigning a negative timing cost when a trade begins in these favorable conditions.

3 Illustration of Approach

Decomposing the price process into its permanent and transitory components is fundamental to our approach. Hasbrouck (2007, Ch.8) provides a detailed discussion of the challenges in identifying the two components. Here we follow an approach developed for analyzing cyclical macroeconomic time series. This approach puts enough structure on the persistence of the transitory price component to identify the two components. Morley, Nelson, and Zivot (2003, p.240) show that the most parsimonious allowable specification for the temporary component is an AR(2):

> "...the order condition for identification of the unrestricted UC-ARMA(p,q) model, in the sense of having at least as many moment equations as parameters, is $p > 0$, $p > q + 2$, and it is just satisfied with $p = 2$, $q = 0$."

In the state space representation, the observation equation is:

$$p_t = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} m_t \\ s_t \\ s_{t-1} \end{bmatrix}. \tag{12}$$

The state equation is:

$$\begin{bmatrix} m_t \\ s_t \\ s_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \varphi_1 & \varphi_2 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} m_{t-1} \\ s_{t-1} \\ s_{t-2} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_t \\ \varepsilon_t \end{bmatrix}, \tag{13}$$

where the variance-covariance matrix of state innovations is:

$$\Omega = \begin{bmatrix} \sigma_w^2 & \rho\sigma_w\sigma_\varepsilon & 0 \\ \rho\sigma_w\sigma_\varepsilon & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{14}$$

The unknown parameters in the state space model are $(\sigma_w, \sigma_\varepsilon, \rho, \varphi_1, \varphi_2)$. The observed price can be net of any market or industry sector movements. This is appropriate and efficient if trading occurs only in individual securities. Controlling for market and other factor movements is more complicated if the trading is part of a larger portfolio transaction which could possibly impact market or factor prices.

As discussed above, additional information can be utilized in decomposing the price process into its permanent and transitory components. The most common approach is to add additional state variables reflecting publicly available order flow information, such as buy and sell liquidity demand or the imbalance between the two. Brogaard, Hendershott, and Riordan (2012) extend this approach in the state-space context by using nonpublic information from NASDAQ on whether or not the liquidity demander in each trade is a high-frequency proprietary trading firm. Hendershott and Menkveld (2011) use NYSE market-maker inventory data, and Menkveld (2011) uses data from a high-frequency trading firm's inventory positions. The amount of data that can be potentially incorporated into the estimation is enormous. For example, all orders, trades, and news in every related market and security could be utilized. For parsimony in our examples we only use past prices, in one case adjusted for an industry factor.

3.1 Implementation shortfall calculations

To illustrate our approach, we use two different examples with trading data observed at different frequencies: one example with daily trading data, and one example of a parent order where we observe the size, time, and price of the individual child order executions during the trading day. In the daily example, we have two months' worth of trades by the same fund in the same stock, aggregated at the daily level, and we estimate the efficient and transitory price components at a daily frequency. This approach is most relevant to investors that make each day's trading decisions overnight while the market is closed, because in that trading environment implementation shortfall is naturally calculated relative to the previous closing price.

It is worth noting that the decomposition in (11) requires the efficient price estimate at the time of each transaction, $\hat{m}_{ij}$. In the daily example, however, we only calculate end-of-day efficient price estimates because we do not know when the trades actually take place during the day. This timing mismatch reduces the precision of the implementation shortfall decomposition and may also introduce bias. The main issue is the allocation of the shortfall between the first two terms of (11), the liquidity and information costs. These two components can be thought of as corresponding to temporary and permanent price impacts, respectively. If there is positive correlation between the direction of trading and the movement in the efficient price, then using an estimate of the efficient price prior to transaction j will overestimate the liquidity cost and underestimate the information cost. Conversely, using an estimate of the efficient price after transaction j will underestimate the liquidity cost and overestimate the information cost. If only coarse execution data are available and temporary components are sufficiently persistent, however, the decomposition may still prove useful.

For the intraday example, we obtain an efficient price estimate for each minute of the trading day. We use these efficient price estimates to evaluate the execution of a single parent order that is gradually executed over the course of about 30 minutes. The intraday horizon allows for an evaluation of the high-frequency price dynamics during order execution.

To calculate our implementation shortfall decomposition we use equation (11) with the prices at time j modified as follows:

1. the subscript i is dropped as there is only one security;

2. p_j is the average price at which the institution's trades execute at time j;

3. p^b is the quote midpoint prior to beginning execution;

4. $\hat{m}_j$ is the estimate of the efficient price at time j;

5. $\hat{m}^b$ is the estimate of the efficient price prior to beginning execution.

Using these prices the per-share execution costs can be represented as:

$$\begin{aligned}&\Big(\sum \underbrace{(p_j - \hat{m}_j)\text{sign}(t_j)}_{\text{Liquidity Cost } j} |t_j| + \sum \underbrace{(\hat{m}_j - \hat{m}^b)\text{sign}(t_j)}_{\text{Informational Cost } j} |t_j| + \\ &\sum \underbrace{(\hat{m}^b - p^b)\text{sign}(t_j)}_{\text{Timing Cost } j} |t_j|\Big) \quad / \quad \Big(\sum |t_j|\Big)\end{aligned} \tag{15}$$

3.2 Daily estimation

For our first example, the execution data are from a long-short equity hedge fund with approximately \$150 million in assets under management and an average holding period of about one month. For each stock traded by this fund, we know the total number of shares bought and sold each day along with the weighted average execution price. In this case, we do not have information on individual intraday trade executions. This is the standard granularity for institutional trading cost analysis, because this information along with a pre-trade benchmark price (such as the previous closing price, the opening price on the day of execution, or the price at the time the order is released) is sufficient to measure implementation shortfall.

The chosen example is for AEC, which is the ticker symbol for Associated Estates Realty Corporation, a real estate investment trust (REIT) listed on the New York Stock Exchange with a market cap of around \$650 million during the sample period. We examine the fund's trading in AEC during November and December of 2010. The fund traded a total of 559,356 shares of AEC during this time period on 20 separate trading days. The stock has an average daily volume of roughly 460,000 shares over these two months, so the analyzed trades constitute about 2.8 percent of the total trading volume in AEC during this interval.

The implementation shortfall decomposition is illustrated based on daily data and one investor's trades in a single security. The index j runs over days and the price snapshot is taken at the end-of-day (closing) price, i.e., the bid-ask midpoint at the end of the trading day.

The state space model characterized in (12)-(14) is estimated on daily data from Jan 4, 2010 through Sep 28, 2012. We use daily closing stock prices to calculate excess returns over the MSCI US REIT index, which is commonly referred to by its ticker symbol RMZ. To be precise, the observed p_t is the log closing price of AEC adjusted for dividends less the log RMZ index level. The parameter estimates are:

parameter	estimate	description
σ_w	91 basis points	stdev efficient price innovation
σ_ε	44 basis points	stdev pricing error residual
ρ	0.38	corr(w, ε)
φ_1	0.65	AR1 coefficient pricing error
φ_2	-0.05	AR2 coefficient pricing error

Figure 1 illustrates the estimates by plotting the observed end-of-day (closing) midquote, the efficient price estimate, and the investor's trades each day for the trading period November 2, 2010 through December 31, 2010. Because the pricing error follows a somewhat persistent AR(2) process, the daily pricing error innovation of 44 basis points scales up to a 71 basis point standard deviation for the pricing error itself. This means that the typical temporary component is estimated to account for 11 cents on this $15 stock. This is roughly five times the typical bid-ask spread for this stock over our sample period. The temporary component is of the same order of magnitude as the standard deviation of daily innovations on the efficient price (91 basis points).

Figure 1 here.

Based on the resulting estimates of efficient prices, the total implementation shortfall of 51.6 basis points can be decomposed as follows:

Avg Liquidity Cost	7.0 basis points
Avg Information Cost	65.1 basis points
Avg Timing Cost	-20.5 basis points
Avg Total Cost	51.6 basis points

The negative timing cost component of -20.5 basis points measures the contribution to fund performance from following a mean-reversion trading strategy that takes advantage of temporary pricing errors. The other notable quantity is the liquidity cost component, which is a modest 7.0 basis points. Recall that when the model is implemented at the daily horizon, the liquidity cost component measures the average difference between execution prices and the post-trade efficient price at the close. The gap between trade time and measurement of the efficient price argues against making direct use of the numbers as estimates of the cost of temporary price moves when the price decomposition is performed at the daily horizon. Instead, we advocate using this component on a relative basis to compare executions across brokers, across stocks, and over time.

To illustrate the breakdown of execution costs across days, Figure 2 plots the size of the total costs and each of its components for each day's trading. As in Figure 1, the size of the dot is proportional to the amount traded ($(|t_j|)$).[7]

Figure 2 here.

As is often the case with execution cost measurement, there is substantial variation in the costs. Daily implementation shortfalls in this case are between -2.5 and 3.3%. The total costs are highest in the beginning of the sample, especially for the first few large orders, suggesting that the fund quickly became aware of its price impact and subsequently traded in smaller sizes. For these first few large orders, the timing costs are negative, indicating that these orders began when prices were relatively attractive, but the large informational costs quickly swamped the timing benefit. Because we are using an end-of-day post-trade efficient price estimate to split the price impact into liquidity (temporary) and informational (temporary) components, we do not want to overinterpret this part of the decomposition. However, because it is a post-trade price, our liquidity component bears a strong resemblance to the traditional measure of the temporary component discussed earlier. In fact, some traders regularly measure trading costs against a post-trade price. Our innovation is to gain additional insight by using a post-trade efficient price from the state space model rather than use a closing quote or closing auction price.

3.2.1 Re-calculation based on filtered estimates

It is also possible to decompose the implementation shortfall using filtered estimates of the efficient price instead of smoothed estimates by substituting $\tilde{m}_j$

[7] On most days the fund traded in only one direction. However, on three days the fund bought and sold shares. On those days, only the net trade enters the analysis along with the average price across all trades that day. For example, if the fund bought 35,000 shares at $15 and sold 5,000 shares at $16, then the net trade that day was a buy of 30,000 shares at a price of (35,000*$15-5,000*$16)/30,000 = $14.83.

for $\hat{m}_j$ in equation (15). The filtered estimates yield:

Avg Liquidity Cost	10.4 basis points
Avg Information Cost	69.4 basis points
Avg Timing Cost	-28.2 basis points
Avg Total Cost	51.6 basis points

Of course, the total implementation shortfall is calculated using observed prices, so it remains unchanged. The timing cost component using filtered estimates is of particular interest, because it has a natural interpretation as the gross short-term alpha conditional on the subset of information included in the model available at the designated pre-trade time (the previous close in this case). Using filtered estimates, the timing cost component for this example is more negative at -28.2 basis points, indicating that an important source of overall return for this investor (or equivalently, an important source of trading cost minimization) is trading against temporary pricing errors.

3.3 Intraday estimation

Our second example uses data from a well-known firm that provides equity transactions cost analysis to institutional clients. We know the size and release time of the parent order, and the size, price, and time of each child order execution. To illustrate the method, we choose one such parent order arbitrarily from a set of recent large orders in less active mid-cap stocks. We also require the order to be executed in one day. The chosen example is a December 13, 2012 sell order in HMST, which is the symbol for Homestreet, Inc., a Nasdaq-listed community bank on the west coast of the U.S. with a market cap of around \$360 million. The sell order is for 6,365 shares, and the stock has an average daily volume of 119,000 shares during December 2012.

The order is released right around 11:00am, and it is fully completed in 50 child order executions over the space of about 30 minutes. During the half hour from 11:00am to 11:30am, total trading volume in this symbol was 34,192 shares, so this client ended up trading 18.6% of the total volume during this interval.[8]

We estimate the state space model using NBBO midpoints at each minute during regular trading hours for 15 trading days from December 1, 2012 through

[8] There was no news released on HMST that day, and during the 11am-11:30am period, the S&P500 fell by 0.2%, compared to a share price drop of about 2% over the same interval in HMST. Thus, it appears that most of the price moves documented here are idiosyncratic.

December 21, 2012.[9] We discard quote midpoints for the first five minutes from 9:30am to 9:35am, as we find that prices right around the open exhibit a different pattern of persistence and are much more volatile. Thus, the state space model is designed to model share price behavior after the beginning of the trading day and, at least in this case, the resulting implementation shortfall decomposition is best applied to trading that avoids the opening 5-minute period.

The parameter estimates from the 1-minute state space model are as follows:

parameter	estimate	description
σ_w	11 basis points	stdev efficient price innovation
σ_ε	6 basis points	stdev pricing error residual
ρ	0.15	corr(w, ε)
φ_1	0.76	AR1 coefficient pricing error
φ_2	0.19	AR2 coefficient pricing error

As noted in the earlier example, the average size of the temporary component is much bigger than the standard deviation of the innovation due to the substantial persistence implied by the AR(2) specification. In this case, the standard deviation of the temporary component innovation is 5.9 basis points, and the temporary component itself has a standard deviation of 51 basis points, or about 12.5 cents on this $25 stock. The AR coefficients imply a slow decay of the temporary component, with an estimated half-life of 14 minutes. As in the earlier example, the correlation between innovations to the two unobserved components continues to be positive, though it is somewhat smaller here. The standard deviation of the random walk component is 11 basis points over the 1-minute interval, which scales up as the square root of t to 216 basis points per trading day.

Combining the smoothed estimates of the efficient price with the child order executions, we obtain the following decomposition of the implementation shortfall:

[9] We also experimented with estimating the state-space model trade by trade rather than in calendar time. We find relatively little persistence in the temporary component when the model is estimated in trade time, most likely because the specification imposes an exponential decay on the temporary component that does not seem to fit the trade-by-trade time series. In addition, the results are very sensitive to how one aggregates trades that are within a few milliseconds of each other but are not exactly simultaneous.

Avg Liquidity Cost	48 basis points
Avg Information Cost	219 basis points
Avg Timing Cost	-5 basis points
Avg Total Cost	262 basis points

The overall implementation shortfall is 262 basis points, and the large information cost component reflects the fact that this order is selling as the estimated efficient price is falling. The negative timing cost component of -5 basis points simply reflects the fact that the sell parent order was released at a time when the observed midpoint was slightly above the estimated efficient price.

Perhaps the most interesting component of our decomposition is the liquidity cost, and it is particularly useful to compare our implementation shortfall decomposition to a more traditional one. Recall that the liquidity cost component measures the average difference between execution prices and the estimated efficient price in effect at the time. While the child orders here execute an average of 48 basis points below the estimated efficient price, the liquidity cost would only be 9 bps if we compare trades to quote midpoints in effect at the time of the child order execution. This is a substantial difference and highlights that the temporary component in prices clearly contributes to the overall trading costs for this order.

Figure 3 illustrates the estimates by plotting the observed end-of-minute NBBO midquote, the efficient price estimate, and the investor's trades each minute. An initial burst of selling coincides with a sharp price decline. We cannot make causal statements, but it is certainly possible that the selling pressure from this parent order caused the price decline. Much of the decline appears to be temporary. The share price bounces back by noon once this order is completed and the selling pressure abates. This armchair empiricism is confirmed by the efficient price estimate, which never moves down as far as the observed quote midpoint and is as much as 14 cents above the midquote during this order execution. The deviation between the observed midquote and efficient price begins to appear as child orders begin to execute. After selling 4,365 shares in the space of five minutes from 11:05 to 11:09 (or 23% of the 19,096 HMST shares that trade in this interval), the transitory component reaches its maximum deviation. Thereafter, execution slows and the transitory component gradually shrinks.

Figure 3 here.

To illustrate the minute-by-minute breakdown of execution costs Figure 4 plots the size of the total costs and each of its components for trades in each minute. As in Figure 3, the size of the dot is proportional to the number of shares filled in each minute. As noted earlier, the efficient price moves down sharply as the first few minutes of selling unfold. This is reflected in the initial upward trend in the informational cost component. The liquidity component increases rapidly from 39 basis points for executions at 11:05am to 64 basis points for the 11:11am fills. Thereafter, the liquidity component generally declines, although the scaling of the graph makes this difficult to see. The timing component is constant at -5 basis points, as this illustration is for a single parent order. Because the informational costs are by far the largest component of the implementation shortfall, the pattern for total costs closely tracks the informational cost component.

Figure 4 here.

4 Conclusion

In this chapter, we decompose a sequence of observed asset prices into a permanent and temporary component. We use this price process decomposition to provide a novel and useful decomposition of the standard implementation shortfall transaction cost measure.

Investors often think in terms of earning the spread, evaluating individual executions vs. the prevailing quote midpoint. Our methodology provides an alternative benchmark. Individual executions should be evaluated against the estimated efficient price, which can be far from the current quote midpoint (a root-mean-squared average of 51 basis points in the case of HMST, our intraday example).

Our methodology also captures the fact that a sequence of trades in the same direction can generate or contribute to a temporary component, and it allows an institutional investor to measure how much its own trading has done so. This seems particularly important in the current automated equity market structure, where transitory price impact may be due to some traders following order anticipation strategies. An institutional investor or algorithm provider can use these empirical techniques to discern whether its algorithms or trading practices minimize these temporary price moves. The empirical examples indicate that the temporary component could be an important contributor to overall trading costs: 48 basis points out of a total of 262 basis points for the intraday example that we study.

We have provided two simple applications of the methodology here. While we only use past prices, we want to reiterate that additional variables can and probably should be added to the filtration. Signed order flow, information on short sales, and position data can all be valuable in determining the latent efficient price.

Finally, our decomposition may be useful in implementing the optimal trading strategy in Gârleanu and Pedersen (2012). They derive an elegant and insightful closed form solution for optimal dynamic execution in the presence of quadratic costs and decaying sources of alpha. Their model draws a distinction between temporary and permanent price impact, and our estimates of the permanent and temporary components of transaction costs can be used to operationalize their results.

Appendix: Implementation details

A useful general reference on state space models (SSM) is Durbin and Koopman (2001). One standard way to the estimate parameters of a state space model is maximum likelihood. The Kalman filter is used to calculate the likelihood given a particular set of parameters.

One standard approach to implement maximum likelihood is to use the Expectation-Maximization (EM) algorithm (see Dempster, Laird, and Rubin, 1977 for EM and Shumway and Stoffer, 1982 for EM and SSM). Its appeal relative to Newton-Raphson type approaches is (i) that it avoids a numerically expensive calculation of the inverse of the matrix of second order partials and (ii) with each step the likelihood is guaranteed to increase. Its relative disadvantage is that convergence is relatively slow in the latter stages. Both approaches, however, could converge to a local maximum. One way to avoid local maxima is to search over a parameter grid.

We use two different estimation methods for the two examples presented in Section 3. The intraday example employs the state space estimation commands in Stata. To investigate robustness, we experimented with different hill-climbing algorithms, starting values, and convergence tolerances. In every case, we end up with the same estimates, suggesting that the likelihood function is well-behaved.

The likelihood optimization for the daily example is implemented in python and uses the pykalman package[10]. The EM algorithm is combined with a parameter grid search for the AR parameters of the pricing error process: φ_1 and φ_2. The choice for a grid on this subset of model parameters is informed by studying convergence based on random sets of starting values. It turns out that the parameters at the optimum are particularly sensitive to starting values of φ_1 and φ_2. Grid search involved a grid over $[-0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8]^2$ and finer grids with step sizes down to 0.05 around the optimum.

References

Almgren, Robert, Chee Thum, Emmanuel Hauptmann, and Hong Li. 2005. "Equity market impact." *Journal of Risk* July:57–62.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2012. "High Frequency Trading and Price Discovery." Manuscript, University of California, Berkeley.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society* 39:1–38.

Duffie, Darrell. 2010. "Presidential Address: Asset Price Dynamics with Slow-Moving Capital." *Journal of Finance* 65:1237–1267.

Durbin, Jim and Siem Jan Koopman. 2001. *Time Series Analysis by State Space Models*. Oxford: Oxford University Press.

Engle, Robert F. and Robert Ferstenberg. 2007. "Execution Risk." *Journal of Portfolio Management* :34–44.

Gârleanu, Nicolae and Lasse H. Pedersen. 2012. "Dynamic Trading with Predictable Returns and Transaction Costs." *Journal of Finance (forthcoming)* .

Glosten, Lawrence and Paul Milgrom. 1985. "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents." *Journal of Financial Economics* 14:71–100.

Hasbrouck, Joel. 1995. "One Security, Many Markets: Determining the Contributions to Price Discovery." *Journal of Finance* 50:1175–1199.

[10] Package documentation is at http://pykalman.github.com/.

———. 2007. *Empirical Market Microstructure*. New York: Oxford University Press.

Hendershott, Terrence and Albert J. Menkveld. 2011. "Price Pressures." Manuscript, VU University Amsterdam.

Kraus, Alan and Hans Stoll. 1972. "Price Impacts of Block Trading on the New York Stock Exchange." *Journal of Finance* 27:569–588.

Menkveld, Albert J. 2011. "High Frequency Trading and the New-Market Makers." Manuscript, VU University Amsterdam.

Menkveld, Albert J., Siem Jan Koopman, and André Lucas. 2007. "Modelling Round-the-Clock Price Discovery for Cross-Listed Stocks using State Space Methods." *Journal of Business & Economic Statistics* 25:213–225.

Morley, James C., Charles R. Nelson, and Eric Zivot. 2003. "Why are the Beveridge-Nelson and Unobserved-Components Decompositions of GDP so Different?" *The Review of Economics and Statistics* 2:235–243.

Perold, André F. 1988. "The Implementation Shortfall: Paper versus Reality." *Journal of Portfolio Management* 14:4–9.

Shumway, Robert H. and David S. Stoffer. 1982. "An Approach to time series smoothing and forecasting using the EM algorithm." *Journal of Time Series Analysis* 3:253–264.

Stock, James H. and Mark W. Watson. 1988. "Testing for Common Trends." *Journal of the American Statistical Association* 83:1097–1107.

Figure 1: This graph depicts the end-of-day midquote, the efficient price estimate, and the average execution price of the investor's (parent) orders for each day in the sample. The efficient price estimate is based on a state-space model that was estimated for the entire sample: January 4, 2010 through September 28, 2012. The price estimate is based on the entire sample to obtain maximum efficiency.

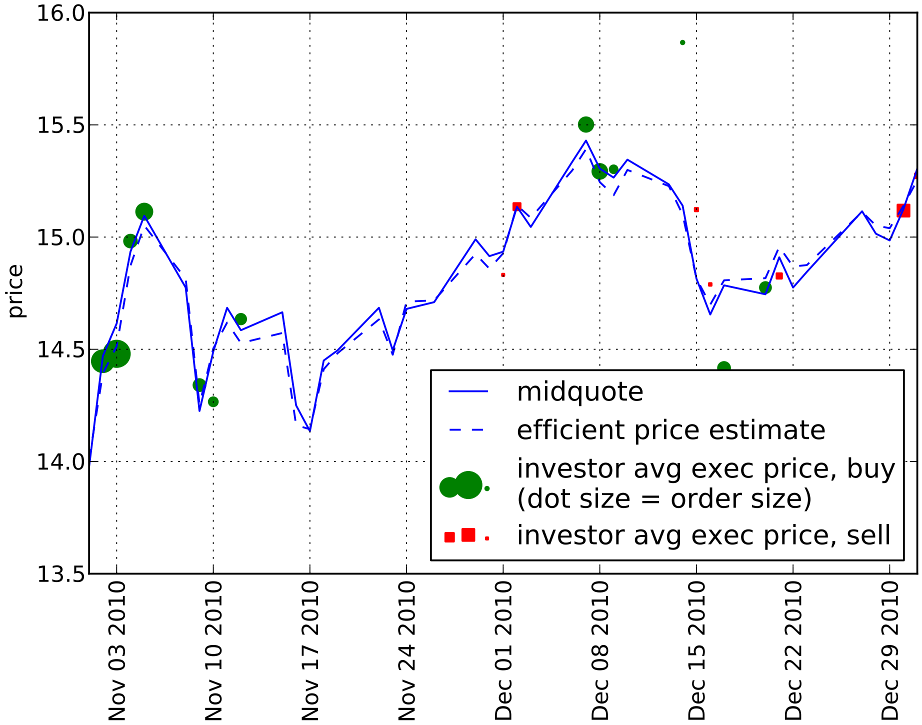


Figure 2: This plot graphs the various components of the implementation shortfall on investor trades each day. They are based on efficient price estimates obtained from a state-space model and based on the entire sample: Jan 4, 2010 through Sep 28, 2012. The components are defined in equation (15).

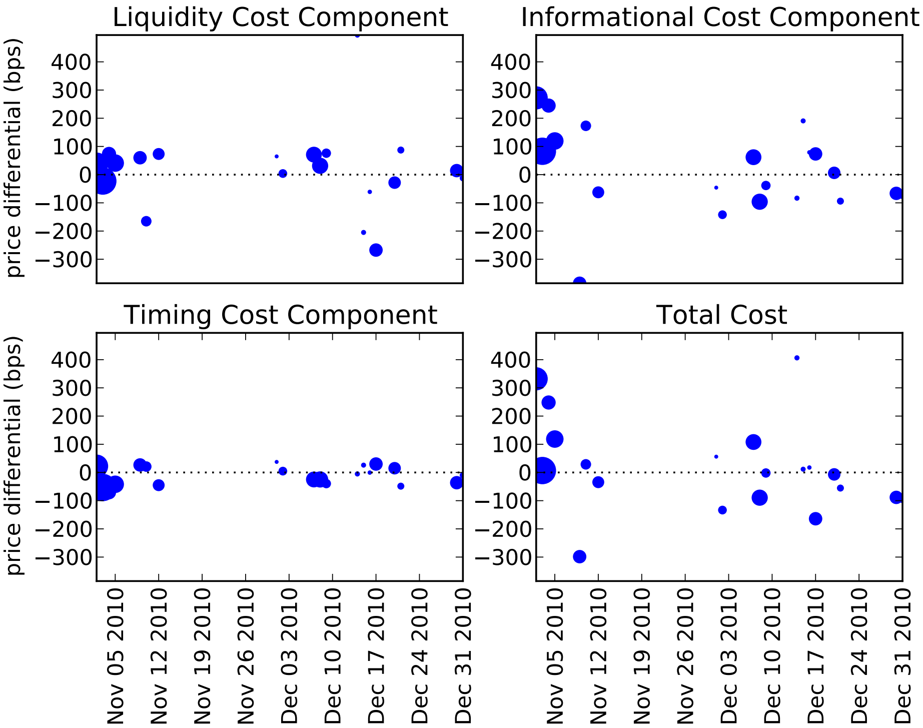


Figure 3: This graph depicts the end-of-minute midquote, the efficient price estimate, and the average execution price of the investor's trades for each minute in the sample. The efficient price estimate is obtained from a state-space model using NBBO midpoints at each minute during regular trading hours for 15 trading days from December 1, 2012 through December 21, 2012, discarding quote midpoints for the first five minutes of trading (9:30am to 9:35am).

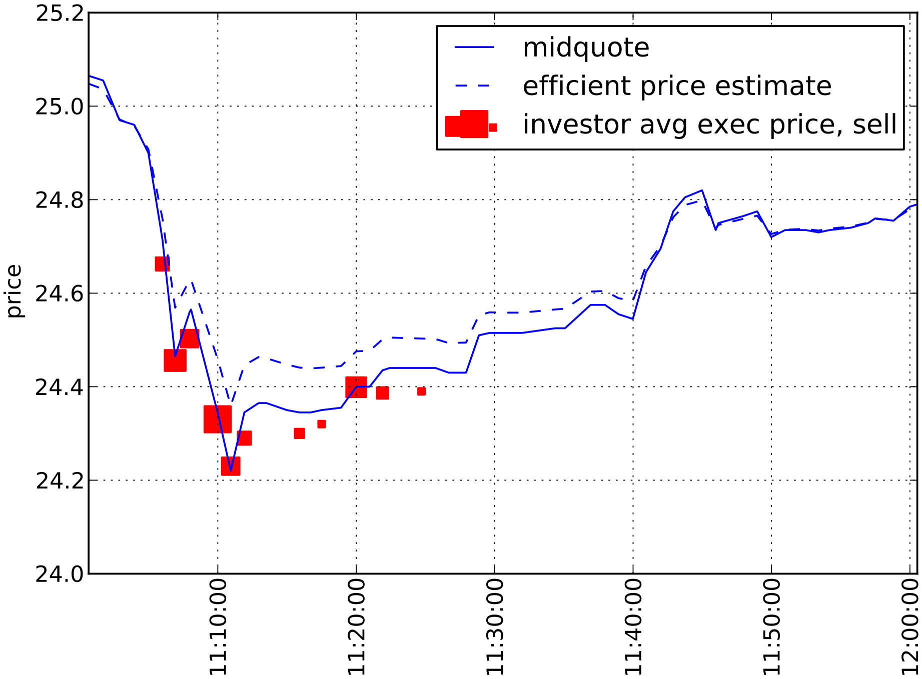
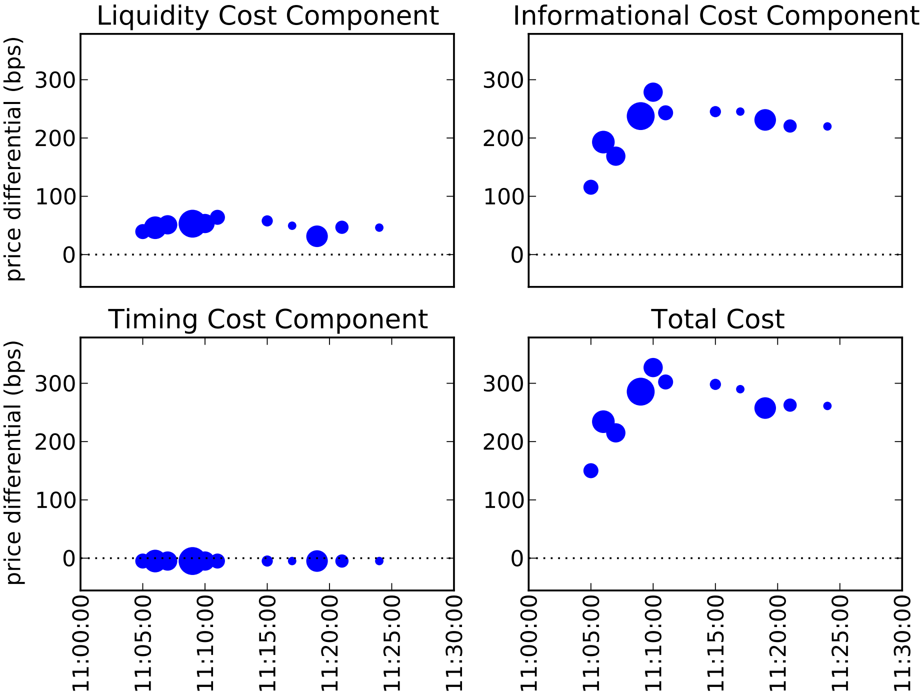


Figure 4: This plot graphs the various components of the implementation shortfall for trades aggregated within each minute. They are based on efficient price estimates obtained from a state-space model using NBBO midpoints at each minute during regular trading hours for 15 trading days from December 1, 2012 through December 21, 2012, discarding quote midpoints for the first five minutes of trading (9:30am to 9:35am). The components are defined in equation (15).


Liquidity Cost Component
Informational Cost Component
Timing Cost Component
Total Cost
price differential (bps)
price differential (bps)
11:00:00
11:05:00
11:10:00
11:15:00
11:20:00
11:25:00
11:30:00