A piecewise linear model for trade sign inference

Adam Blazejewski [a,*], Richard Coggins [b]

[a] *School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia*

[b] *Discipline of Finance, School of Business, University of Sydney, NSW 2006, Australia*

Abstract

We use transaction level data for twelve stocks with large market capitalization on the Australian Stock Exchange to develop an empirical model for trade sign (trade initiator) inference. The new model is a piecewise linear parameterization of the model proposed recently in Ref. [1]. The space of the predictor variables is partitioned into six regions. Signs of individual trades within the regions are inferred according to simple and interpretable rules. Across the 12 stocks the new model achieves an average out-of-sample classification accuracy of 74.38% (SD=4.25%), which is 2.98% above the corresponding accuracy reported in Ref. [1]. Two of the model's regions, together accounting for 16.79% of the total number of daily trades, have each an average classification accuracy exceeding 91.50%. The results indicate a strong dependence between the predictor variables and the trade sign, and provide evidence for an endogenous component in the order flow. An interpretation of the trade sign classification accuracy within the model's regions offers new insights into a relationship between two regularities observed in the markets with a limit order book, competition for order execution and transaction cost minimization.

Key words: Order submission, Trade classification, Piecewise linear, Competition for order execution, Transaction cost minimization
PACS: 89.65.Gh, 89.75.Kd, 05.45.Tp, 05.45.-a

* Corresponding author. Tel.: +61-2-9351-3229; fax: +61-2-9351-3847.
E-mail address: adamb@sedal.usyd.edu.au (A. Blazejewski).

1 INTRODUCTION

The Australian Stock Exchange (ASX) is a limit order market without specialists. It implements a continuous double auction via an electronic limit order book. Every type of transaction on the stock exchange is recorded in an electronic file. Similar electronic systems are employed by other stock exchanges with a limit order book, for example the Paris Bourse, the London Stock Exchange, and the Tokyo Stock Exchange. The recorded single event data allow researchers to analyze the trading process at an ultra-high frequency. In particular, it becomes possible to reconstruct the complete order flow, which in turn enables formulation and testing of event level models of order submission strategies and price formation.

Recent empirical studies have shown that the order flow depends on the state of the limit order book [2–11]. Most of the observed dependence seems to be caused by two behavioral regularities in order submission strategies. The first regularity concerns competition for order execution. If one side of the book is dominant, where the dominant side is the one with more depth [1] , then there is an imbalance between supply and demand, and limit orders on the dominant side face a longer time to execution [3] and a higher risk of an adverse price movement leading to non-execution. Consequently, traders on the same side of the market as the dominant side of the book are more likely to submit market orders to achieve an immediate execution [4, 8–11]. The behavior of buyers and sellers, however, may not be perfectly symmetrical [4, 12]. The second regularity is called by us transaction cost minimization. It is observed that a majority of individual market orders consume only a part or the whole volume available at the best price in the order book [2–8, 10]. Traders try to minimize their transaction costs and by following this regularity ensure that the price per share of their trades will differ from the pre-trade mid-point price by the value of half spread only. Apart from the two regularities other empirical studies found an autocorrelation in the unconditional and conditional order flow, where similar events tend to follow one another [1–3, 7, 8, 13–15]. In particular, Refs. [13–16] reported an autocorrelation in the trade sign [2] , Ref. [13] presented evidence for the long memory in the trade sign, while Ref. [16] found the long memory in the order flow. The works of Porter [17] and Aitken et al. [18], on the other hand, detected temporal patterns in the probability of trading at the best ask, and represent the closest

[1] The depth is measured as the volume (total number of shares) on a given side of the limit order book, usually at a single price (best price) or at a number of prices closest to the mid-point price. The mid-point price is an average of the best bid price (best bid) and the best ask price (best ask).

[2] The trade sign has also been referred to by various authors as a trade initiator, trade indicator, trade direction, or buy/sell indicator. Similar synonyms exist for the market order sign.

prior work to the study in Ref. [1], which is the starting point of our present research.

In this paper we develop a piecewise linear parameterization of the trade sign inference model proposed recently by Blazejewski and Coggins [1]. Those authors reported that a k-nearest-neighbor classifier can infer the trade sign with an average accuracy of over 71%, for a set of 12 stocks on the ASX. The classifier used three predictor variables, the volume at the best bid and at the best ask just before a trade, and the trade size. Across the 12 stocks the highest classification accuracy was achieved for a training interval of 30 days. Our new model is piecewise linear and employs the same set of the three predictor variables. We do not use trade and quote (bid and ask) prices, and our purpose is different from that of the trade classification algorithms [19–22]. Those algorithms were designed for markets where full, correctly time-stamped limit order book data, and the trade sign in particular, are not available. Our empirical model is constructed to demonstrate a strong dependence between the three predictor variables and the trade sign as evidence for an endogenous component in the order flow.

The space of the three predictor variables is partitioned into six regions. The trades within each region are signed according to rules derived from the two regularities in the order flow discussed earlier. The boundaries between the regions form a set of three partitioning planes. The coefficients of these planes are estimated over the first 30 days in the data set. The estimation procedure employs three different methods to produce three corresponding coefficient vectors. The mean in-sample daily classification accuracy is then calculated over the first 30 days. The out-of-sample estimate of this accuracy is determined over the remaining 169 days. The calculations are performed separately for each stock and each coefficient vector. We show that the new model with an intuitively interpretable set of coefficients outperforms the k-NN classifier and achieves an average out-of-sample accuracy of 74.38% (SD=4.25%). We also find that two of the six regions which represent, on average, 16.79% of the total number of daily trades, have each an average classification accuracy exceeding 91.50%.

2 DATA SET

We use a data set for 12 stocks on the Australian Stock Exchange (ASX). The same data set has been used previously in Ref. [1] to develop a local non-parametric model for trade sign inference. The data were collected over 199 trading days, between 11 November 2002 and 27 August 2003. The twelve stocks were selected from the 30 stocks with the largest market capitalization during the period considered, while 8 of our stocks belonged to the top 10.

All of the selected stocks were traded on each day in the data set and they did not experience any major price revisions. The three letter institutional codes of the stocks, ordered by decreasing market capitalization, are: NAB, BHP, CBA, ANZ, WBC, NCP, RIO, WOW, FGL, SUN, SGB, MIG. Our data set contains information on all orders entered, amended, and deleted in the limit order book, as well as all trades transacted through the order book [3], as recorded by the ASX for the selected stocks during the investigated period. Each order has three main attributes: side (buy or sell), size (volume), and price. The true size of orders with an undisclosed volume [23] is included in the data set. The three main attributes of trades are: size, price, and trade sign. All transactions have a correct time-stamp and are ordered chronologically. The data set contains 2,355,334 trades. Our analysis is restricted to buyer-initiated and seller-initiated trades only. There are 2,184,046 such trades in the data set (92.73%), out of which 50.44% are buyer-initiated. Trades resulting from the same market order are aggregated together as in Ref. [1], whereby an aggregated trade becomes a proxy for that market order [4]. After aggregation we have 1,542,205 buyer-initiated and seller-initiated trades, with 51.78% of them being buyer-initiated. The first five aggregated trades on each day are omitted in order to use the same set of trades as in Ref. [1] [5].

3 METHODS

The trade sign inference model proposed in Ref. [1] is based on the k-nearest-neighbor, which is a local, non-parametric, memory-based classifier. That model is a starting point for the development of our parametric model. In particular, we use the same set of predictor variables as was employed in that study for the k-NN classifier with the highest predictive accuracy. This set contains the following three variables:

- a_{n} - total volume at the best ask in the limit order book, recorded just before an order which caused the n-th trade; n is an index of aggregated trades over a single day.
- b_{n} - total volume at the best bid in the limit order book, recorded just before an order which caused the n-th trade.
- s_{n} - size of the n-th trade.

[3] Trades transacted outside the order book, called off-market, are excluded from our analysis.

[4] We do not make a distinction between market orders and marketable limit orders. The latter are limit orders priced for immediate execution.

[5] That study used the omitted trades to obtain lagged values of model variables.

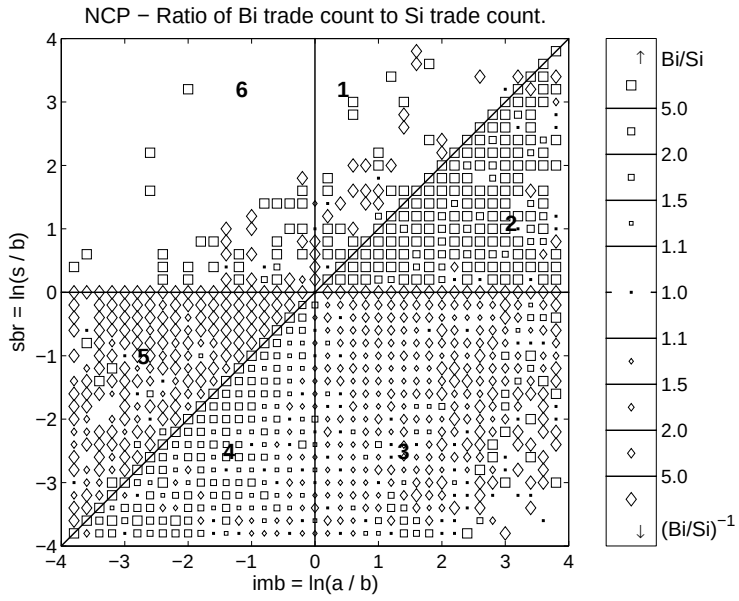

Fig. 1. Ratio of trade counts for buyer-initiated (Bi) and seller-initiated (Si) trades, for the first 30 trading days of the NCP stock. Six classification regions r_i, $i = 1 \ldots 6$, are numbered from 1 to 6. Bin size is 0.2x0.2.

There is one target variable (inference target), the trade sign $\epsilon_{\rm n}$. To simplify the notation the trade index n will be omitted in the remainder of the paper. The predictor variables will be denoted as a, b, and s, respectively, while ϵ will stand for the target variable. All three predictor variables are measured in the same units, number of shares. The trade sign is a binary variable which represents buyer-initiated and seller-initiated trades as $+1$ and -1, respectively.

To obtain some insight into the relationship between the three predictor variables and the target variable we constructed two types of histograms using the first 30 days in the data set, for each stock separately. The two histograms for the NCP stock in Figs. 1 and 2 qualitatively agree with the corresponding histograms for most of our stocks. The first type of histogram, shown in Fig. 1, depicts a trade count ratio. The trade count ratio is defined as a ratio of a buyer-initiated bin trade count to a seller-initiated bin trade count. A bin trade count is as a total number of trades in a given histogram bin. A trade count for buyer(seller)-initiated trades counts the trades with the specified sign only. The trade count ratio is shown as a function of the imbalance in the order book *imb* and the ratio *sbr* of the trade size s to the total volume at the best bid b. The presence of a square in a given histogram bin indicates a dominance of buyer-initiated trades, while a diamond stands for a majority of seller-initiated trades. A small dot represents approximately equal ($\pm 10\%$) trade counts. Bins without any trades are marked by a blank space without any symbol. It can be seen that there are several well defined regions, each dominated by a particular trade sign. The boundaries between the regions seem to form three straight lines, horizontal, diagonal, and vertical. To highlight

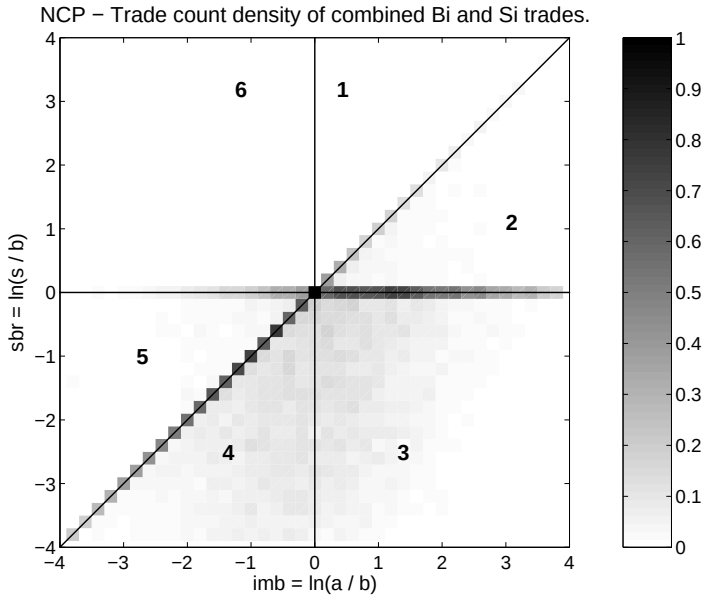

Fig. 2. Trade count density of combined buyer-initiated (Bi) and seller-initiated (Si) trades, for the first 30 trading days of the NCP stock. Six classification regions r_i, $i = 1 \ldots 6$, are numbered from 1 to 6. Bin size is 0.2x0.2.

these features three thin lines have been added to the figure. The horizontal line represents the condition $sbr = 0$, which means that $s = b$. The diagonal line can be shown to correspond to the condition $s = a$. The vertical line, on the other hand, represents the situation where $imb = 0$, which is equivalent to $a = b$. The regions delineated by the three lines have been numbered from one to six and will be denoted as r_i, $i = 1 \ldots 6$. Regions r_2 and r_4 seem to be dominated by buyer-initiated trades, while regions r_3 and r_5 both have a majority of seller-initiated trades. The other two regions, r_1 and r_6, do not have an obvious dominant trade sign. They are sparsely occupied and have approximately similar numbers of squares and diamonds. We note that the lower right corner of region r_3 does not show a clear majority either.

The second type of histogram, shown in Fig. 2, depicts a trade count density for combined buyer-initiated and seller-initiated trades. The density was calculated by dividing a given bin trade count by a total number of trades in all bins. The axes in the figure, as well as the lines separating the regions, and the region numbers are the same as in Fig. 1. The trade count density values were transformed to a relative scale from 0 to 1, and mapped to shades of grey, from white to black. The white color indicates very few or no trades, while the black stands for the highest trade count density. It is evident that parts of the horizontal and diagonal separating lines are located over areas of high trade concentration. Those areas seem to closely follow the course of the lines and do not reach further than half of the bin size away from the lines, in either axial direction. A large cluster of trades can be seen in regions r_3 and r_4. The bin density in that cluster is not as high as the density over the sep-

arating lines described above but the cluster covers a larger area. Regions r_1, r_2, r_5, and r_6 appear to have few trades beyond the areas under the horizontal and diagonal separating lines. The lower right corner of region r_3, mentioned earlier in the context of Fig. 1, has a very low trade count density too. As far as trades within bins which the separating lines cut in half are concerned it is unclear if they have a preference for which side of a given line they are on.

The analysis of the histograms revealed the existence of regions dominated by particular trade signs. The boundaries of the regions are clearly delineated and form a set of three separating lines. The only exceptions are regions r_1 and r_6. It is not obvious where the boundary line between them is located and what their dominant trade signs are in areas outside the horizontal and diagonal separating lines. However, the trade count density in these areas is very low and, consequently, their contribution to the total trade sign classification accuracy should be minimal. The discovered features and the two regularities discussed in the introduction lead us to propose the following trade sign inference model:

$$
\begin{aligned}
\epsilon &= p(\mathbf{c}) \qquad (1)\\
p(\mathbf{c}) &= \sum_{i=1}^{6} \epsilon_{\mathrm{r_i}},\ \mathbf{c} = (\alpha_a, \beta_a, \alpha_b, \beta_b)\\
\epsilon_{\mathrm{r_1}} &= \begin{cases} -1 & \text{if } s > f_D(a) \text{ and } a > f_V(b) \\ 0 & \text{otherwise} \end{cases}\\
\epsilon_{\mathrm{r_2}} &= \begin{cases} +1 & \text{if } s \leq f_D(a) \text{ and } s > f_H(b) \\ 0 & \text{otherwise} \end{cases}\\
\epsilon_{\mathrm{r_3}} &= \begin{cases} -1 & \text{if } s \leq f_H(b) \text{ and } a > f_V(b) \\ 0 & \text{otherwise} \end{cases}\\
\epsilon_{\mathrm{r_4}} &= \begin{cases} +1 & \text{if } s \leq f_D(a) \text{ and } a \leq f_V(b) \\ 0 & \text{otherwise} \end{cases}\\
\epsilon_{\mathrm{r_5}} &= \begin{cases} -1 & \text{if } s > f_D(a) \text{ and } s \leq f_H(b) \\ 0 & \text{otherwise} \end{cases}\\
\epsilon_{\mathrm{r_6}} &= \begin{cases} +1 & \text{if } s > f_H(b) \text{ and } a \leq f_V(b) \\ 0 & \text{otherwise} \end{cases}\\
f_D(a) &= \alpha_a a + \beta_a,\ \alpha_a \neq 0\\
f_H(b) &= \alpha_b b + \beta_b,\ \alpha_b \neq 0\\
f_V(b) &= \frac{\alpha_b}{\alpha_a} b + \frac{\beta_b - \beta_a}{\alpha_a}
\end{aligned}
$$

The six functions $\epsilon_{\mathrm{r_i}}$, $i = 1 \ldots 6$, define six disjoint regions in the space of the three predictor variables. Those six regions are generalizations of the regions shown in Figs. 1 and 2, and are denoted in the same way, i.e. r_i, $i = 1 \ldots 6$. Each function can assume only two values: 0 and either $+1$ or -1. For a given combination of the three predictor variables a, b, and s only one function $\epsilon_{\mathrm{r_i}}$ assumes a non-zero value due to mutually exclusive sets of conditions imposed on the predictor variables. The boundaries of the regions form a set of three boundary planes which are generalizations of the diagonal, horizontal, and vertical separating lines in the histograms. The boundary planes are defined by the following conditions: $s = f_D(a)$, $s = f_H(b)$, and $a = f_V(b)$. The conditions employ three boundary functions, $f_D(a)$, $f_H(b)$, and $f_V(b)$. The four coefficients α_a, β_a, α_b, and β_b form a coefficient vector $\mathbf{c}$ and allow us to search for the optimal locations of the separating planes. The third boundary function $f_V(b)$ has its own coefficients defined entirely in terms of the coefficients of the other two boundary functions. That constraint ensures that the three separating planes will always intersect along a single line. Consequently, the number of regions will stay fixed at six, irrespective of the values of the four coefficients, as long as $\alpha_a \neq 0$ and $\alpha_b \neq 0$.

To estimate the proposed model we will look for a coefficient vector $\mathbf{c}$ which maximizes A_p, where A_p denotes the classification accuracy of the function $p(\mathbf{c})$. We employ three methods for this purpose. The simple method is derived from the exact arrangement of the three separating lines shown in Figs. 1 and 2. Its coefficient vector $\mathbf{c_{smp}}$ is set arbitrarily to $(1, 0, 1, 0)$, resulting in the simplified boundary functions: $f_D(a) = a$, $f_H(b) = b$, $f_V(b) = b$. The second method uses the Nelder-Mead local optimizer [24–26] to maximize A_p. This algorithm performs a local search in the space of the four coefficients, starting from $\mathbf{c_{smp}}$. The result of the search is referred to as the Nelder-Mead optimized coefficient vector $\mathbf{c_{nm}}$. The third method is a global optimizer based on a recently developed particle swarm optimization algorithm [27–29]. The optimizer searches the neighborhood of $\mathbf{c_{smp}}$ with a swarm of virtual particles. The particles fly through the coefficient space and home in on the maxima of A_p. The number of particles is set to 200. The initial neighborhood size is set to $[0.5, 1.5]$ for α_a and α_b, and to $[-1, 1]$ for β_a and β_b. As the search progresses a series of new, smaller neighborhoods are constructed around the latest locally optimal vector with the highest A_p among the visited points. The process is continued until the classification accuracy can no longer be improved by a specified increment value. The best solution found is selected as the PSO optimized coefficient vector $\mathbf{c_{pso}}$.

The optimized coefficient vectors $\mathbf{c_{nm}}$ and $\mathbf{c_{pso}}$ are estimated separately by following the same procedure described below. The estimation is performed over the subset $\mathbf{E}$, which comprises the first 30 days in the data set. To prevent overfitting, where an estimated solution does not generalize to unseen data, we adopt an approach based on the ten-fold cross-validation [30, 31]. The whole

period of 30 days is divided into 10 consecutive subperiods, called folds, of the same length of 3 days. The estimation is conducted 10 times, each time on a different subset $\mathbf{E}_i$ with 27 days. During the i-th estimation, where $i = 1...10$, the i-th fold is omitted. An optimization algorithm looks for a vector $\mathbf{c}$ which maximizes A_p over a given subset $\mathbf{E}_i$. The classification accuracy is calculated as a single value over all days in $\mathbf{E}_i$, which is equivalent to calculating daily classification accuracy values and weighting them by the number of trades on corresponding days. We chose to calculate the accuracy this way because the k-NN classifier constructed in Ref. [1] achieved the best results with the training interval of 30 days. The whole process produces 10 estimates of an optimized coefficient vector, and the average optimized vector is determined by computing mean coefficient values across those 10 estimates.

Subsequently we use the subset $\mathbf{E}$ and the three coefficient vectors $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, and $\mathbf{c_{pso}}$ to calculate mean values of the in-sample daily classification accuracy, separately for each stock and each coefficient vector. The mean daily accuracy is computed as a mean of daily classification accuracy values of A_p. The same procedure is performed over the evaluation period, comprising the remaining 169 days in the data set, in order to determine the out-of-sample daily classification accuracy.

The data processing and statistical calculations in our experiment were implemented in the proprietary market surveillance and trading system called SMARTS® and on the Matlab® computing platform. Two freely available Matlab toolboxes, NETLAB [32] and PSOt [33], were also used.

4 RESULTS

The in-sample classification accuracy statistics for the piecewise linear model with the three coefficient vectors $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, and $\mathbf{c_{pso}}$ are presented in Table 1. The table also shows statistics for the coefficients of the PSO optimized vector $\mathbf{c_{pso}}$. The results were calculated using the first 30 days in the data set. The table reports the PSO optimized vectors only because all but a few of their mean coefficients are located further away from the corresponding mean coefficients of the simple vector $\mathbf{c_{smp}}$ than the respective mean coefficients of the optimized vectors found by the Nelder-Mead algorithm. In other words, the coefficients produced by the local optimizer are located closer to the coefficients of $\mathbf{c_{smp}}$. The distance between the coefficients was measured with a one-dimensional Euclidean distance metric, for each stock and coefficient separately. The same relationship exists for the average optimized vectors across the 12 stocks, where the average $\mathbf{c_{pso}}$ is (0.997379, 0.032122, 0.997895, 0.044311), while the average $\mathbf{c_{nm}}$ is (1.000255, 0.000125, 1.001402, 0.000043). The coefficients are reported with six digits after the decimal point to emphasize the differences between

Table 1
In-sample daily classification accuracy (%) and coefficients of the PSO optimized vector $\mathbf{c_{pso}}$ for individual stocks - 30 estimation days.[1]

Stock	$\mathbf{c_{smp}}$	$\mathbf{c_{nm}}$	$\mathbf{c_{pso}}$	α_a	β_a	α_b	β_b
NAB	79.65	79.70	79.69	0.999659	0.004462	0.999933	0.001693
	2.28	2.27	2.27	0.000460	0.005209	0.000090	0.001818
BHP	75.56	75.60	75.55	0.999166	0.011990	0.999938	0.010441
	3.53	3.55	3.57	0.002346	0.027007	0.001185	0.009774
CBA	72.71	72.68	72.66	1.000042	0.000883	1.000145	0.002786
	3.59	3.60	3.58	0.000203	0.001387	0.000785	0.003218
ANZ	77.83	77.84	77.86	0.999973	0.000461	1.000350	0.001773
	2.75	2.79	2.79	0.000099	0.001106	0.000451	0.005081
WBC	76.22	76.22	76.20	0.999855	0.002027	1.000309	0.001833
	3.09	3.11	3.09	0.000304	0.003714	0.000403	0.002095
NCP	76.31	76.33	76.32	0.999943	0.001015	1.000003	0.000024
	2.91	2.90	2.90	0.000201	0.002350	0.000011	0.000046
RIO	80.08	80.07	80.02	0.999712	0.006057	0.998399	0.016841
	2.86	2.80	2.92	0.000663	0.007617	0.002133	0.021967
WOW	75.70	75.74	75.72	0.999627	0.007419	0.998742	0.014167
	3.20	3.14	3.13	0.001004	0.011051	0.001879	0.021138
FGL	67.67	67.74	67.75	0.979487	0.246303	0.977927	0.416271
	6.28	6.33	6.19	0.033823	0.401725	0.033676	0.392509
SGB	77.31	77.49	77.52	0.999834	0.003600	0.999668	0.004165
	3.14	3.15	3.15	0.000675	0.006450	0.000542	0.005432
SUN	73.53	73.48	73.72	0.999999	0.000076	1.000843	0.012721
	3.68	3.71	3.87	0.000001	0.000159	0.001012	0.007791
MIG	71.14	71.29	71.13	0.991248	0.101173	0.998482	0.049018
	6.53	6.39	6.63	0.027406	0.315834	0.011447	0.119704
Av.	75.31	75.35	75.34	0.997379	0.032122	0.997895	0.044311
value	3.65	3.65	3.67	0.005599	0.065301	0.004468	0.049214

[1]For $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, and $\mathbf{c_{pso}}$ the first and the second line for each stock show means and standard deviations of the classification accuracy, respectively. For coefficients α_a, β_a, α_b, and β_b the first and the second line for each stock show means and standard deviations of the PSO optimized coefficients, respectively. The optimized coefficients were calculated across the 10 subsets with 27 days, over the 30 estimation days. The first and the second line of averages show average values of the corresponding stock specific statistics above them, calculated across the 12 stocks. Abbreviated headings: $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, $\mathbf{c_{pso}}$ - piecewise linear model with the simple, Nelder-Mead optimized, and the PSO optimized coefficients, respectively; α_a, β_a, α_b, β_b - PSO optimized coefficients.

them.

The optimized vectors $\mathbf{c_{nm}}$ and $\mathbf{c_{pso}}$ do not result in a substantially improved in-sample classification accuracy when compared to the simple vector $\mathbf{c_{smp}}$. This is the case for the stock specific as well as the average (across the 12 stocks) mean accuracy. The average mean accuracy for the piecewise linear

Table 2
Out-of-sample daily classification accuracy (%) and fraction of the daily trade count (%) for individual stocks - 169 evaluation days.[1]

Stock	k-NN	$\mathbf{c_{smp}}$	$\mathbf{c_{nm}}$	$\mathbf{c_{pso}}$	r_1	r_2	r_3	r_4	r_5	r_6
NAB	74.87	*78.54	78.54	78.52	60.79	93.22	74.29	73.73	94.27	56.38
	3.08	3.09	3.09	3.09	0.53	12.16	36.94	37.81	12.02	0.53
BHP	70.08	*72.27	72.26	72.28	50.30	88.85	69.16	72.71	88.90	59.52
	4.25	3.75	3.75	3.72	0.04	2.95	47.89	46.25	2.81	0.05
CBA	73.48	*75.59	75.60	75.56	58.05	92.10	71.04	71.93	90.91	52.02
	2.25	2.74	2.75	2.74	0.55	11.49	38.35	38.14	10.96	0.52
ANZ	73.72	*76.44	76.43	76.42	57.00	94.30	69.47	76.19	92.83	56.95
	3.27	3.42	3.42	3.42	0.28	10.04	41.41	39.09	8.86	0.32
WBC	73.41	*77.31	77.30	77.28	59.82	93.21	73.21	74.40	94.08	59.63
	3.58	3.22	3.20	3.22	0.28	9.69	40.34	40.16	9.23	0.29
NCP	70.97	*75.17	75.18	75.18	54.87	92.57	71.87	74.10	91.66	51.87
	3.66	4.03	4.02	4.02	0.15	6.27	43.60	43.79	6.03	0.16
RIO	76.34	*79.13	79.09	79.06	60.17	93.87	70.78	78.59	92.58	58.98
	2.92	3.63	3.60	3.57	0.74	14.77	37.49	34.00	12.28	0.73
WOW	72.07	*75.11	75.09	75.04	58.86	92.81	75.26	68.51	93.32	57.41
	4.15	4.04	4.02	4.03	0.25	8.44	39.74	42.19	9.12	0.26
FGL	62.69	63.84	64.20	64.12	83.33	92.81	56.75	71.68	85.18	42.86
	6.08	6.71	6.67	6.67	0.01	1.57	50.76	46.26	1.38	0.01
SGB	72.58	*75.63	75.65	75.64	63.82	92.11	76.69	66.63	93.44	48.66
	4.91	5.04	5.07	5.09	0.52	10.83	35.33	40.50	12.25	0.57
SUN	69.24	*73.70	73.71	73.55	60.93	90.73	69.77	69.66	92.01	51.49
	3.87	4.68	4.63	4.63	0.51	10.98	39.49	38.05	10.44	0.54
MIG	67.39	*69.79	69.80	69.62	54.55	90.64	63.71	73.69	90.02	50.00
	6.08	6.67	6.71	6.68	0.04	3.66	46.67	46.31	3.26	0.06
Av.	71.40	74.38	74.40	74.36	60.21	92.27	70.17	72.65	91.60	53.81
value	4.01	4.25	4.24	4.24	0.33	8.57	41.50	41.05	8.22	0.34
					38.69	6.91	10.00	9.43	8.15	39.46
					0.29	2.99	7.09	6.92	2.93	0.30

[1]For k-NN, $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, and $\mathbf{c_{pso}}$ the first and the second line for each stock show means and standard deviations of the classification accuracy, respectively. For regions r_i, $i = 1 \dots 6$, the first and the second line for each stock show means of the classification accuracy and means of the fraction of the daily trade count, respectively. The first and the second line of averages show average values of the corresponding stock specific statistics above them. The third and the fourth line of averages show average standard deviations of the regional classification accuracy and average standard deviations of the regional fraction of the daily trade count, respectively. All average values were calculated across the 12 stocks.
Abbreviated headings: k-NN - k-nearest-neighbor ($k = 9$), $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, $\mathbf{c_{pso}}$ - piecewise linear model with the simple, Nelder-Mead optimized, and the PSO optimized coefficients, respectively; r_i, $i = 1 \dots 6$ - six classification regions of the piecewise linear model with $\mathbf{c_{smp}}$.
* - statistically significant at the level of 0.01.

model with the vectors $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, and $\mathbf{c_{pso}}$ is equal to 75.31%, 75.35%, and 75.34%, respectively. The corresponding standard deviations are 3.65%, 3.65%, and 3.67%. The differences between the mean values, and between the standard deviations, are minimal. Consequently, when measured by the in-sample classification accuracy, the three coefficient vectors seem to be equivalent and the simple vector $\mathbf{c_{smp}}$ is the preferable one due to its intuitive interpretation.

Table 2 shows the out-of-sample classification accuracy statistics for the best k-NN ($k = 9$) classifier found by Blazejewski and Coggins [1], and for the piecewise linear model with the three coefficient vectors $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, and $\mathbf{c_{pso}}$. The statistics were calculated over the 169 evaluation days. The differences between the mean classification accuracies for the coefficient vectors $\mathbf{c_{smp}}$, $\mathbf{c_{nm}}$, and $\mathbf{c_{pso}}$ are very small and statistically not significant. Furthermore, for all stocks in our data set the mean accuracy for $\mathbf{c_{smp}}$ is substantially greater than the mean accuracy for the best k-NN ($k = 9$) classifier reported in Ref. [1]. The differences in their stock specific means, tested with the one tailed paired t-test, are statistically significant for 11 stocks at the level of 0.01, after the Bonferroni adjustment [34] to account for multiple comparisons [6]. The only exception is the FGL stock. The statistical significance indicated in Table 2 refers only to tests between the piecewise linear model with the simple coefficient vector $\mathbf{c_{smp}}$ and the k-NN ($k = 9$) classifier. The average mean out-of-sample classification accuracy for $\mathbf{c_{smp}}$ is 74.38% (SD=4.25%). The corresponding average for the k-nearest-neighbor is 71.40% (SD=4.01%).

Table 2 also reports statistics for the classification accuracy and the fraction of the daily trade count for each of the six regions r_i, $i = 1 \ldots 6$, of the piecewise linear model. The statistics were calculated for the simple vector $\mathbf{c_{smp}}$ over the 169 evaluation days in the data set. The presented results quantitatively confirm the observations made during the analysis of Figs. 1 and 2. In particular, regions r_1 and r_6 have the lowest average mean classification accuracy out of the six regions. The average mean accuracy for the two regions is 60.21% and 53.81%, respectively. These low values, however, do not have much influence on the total classification accuracy A_p because the two regions together represent, across the 12 stocks, only 0.66% [7] of the total number of daily trades. The majority of trades occupy regions r_3 and r_4. The average mean accuracy

[6] For each stock the total number of new and prior comparisons was 242. The piecewise linear model with the simple coefficient vector was compared against the 96 k-NN models constructed in Ref. [1]. The two other piecewise linear models as well as the 144 prior comparisons in Ref. [1] were also accounted for.

[7] The apparent discrepancy of 0.01% between this value and 0.67%, the latter being the sum of the respective averages for the two regions in Table 2, is caused by rounding to two digits after the decimal point. The reported sum of 0.66% was calculated using the full precision mean fractions for r_1 and r_6, and rounding the result. Similar discrepancies can occur for other aggregated values.

for each of these regions is above 70%, while their combined [8] average mean fraction of the daily trade count is 82.55%. As far as individual stocks are concerned, the mean classification accuracy can be as low as 56.75% in r_3 (FGL) and as high as 78.59% in r_4 (RIO). The stock-specific combined mean fraction of the daily trade count varies between 71.49% (RIO) and 97.02% (FGL).

Two regions, r_3 and r_4, contain all trades for which $s \leq a$ and $s \leq b$. This means that the second regularity, transaction cost minimization, is satisfied in these regions by design. The sign of the trades reflects the imbalance in the order book and is determined by finding the dominant side. In our model the dominant side in the book is the one with more volume at the best price, either bid or ask. The trade sign is set to buyer-initiated if $a \leq b$, and to seller-initiated otherwise. This signing rule and the achieved classification accuracy for r_3 and r_4 indicate that the first regularity, competition for order execution, is satisfied too (statistically). Interestingly, the other two regions, r_2 and r_5, sign trades in the opposite direction than the last signing rule. Contrary to the first regularity, their trade sign reflects the non-dominant side in the order book. The cause of this reversal lies in the constraint imposed on the trade size. Trades must satisfy the condition $b < s \leq a$ to belong to r_2 and be classified as buyer-initiated, or the condition $a < s \leq b$ to belong to r_5 and be classified as seller-initiated. The average mean accuracy for each of these two regions is above 91.50%, and together they represent, on average, 16.79% of the daily trade count. On an individual stock basis the mean accuracy has the lowest value of 85.18% in r_5 (FGL), and the highest value of 94.30% in r_2 (ANZ). The stock-specific combined mean fraction of the daily trade count varies between 2.96% (FGL) and 27.04% (RIO). The trade size constraints and the achieved classification accuracy for r_2 and r_5 indicate that if there is a conflict between the two regularities then traders will follow the second one, transaction cost minimization. Furthermore, due to the high classification accuracy in r_2 and r_5, the overall classification accuracy A_p tends to be higher for stocks with a relatively high fraction of the daily trade count in these two regions.

The above results suggest that the piecewise linear model with the simple coefficient vector $\mathbf{c_{smp}}$ is superior to the local non-parametric model proposed in Ref. [1], due to the higher classification accuracy and the intuitive interpretation. We note that the average mean classification accuracy across the 12 stocks, for each of the three coefficients vectors in Table 2, is approximately 1% smaller than the corresponding average in Table 1. The average standard deviations, on the other hand, are larger in Table 2. The differences between mean accuracies for the same individual stocks in the two tables are even more pronounced, varying in value and sign. The observed differences are most probably due to the short estimation period of 30 trading days, relative

[8] The two regions are combined together.

to the 169 trading days of the evaluation period.

5 CONCLUSIONS

We developed an empirical model for trade sign inference. Our model is a piecewise linear parameterization of the model proposed recently in Ref. [1]. The model employs three predictor variables, the volume at the best bid and at the best ask just before a trade, and the trade size. There are four parameters which serve as coefficients of the boundary planes that partition the space of the three predictor variables into six regions. Each region has a dominant trade sign associated with it. All trades belonging to a given region are classified as having the dominant sign of that region. The best values of the four coefficients, in terms of the classification accuracy and parsimony, were found by constructing and evaluating the piecewise linear model with three different coefficient vectors. The simple coefficient vector $\mathbf{c_{smp}}$, equal to $(1, 0, 1, 0)$, was shown to perform equally well as the locally optimized (Nelder-Mead) vector $\mathbf{c_{nm}}$ and the globally optimized (PSO) vector $\mathbf{c_{pso}}$. The simple vector $\mathbf{c_{smp}}$ was selected as the best vector because of its intuitive interpretation. Our piecewise linear model outperforms the k-NN ($k = 9$) classifier developed in Ref. [1], on a stock specific basis (11 out of 12 stocks) as well as across the 12 stocks. The out-of-sample statistics for individual stocks were calculated over the 169 trading days and are significant at the level of 0.01. The average mean classification accuracy for the new model with the simple vector $\mathbf{c_{smp}}$ is 74.38% (SD=4.25%). This value is 2.98% above the average mean of 71.40% (SD=4.01%) reported in Ref. [1]. The overall classification performance of our new model indicates a strong dependence between the trade sign and the three predictor variables, and provides evidence for an endogenous component in the order flow.

The proposed piecewise linear model with the simple coefficient vector $\mathbf{c_{smp}}$ partitions the space of the three predictor variables into six regions. The classification accuracy and the fraction of the daily trade count vary between the regions. Two regions for which $s \leq a$ and $s \leq b$ have a combined average mean fraction of the daily trade count of 82.55%, while each of them has an average mean classification accuracy above 70%. The trade sign within these regions reflects the dominant side in the order book. These results suggest that most of the trades within the two regions satisfy the first regularity, competition for order execution. The second regularity, on the other hand, is satisfied by all trades in these regions by design. Two other regions, where either $b < s \leq a$ or $a < s \leq b$, together represent, on average, 16.79% of the total number of daily trades. They both have an average mean classification accuracy exceeding 91.50%, while their trade sign reflects the non-dominant side in the order book. The results for these two regions indicate that when

there is a conflict between the two regularities then the second one, transaction cost minimization, prevails. The remaining two regions contain a small number of trades whose size is larger than the volume on both sides of the order book. Their combined average of the daily trade count is only 0.66%, which further reinforces the evidence for the second regularity. Consequently, the regions' influence on the overall trade sign classification accuracy of our model is negligible.

The daily classification accuracy of the piecewise linear model developed in this paper could be used as a new order flow metric. The temporal evolution of the metric and a question of its privileged timescale are good topics for future research. The new model captures the two regularities discussed in the introduction, competition for order execution and transaction cost minimization. These regularities are reflected in the relationship between the predictor variables and the trade sign, and can probably explain the monotonically increasing accuracy of the model proposed in Ref. [1]. That study reported that the classification accuracy of the k-NN classifier increased with the length of the training interval, which indicated a memory of a corresponding length. It appears that the increase in the training interval length provided more data points (trades) for the estimation, which in turn allowed for a better approximation of the relationship between the variables concerned. The observed memory could therefore be a consequence of the two regularities operating on a single trade level, as discussed in this paper. We also believe that the same two regularities are involved in the interplay [16] between the long memory in the market order sign [13, 16], the long memory in the trade size and the long memory in the volume in the limit order book [16]. A rigorous investigation of this idea is another possible direction for further work.

6 ACKNOWLEDGEMENTS

We gratefully acknowledge that the Capital Markets Cooperative Research Centre (CMCRC) and its industry partners provided the ASX data and the software for data extraction. A. Blazejewski wishes to thank the CMCRC for a scholarship.

References

[1] A. Blazejewski, R. Coggins, A local non-parametric model for trade sign inference, Physica A: Statistical Mechanics and its Applications (2004) in press.

[2] B. Biais, P. Hillion, C. Spatt, An empirical analysis of the limit order

book and the order flow in the Paris Bourse, Journal of Finance 50 (5) (1995) 1655–1689.

[3] M. Al-Suhaibani, L. Kryzanowski, An exploratory analysis of the order book, and order flow and execution on the Saudi stock market, Journal of Banking & Finance 24 (8) (2000) 1323–1357.

[4] A. Ranaldo, Order aggressiveness in limit order book markets, Journal of Financial Markets 7 (1) (2004) 53–74.

[5] M. Potters, J.-P. Bouchaud, More statistical properties of order books and price impact, Physica A: Statistical Mechanics and its Applications 324 (1-2) (2003) 133–140.

[6] J. D. Farmer, L. Gillemot, F. Lillo, S. Mike, A. Sen, What really causes large price changes?, preprint cond-mat/0312703, 2003.

[7] H. Degryse, F. de Jong, M. van Ravenswaaij, G. Wuyts, Aggressive orders and the resiliency of a limit order market, Discussion Paper 80, Tilburg University, Center for Economic Research, 2002.

[8] M. D. Griffiths, B. F. Smith, D. A. S. Turnbull, R. W. White, The costs and determinants of order aggressiveness, Journal of Financial Economics 56 (1) (2000) 65–88.

[9] C. Cao, O. Hansch, X. Wang, The informational content of an open limit order book, EFA 2004 Maastricht Meetings Paper No. 4311, http://ssrn.com/abstract=565324, 2004.

[10] P. Verhoeven, S. Ching, H. G. Ng, Determinants of the decision to submit market or limit orders on the ASX, Pacific-Basin Finance Journal 12 (1) (2004) 1–18.

[11] K. Omura, Y. Tanigawa, J. Uno, Execution probability of limit orders on the Tokyo Stock Exchange, Working Paper (November 2000), http://ssrn.com/abstract=252588.

[12] K. Hedvall, J. Niemeyer, G. Rosenqvist, Do buyers and sellers behave similarly in a limit order book? A high-frequency data examination of the Finnish stock exchange, Journal of Empirical Finance 4 (2-3) (1997) 279–293.

[13] J.-P. Bouchaud, Y. Gefen, M. Potters, M. Wyart, Fluctuations and response in financial markets: the subtle nature of 'random' price changes, Quantitative Finance 4 (2) (2004) 176–190.

[14] Y. Hamao, J. Hasbrouck, Securities trading in the absence of dealers: trades and quotes on the Tokyo Stock Exchange, Review of Financial Studies 8 (3) (1995) 849–878.

[15] F. de Jong, T. Nijman, A. Röell, Price effects of trading and components of the bid-ask spread on the Paris Bourse, Journal of Empirical Finance 3 (2) (1996) 193–213.

[16] F. Lillo, J. D. Farmer, The long memory of the efficient market, Studies in Nonlinear Dynamics & Econometrics 8 (3) (2004) Article 1, http://www.bepress.com/snde.

[17] D. C. Porter, The probability of a trade at the ask - An examination of interday and intraday behavior, Journal of Financial and Quantitative

Analysis 27 (2) (1992) 209–227.
[18] M. Aitken, P. Brown, H. Y. Izan, A. Kua, T. Walter, An intraday analysis of the probability of trading on the ASX at the asking price, Australian Journal of Management 20 (2) (1995) 115–154.
[19] C. M. C. Lee, M. J. Ready, Inferring trade direction from intraday data, Journal of Finance 46 (2) (1991) 733–746.
[20] M. Aitken, A. Frino, The accuracy of the tick test: Evidence from the Australian stock exchange, Journal of Banking & Finance 20 (10) (1996) 1715–1729.
[21] K. Ellis, R. Michaely, M. O'Hara, The accuracy of trade classification rules: Evidence from Nasdaq, Journal of Financial and Quantitative Analysis 35 (4) (2000) 529–551.
[22] E. R. Odders-White, On the occurence and consequences of inaccurate trade classification, Journal of Financial Markets 3 (3) (2000) 259–286.
[23] M. J. Aitken, H. Berkman, D. Mak, The use of undisclosed limit orders on the Australian Stock Exchange, Journal of Banking & Finance 25 (8) (2001) 1589–1603.
[24] J. C. Lagarias, J. A. Reeds, M. H. Wright, P. E. Wright, Convergence properties of the Nelder-Mead simplex method in low dimensions, SIAM Journal on Optimization 9 (1) (1998) 112–147.
[25] J. A. Nelder, R. Mead, A simplex method for function minimization, Computer Journal 7 (4) (1965) 308–313.
[26] C. J. Price, I. D. Coope, D. Byatt, A convergent variant of the Nelder-Mead algorithm, Journal of Optimization Theory and Applications 113 (1) (2002) 5–19.
[27] H. Xiaohui, S. Yuhui, R. Eberhart, Recent advances in particle swarm, in: Proceedings of the IEEE Congress on Evolutionary Computation (CEC2004), 2004, Portland, Oregon, USA, Vol. 1, 2004, pp. 90–97.
[28] K. E. Parsopoulos, M. N. Vrahatis, Recent approaches to global optimization problems through Particle Swarm Optimization, Natural Computing 1 (2-3) (2002) 235–306.
[29] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of the IEEE International Conference on Neural Networks, 1995, Perth, WA, Australia, Vol. 4, 1995, pp. 1942–1948.
[30] B. D. Ripley, Pattern recognition and neural networks, Cambridge University Press, Cambridge, 1996.
[31] C. M. Bishop, Neural networks for pattern recognition, Oxford University Press, New York, 1995.
[32] I. T. Nabney, NETLAB: Algorithms for Pattern Recognition, Springer, 2001.
[33] B. Birge, PSOt: a particle swarm optimization toolbox for use with Matlab, in: Proceedings of the IEEE Swarm Intelligence Symposium (SIS '03), 2003, Indianapolis, Indiana, USA, 2003, pp. 182–186.
[34] J. A. Rafter, M. L. Abell, J. P. Braselton, Multiple comparison methods for means, SIAM Review 44 (2) (2002) 259–278.