Quantitative Strategy
North America United States

Deutsche Bank

10 November 2010

Signal Processing

Frequency arbitrage

Research Summary
In this report we bridge the gap between high and low frequency quant. We find that factors derived from high frequency data do have predictive power even for "traditional", lower-frequency quant investors.

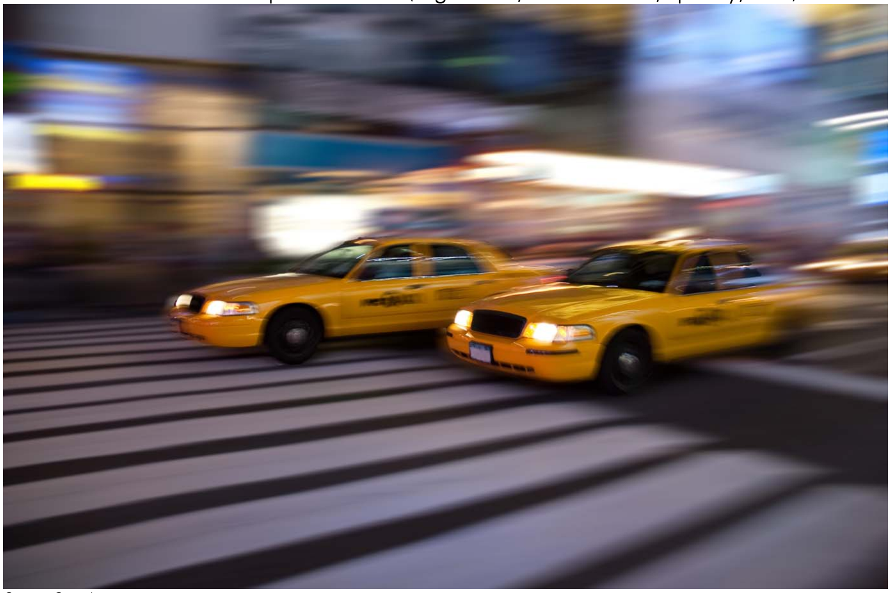
High frequency signals for low frequency investors

Bridging the frequency gap
In both academic and practitioner quantitative research there is a wide gulf between the traditional, low frequency asset pricing research and high frequency, market microstructure research. In this report we try to bridge this gap by showing that quant signals derived from high frequency data can add value even in a low frequency investment strategy.

Three high frequency factors
Specifically, we use the Tick and Quote (TAQ) database to construct three new factors for low frequency investors:

- Order Imbalance
- Probability of Informed Trading
- Abnormal Volume in Large Trades

Avoiding information risk
Of these factors, we find the Probability of Informed Trading (*PIN*) to be the most promising. We show that a variant of *PIN* – where we adjust for size, liquidity, and volatility biases – performs very well as a stand alone factor. More importantly, we find that this factor, which we call *RPIN*, is on average negatively correlated with most of the "standard" quant factors (e.g. value, momentum, quality, etc.).



Source: Getty Images

Team Contacts

Rochester Cahan, CFA
Strategist
(+1) 212 250-8983
rochester.cahan@db.com

Yin Luo, CFA
Strategist
(+1) 212 250-8983
yin.luo@db.com

Javed Jussa
Strategist
(+1) 212 250-4117
javed.jussa@db.com

Miguel-A Alvarez
Strategist
(+1) 212 250-8983
miguel-a.alvarez@db.com

Table of Contents

A letter to our readers

High frequency signals for low frequency investors

In this report, we continue our research into new and innovative data sources

This research report is the fourth in a series of studies looking at how we can use new databases to build less crowded quant factors. In our past research we have looked at options data, industry specific data, and news sentiment data.[1] With all three databases we found that it is possible to construct stock selection signals that perform well in their own right, but more importantly are relatively orthogonal to the traditional quant factor library of value, momentum, quality, etc.

We look at tick and quote (TAQ) data to see if we can construct low frequency signals from high frequency data

In this report we continue the theme by diving into a database that we think is the next frontier for lower frequency quant investors: intraday tick-by-tick data.[2] On face value this statement is a bit of a paradox. Why would we, as relatively low frequency investors, be interested in high frequency data? The fact that this is the first question that springs to mind is precisely why there *is* value in high frequency data. Like the other innovative databases we have studied recently, high frequency data is rarely used by traditional quants and as a result there is a better chance that signals from this database will be less crowded and less correlated with the rest of the factors in our models.

Bridging the frequency divide

We find there are useful signals based on high frequency data that low frequency investors can use

In both academic and practitioner research there is a wide gulf between the traditional, low frequency asset pricing research and high frequency, market microstructure research. However, there are some signals that bridge the gap. In this paper we study three such factors – Order Imbalance, the Probability of Informed Trading (*PIN*), and Abnormal Volume in Large Trades. We find that one signal in particular, a modified version of *PIN* which we call *RPIN*, performs very well on a standalone basis, and more importantly has a negative correlation with most of the typical quant factors. *RPIN* is designed to avoid stocks with high information risk, while at the same time controlling for inherent exposures to volatility, size, and liquidity.

There is no free lunch… but we can help

Intraday data is expensive and hard to use, but we can help by providing data feeds

Of course, high frequency data is not a magic bullet. It is extremely expensive and the technological learning curve required to use it is steep. However, keep in mind that we are more than happy to work with you to set up data feeds or help on the technology side if you would like to test high frequency data within your own investment process. Hopefully we can help make this formidable but promising data set a little easier to use.

Regards,

Yin, Rocky, Miguel, Javed, and John
Deutsche Bank North American Equity Quantitative Strategy

[1] See Cahan et al. [2010a], Luo et al. [2010a], and Cahan et al. [2010b] respectively for details on each of these databases. Complete references for all papers mentioned are available in the "References" section at the back of this report.

[2] When we say "low" frequency in this paper, we primarily mean "traditional" quant investors who are running multifactor models and rebalancing their portfolios at a weekly to quarterly frequency.

Stock screen

We screen for stocks in the S&P 500 with high and low information risk

Below we present two stock screens based on the ideas in this research report. We look for stocks from the S&P 500 universe that have high or low information risk. Our results in this study show that stocks with low information risk tend to outperform on average, while stocks with high information risk tend to underperform.

In these screens we assess information risk using a factor we call *RPIN*. This factor is designed to measure the probability that a stock has heavy informed trading, after controlling for volatility, size, and liquidity. The complete details for this factor can be found in the body of this report.

Long ideas: Screening for stocks with low information risk

Figure 1: Lowest information risk stocks, S&P 500 (long ideas)

Ticker	Name	GICS Sector	Information Risk (*lower* number is better)
IRM	IRON MOUNTAIN INC	Industrials	-2.51
OKE	ONEOK INC	Utilities	-2.39
GAS	NICOR INC	Utilities	-2.27
SJM	SMUCKER (JM) CO	Consumer Staples	-2.18
GT	GOODYEAR TIRE & RUBBER CO	Consumer Discretionary	-2.08
GPC	GENUINE PARTS CO	Consumer Discretionary	-1.94
CTAS	CINTAS CORP	Industrials	-1.84
LUK	LEUCADIA NATIONAL CORP	Financials	-1.84
BMS	BEMIS CO INC	Materials	-1.75
MWV	MEADWESTVACO CORP	Materials	-1.73

Note: Information risk is measured using our 12M average RPIN factor. A lower score is better. For a complete description of this factor, see the body of this report.
Source: TAQ, Deutsche Bank

Short ideas: Screening for stocks with high information risk

Figure 2: Highest information risk stocks, S&P 500 (short ideas)

Ticker	Name	GICS Sector	Information Risk (*lower* number is better)
C	CITIGROUP INC	Financials	6.79
Q	QWEST COMMUNICATION INTL INC	Telecommunication Services	4.38
AIG	AMERICAN INTERNATIONAL GROUP	Financials	3.57
S	SPRINT NEXTEL CORP	Telecommunication Services	3.03
AAPL	APPLE INC	Information Technology	2.90
VIA.B	VIACOM INC	Consumer Discretionary	2.86
GS	GOLDMAN SACHS GROUP INC	Financials	2.70
F	FORD MOTOR CO	Consumer Discretionary	2.66
V	VISA INC	Information Technology	2.64
MA	MASTERCARD INC	Information Technology	2.63

Note: Information risk is measured using our 12M average RPIN factor. A lower score is better. For a complete description of this factor, see the body of this report.
Source: TAQ, Deutsche Bank

Setting the scene

Introducing the TAQ database

We use the NYSE TAQ database for this research

For this study we use the NYSE Tick and Quote (TAQ) database. This database contains intraday transaction data for all NYSE, Amex, and Nasdaq listed securities. At Deutsche Bank we have access to historical data that extends back to 2003. The two most important aspects of the database are transaction level data for every trade conducted in each security (Figure 3) and quote data for all securities (Figure 4).[3]

Figure 3: Example of Deutsche Bank's TAQ database – trade data

	sym	date	time	price	size	cond	cancelled
0	IBM.N	2009.12.09	09:30:05.411	126.5	118600	O	0
1	IBM.N	2009.12.09	09:30:05.455	126.6999969	200	F	0
2	IBM.N	2009.12.09	09:30:06.485	126.5500031	100	F	0
3	IBM.N	2009.12.09	09:30:06.977	126.6699982	100	F	0
4	IBM.N	2009.12.09	09:30:08.462	126.4700012	100	F	0
5	IBM.N	2009.12.09	09:30:11.450	126.5100021	100	F	0
6	IBM.N	2009.12.09	09:30:13.613	126.5199966	600	F	0
7	IBM.N	2009.12.09	09:30:13.619	126.5199966	300	@	0
8	IBM.N	2009.12.09	09:30:13.625	126.5199966	200	@	0
9	IBM.N	2009.12.09	09:30:15.715	126.4800034	100	@	0
10	IBM.N	2009.12.09	09:30:23.918	126.6100006	100	@	0
11	IBM.N	2009.12.09	09:30:26.328	126.5599976	100	@	0
12	IBM.N	2009.12.09	09:30:40.033	126.5699997	100	@	0
13	IBM.N	2009.12.09	09:30:55.360	126.5299988	100	F	0
14	IBM.N	2009.12.09	09:30:55.366	126.5	300	F	0

Source: TAQ, KDB+, Deutsche Bank

Figure 4: Example of Deutsche Bank's TAQ database – quote data

	sym	date	time	bidPrice	bidSize	askPrice	askSize	cond
0	IBM.N	2009.12.09	09:30:05.456	126.4599991	1	126.7300034	1	R
1	IBM.N	2009.12.09	09:30:05.458	126.5	1	126.7300034	1	R
2	IBM.N	2009.12.09	09:30:05.462	126.4599991	1	126.7300034	1	R
3	IBM.N	2009.12.09	09:30:05.653	126.4599991	1	126.6900024	1	R
4	IBM.N	2009.12.09	09:30:06.461	126.5500031	1	126.6900024	2	R
5	IBM.N	2009.12.09	09:30:06.487	126.4599991	1	126.6900024	2	R
6	IBM.N	2009.12.09	09:30:08.111	126.4700012	1	126.6900024	2	R
7	IBM.N	2009.12.09	09:30:08.465	126.4499969	1	126.5699997	11	R
8	IBM.N	2009.12.09	09:30:11.452	126.4700012	1	126.5699997	11	R
9	IBM.N	2009.12.09	09:30:13.609	126.4800034	1	126.5199966	9	R
10	IBM.N	2009.12.09	09:30:13.615	126.4800034	1	126.5199966	3	R
11	IBM.N	2009.12.09	09:30:13.617	126.4800034	1	126.5199966	4	R
12	IBM.N	2009.12.09	09:30:13.618	126.4800034	1	126.5199966	5	R
13	IBM.N	2009.12.09	09:30:13.622	126.4800034	1	126.5199966	2	R
14	IBM.N	2009.12.09	09:30:13.624	126.4800034	1	126.6500015	1	R
15	IBM.N	2009.12.09	09:30:15.718	126.4599991	1	126.6399994	1	R
16	IBM.N	2009.12.09	09:30:15.787	126.4700012	10	126.6399994	1	R
17	IBM.N	2009.12.09	09:30:17.986	126.4400024	2	126.5999985	1	R

Source: TAQ, KDB+, Deutsche Bank

Traditional relational databases are ill suited to tick-by-tick data; we use an advanced in-memory database called KDB+

Needless to say, the volume of data in this database is enormous and it is almost impossible to manage it using a traditional relational database (e.g. Microsoft SQL Server, Oracle). Instead, we use a database called KDB+, which has become something of an industry standard for handling tick-by-tick data. KDB+ is one of a new breed of databases specifically designed to hold vast volumes of data in a column-based, in-memory format. The biggest advantage of the database is speed of access; it comes with its own query language called Q which is able to extract data extremely quickly, and is specifically designed to handle time-series manipulations.[4]

The steep learning curve: A blessing in disguise?

The difficulty of using intraday data might be a good thing – signals derived from it may not be arbitraged as quickly

In fact, we think the steep technological learning curve is the main reason why there is such a gap between high frequency and low frequency research. Traditional asset pricing researchers are usually more familiar with using standard database packages like SAS and SQL to manage data, and then statistical packages like MATLAB or R to do the manipulation. Before this project, we would put ourselves firmly in that camp. However, after tackling the TAQ data, we believe that it is the natural next frontier for quantitative investors looking for fresh factor ideas. In our view, the difficulty in using the data is a positive – it means signals derived from it are less likely to be arbitraged away quickly.

The Deutsche Bank setup

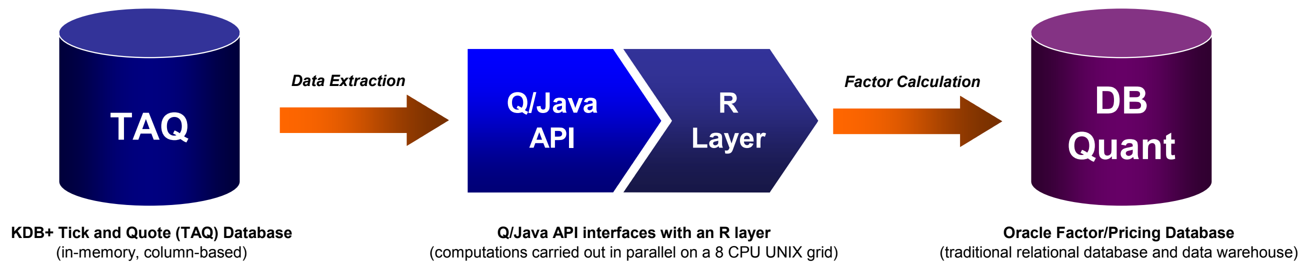
Figure 5 shows the technology framework we use to harness the TAQ data. The key feature is a proprietary API (built in Java and Q) that dramatically simplifies access to the raw tick data. The API is designed to give researchers a set of tools to do low level data manipulation (e.g. aggregating volume by, say, five minute intervals) without having to write the Q code

[3] See http://www.nyxdata.com/Data-Products/Daily-TAQ for more details on the TAQ database.

[4] For more details on KDB+ and Q, see http://kx.com/Products/kdb+.php.

themselves. Another key feature of the API is the ability to call it from R. This means that more complicated statistical procedures that might be difficult in Q can easily be coded in R.

Figure 5: Technology infrastructure for extracting TAQ data into the DB Quant factor database



Source: TAQ, KDB+, Deutsche Bank

At DB, we use R to call a Q/Java API and then do our processing in parallel on a UNIX grid

To speed up the R step, we run the computations on an eight CPU UNIX grid with 16 GB of RAM, which allows us to take advantage of the latest R packages for parallel computing. For example, when we are computing a factor one stock at a time, we can easily parallelize the calculations to do many stocks at once, one on each core. However, even with this cutting edge technology, using intraday data is still a tedious exercise even on a good day. For example, to compute a factor called *PIN* we need to process 1 GB of data per stock, across the 5,000 stocks that have been in the Russell 3000 at one time or another since the start of our data. This roughly equates to 5 terabytes of data that need to be processed. In all, computing the back history of the factor for this universe at a monthly frequency took 10 days of 24/7 computing on our UNIX grid.[5] More than enough time to make a coffee or two between pushing the button and getting the results.

The tricky business of classifying trades

One of the most common tasks with tick-by-tick data is classifying trades as buyer or seller initiated

One of the most common requirements when dealing with TAQ data is a method for classifying trades as either buyer or seller initiated. Many, if not most, of the quant factors we could conceivably construct from TAQ data require that we know whether a particular trade was buyer or seller initiated (also known as the "sign" of the trade). Unfortunately, in TAQ databases we cannot observe this directly, since market participants are of course not required to disclose such information.

To get around this limitation one could try to obtain actual flow data, for example from a broker-dealer, which will have trades tagged as buys or sells. However, there are numerous drawbacks to this approach. First, such flow data is unlikely to represent the whole market, since even a large broker-dealer will only execute a fraction of daily volume; second, such data is rarely available in a timely fashion; and third, from an asset manager's perspective it would be difficult to obtain the data at all on an ongoing basis since few broker-dealers would allow their actual transaction data to be distributed regularly to a buy-side firm.

Thus most investors must rely on statistical algorithms to try to classify trades from a TAQ database into buys and sells. The simplest of these algorithms is the so called "tick test", while the most common is probably the Lee-Ready algorithm.

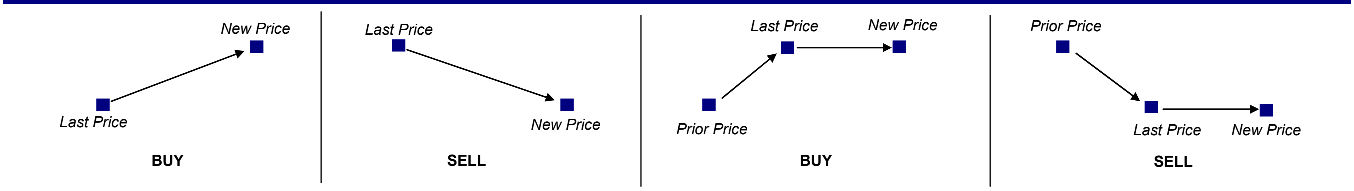
[5] However, keep in mind that the ongoing monthly update of the factors is in the order of one to two hours, so we are not introducing look-ahead bias by using factors that would have been impossible to calculate on a timely basis at each point in time.

There are two main algorithms: the tick test and the Lee-Ready algorithm

The tick test

The tick test is extremely simple (Figure 6). If a trade occurs at a higher price than the last trade, then it is buyer initiated; if the trade is at a lower price than it is seller initiated. If the trade occurs at the same price, then it is a buy if the prior trade was a buy, and a sell if the prior trade was a sell. The advantage of the tick test is that it is extremely easy to compute, which is a non-trivial consideration when dealing with intraday data, because it can speed up processing time significantly.

Figure 6: Tick test classification scenarios

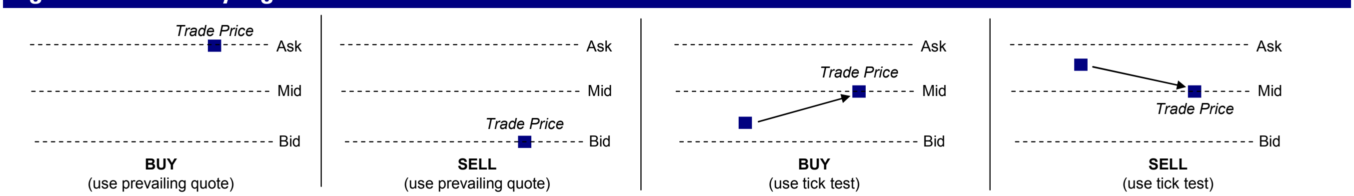


Source: Deutsche Bank

The Lee-Ready algorithm

The Lee-Ready algorithm is named after a paper by Lee and Ready [1991]. Their paper has become something of a seminal paper in the space, because almost all subsequent academic publications use this algorithm. The idea behind the Lee-Ready algorithm is that trade prices in their own right are not enough to accurately classify trades. Instead the Lee and Ready propose joining trade data with the prevailing bid and ask quote for each trade. Trades occurring above the midpoint are classified as buyer initiated and those below the midpoint are classified as seller initiated. For trades occurring at the midpoint, the tick test is used (Figure 7).

Figure 7: Lee-Ready algorithm classification scenarios



Source: Lee and Ready [1991], Deutsche Bank

The Lee-Ready algorithm is computationally slower, and relies on assumptions about quote lag

This sounds almost as simple as the tick test, but in reality it is an order of magnitude more complicated. The difficulty lies in joining the trade data to the quote data, because the time stamps on both data sources can be misleading. The problem is that historically, when trading was less electronic and more manual, quotes and trades were recorded using different systems. For example, Lee and Ready give the example of a floor specialist on the NYSE who calls out the details of a just completed trade and his new quotes. Historically the quote changes would be entered by the specialist's clerk into an electronic workstation, while the trade would be recorded by a stock exchange employee on a separate system. If the specialist's clerk happened to enter the new quotes before the trade was entered, then the timestamps on the two data points would be out of order. This would mean that if one tried to use the timestamps to identify the quote that prevailed when the trade was executed, one could potentially use the quote that actually occurred *after* the trade, not the quote that really existed at the time of the trade.

Lee and Ready suggest a lag of five seconds, but this is probably too long for today's markets

The way Lee and Ready deal with this problem is to lag quotes by five seconds. Their analysis suggests using this lag will eliminate almost all look-ahead bias where quotes from after a trade are recorded ahead of the trade. However, using a lag introduces its own problems, particularly with determining what lag to use. Five seconds was a good rule of thumb in 1991 when Lee and Ready's paper was published, but is almost certainly too long now, given the dramatic increase in electronic trading (indeed a paper by Bessembinder [2003] argues one should use no quote lag at all).

Problems with trade classification algorithms

Trade classification algorithms have a number of weaknesses…

Unfortunately, the issues with trade classification algorithms are not limited to mismatched timestamps. There are a number of other issues that also impact their accuracy:

- **Short sales:** Both the tick test and Lee-Ready algorithm can be inaccurate at classifying short sales, particularly pre 2007. This is because before 2007 the uptick rule was in place, meaning a short sale could only follow a price rise. As a result, short sales would tend to be incorrectly classified as buys by both the tick test and the Lee-Ready algorithm. Indeed, a paper by Asquith, Oman, and Safaya [2010] finds the misclassification rate for short sales to be extremely high regardless of which trade classification algorithm is used.
- **Nasdaq trades:** Another paper, by Ellis, Michaely, and O'Hara [2000], finds that Nasdaq stocks can also be problematic for trade classification algorithms. The authors find that while the algorithms work reasonably well overall, trades inside the quote have a high error rate. They further argue that trades executed on ECNs are more likely to be between the quotes, and hence misclassified. This raises potential concerns given the dramatic rise in trading on these alternative venues.
- **Narrower bid-ask spreads:** Post decimalization, bid-ask spreads have contracted dramatically. This can also hinder trade classification algorithms, because now the midpoint is much closer to the bid and ask. Asquith et al. [2010] point out that most trade classification algorithms are more accurate when trades are at the bid and ask, and less accurate when they are at the midpoint. With a narrower spread, it may be more difficult to identify whether a trade is at the bid, the ask, or the midpoint, and hence accuracy may suffer.
- **High frequency trading:** Ellis et al. [2000] also show that trade classification becomes less accurate as trading frequency increases. This is problematic when looking at recent data, given the exponential increase in high frequency trading in U.S. equity markets.

…but unfortunately there aren't any good alternatives

All these reasons suggest a good deal of caution is warranted when using trade classification algorithms. However, there is a reason why algorithms like Lee-Ready continue to be used almost 20 years after publication: there just aren't any good alternatives. Thus, while it is important to be cognizant of the potential shortcomings with these metrics, the lack of a compelling alternative means we are somewhat stuck with these imperfect measures.

How important is the trade classification algorithm?

Here we assess the differences between the tick test and the Lee-Ready algorithm

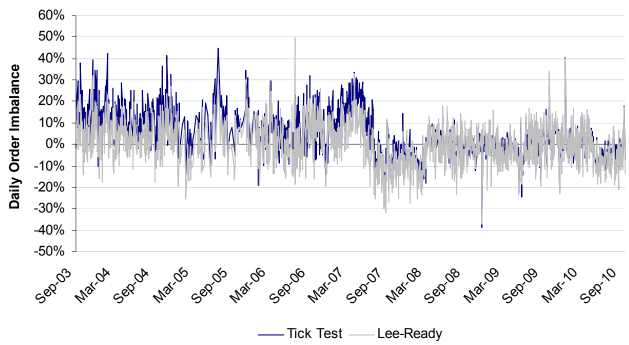
From our perspective, accuracy is more a question of economic impact rather than the academic pursuit of the "perfect" trade classification method. We also have to deal with practical considerations. Most of the academic studies that use intraday data in asset pricing tests only form portfolios once a year and do not require timely implementation, whereas we typically construct our signals at least monthly and need to be able to calculate the factor score quickly so we can implement the trades. This means we have a much higher computational burden, and hence the speed of the algorithm is a critical factor for us.

As already mentioned, from a computational perspective we favor the tick test – it is much faster because it saves us from having to join the tick data to the quote data which is slow over millions of iterations. So the question is whether it is worth sacrificing speed for the better accuracy of the Lee-Ready algorithm.[6] To get a sense for the impact from making this

[6] Accuracy tests between trade classification algorithms are not clear cut, for the same reason that we need them in the first place; if there was enough actual trade data to test the algorithms, then we wouldn't need to bother with them, we

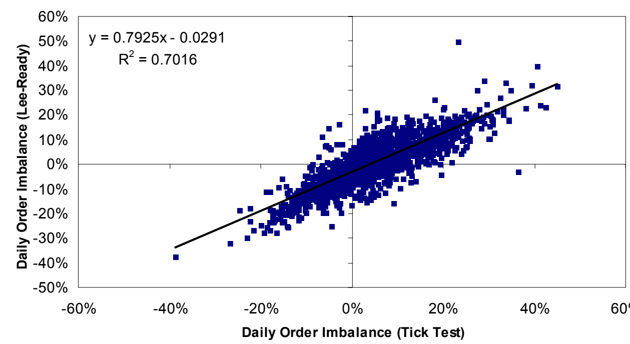
trade-off, we compute some of our factors using both methods. In Figure 8 we show the time-series of an order imbalance factor, computed daily using the tick test and the Lee-Ready algorithm for a single stock (we will more precisely define our factors in the next section). We find that overall the differences are small, particularly in recent years. This is confirmed in Figure 9 where we show a scatter plot of the same data series.

Figure 8: Daily order imbalance for IBM, computed using Tick Test and Lee-Ready algorithm



Source: TAQ, Deutsche Bank

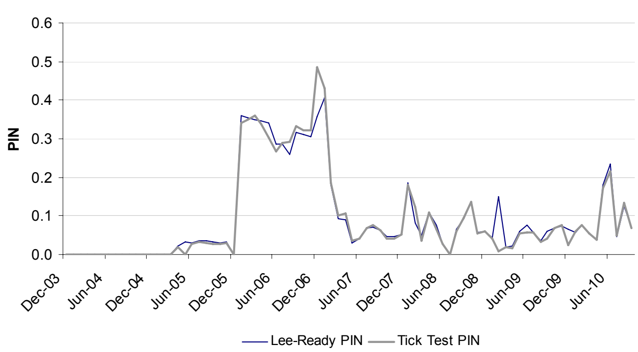
Figure 9: Scatter plot of daily order imbalance for IBM, computed using Tick Test and Lee-Ready algorithm



Source: TAQ, Deutsche Bank

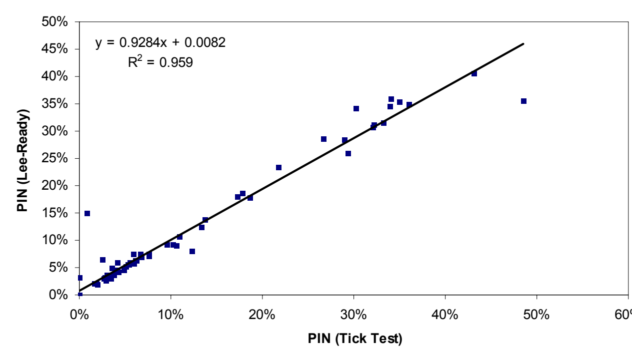
Figure 10, below, shows the same analysis for another factor we consider – the probability of informed trading, or *PIN*. Again, we find the differences between the two versions of the factor – one computed using tick-test signed trades and one computed using Lee-Ready signed trades – is muted. Figure 11 further reinforces that the differences are small.

Figure 10: Monthly PIN for IBM, computed using Tick Test and Lee-Ready algorithm



Source: TAQ, Deutsche Bank

Figure 11: Scatter plot of monthly PIN for IBM, computed using Tick Test and Lee-Ready algorithm



Source: TAQ, Deutsche Bank

We find the difference in final factor scores calculated with each algorithm is marginal

Based on these results, we believe the computational gains from using the tick test outweigh any potential loss of accuracy. However, one thing to keep in mind is that here we are only comparing two alternative trade classification schemes against each other; these results say nothing about whether *both* metrics might be biased in one direction or another. Indeed a paper by Boehmer, Grammig, and Theissen [2006] addresses this very question in the context of *PIN*, and finds that inaccurate trade classification can lead to a downward bias in *PIN*

---

could just use the trade data directly. Nonetheless, the general consensus in the academic literature is that Lee-Ready is more accurate. For example, Ellis et al. [2000] find the Lee-Ready correctly signs 81% of trades, compared to 78% for the tick test, for a selection of Nasdaq trades. Finucane [2000] tests NYSE data and finds Lee-Ready accurate 84% of the time compared to 83% for the tick test.

estimates. This in itself may not be a disaster for us since we are ranking stocks cross-sectionally, so as long as the bias is consistent across all stocks in the universe then the impact on portfolio performance will be limited. However, more worryingly, Boehmer et al. find that the downward bias is related to the intensity of trading in each stock. This is much more problematic when constructing cross-sectional factors. We suggest potential corrections for this issue in the following section, where we introduce *PIN* in more detail.

High frequency factors

Order imbalance (*IMBAL*)

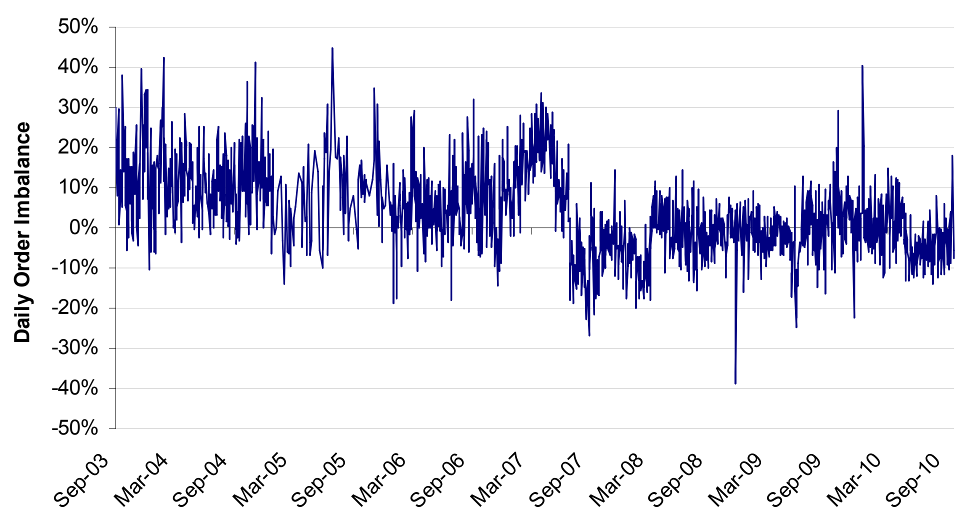
The first factor we consider is Order Imbalance

The first potential factor we consider is order imbalance. This is probably the simplest factor we could construct, and simply involves signing all trades each day, and then computing the difference between buyer initiated and seller initiated trades. Order imbalance on day *t* is simply

$$IMBAL_t = \frac{\sum_{b=1}^{B} VOL_{b,t} - \sum_{s=1}^{S} VOL_{s,t}}{\sum_{b=1}^{B} VOL_{b,t} + \sum_{s=1}^{S} VOL_{s,t}}$$

where $VOL_{b,t}$ is the volume (in number of shares) for the b th buy trade on day *t*, $VOL_{s,t}$ is the volume for the s th sell trade on day *t*, B is the total number of buyer initiated trades on day *t*, and S is the total number of seller initiated trades on day *t*. In other words, we just compute the difference between the total number of shares from buyer and seller initiated trades on a given day, and divide by total number of shares traded on that day. This is the standard definition in the academic literature, for example see Chung and Kim [2010]. Figure 12 shows an example of the daily order imbalance for IBM, computed using this methodology.

Figure 12: Daily order imbalance for IBM



Source: TAQ, Deutsche Bank

In our backtesting (see the next section), we test various moving averages of this daily metric as our monthly factor score.

Probability of Informed Trading (*PIN*)

The second factor we consider is PIN

The concept of Probability of Informed Trading (*PIN*) was first introduced in Easley , Keifer, and O'Hara [1997], but from an asset pricing perspective the more relevant papers are Easley, Hvidkjaer, and O'Hara [2002] in *Journal of Finance* and Easley, Hvidkjaer, and O'Hara [2010] in *Journal of Financial and Quantitative Analysis*.

Definition

PIN is derived from a market microstructure model in which there are three players: market makers, informed traders, and uninformed traders

Because we cannot observe the probability that trades are informed, we need a model to make inferences about what this probability might be. In their series of papers, Easley et al. develop a market microstructure model in which market makers watch market data and use their observations to infer the probability that trades are based on private information. For a complete description of the economic and theoretical rationale behind their framework, Easley et al. [1997] is an excellent starting point. The model proposed in that paper is developed further in Easley et al. [2002] and Easley et al. [2010]. In our research, we use the specification in the latter paper. For a complete description of the model we refer the reader to these papers; here we present only a high level summary of the salient features of the model.

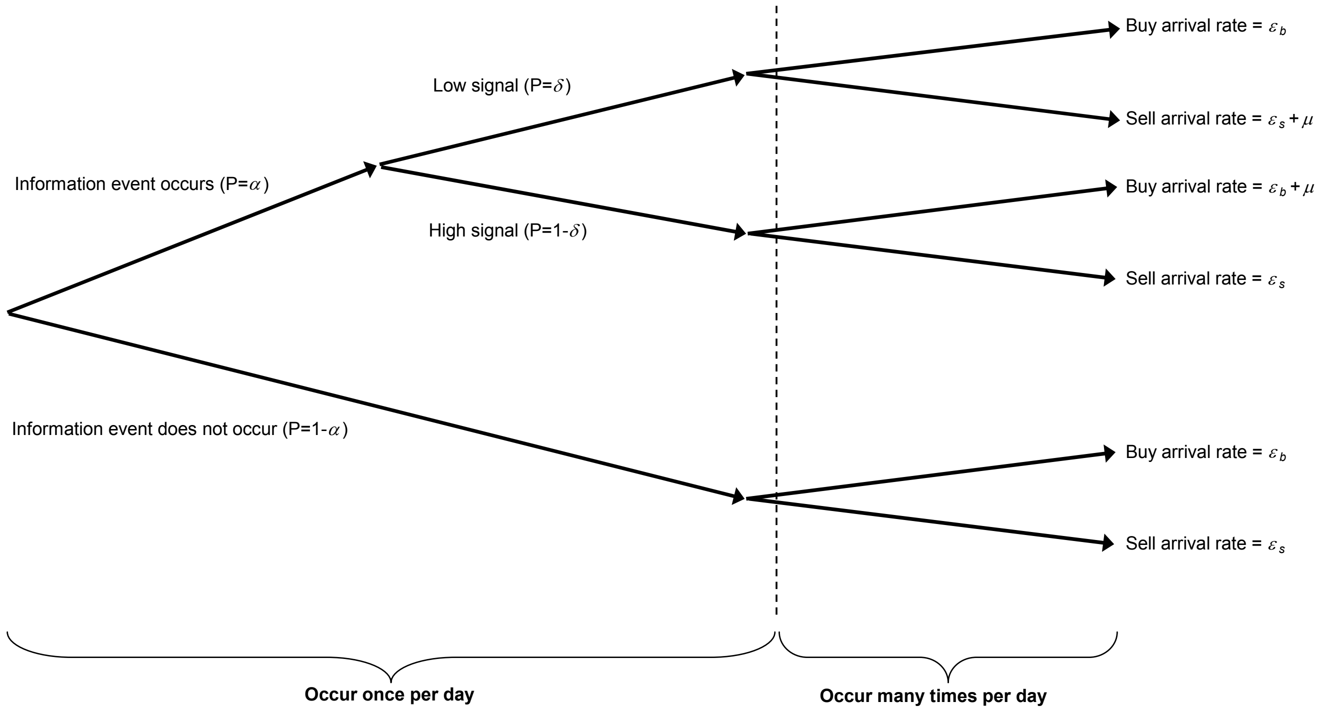
The basic idea is that trading is a game between three players: a competitive market maker, informed traders, and uninformed traders. The market maker observes the sequence of trades and from it tries to infer the probability that trading is being driven by informed or uniformed traders. He or she then uses this information in setting new quotes. The process is captured by a series of probabilities:

- At the start of a trading day, the probability that an information event occurs is α. An information event is an event that gives only the informed traders a signal about the future price of the stock (i.e. will it be higher or lower in the future). If no information event occurs, then every trader will be an uninformed trader.
- Given an information event has occurred, the probability that it is bad news, i.e. signals a lower price, is δ. The probability that it is good news is 1-δ.
- The market maker sets bid and ask quotes at each point in time t during the day. On information days, orders from informed traders arrive at a rate called μ, while buy and sell orders from uninformed traders arrive at rates ε_b and ε_s respectively. On non-information days, all orders are uniformed and arrive at rates ε_b and ε_s respectively. Figure 13 shows this process diagrammatically.

The market maker uses the pattern of buys and sells to infer whether a trade is informed

The market maker of course cannot know whether a trade is informed or uninformed. However, the market maker can use the pattern of trades to estimate the probability that a buy or sell order is information driven. In other words, the market maker can infer where on the tree in Figure 13 she is. For example, if the market maker is observing roughly equal numbers of buy and sell orders, then she might infer she is at the bottom branches of the tree and no information event has occurred. However, if buys are outnumbering sells then perhaps she is in the middle two branches of the tree, which implies an information event has occurred and most likely that event conveyed positive information to informed traders.

Figure 13: Tree diagram of Easley et al. trading model



Source: Easley et al. [2002] , Deutsche Bank

The probability is obtained by estimating a set of parameters via maximum likelihood

When put this way, the model seems quite simple, and not particularly different from a simple order imbalance statistic. However, the beauty of having a model is that we can back out the implied probability of informed trading, based on the observed buy and sell orders over a period of time (we use a 60-day trailing window). If we make the assumption that the arrival of buy and sell orders over the day from uninformed traders follow independent Poisson processes, then Easley et al. [2010] show that the log likelihood function is given by

$$L((B_t, S_t)_{t=1}^{T} \mid \theta) = \sum_{t=1}^{T} [-\varepsilon_b - \varepsilon_s + M_t(\ln x_b + \ln x_s) + B_t \ln(\mu + \varepsilon_b) + S_t \ln(\mu + \varepsilon_s)]$$
$$+ \sum_{t=1}^{T} \ln[\alpha(1-\delta)e^{-\mu} x_s^{S_t - M_t} x_b^{-M_t} + \alpha\delta e^{-\mu} x_b^{B_t - M_t} x_x^{-M_t} + (1-\alpha) x_s^{S_t - M_t} x_b^{B_t - M_t}] ,$$

where B_t is the number of buyer initiated trades on day t, S_t is the number of seller initiated trades on day t, $M_t = \min(B_t, S_t) + \max(B_t, S_t)/2$, $x_s = \varepsilon_s / (\mu + \varepsilon_s)$, $x_b = \varepsilon_b / (\mu + \varepsilon_b)$, and $\theta = (\mu, \varepsilon_b, \varepsilon_s, \alpha, \delta)$. Note that we are summing across days $t = 1$ to $T = 60$. Using this equation, we can estimate the five parameters in the model via maximum likelihood. Recall these five parameters are:

δ = Probability of bad news

μ = Daily arrival rate of orders from informed traders

ε_b = Daily arrival rate of buy orders from uninformed traders

ε_s = Daily arrival rate of sell orders from uniformed traders

α = Probability that an information event occurs

Once the parameters are estimated, the PIN calculation is easy

Easley et al. then go on to show that *PIN*, the probability of informed trading, is given by

$$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s},$$

where $\alpha\mu + \varepsilon_b + \varepsilon_s$ is the arrival rate for all orders and $\alpha\mu$ is the arrival rate for informed orders. To estimate *PIN* in practice, one chooses a trailing window (e.g. 60 days) to watch trades over. Each trade in this window needs to be classified as buyer initiated or seller initiated. As mentioned previously, there are a number of algorithms that can be used to do so, but we use the simplest – the tick test. Once one has the list of buyer and seller initiated trades, one can estimate the five model parameters using maximum likelihood, and then compute *PIN* for that stock at that point in time.

A simple example

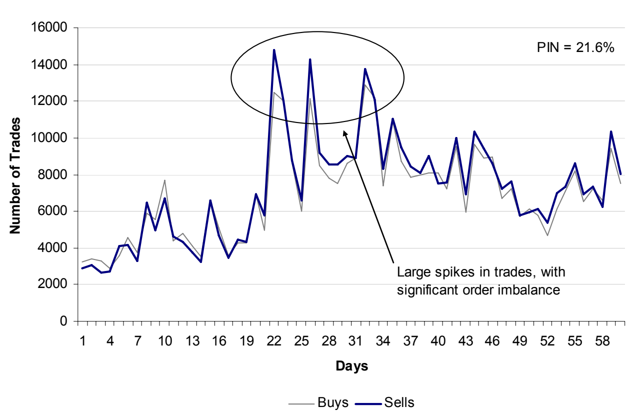
A simple example can help make PIN more transparent

This may seem a little opaque, but a simple example presented in Easley et al. [2002] makes things a little clearer. Suppose on 20% of days a stock has 90 buy trades and 40 sell trades, and on another 20% of days it has 40 buys and 90 sells. For the other 60% of days the stock has 40 buys and 40 sells. If we use this information and estimate our parameters via maximum likelihood, we would obtain $\varepsilon_b = \varepsilon_s = 40$, $\mu = 50$, $\alpha = 0.4$, and $\delta = 0.5$. From this, we would estimate *PIN* as 20%. This is somewhat intuitive. In this example, the "natural" level of buy and sell orders appears to be 40. When we have a deviation from this, i.e. 90 buys or 90 sells, it makes sense that the difference of 50 might represent informed trading. The results for α and δ, which represent the probability of an information event and probability of bad news respectively, also make sense: 40% of days seem to have abnormal trading which might signal an information event, and that abnormal trading is split 50/50 between abnormal buying and abnormal selling.

High PIN is driven by spikes in order flow that occur with a large imbalance in buyer versus seller initiated trades

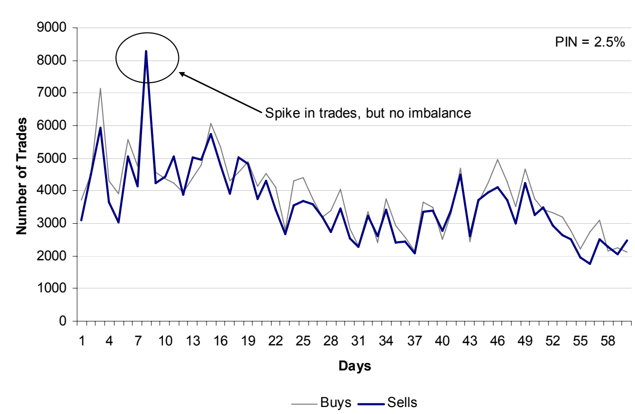
It is also useful to look at *PIN* visually. In essence *PIN* is designed to capture an imbalance between buy and sell orders over some time interval, relative to the "normal" level for that stock. Figure 14 shows an example of an actual 60-day sequence of trades that generated a high *PIN* estimate of 21.6%. Figure 15 shows a sequence that led to a low *PIN* estimate. Roughly speaking, the key difference between the high and low *PIN* sequence is the imbalance at the peak of the large trading spikes in the left hand chart. Based on the *PIN* methodology, this suggests 1) an information event on those days, and 2) informed trading on the back of those events.

Figure 14: Example of high PIN trade sequence



Source: TAQ, Deutsche Bank

Figure 15: Example of low PIN trade sequence



Source: TAQ, Deutsche Bank

Adjustments for size, liquidity, and volatility

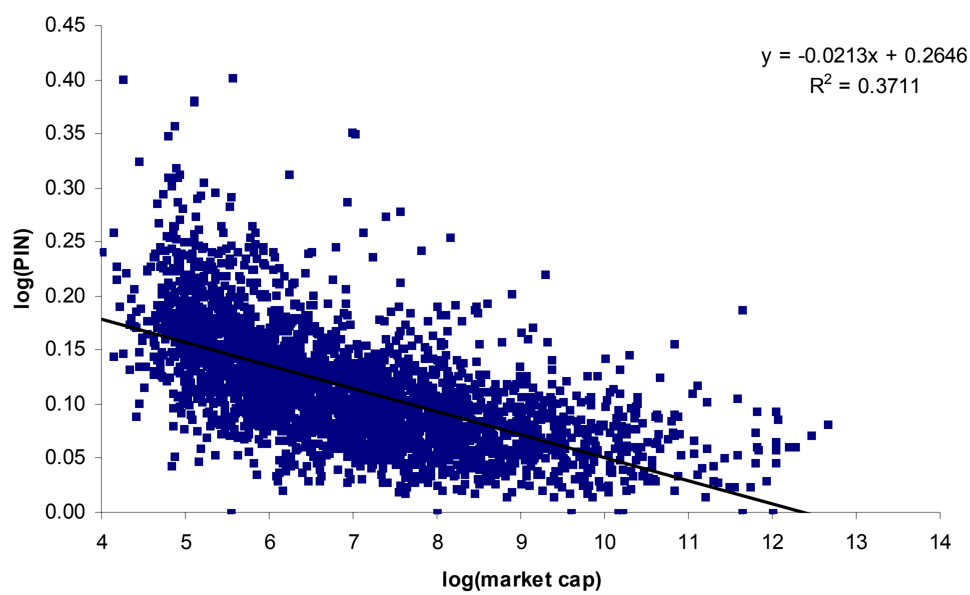
PIN has inherent biases to size, liquidity, and volatility

PIN sounds promising in theory, but it does have weaknesses. For quant investors, the most problematic is that *PIN* tends to be related to size, liquidity, and volatility (see Easley et al. [2002], Hwang and Qian [2010]). Which raises the question: is *PIN* capturing something new, or is it just a proxy for a combination of other well known factors?

PIN is negatively correlated with size

Figure 16 shows the cross-sectional correlation between *PIN* and market cap at a point in time. A very clear negative and convex relationship is apparent, indicating that *PIN* tends to be higher for small cap stocks, which is intuitive.

Figure 16: Cross-sectional correlation: PIN vs. Size, as at 30-Sep-2010

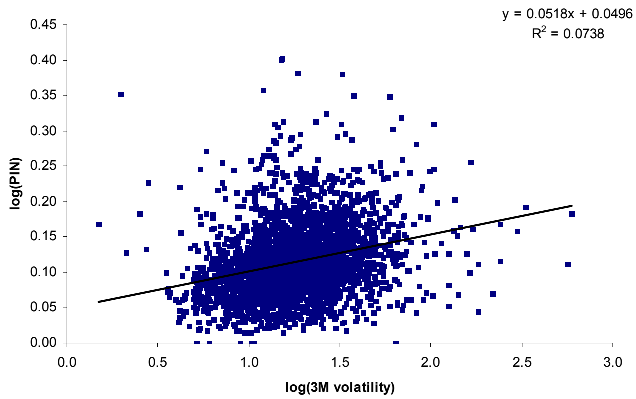


Source: TAQ, Deutsche Bank

PIN is positively correlated with volatility and negative correlated with liquidity

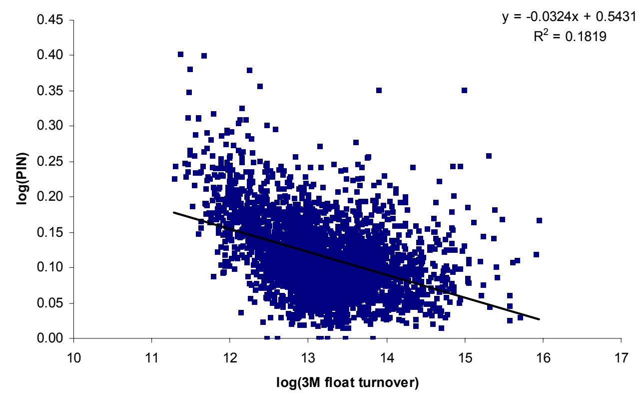
Similarly, Figure 17 shows the relationship of *PIN* with volatility (computed using three months of trailing daily returns) and Figure 18 shows *PIN* versus turnover (measured as the percent of total shares turned over in three months). From the charts it is clear that *PIN* is somewhat related to both of the variables – *PIN* tends to be positively correlated with volatility and has a negative and convex correlation with turnover.

Figure 17: Cross-sectional correlation: PIN vs. Volatility, as at 30-Sep-2010



Source: TAQ, Deutsche Bank

Figure 18: Cross-sectional correlation: PIN vs. Turnover, as at 30-Sep-2010



Source: TAQ, Deutsche Bank

In other words, buying high *PIN* stocks is akin to buying high volatility, low liquidity, small cap names. This is problematic, because it suggests any returns to *PIN* may just be compensation for these well known risk factors. In our research, we address this issue by proposing a modified version of *PIN* that we call residual *PIN*, or *RPIN* for short. The idea is simple. At each point in time we use a cross-sectional regression where we regress *PIN* scores onto size, volatility, and turnover factors, effectively stripping out the correlation to these factors. Mathematically, the *RPIN* factor score for stock i at time t is given by

$$RPIN_{i,t} = \varepsilon_{i,t}$$

We remove the biases in PIN using a cross-sectional regression at each point in time

where $\varepsilon_{i,t}$ is the residual for the i th stock from the cross-sectional regression

$$\log(PIN_t) = c + \log(SIZE_t) + \log(\sigma_t) + \log(LIQ_t) + \varepsilon_t$$

where $SIZE_t$ is market cap, σ_t is the standard deviation of daily returns over the past three months, and LIQ_t is the percent of total shares on issue traded in the past three months.

This regression helps reduce the inherent bias in *PIN* towards small, high volatility, low liquidity stocks. However, the regression only removes the cross-sectional exposure of the factor scores to these factors; it does not preclude the possibility that the returns to the factor are still driven by these types of stocks. We address this issue in more detail in our backtesting section.

Abnormal Volume in Large Trades (*ALT*)

The third factor we consider is Abnormal Volume in Large Trades

Another factor we consider is the abnormal volume in large trades (*ALT*). This factor is inspired by a paper by Tong [2009], and is based on the idea that informed traders who have compelling private information are likely to trade more aggressively. Tong argues that a fingerprint of this type of trading is higher volume in "large" trades. The definition of the factor is simple:

- At the start of each month, compute the 30%, 60%, and 90% fractiles using one year of trailing trade data. The fractiles are computed over volume (i.e. number of shares).
- For this month, classify all trades with volume greater than the 90% cutoff as large trades.
- Sum the volume for all the large trades in this month and compute

$$ALT_t = \frac{\text{sum large trade volume in month } t}{\text{sum large trade volume in last 12 months}}.$$

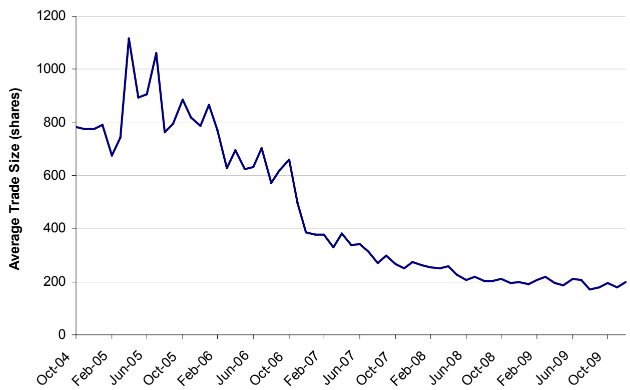
ALT has been shown to dominate PIN in academic asset pricing tests

Tong specifically runs a horse race between *PIN* and *ALT*, and finds that *ALT* has a more robust relationship with future returns than *PIN*. The *ALT* factor also has another advantage in that it is not dependent on signing trades, hence it avoids many of the drawbacks of imbalance and *PIN*. However, a potential weakness of *ALT* is its dependence on the premise that large trades represent informed trading. The rise of algorithmic execution, direct market access, and alternative trading venues (e.g. dark pools) means that investors are now much better at disguising their trading activities.

But ALT may be affected by the rise in algorithmic trading and alternative venues

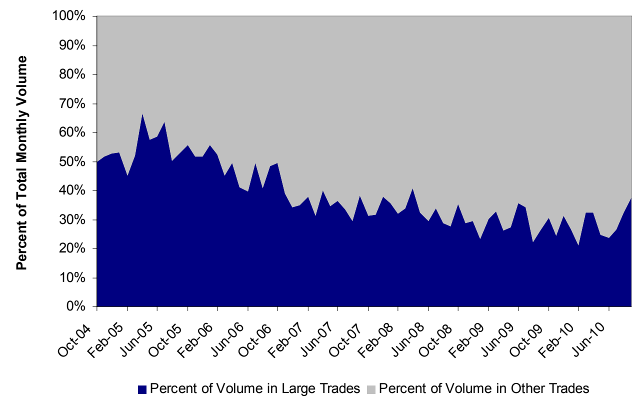
Figure 19 shows the average size of each individual trade for IBM over time. Clearly there has been a dramatic decline, even though we are only looking at the past six years. Similarly, Figure 20 shows the percent of total volume that is classified as large trades, i.e. what percent of volume is contained in the top 10% of largest trades. Again there has been a dramatic decline since 2004. Both charts support our argument that the average trade size is becoming much smaller as more and more trading shifts to machines. This could be a problem for *ALT*, or it could be a good thing if it means that large trades are now even more meaningful because of their increasing rarity. The only way to find out is to do the backtest.

Figure 19: Average size (number of shares) per trade for IBM by month



Source: TAQ, Deutsche Bank

Figure 20: Percent of trades for IBM classified as large trades, by monthly volume



Source: TAQ, Deutsche Bank

We also calculate two alternative definitions of ALT – the Percent of Large Trades and residual ALT

Alternative ALT measures

In addition to the basic *ALT* factor, we also compute two other variants. The first, which we call Percent of Large Trades (*PLT*) is simply the monthly volume in large trades in a given month, divided by the total monthly volume in the same month, i.e.

$$PLT_t = \frac{\text{sum large trade volume in month } t}{\text{sum all trade volume in month } t}.$$

The second metric we consider is residual *ALT*, or (you guessed it) *RALT*. This is constructed in exactly the same way as *RPIN*, i.e. at each point in time we regress *ALT* cross-sectionally onto size, volatility, and liquidity factors. We then define *RALT* as the residual from that regression. Constructing this factor allows for a fairer comparison with *RPIN* when backtesting.

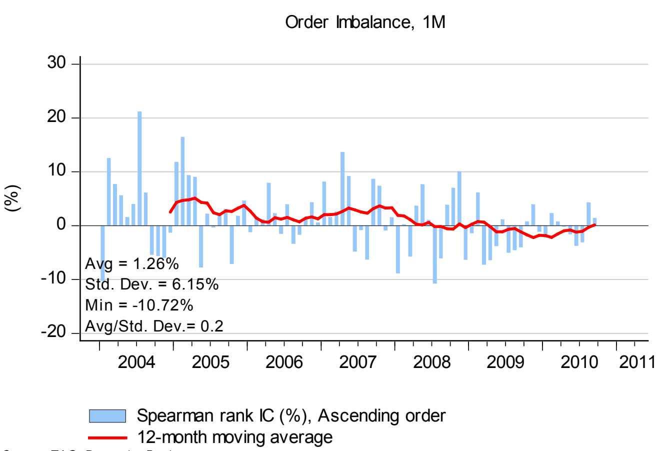
Backtesting results

Order Imbalance

We find order imbalance is a relatively weak factor

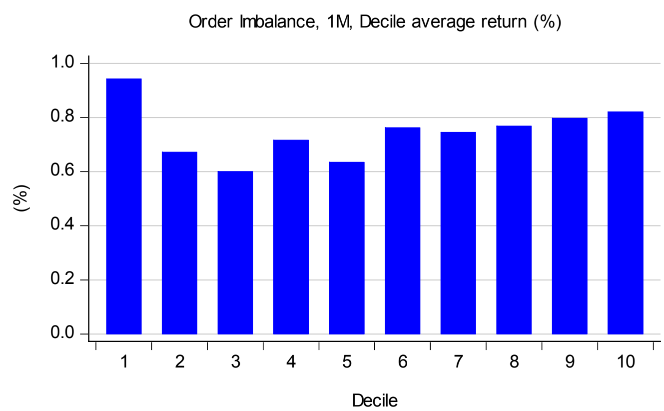
The first factor we backtest is order imbalance. We try two variations using one month (1M) and three month (3M) trailing order imbalance. Figure 21 and Figure 22 show the monthly rank information coefficient (IC) and the average monthly decile returns, respectively, for the 1M factor. Both charts suggest that this factor is not particularly effective – the average rank IC is only 1.26% over the backtest period and recent performance is particularly poor. Furthermore, the decile returns are not particularly monotonic.

Figure 21: 1M order imbalance, rank IC



Source: TAQ, Deutsche Bank

Figure 22: 1M order imbalance, average decile returns

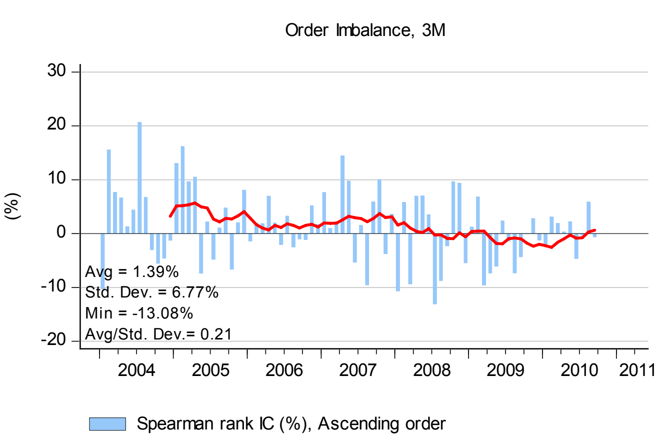


Source: TAQ, Deutsche Bank

The poor performance is not surprising since order imbalance is widely reported in the media

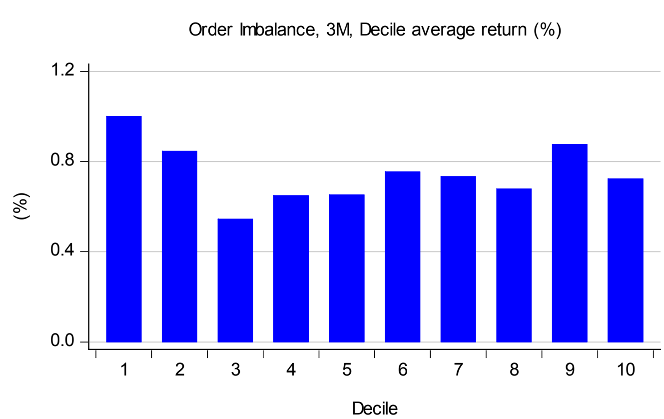
The 3M order imbalance factor does not fare much better. The average rank IC is only marginally better (Figure 23) and again the decile returns are not very monotonic (Figure 24). The poor performance of this factor is not unexpected – order imbalance data is regularly reported by major financial news organizations and as a result we would not expect there to be too much alpha left in the signal. Given our findings, we do not pursue the order imbalance factor further in this report.

Figure 23: 3M order imbalance, rank IC



Source: TAQ, Deutsche Bank

Figure 24: 3M order imbalance, average decile returns



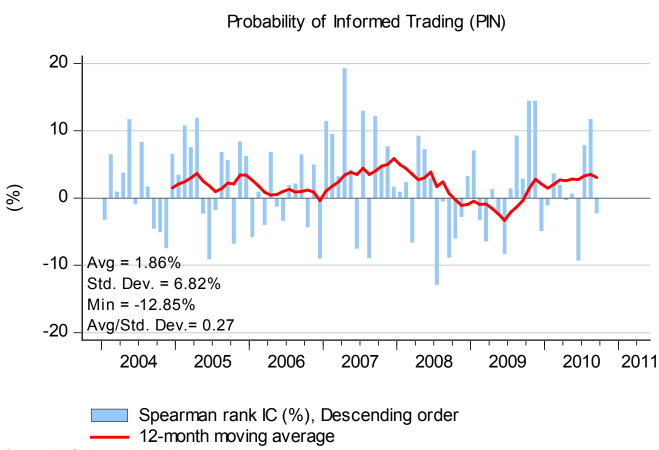
Source: TAQ, Deutsche Bank

Probability of Informed Trading

In our initial backtests PIN shows some promise

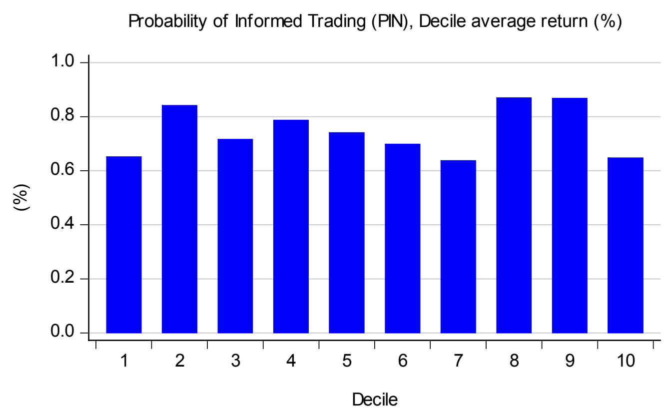
We start by backtesting the simple *PIN* factor, without the regression adjustments described in the previous section. The average rank IC of the factor is actually quite promising at 1.86% (Figure 25), particularly considering that the latter half of the backtest period was particularly challenging for most the traditional factors. However, if we look at the average monthly returns to decile portfolios, in Figure 26, we see that the factor lacks a strong monotonic pattern. This suggests that while the factor does reasonably well in ranking stocks, this efficacy is not borne out in return space.

Figure 25: PIN, rank IC



Source: TAQ, Deutsche Bank

Figure 26: PIN, average decile returns

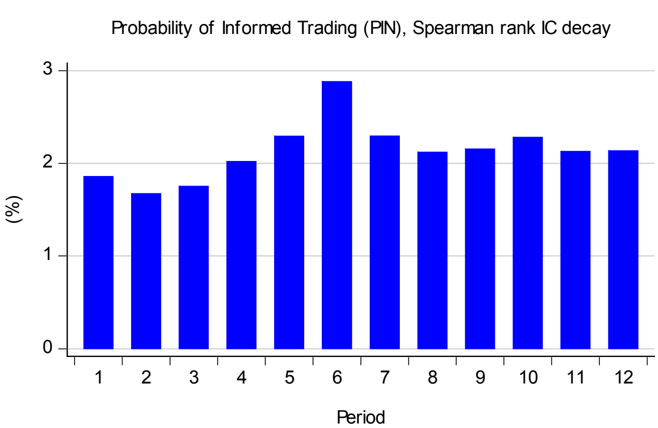


Source:TAQ, Deutsche Bank

But the problem is a lack of monotonicity in returns, and an irregular information decay profile

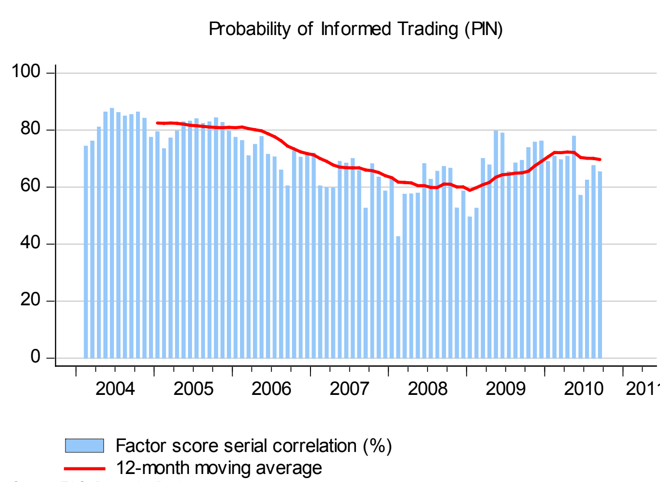
Another problematic feature with the basic *PIN* factor is shown in Figure 27. The IC decay profile actually shows that the factor works better at a longer horizon, peaking at a six month lag. This is in line with the academic research (e.g. Easley et al. [2002, 2008]) who find predictive power at a one-year holding period), and suggests that using *PIN* with monthly rebalancing may not be optimal. On the positive side, *PIN* is a relatively low turnover factor, which may come as a surprise to those who automatically assume that high frequency data will only yield high frequency factors (Figure 28).

Figure 27: PIN, rank IC decay



Source: TAQ, Deutsche Bank

Figure 28: PIN, autocorrelation



Source: TAQ, Deutsche Bank

We find high PIN stocks underperform on average, which is opposite the academic literature

Higher PIN is bad?

In discussing the statistical results, we have to this point glossed over what we think is the most interesting finding: stocks with high *PIN* tend to *underperform* on average. This is exactly opposite the academic evidence. Indeed, the standard academic argument is that higher *PIN* equals higher risk (since to trade these stocks one takes on the risk of trading against someone with "better" information), and consequently one should be compensated for this with higher returns. However, we argue that our *PIN* results are consistent with what we find for all risk measures, not just *PIN*. In our research, we consistently find that it is actually *low* risk stocks that tend to outperform. When we backtest a wide range of risk metrics – for example realized volatility, realized skewness, realized kurtosis, beta, CT-risk [7] – we consistently find that for the U.S. market it is low risk stocks that outperform on average (Figure 29). In this light, we would argue that if *PIN* does indeed proxy for information risk, then like the other forms of risk we look at we would expect low risk stocks to outperform high risk stocks.

Figure 29: Backtesting performance of common risk factors, Russell 3000, 1988-present

Factor	Direction	Average Monthly Rank IC
CAPM beta, 5Y monthly	Descending	0.76
CAPM idosyncratic vol, 1Y daily	Descending	4.68
Realized vol, 1Y daily	Descending	4.58
Skewness, 1Y daily	Descending	1.15
Kurtosis, 1Y daily	Descending	1.31

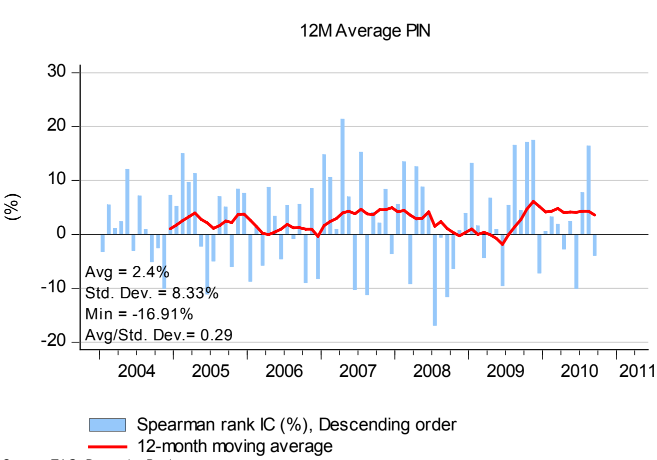
Note: "Descending" means that a lower factor score is better, i.e. in all cases stocks with lower risk outperform those with higher risk.
Source: Bloomberg, Compustat, Haver, Russell, S&P, Thomson Reuters, Deutsche Bank

A rolling average of PIN works better than spot PIN

12-month average PIN

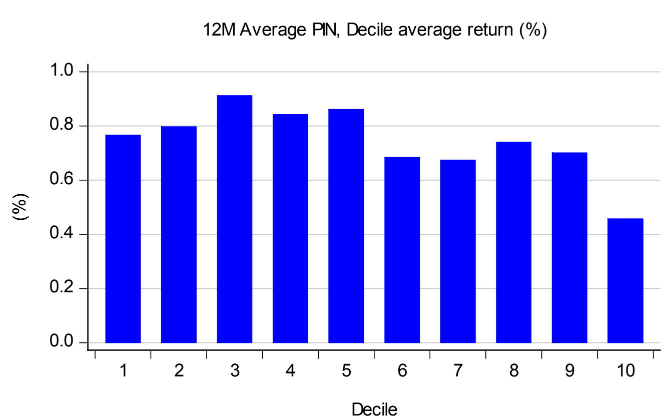
To explore the idea that *PIN* may be better used as a "slow burn" factor, we also backtest a simple 12-month (12M) average *PIN* factor. This factor is computed by using a 12-month rolling average of monthly *PIN* scores for each stock at each point in time. In effect, we are smoothing out some of the month-to-month volatility in *PIN* at the stock level. Figure 30 shows that doing this actually improves the IC of the factor considerably, raising the average to 2.4% (Figure 30). The decile returns also show a more monotonic pattern (Figure 31).

Figure 30: 12M average PIN, rank IC



Source: TAQ, Deutsche Bank

Figure 31: 12M average PIN, average decile returns



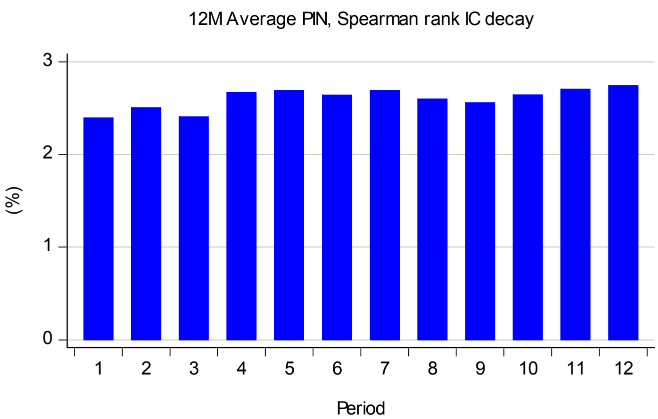
Source: TAQ, Deutsche Bank

[7] CT-risk, or Contribution to Risk, is an interesting new risk factor that we propose in our *Portfolios Under Construction* research series. The factor considers not only a stock's own volatility, but also its co-movement with other stocks in the universe. For further details, see Luo, Cahan, Jussa, and Alvarez [2010b].

Using an average improves information decay and also reduces turnover

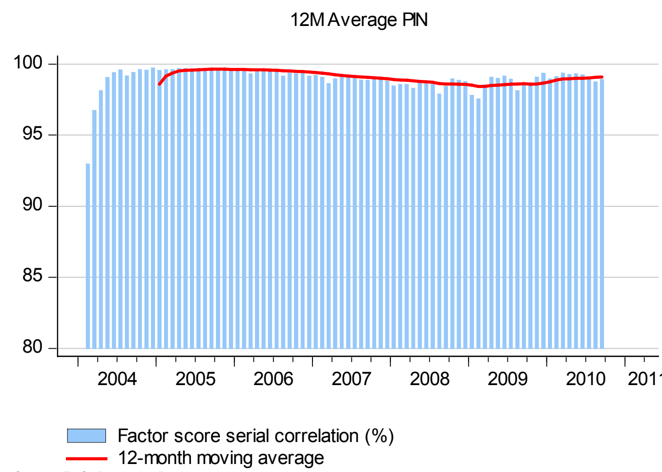
Smoothing the factor also improves the decay profile (Figure 32) and makes the factor turnover very moderate (Figure 33). A month-to-month autocorrelation of greater than 95% is in line with slow burn factors like value. Given the improved performance that comes from averaging the factor, we use the 12M average *PIN* factor as our preferred *PIN* metric going forward.[8]

Figure 32: 12M average PIN, rank IC decay



Source: TAQ, Deutsche Bank

Figure 33: 12M average PIN, autocorrelation



Source: TAQ, Deutsche Bank

Residual PIN

Next we address the biases in PIN by testing residual PIN

So far we have only considered simple *PIN*, as defined in the academic literature. As mentioned in the previous section, *PIN* has the shortcoming that it is skewed towards high volatility, low liquidity, small cap stocks. Therefore, the returns highlighted above may simply be the result of taking exposure to these factors. To test this, we backtest the residual *PIN* factor (*RPIN*) we described previously. As explained previously, this factor essentially controls for the inherent size, volatility, and liquidity biases in *PIN*.

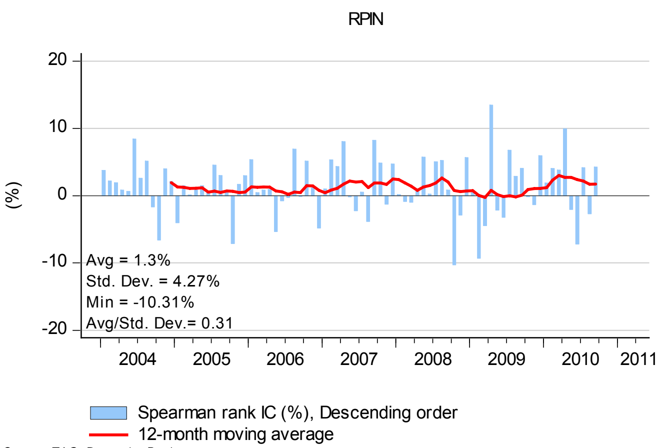
We find RPIN improves risk adjusted performance

Figure 34 shows the rank IC for *RPIN*. As expected, the average IC drops – from 1.86% to 1.30% - compared to the basic *PIN* factor (recall Figure 25). However, in risk-adjusted terms, the performance of the *RPIN* factor is actually better: 0.31 versus 0.27. As well, the average decile returns to *RPIN* show a more monotonic pattern compared to basic *PIN* (Figure 35 versus Figure 35).

---

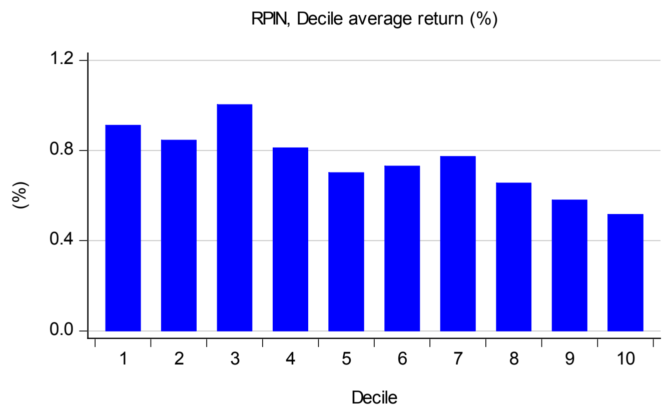
[8] Note we also tried a three-month averaging window, which yielded results in between the one-month and 12-month results.

Figure 34: RPIN, rank IC



Source: TAQ, Deutsche Bank

Figure 35: RPIN, average decile returns

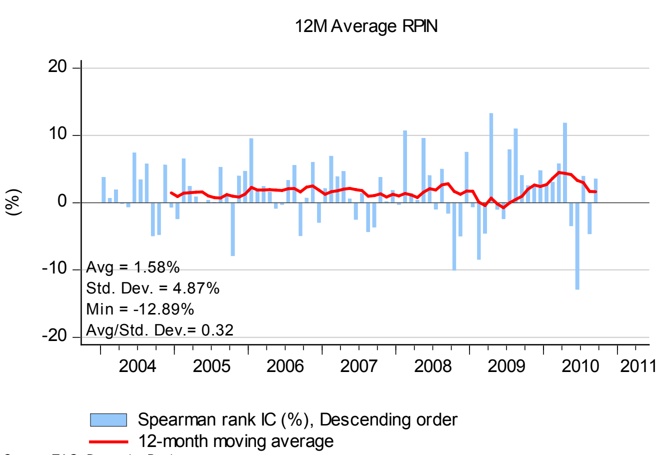


Source: TAQ, Deutsche Bank

Using a rolling average improves RPIN

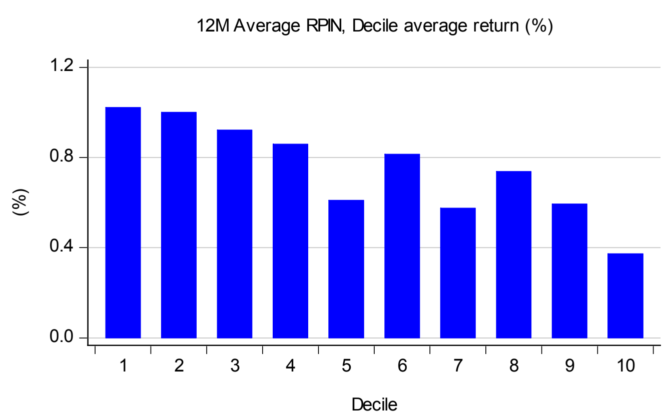
If we consider 12M average *RPIN* and compare it to 12M average basic *PIN*, we see a similar drop in performance in absolute terms, but an improvement in risk-adjusted terms (Figure 36). In fact, the rank IC chart shows a pleasing consistency of performance over time, with the 12-month rolling average rank IC almost never dropping below zero.

Figure 36: 12M average RPIN, rank IC



Source: TAQ, Deutsche Bank

Figure 37: 12M average RPIN, average decile returns



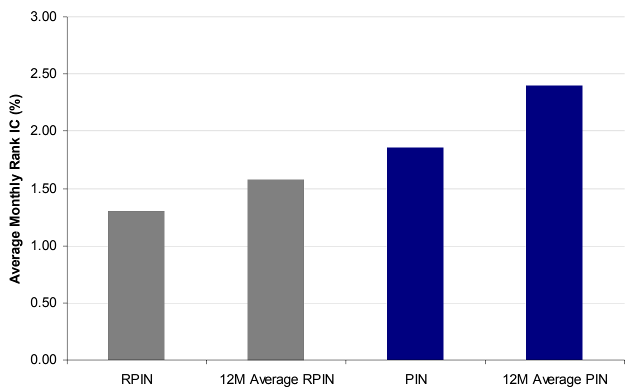
Source: TAQ, Deutsche Bank

We find RPIN works better in risk-adjusted terms, but is worse in absolute terms

The two charts below compare the average rank IC (Figure 38) and risk-adjusted rank IC (Figure 39) for our four *PIN* measures. Broadly speaking we can draw two conclusions:

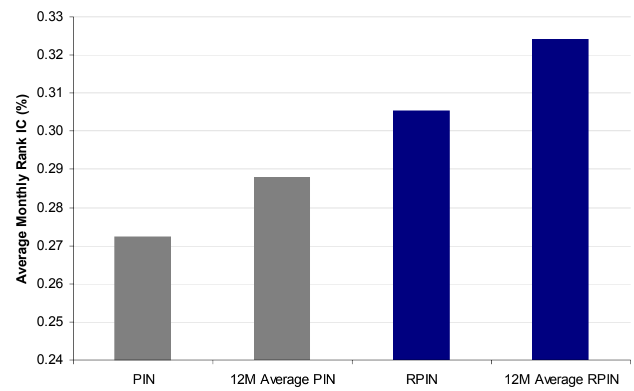
- Using a 12M average of the factor is beneficial for both basic *PIN* and *RPIN*. In both absolute and risk-adjusted terms the 12M average version of each factor performs better over the backtest.
- *RPIN* reduces performance in absolute terms, but improved performance in risk-adjusted terms.

Figure 38: Summary of PIN rank ICs



Source: TAQ, Deutsche Bank

Figure 39: Summary of PIN risk-adjusted performance



Source: TAQ, Deutsche Bank

We prefer 12M average RPIN as our PIN factor

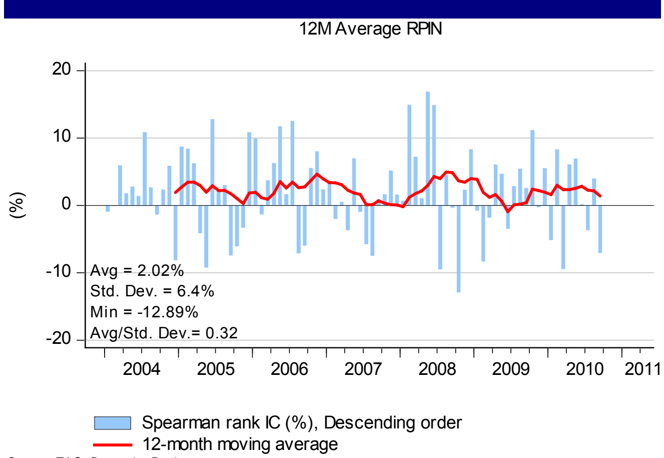
Based on these findings, we will use the 12M average *RPIN* factor as our preferred *PIN* factor in the rest of this report. This choice is a little prone to data mining, in the sense that we have tested a number of factors and picked the best one in risk-adjusted terms. However, we do note that the rest of the results in this paper are not particularly sensitive to our specific choice of *PIN* factor.

Results by size segment

We find the performance of RPIN actually improves when we use the S&P 500 universe

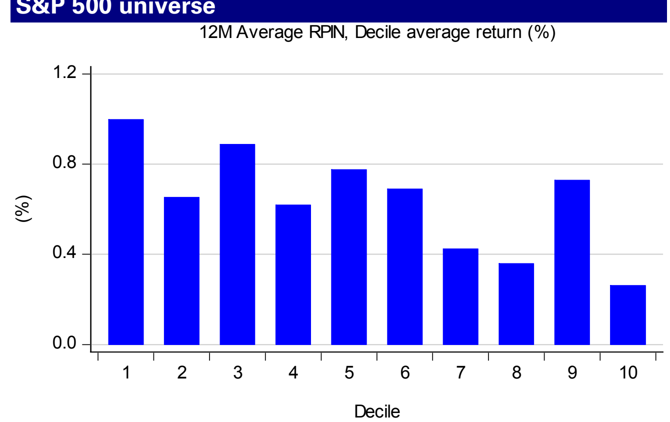
Even after controlling for biases in the *PIN* factor score via our *RPIN* factor, there is still the risk that the bulk of the performance is being driven by the small, high volatility, low liquidity subset of the market. Our first and simplest test is to re-run our backtesting in the S&P 500 universe, rather than the Russell 3000. As shown in Figure 40, we find a surprising result – the average rank IC actually increases in the S&P 500 universe (from 1.58% to 2.02%). This is a good result, because the vast majority of quant factors tend to do worse for large caps compared to small caps. The average decile returns also continue to show a reasonably consistent monotonic pattern (Figure 41). These results give us comfort that the performance of *RPIN* is not exclusively a small cap phenomenon.

Figure 40: 12M average RPIN, rank IC, S&P 500 universe



Source: TAQ, Deutsche Bank

Figure 41: 12M average RPIN, average decile returns, S&P 500 universe

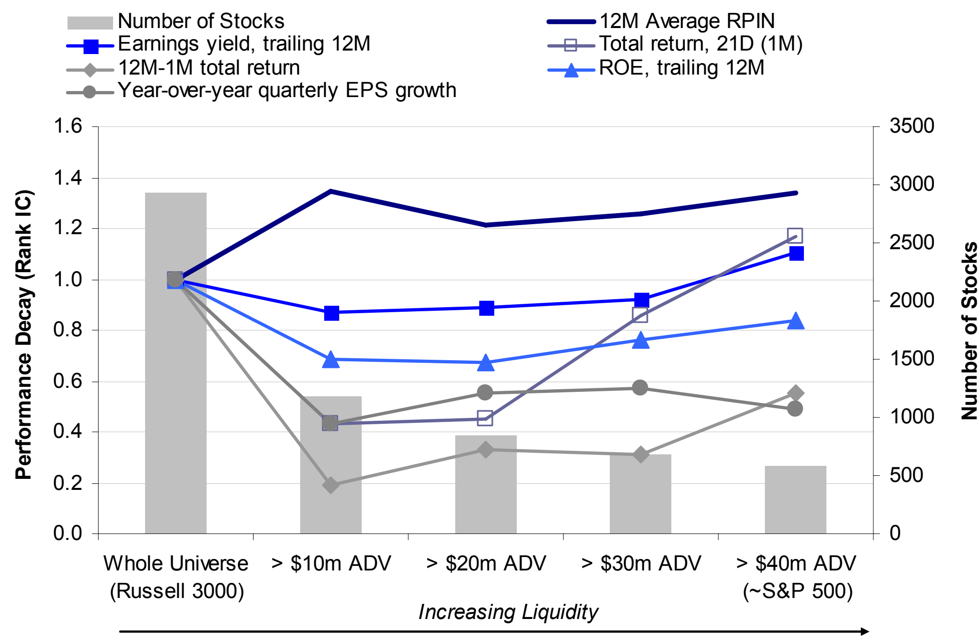


Source: TAQ, Deutsche Bank

Are we just buying illiquid stocks?

To further explore potential biases in *PIN* performance, we also look at how the performance of the factor decays as we move towards a more and more liquid universe. Figure 42 shows how the average rank IC for a number of quant factors changes as we add increasingly tight liquidity bands to the universe.

Figure 42: Factor performance decay as liquidity requirement is tightened, 2004-present (note rank IC for all factors normalized to 1 at zero liquidity constraint)



Source: TAQ, Bloomberg, Compustat, Haver, Russell, S&P, Thomson Reuters, Deutsche Bank

We find RPIN performance is not limited to illiquid stocks; in fact it actually works better in a high liquidity universe

To generate the first data point in the chart, we backtest each factor over the whole Russell 3000 universe, and then normalize the average rank IC for each factor to 1. To generate the second data point, we re-backtest each factor in a smaller universe where we only include stocks with an average daily volume (ADV) greater the $10 million. We repeat this process for constraints of $20m, $30m, and $40m. As the chart shows, adding these liquidity constraints reduces the average number of stocks in the universe from 3,000 (no constraint) to 500 (> $40m constraint).

The interesting result is that while most common factors tend to lose efficacy as we move towards a more liquid universe, *RPIN* actually improves. This is a very promising finding, because it is extremely difficult to find factors that work better for large caps than small caps. The results also confirm that our *RPIN* factor is doing a reasonably good job of generating returns across the investment universe, not just in small cap, illiquid names. This in turn suggests *RPIN* is capturing an underlying anomaly, and is not just proxying for illiquidity or size.

Correlation analysis

The most important question is whether PIN proxies for information already captured by other factors

As always, one of the biggest questions with any new factor is how it correlates with existing factors. Even the most exciting new factor is redundant if it just captures information already contained in the standard set of quant factors. Figure 43 shows the biggest negative and positive correlations with 12M average *RPIN*, where correlation is measured as the time-series correlation of monthly rank ICs.

The results are quite interesting. We find a strong negative correlation with beta and Merton's distance to default model. Both these factors on average buy low volatility stocks and short high volatility stocks, so this finding is attractive because it suggests that information risk, as captured by *PIN*, is different to the way we usually think about risk, i.e. in terms of volatility. Put another way, *PIN* is not just another way to measure volatility.

We find PIN measures something different from traditional risk as measured by volatility

The largest positive correlations tend to be with value factors. In other words, buying low *PIN* and buying cheap stocks are somewhat similar in terms of performance. This is interesting, because it suggests that more expensive stocks tend to have higher *PIN*. We don't have a perfect explanation for why this might be – perhaps expensive, "glamour" type stocks are more likely to be the focus of those trading on private information, or indeed have more scope to generate private information in the first place.

Figure 43: Biggest positive and negative time-series rank IC correlations with 12M average RPIN

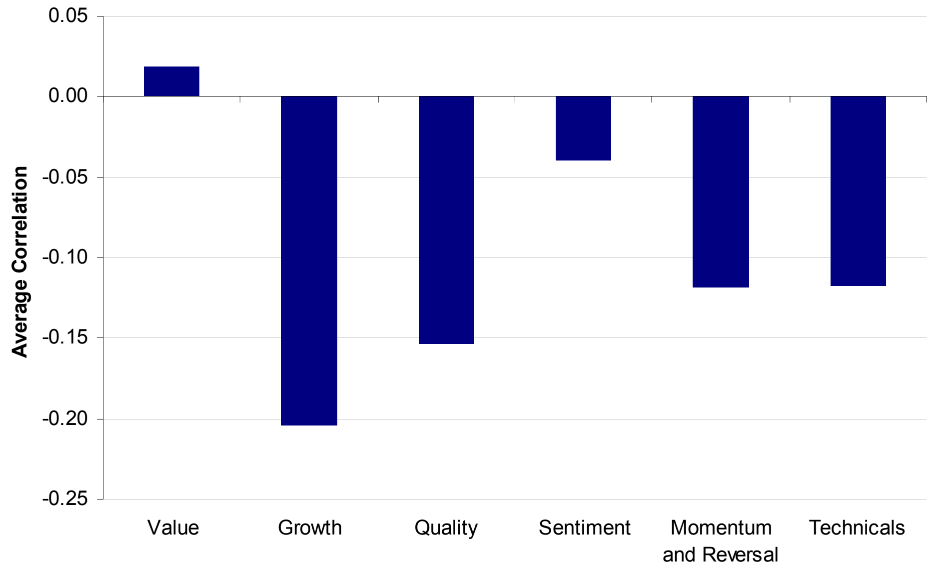
BIGGEST NEGATIVE CORRELATIONS		BIGGEST POSITIVE CORRELATIONS	
Factor	**Time-Series Rank IC Correlation**	**Factor**	**Time-Series Rank IC Correlation**
Operating profit margin	-0.63	Altman's z-score	0.50
Mohanram's G-score	-0.57	Price-to-sales, trailing 12M	0.47
IBES FY1 EPS dispersion	-0.55	Cash flow yield, FY1 mean	0.41
IBES FY2 mean DPS growth	-0.54	Target price implied return	0.33
Price-to-book adj for ROE, sector adj	-0.53	Sales to total assets (asset turnover)	0.30
IBES 5Y EPS stability	-0.51	Price-to-book	0.30
CAPM beta, 5Y monthly	-0.50	YoY change in debt outstanding	0.29
IBES 5Y EPS growth/stability	-0.47	Long-term debt/equity	0.26
Ohlson default model	-0.42	Earnings yield x IBES 5Y growth	0.23
Merton's distance to default	-0.41	# of month in the database	0.21

Source: TAQ, Bloomberg, Compustat, Haver, Russell, S&P, Thomson Reuters, Deutsche Bank

PIN has a negative correlation on average with five of our six style buckets

At a broader level, we find that our *PIN* factor tends to have a reasonably low (and indeed negative) correlation with most factors in our "standard" library. Figure 44 shows the average correlation of 12M average *RPIN* with every other factor in each broad style bucket. Interestingly, the correlation is negative for five of the six styles, and only marginally positive for value. Of course, some of this negative correlation is a reflection of the fact that over the shorter backtesting period we look at in this study (we are limited by our intraday data history), most of the common quant styles underperformed whereas *RPIN* outperformed. Nonetheless, we do think this negative correlation is promising because it does suggest we can build somewhat orthogonal factors from intraday data.

Figure 44: Average correlation with all factors in each style bucket



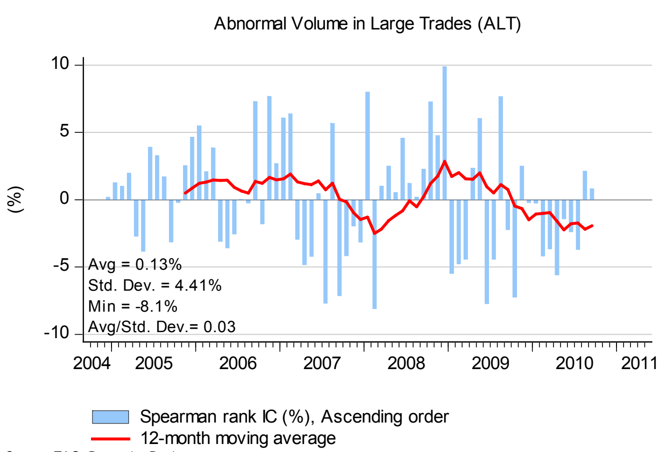
Source: TAQ, Bloomberg, Compustat, Haver, Russell, S&P, Thomson Reuters, Deutsche Bank

Abnormal Volume in Large Trades

All three ALT factors we test perform relatively poorly in backtesting

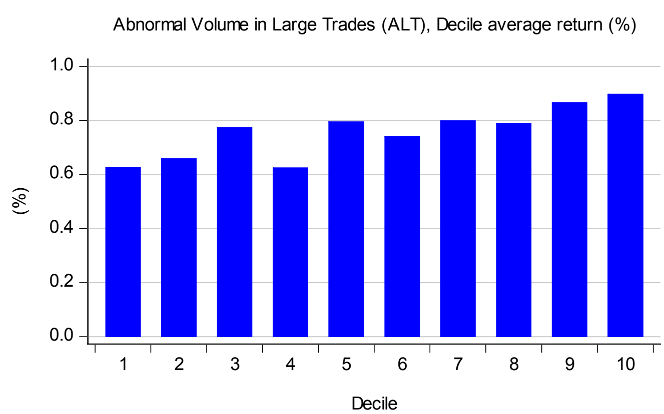
The third factor we test in this paper is the Abnormal Volume in Large Trades, or *ALT*. Figure 45 shows the monthly rank IC for the factor, and Figure 46 shows the average monthly decile returns. Overall, *ALT* does not appear to be a particularly good factor. The long-term average IC is only marginally above zero. The average decile portfolio returns are slightly more promising, with a reasonably consistent monotonic pattern, but the poor rank IC suggests these are being driven by a limited number of outlier returns.

Figure 45: ALT, rank IC



Source: TAQ, Deutsche Bank

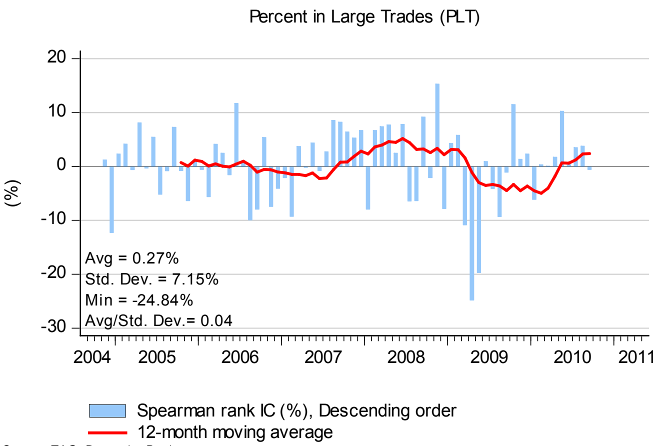
Figure 46: ALT, average decile returns



Source: TAQ, Deutsche Bank

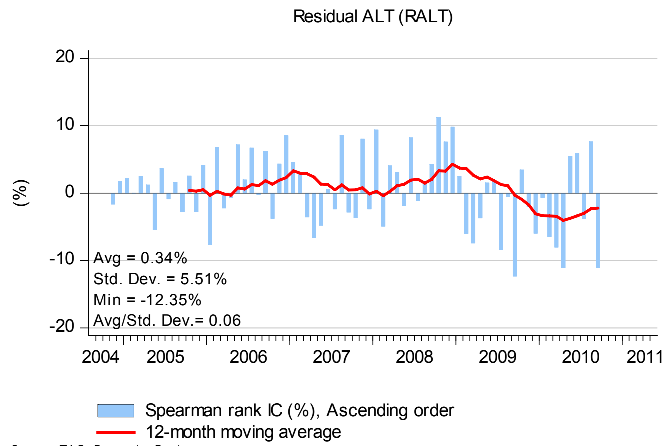
We also backtest the two alternative definitions of *ALT*, the Percent of Large Trades (*PLT*) and residual *ALT* (*RALT*). Unfortunately using these alternative definitions does not improve performance significantly. Of the two variations, *RALT* does the best, but even so the average rank IC of 0.34% is poor even when judged against the fairly weak performance of most traditional factors in recent years.

Figure 47: PLT, rank IC



Source: TAQ, Deutsche Bank

Figure 48: RALT, rank IC

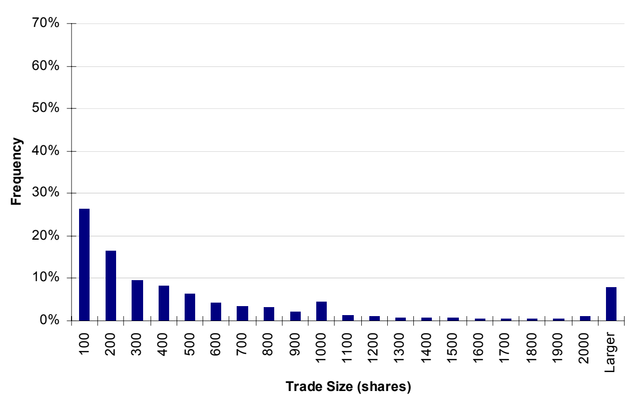


Source: TAQ, Deutsche Bank

The biggest problem with ALT is the shift to smaller, more homogeneous trade sizes

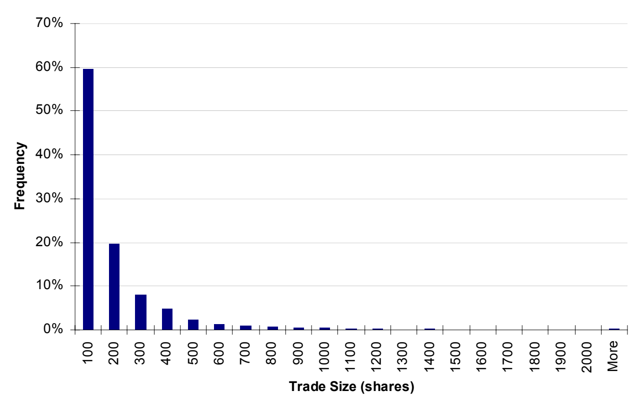
We think the biggest problem with *ALT* is the big shift towards electronic trading and alternative venues (e.g. dark pools). If we look at the distribution of trade sizes for a large cap stock (in this case IBM) for one week at the start and end of our sample period, we see a dramatic shift. Even back in 2005 (Figure 49) there was a reasonable distribution of trade sizes. Compare this to 2010 (Figure 50). Now the vast majority of trades are in 100 share blocks, and there are almost no trades in blocks greater than 500 shares. This makes *ALT* a somewhat meaningless measure, because it suggests even informed traders can now trade without revealing themselves through large trades.

Figure 49: Distribution of trade sizes, IBM, 1 week period at end of September 2005



Source: TAQ, Deutsche Bank

Figure 50: Distribution of trade sizes, IBM, 1 week period at end of September 2010



Source: TAQ, Deutsche Bank

Given these results, we do not pursue *ALT* further in this report, and instead turn our attention to testing *RPIN* in a real-world portfolio setting.

Real-world portfolio simulation

We conduct a real-world portfolio simulation to assess the efficacy of the RPIN factor

As a final test of our high frequency signals in a more real-world setting, we carry out a portfolio simulation with realistic constraints and transaction costs. Given the results from our univariate backtesting, we focus on the 12M average *RPIN* factor in this analysis. Our basic framework is to take a generic quant alpha model and assess the incremental performance gain from adding the *RPIN* signal as an additional factor in the alpha model.

We compare the performance of a multifactor model with and without including RPIN as one of the factors

Our generic alpha model is a five-factor model with the following factors: trailing earnings yield, 1M reversal, 12M-1M price momentum, year-on-year EPS growth, and ROE. We equally weight each factor to construct the final alpha signal. Each month we use the Axioma portfolio optimizer to construct a long/short portfolio targeting 5% tracking error, with reasonable sector-neutrality constraints and a beta neutrality constraint. In optimizing the portfolio, we seek to maximize expected returns with a transaction cost penalty, and in measuring the performance we also charge transaction costs. We use a simple linear costs assumption of 20bps one-way (i.e. we charge this twice for a rebalance, once for the sale of the old position, and once for the purchase of the new position). We constraint turnover to be no more than 600% p.a. two-way.

In addition to the generic model, we test a six-factor model where we add in the 12M average *RPIN* factor as a sixth factor in the model. All other backtesting parameters remain the same. Figure 51 shows the after-cost performance statistics for each backtest over the Russell 3000 universe from 2004-present.

Figure 51: Performance statistics for market neutral optimized portfolios, 2004-present, Russell 3000 universe

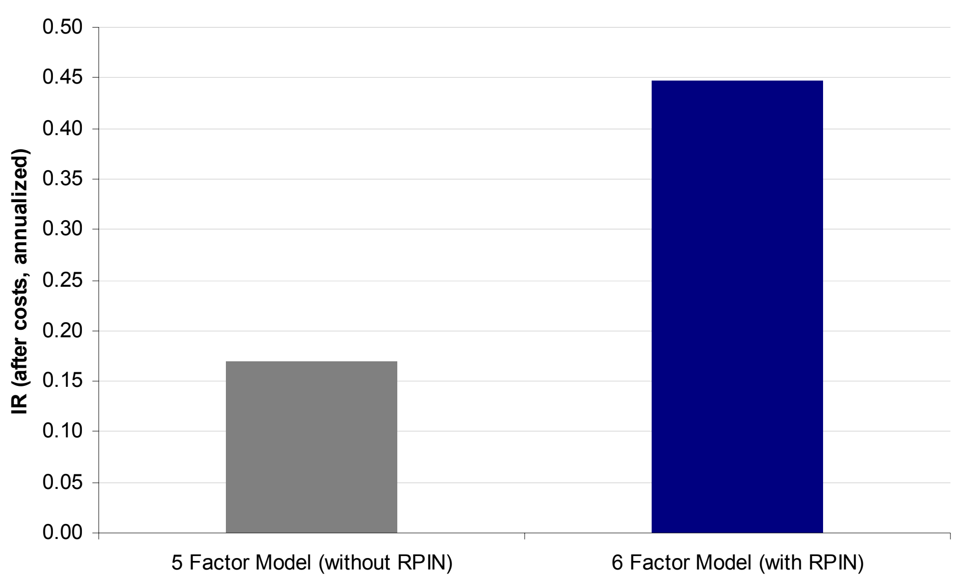
	Return (annualized, after costs)	Standard Deviation (annualized, after costs)	Information Ratio (annualized, after costs)	Turnover (annualized, two-way)	Transfer Coefficient (average)
5 Factor Model (without RPIN)	1.08%	6.40%	0.17	600%	0.50
6 Factor Model (with RPIN)	2.67%	5.97%	0.45	600%	0.47

Source: TAQ, Bloomberg, Compustat, Haver, Russell, S&P, Thomson Reuters, Deutsche Bank

We find the model with RPIN performance better in both absolute and risk-adjusted terms

Overall, adding *RPIN* to the model significantly improved performance in both absolute and risk-adjusted returns. The annualized information ratio (after costs) goes from 0.17 to 0.45 (Figure 52). Admittedly, over this period the generic five-factor model is a low hurdle because these five factors – like most traditional quant factors – underperformed severely over the latter part of the backtest period.

Figure 52: Information ratio (after costs, annualized), 2004-present, Russell 3000 universe



Source: Deutsche Bank

Indeed, the short backtest period is one of the biggest problems with assessing the efficacy of the *RPIN* factor relative to the standard quant library. Should we assume that the traditional factors will make a comeback, in which case *RPIN* will face a much higher hurdle rate in the future? Or are the "good old days" of quant gone forever, in which case *RPIN* appears to measure up reasonably well compared to the deteriorating performance of the traditional factors? We tend to subscribe to the latter point of view, and as a result we think that high frequency data is well worth a look for those seeking new and differentiated alpha sources.

Further analysis and future research

Abnormal options volume as a proxy for informed trading

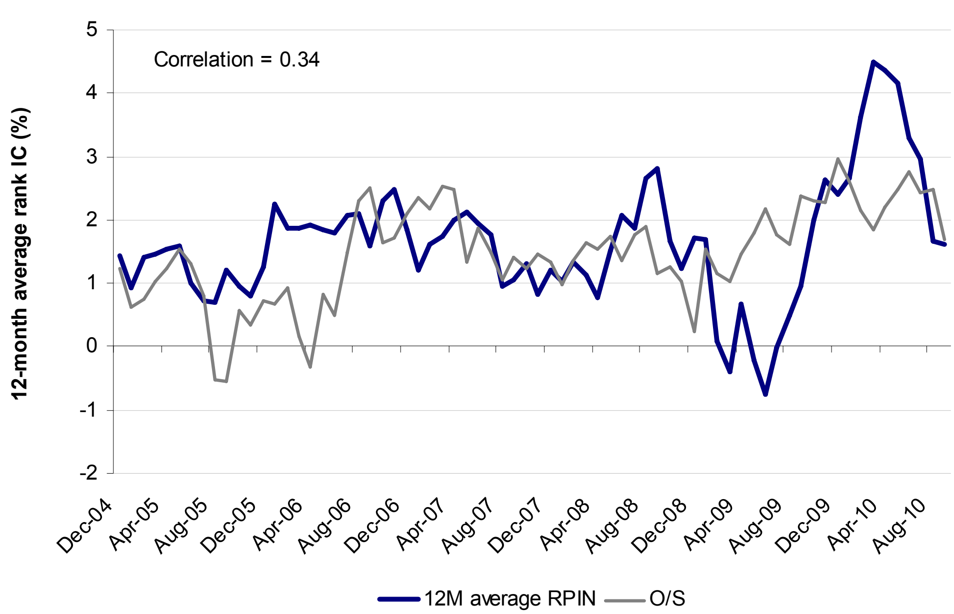
Is PIN related to abnormal options volume, which we hypothesize is another way to measure information risk?

In our recent research on options data (Cahan et al. [2010a]) we looked at an interesting factor called the O/S ratio. This ratio is simply the dollar value of options traded on a given day, divided by the dollar value of stock traded on the same day.[9] We found that the O/S ratio is a good negative predicator of one-month-ahead stock returns. One of our hypotheses was that the O/S ratio is a proxy for information risk. We argued that stocks with high abnormal options volume are potentially stocks with heavy information-based trading (since it is often argued that options traders tend to be more informed on average than stock traders). Hence we concluded that the underperformance of stocks with heavy options volume could be the same thing as the underperformance of stocks with high information risk. Now that we have computed *PIN*, we have a direct way of testing this hypothesis.

We find a correlation of 0.34, which suggest the two factors are similar but not identical

Figure 53 shows the 12-month average rank IC for our 12M average *RPIN* factor and our O/S factor. The time-series correlation is around 0.34, which is high enough to suggest that *PIN* and O/S do capture some of the same information, but low enough to suggest we could use both factors in a model without too much multicollinearity. However, for those quantitative investors without the resources to integrate intraday data, the O/S ratio may be a lower-cost alternative for capturing information risk.

Figure 53: 12-month average rank IC for 12M average RPIN factor and O/S factor



Source: TAQ, eDerivatives, Deutsche Bank

[9] Our specific definition takes the average of the O/S ratio over the past 21 trading days, and then normalizes that by the average of the O/S ratio over the past 252 trading days. Essentially this captures "abnormal" options volume in the last month.

Future research: PIN and the news

The interaction of news and PIN is something we would like to explore in the future

Another interesting set of factors we considered recently were those derived from news sentiment. In Cahan et al. [2010b] we showed how to use advanced non-linear models to extract short-term alpha out of news sentiment. However, we also believe news sentiment can be a useful conditioning tool for other factors, and we think there is a potential interesting cross-over between news sentiment and *PIN*.

For example, *PIN* is tied to the idea of information events which cause potential imbalances in order flow depending on whether they are private or public knowledge. This raises an interesting question – can we use news sentiment to determine directly which information events are public (presumably events where we have a news story on the day with sentiment in the direction of the order imbalance) versus those that are private (perhaps days where there is an order imbalance but no news on the day). Perhaps such techniques would allow us to construct a more accurate *PIN* measure? This is definitely an area for future research.

Future research: PIN as a risk management tool

A fascinating new paper by Easley, Lopez de Prado, and O'Hara [2010] uses a modified version of *PIN*, called Volume-Synchronized Probability of Informed Trading (*VPIN*), to analyze the so-called "flash crash" on May 6th, 2010. They show how *VPIN* can be used to measure order flow "toxicity" from the perspective of a liquidity provider. As a result, they argue that *VPIN* could be used to predict when liquidity providers are likely to withdraw from the market, an action that could lead to the type of rapid collapse in prices seen on May 6th. The authors show compelling evidence that *VPIN* did indeed peak at extremely high levels *before* the crash actually started, and hence could serve as an early warning sign for future such events. We think this is an excellent illustration of how data from the high frequency world can be useful even for lower frequency investors, whom are equally likely to be affected by events like those on May 6th.

References

Asquith, P., R. Oman, and C. Safaya, 2010, "Short sales and trade classification algorithms", *Journal of Financial Markets*, Volume 13, Issue 1

Bessembinder, H., 2003, "Issues in assessing trade execution costs", *Journal of Financial Markets*, Volume 6, Number 3

Boehmer, E., J. Grammig, and E. Theissen, 2006, "Estimating the probability of informed trading – Does trade misclassification matter?", *Journal of Financial Markets*, Volume 10, Issue 1

Cahan, R., Y. Luo, J. Jussa, and M. Alvarez, 2010a, "Signal Processing: The options issue", *Deutsche Bank Quantitative Strategy*, 12 May 2010

Cahan, R., Y. Luo, J. Jussa, and M. Alvarez, 2010b, "Signal Processing: Beyond the headlines", *Deutsche Bank Quantitative Strategy*, 19 July 2010

Chung, Y. P. and T. Kim, 2010, "Why do stocks move together? Evidence from commonality in order imbalances", SSRN working paper, available at http://ssrn.com/abstract=1678151

Easley, D., S. Hvidkjaer, and M. O'Hara, 2002, "Is information risk a determinant of asset returns?", *Journal of Finance*, Volume 57, Number 5

Easley, D., S. Hvidkjaer, and M. O'Hara, 2010, "Factoring information into returns", *Journal of Financial and Quantitative Analysis*, Volume 45, Number 2

Easley, D, M. Lopez de Prado, and M. O'Hara, 2010, "The microstructure of the 'flash crash': Flow toxicity, liquidity crashes and the probability of informed trading", SSRN working paper, available at http://ssrn.com/abstract=1695041

Easley, D., N. Kiefer, and M. O'Hara, 1997, "One day in the life of a very common stock", *Review of Financial Studies*, Volume 10, Number 3

Ellis, K. R. Michaely, and M. O'Hara, 2000, "The accuracy of trade classification rules: Evidence from the Nasdaq",*Journal of Financial and Quantitative Analysis*, Volume 35, Number 4

Ficucane, T., 2000, "A direct test of methods for inferring trade direction from intra-day data", *Journal of Financial and Quantitative Analysis*, Volume 35, Issue 4

Hwang, C. and X. Qian, 2010, "Is information risk priced? Evidence from the price discovery of large trades", SSRN working paper, available at http://ssrn.com/abstract=1688229

Lee, C. and M. Ready, 1991, "Inferring trade direction from intraday data", *Journal of Finance*, Volume 46, Number 2

Luo, Y., R. Cahan, J. Jussa, and M. Alvarez, 2010a, "Signal Processing: Industry-specific factors", *Deutsche Bank Quantitative Strategy*, 8 June 2010

Luo, Y., R. Cahan, J. Jussa, and M. Alvarez, 2010b, "Portfolios Under Construction: Volatility = 1/N", *Deutsche Bank Quantitative Strategy*, 16 June 2010

Tong, Q., 2009, "Abnormal volume in large trades and the cross-section of expected stock returns", Emory University working paper, available at http://www.business.smu.edu.sg/disciplines/finance/Campus%20Visit/Papers/QingTong_04Dec09.pdf

Appendix 1

Important Disclosures

Additional information available upon request

For disclosures pertaining to recommendations or estimates made on a security mentioned in this report, please see the most recently published company report or visit our global disclosure look-up page on our website at http://gm.db.com/ger/disclosure/DisclosureDirectory.eqsr**.**

Analyst Certification

The views expressed in this report accurately reflect the personal views of the undersigned lead analyst(s). In addition, the undersigned lead analyst(s) has not and will not receive any compensation for providing a specific recommendation or view in this report. Rochester Cahan/Yin Luo/Javed Jussa

Hypothetical Disclaimer

Backtested, hypothetical or simulated performance results discussed on page 10 herein and after have inherent limitations. Unlike an actual performance record based on trading actual client portfolios, simulated results are achieved by means of the retroactive application of a backtested model itself designed with the benefit of hindsight. Taking into account historical events the backtesting of performance also differs from actual account performance because an actual investment strategy may be adjusted any time, for any reason, including a response to material, economic or market factors. The backtested performance includes hypothetical results that do not reflect the reinvestment of dividends and other earnings or the deduction of advisory fees, brokerage or other commissions, and any other expenses that a client would have paid or actually paid. No representation is made that any trading strategy or account will or is likely to achieve profits or losses similar to those shown. Alternative modeling techniques or assumptions might produce significantly different results and prove to be more appropriate. Past hypothetical backtest results are neither an indicator nor guarantee of future returns. Actual results will vary, perhaps materially, from the analysis.

Regulatory Disclosures

1. Important Additional Conflict Disclosures

Aside from within this report, important conflict disclosures can also be found at https://gm.db.com/equities under the "Disclosures Lookup" and "Legal" tabs. Investors are strongly encouraged to review this information before investing.

2. Short-Term Trade Ideas

Deutsche Bank equity research analysts sometimes have shorter-term trade ideas (known as SOLAR ideas) that are consistent or inconsistent with Deutsche Bank's existing longer term ratings. These trade ideas can be found at the SOLAR link at http://gm.db.com.

3. Country-Specific Disclosures

Australia: This research, and any access to it, is intended only for "wholesale clients" within the meaning of the Australian Corporations Act.
EU countries: Disclosures relating to our obligations under MiFiD can be found at http://globalmarkets.db.com/riskdisclosures.
Japan: Disclosures under the Financial Instruments and Exchange Law: Company name - Deutsche Securities Inc. Registration number - Registered as a financial instruments dealer by the Head of the Kanto Local Finance Bureau (Kinsho) No. 117. Member of associations: JSDA, The Financial Futures Association of Japan. Commissions and risks involved in stock transactions - for stock transactions, we charge stock commissions and consumption tax by multiplying the transaction amount by the commission rate agreed with each customer. Stock transactions can lead to losses as a result of share price fluctuations and other factors. Transactions in foreign stocks can lead to additional losses stemming from foreign exchange fluctuations. "Moody's", "Standard & Poor's", and "Fitch" mentioned in this report are not registered as rating agency in Japan unless specifically indicated as Japan entities of such rating agencies.
New Zealand: This research is not intended for, and should not be given to, "members of the public" within the meaning of the New Zealand Securities Market Act 1988.
Russia: This information, interpretation and opinions submitted herein are not in the context of, and do not constitute, any appraisal or evaluation activity requiring a license in the Russian Federation.

Deutsche Bank Securities Inc.

North American location

Deutsche Bank Securities Inc.
60 Wall Street
New York, NY 10005
Tel: (212) 250 2500

Deutsche Bank Securities Inc.
One International Place
12th Floor
Boston, MA 02110
United States of America
Tel: (1) 617 217 6100

Deutsche Bank Securities Inc.
222 South Riverside Plaza
30th Floor
Chicago, IL 60606
Tel: (312) 537-3758

Deutsche Bank Securities Inc.
3033 East First Avenue
Suite 303, Third Floor
Denver, CO 80206
Tel: (303) 394 6800

Deutsche Bank Securities Inc.
1735 Market Street
24th Floor
Philadelphia, PA 19103
Tel: (215) 854 1546

Deutsche Bank Securities Inc.
101 California Street
46th Floor
San Francisco, CA 94111
Tel: (415) 617 2800

Deutsche Bank Securities Inc.
700 Louisiana Street
Houston, TX 77002
Tel: (832) 239-4600

International Locations

Deutsche Bank Securities Inc.
60 Wall Street
New York, NY 10005
United States of America
Tel: (1) 212 250 2500

Deutsche Bank AG London
1 Great Winchester Street
London EC2N 2EQ
United Kingdom
Tel: (44) 20 7545 8000

Deutsche Bank AG
Große Gallusstraße 10-14
60272 Frankfurt am Main
Germany
Tel: (49) 69 910 00

Deutsche Bank AG
Deutsche Bank Place
Level 16
Corner of Hunter & Phillip Streets
Sydney, NSW 2000
Australia
Tel: (61) 2 8258 1234

Deutsche Bank AG
Level 55
Cheung Kong Center
2 Queen's Road Central
Hong Kong
Tel: (852) 2203 8888

Deutsche Securities Inc.
2-11-1 Nagatacho
Sanno Park Tower
Chiyoda-ku, Tokyo 100-6171
Japan
Tel: (81) 3 5156 6770

Disclaimer

The information and opinions in this report were prepared by Deutsche Bank AG or one of its affiliates (collectively "Deutsche Bank"). The information herein is believed to be reliable and has been obtained from public sources believed to be reliable. Deutsche Bank makes no representation as to the accuracy or completeness of such information.

Deutsche Bank may engage in securities transactions, on a proprietary basis or otherwise, in a manner **inconsistent** with the view taken in this research report. In addition, others within Deutsche Bank, including strategists and sales staff, may take a view that is **inconsistent** with that taken in this research report.

Opinions, estimates and projections in this report constitute the current judgement of the author as of the date of this report. They do not necessarily reflect the opinions of Deutsche Bank and are subject to change without notice. Deutsche Bank has no obligation to update, modify or amend this report or to otherwise notify a recipient thereof in the event that any opinion, forecast or estimate set forth herein, changes or subsequently becomes inaccurate. Prices and availability of financial instruments are subject to change without notice. This report is provided for informational purposes only. It is not an offer or a solicitation of an offer to buy or sell any financial instruments or to participate in any particular trading strategy. Target prices are inherently imprecise and a product of the analyst judgement.

As a result of Deutsche Bank’s recent acquisition of BHF-Bank AG, a security may be covered by more than one analyst within the Deutsche Bank group. Each of these analysts may use differing methodologies to value the security; as a result, the recommendations may differ and the price targets and estimates of each may vary widely.

Deutsche Bank has instituted a new policy whereby analysts may choose not to set or maintain a target price of certain issuers under coverage with a Hold rating. In particular, this will typically occur for "Hold" rated stocks having a market cap smaller than most other companies in its sector or region. We believe that such policy will allow us to make best use of our resources. Please visit our website at http://gm.db.com to determine the target price of any stock.

The financial instruments discussed in this report may not be suitable for all investors and investors must make their own informed investment decisions. Stock transactions can lead to losses as a result of price fluctuations and other factors. If a financial instrument is denominated in a currency other than an investor's currency, a change in exchange rates may adversely affect the investment.

All prices are those current at the end of the previous trading session unless otherwise indicated. Prices are sourced from local exchanges via Reuters, Bloomberg and other vendors. Data is sourced from Deutsche Bank and subject companies.

Past performance is not necessarily indicative of future results. Deutsche Bank may with respect to securities covered by this report, sell to or buy from customers on a principal basis, and consider this report in deciding to trade on a proprietary basis.

Derivative transactions involve numerous risks including, among others, market, counterparty default and illiquidity risk. The appropriateness or otherwise of these products for use by investors is dependent on the investors' own circumstances including their tax position, their regulatory environment and the nature of their other assets and liabilities and as such investors should take expert legal and financial advice before entering into any transaction similar to or inspired by the contents of this publication. Trading in options involves risk and is not suitable for all investors. Prior to buying or selling an option investors must review the "Characteristics and Risks of Standardized Options," at http://www.theocc.com/components/docs/riskstoc.pdf If you are unable to access the website please contact Deutsche Bank AG at +1 (212) 250-7994, for a copy of this important document.

The risk of loss in futures trading, foreign or domestic, can be substantial. As a result of the high degree of leverage obtainable in futures trading, losses may be incurred that are greater than the amount of funds initially deposited.

Unless governing law provides otherwise, all transactions should be executed through the Deutsche Bank entity in the investor's home jurisdiction. In the U.S. this report is approved and/or distributed by Deutsche Bank Securities Inc., a member of the NYSE, the NASD, NFA and SIPC. In Germany this report is approved and/or communicated by Deutsche Bank AG Frankfurt authorized by the BaFin. In the United Kingdom this report is approved and/or communicated by Deutsche Bank AG London, a member of the London Stock Exchange and regulated by the Financial Services Authority for the conduct of investment business in the UK and authorized by the BaFin. This report is distributed in Hong Kong by Deutsche Bank AG, Hong Kong Branch, in Korea by Deutsche Securities Korea Co. This report is distributed in Singapore by Deutsche Bank AG, Singapore Branch, and recipients in Singapore of this report are to contact Deutsche Bank AG, Singapore Branch in respect of any matters arising from, or in connection with, this report. Where this report is issued or promulgated in Singapore to a person who is not an accredited investor, expert investor or institutional investor (as defined in the applicable Singapore laws and regulations), Deutsche Bank AG, Singapore Branch accepts legal responsibility to such person for the contents of this report. In Japan this report is approved and/or distributed by Deutsche Securities Inc. The information contained in this report does not constitute the provision of investment advice. In Australia, retail clients should obtain a copy of a Product Disclosure Statement (PDS) relating to any financial product referred to in this report and consider the PDS before making any decision about whether to acquire the product. Deutsche Bank AG Johannesburg is incorporated in the Federal Republic of Germany (Branch Register Number in South Africa: 1998/003298/10). Additional information relative to securities, other financial products or issuers discussed in this report is available upon request. This report may not be reproduced, distributed or published by any person for any purpose without Deutsche Bank's prior written consent. Please cite source when quoting.