

# ERRORS IN TRADE CLASSIFICATION: CONSEQUENCES AND REMEDIES<sup>1</sup>

Carsten Tanggaard  
The Aarhus School of Business,  
Department of Finance,  
Fuglesangs Alle 4,  
DK 8210 Aarhus V,  
Denmark.  
E-mail: cat@asb.dk

February 2002<sup>2</sup>

<sup>1</sup>Research has been supported by a grant from the *Danish Social Science Research Council*. I thank Amber Anand, Tom Engsted, and Elizabeth Odders-White for comments on the research.

<sup>2</sup>**This copy of the paper was provided for review purposes. Please, do not quote without permission.**

## Abstract

This paper demonstrates that consequences of errors in trade classification (buy/sell indicator) are possibly worse than suggested by previous studies. An errors-in-variables model of classification errors shows that the bias in regression-type microstructure models depends in a complicated way on the probability of trade-reversal in addition to the probability of error. This has implications for empirical research using regression type estimation involving trade-indicators. As an example it also affects VAR models for trades and returns. One pitfall is that trade-reporting procedures, and trading strategies are known to affect the clustering of trades. I propose instrumental variable (IV) estimation as an alternative to OLS, and show that the bias of the IV estimator is typically much smaller and only depends on the probability of classification error. Therefore, the IV estimator is not affected by clustering of trades. IV estimation can also be used to estimate the probability of classification error from a trade-indicator regression. The method is illustrated using 51 stocks from the TORQ database. The empirical analysis shows that even a crude estimator of classification error renders the bias statistically insignificant.

JEL: G12, C22.

Keywords: Bias, effective spread, errors in variables, market impact, TORQ, trade indicator.

## 1 Introduction

Bid-ask classification of trades is a fundamental issue in empirical market microstructure research. The problem is about determining which side of a trade initiated the trade. In markets such as the NYSE auction market, the Nasdaq dealer market, and markets driven by limit orders, one can assume that one participant in the trade is passive (the specialist/dealer/limit order trader) while the other part is active. The classification problem consists of finding out whether the active part is buyer or seller.

The trade classification is not like the price and volume a part of the trade record that comes out of the exchange. Rather it is an integrated part of the modeling process. It can be considered an unknown parameter to be estimated or more like unobserved data. In any case improper estimation of trade direction causes noise and bias in inference.

One type of error arises from insufficient data or an inappropriate estimation method. For example, an exchange may not provide detailed order and quote data, and in extreme cases only a trade record is released. If more detailed quote information is available, it's use may be subject to errors. For example, delays in trade reporting and non-synchronous time stamps are sources of error.

Another type of error arises from inappropriate model assumptions. For example, if some brokerage service is involved in trading then it may be wrong to consider one side as initiating the trade and the whole idea of trade classification may even become dubious.

The literature abounds with empirical research based on microstructure trading data, and classification based on quote information. In particular there are many studies of utilizing data from NYSE and other US markets. Unfortunately, the publicly available data from US markets do not allow a secure trade classification. Depending on the filtering of data the classification error can easily be 15% when using the Lee-Ready algorithm. Given this fact it is surprising how little awareness there is in the empirical literature about the consequences of these errors, let alone methods for correcting biases. To my knowledge there is only a handful or two studies about trade classification and errors (see the literature list for some of the examples). Even fewer papers aim at studying the consequences for inference on spreads etc. (Ellis, Michaely, and O'Hara [2000], Finucane [2000], Odders-White [2000], Theissen [2001]).

The cited papers take an indirect approach to correcting classification error bias.

They simply aim at improving the hit rate. However, this is a troublesome approach, because it requires detailed information about which factors affect classification errors (proximity to quotes and mid-quotes, trade-size, trading intensity etc.).

This paper is the first to study the consequences and possible remedies of trade classification errors using a formal errors-in-variables model. In addition my approach differs from existing research by using an approach which is generally applicable, i.e. without using detailed information about underlying error factors.

I show that the existing evidence on consequences of classification errors seems to underestimate the biases. This may be due to some peculiarities of the TORQ database used in some of these studies or due to an off-setting effect of other biases (price discreteness, for example). I derive analytical expressions for the errors-in-variables bias in two examples, which are representative for many applications of trading data. One example is the trade-indicator model, where the bias in estimation of effective spread and adverse selection component of spread is very substantial. The bias depends on the probability of trade-reversal. This is reason for concern as trading strategies and reporting standards may or may not affect the clustering of trades. Also, small firms with less trading intensity may cluster less than trades in large stocks. The problem is related to any regression-type microstructure model estimated using possibly erroneous trade direction data. In the paper, I show how instrumental variables (IV) regressions can be used to remove the bias and at the least make the bias comparable in magnitude with more robust methods of spread estimation. An examination of 51 stocks from the TORQ database shows that the instrumental variable approach corrects most of the bias for middle to large sized samples.

The paper proceeds with an introduction and a literature review in section 2. Section 3 presents a model of classification errors, and analyses it's consequences for estimators of effective spreads and the adverse selection component of the spread. This section also presents an instrumental variable estimator aimed at removing errors-in-variables bias. Section 4 has some empirical evidence on estimation bias in the spread regression model and on the potential for removing bias using IV-estimation. Finally section 5 has a conclusion and some remarks on possible future research.

## 2 Bid-ask classification methods

Methods for classifying trades can be characterized according to the amount of structure put on data and the need for supplemental information beyond the basic trading record.

The most accurate classification schemes require information on the orders participating in the trade. If there is only one seller and only one buyer and if one order is a market order and the other one is a limit order, then the trade can unambiguously be classified as initiated by the market order. This method can, however, be inaccurate if there is some kind of human interference in the trading process. As an example if trades are executed in a specialist market where orders are stopped for price improvement purposes.

An alternative is to identify the initiating part with the latest arriving order [Aitken and Frino 1996, Odders-White 2000]. This method provides a unique classification when applied to data from automatic order match systems [Aitken and Frino 1996]. In such applications the arrival time for one of the orders will (except in special cases) equal the time of trade execution, and the late arriving order may be regarded as initiating the trade (market order or ELO). Nevertheless, the method can also render imprecise for the very same reasons as stated above. Also improper identification of odd-lot orders, which in some systems do not initiate trades, causes the method to be inaccurate.

In dealer markets, where one part of the trade is a dealer, and the other part is a broker or an investor, there is a unique trade direction, in that investors and brokers (trading on behalf of investors) can be assumed to be trade-initiators [Ellis, Michaely, and O'Hara 2000, Lee and RadhaKrishna 2000].

The quote rule applies information about proximity to prevailing quotes in order to classify trades. If the traded price is at the current ask quote or even just above the mid-quote, the trade may be classified as buyer initiated. Clearly this method is most accurate for trades well above or well below the mid-quote and mid-quote trades represent a special problem, which can be handled by the tick rule (or ignored by removing such trades).

Despite its simplicity the tick rule often works quite well [Finucane 2000]. It works as follows. A trade is called an up-tick (down-tick) if the actual traded price is higher (lower) than the previously traded price. If there was no price change, but the most

recent price changes occurred on an up-tick (down-tick) the trade is called a zero-up-tick (zero-down-tick). The tick rule simply classifies trades occurring at an up-tick or zero-up-tick as buyer initiated and conversely for trades occurring at a down-tick or zero-down-tick.

The quote rule - with the tick-rule used as a tie-breaker for mid-quote trades - is called the Lee-Ready algorithm [Lee and Ready 1991], and it is probably the most widely used method in empirical market microstructure research on US stock markets. When applied to data from the NYSE a 5 seconds time lag is used because of delays in the clerical work in the specialist booth, where quotes are typically updated ahead of trades.

Another econometrically interesting, and purely data-driven method, is the mixture of distribution model applied in Glosten and Harris [1988] and in Harris [1990]. This method utilizes a switching regression model under the assumption that the trade-direction is a random binary process. The distribution of the unobservable trade-direction is concentrated out of the likelihood using suitable techniques, and the MLE estimator for the remaining parameters is obtained by complicated computations (while taking care of price discreteness).

How accurate are methods for classifying trades, and what factors cause them to break down?

If we look at the tick rule, it is likely to break down when applied to markets with a small tick size or with substantial pricing errors and with relative large short term volatility compared with the tick size [Aitken and Frino 1996]. This may not be a problem for a highly liquid market like the NYSE, where the trades often take place with a minimum price change of 0 or 1 tick, but for less liquid markets, it may be completely useless.

Lee and Ready [1991] was the first article to thoroughly study the properties of current methods for bid-ask classification on the NYSE. They applied intra-daily observations on trades and quotes and outline a scheme to match trades and quotes to obtain the trade direction (the Lee-ready method). Lee and Ready [1991] assess the reliability of the algorithm and find that it performs quite well, classifying 90% of the trades correctly. They also address the problem of quote changes preceding trades and find empirically that comparing trades to quotes posted at least 5 seconds prior to the trade in question seem to circumvent the problem.

Odders-White [2000] conducted a detailed analysis of the sources and consequences

of inaccurate classification using the TORQ database. It is found that in general the accuracy of the Lee-Ready method is 85%. Another result is that the tick rule systematically mis-classifies trades executed at the mid-quote of the bid-ask spread, small trades, and trades in highly liquid stocks. In fact, the mis-classification problem as presented by Odders-White is strongly related to the probability of trades executing at the mid-quote spread since small trades and trades in very frequently traded stocks show a higher tendency of having these mid-quote trades (as also noted by Odders-White). Odders-White shows that when applying the data classified by the Lee and Ready algorithm in empirical studies, the mis-classifications can bias the estimates. Two examples (earnings announcements, components of the bid-ask spread) of such problems are discussed in Odders-White [2000].

Lee and Radhakrishna analyze the performance of the Lee and Ready algorithm, also using the TORQ data [Lee and RadhaKrishna 2000]. They show that 40% of the trades cannot be classified unambiguously as either buyer or seller initiated, due to complexities of the NYSE auction process. However, of the trades that can be buyer/seller initiated, the Lee and Ready algorithm classifies 93% of the trades correctly (see the papers for a discussion of the somewhat differing conclusions between Lee and RadhaKrishna [2000] and Odders-White [2000]).

Introducing data from the Australian market from June 1992 to July 1994 Aitken and Frino [Aitken and Frino 1996] show that the Lee and Ready algorithm classifies approximately 74% of the trades correctly. This is a significantly lower hit rate than the 90% reported in Lee and Ready [1991], and also lower than the results of Odders-White [2000], Lee and RadhaKrishna [2000]. One feature that distinguishes the Australian market from the NYSE is that most stocks are traded with a smaller tick size, where the TORQ data represents the case of  $\$1/8$  tick size, the minimum tick size for most Australian stocks is 1 cent.

In summary, the quality of the Lee and Ready tick rule will depend on two factors: tick size and short term price volatility. The larger the tick size and the smaller the short term fluctuations, the greater is the likelihood that a non-zero tick is due to reversal of the trade direction. Thus, the apparent success of the tick rule, when applied to NYSE data, and the failure of the algorithm when applied to small tick markets is to be expected.

Recently ? has documented a 72.8% success rate for the Lee-Ready algorithm when applied to data from the Frankfurt Stock Exchange. This study also points to

the consequences of inaccurate trade classification.

Ellis, Michaely, and O'Hara [2000] utilizing Nasdaq data proposes a modification of the Lee-Ready method. The quote rule should only be applied to trades at the current quote (after possible correction for the 5 seconds delay). All other trades are classified by the tick rule.

### 3 Analysis of classification errors

In this section I will conduct a formal examination of the consequences and a possible remedy of errors in trade classification.

The problem can be stated as follows. We will assume that trade direction can take two values 1 (for buyer initiated trades) and -1 (for seller initiated trades). The observed trade-indicator is called  $Q_t$ . If some classification error occurs then

$$Q_t = (1 - 2U_t)Q_t^*, \quad (1)$$

where  $Q^*$  denotes the underlying true trade-indicator, and  $U$  is an error indicator taking the values 0 and 1 (1 if there is an error, 0 otherwise). As such (1) is just a definition; it imposes no restriction on the possible kind of errors. Therefore, more structure is needed in order to formally analyze the problem.

Assume for now that  $Q_t^*$  is a two-state markov chain with probability of trade reversal equal to  $\pi$ , i.e.  $\pi = P(Q_t^* = j | Q_{t-1}^* = i)$  for  $i \neq j$ , and that  $U_t$  is serially uncorrelated and independent of  $Q_t^*$  (at all lags):

$$P(U_t = u) = \begin{cases} \delta & u = 1 \\ 0 & u = 0 \end{cases} .$$

The markov assumption on  $Q_t^*$  is actually quite innocent, as it coincides with model assumptions in standard regression type microstructure models [Glosten and Harris 1988, Harris 1990, Huang and Stoll 1997]. The assumptions on  $U_t$  are more critical and in fact there is some evidence that the serial uncorrelatedness is not fulfilled in the TORQ database used in some empirical studies of trade classification errors. There is also some empirical evidence that classification errors are predictable from observed information [Ellis, Michaely, and O'Hara 2000, Finucane 2000]. I will use it as a first-order approximation. Below I will discuss some more general assumptions and how they will affect inference and conclusions.

### 3.1 Consequences for effective and realized spreads

In the first example I will study the effects on effective and realized spreads. The effective spread at time  $t$  is defined as [Ellis, Michaely, and O'Hara 2000]:

$$S_t = 2Q_t(P_t - M_t),$$

where  $P_t$  is the traded price at time  $t$  and  $M_t$  is the mid-quote at the time of trade. An estimate,  $\bar{S}$ , of the expected effective spread can be found by simply averaging  $S_t$  across trades. If the trade-indicator is subject to noise then  $\text{P lim } \bar{S} = (1 - 2\delta)S$ , where  $S$  is the true average spread. Thus, the average effective spread becomes downward biased by classification errors. The intuition is clear: if buys (occurring at the ask) are mistaken for sells then this will underestimate trading costs and the effective spread will become downward biased. This logic is true if all buys occur above the mid-quote and all sells occur below the mid-quote. If this is not the case, the sign of the bias may reverse [Ellis, Michaely, and O'Hara 2000, Finucane 2000, Odders-White 2000]. Similar comments on the realized spread,  $E_t = 2Q_t(P_t - M_{t+\tau})$ , where  $M_{t+\tau}$  is the mid-quote at some future point in time<sup>1</sup>. The realized spread will be downward biased by the same factor  $(1 - 2\delta)$ .

### 3.2 Consequences for regression type models

Consider the following model involving trade-indicators [Huang and Stoll 1997]:

$$\Delta P_t = \frac{S}{2}\Delta Q_t + \alpha\frac{S}{2}Q_{t-1} + \varepsilon_t, \quad (2)$$

where  $P_t$  is the traded price,  $S$  is the effective spread,  $\alpha$  is the adverse selection component of the spread. I will occasionally call  $\alpha$  as well as  $c = \alpha\frac{S}{2}$  for market impact (sometimes I will also call them for relative and absolute market impact). The error term captures value innovations as well as pricing errors. In general it will be serially correlated in the presence of discreteness-induced errors. We will assume that  $\varepsilon$  is uncorrelated with  $Q_t$  in which case OLS-estimates will be consistent in the absence of classification errors.

In the presence of classification errors, the standard textbook model of errors-in-variables suggests that  $S$  and  $c = \alpha\frac{S}{2}$  are biased due to the effect of noise on the regressor covariance matrix. This is in fact the case as

---

<sup>1</sup>The NYSE *Disclosure of Order Execution* reports apply a 5 minutes lead time.

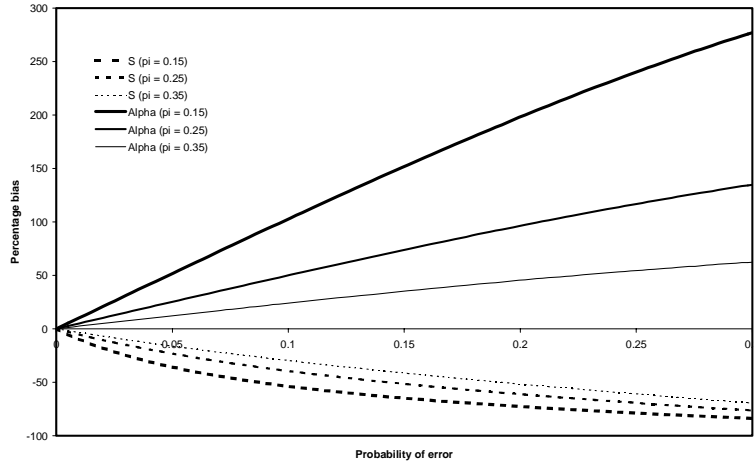


Figure 1: Bias in effective spread and relative market impact

$$P \lim \begin{bmatrix} \widehat{S/2} \\ c \end{bmatrix} = \begin{bmatrix} \frac{1}{2}kS \frac{-1+h^2k^2-h\alpha+h+hk^2\alpha-hk^2}{h^2k^4-1} \\ \frac{1}{2}kS \frac{-h+hk^2-\alpha+1+h^2k^2\alpha-h^2k^2}{h^2k^4-1} \end{bmatrix},$$

where  $k = (1 - 2\delta)$  and  $h = (1 - 2\pi)$ . This shows that the OLS-estimator is asymptotically biased unless  $k = 1$  (i.e. unless  $\delta = 0$ ). Thus, errors in trade classification does not only cause noise to inference, it also causes a bias which does not go away in large samples.

Although it is not hard to conjecture the sign of the bias on spreads, it is more difficult to assess it's magnitude and it's effect on  $\alpha$ .

Let us illustrate this by an example. Huang and Stoll [1997] presents the following cross-sectional average estimates  $(\widehat{S}, \widehat{\alpha}, \widehat{\pi}) = (0.1222, 0.1135, 0.1605)^2$ . Ignoring the possible classification error biases and assuming that these estimates were the true parameters, it is a simple calculation to calculate the bias for different error rates. This is illustrated in figure 1. The bias in  $S/2$  and  $\alpha$  is graphed for different values of  $\delta$  (from 0 to 0.3). For a value of  $\pi = 0.15$  the percentage bias in  $S$  is roughly -65% when the error rate is 15% (which corresponds roughly to the error rate reported by Odders-White [2000] for the Lee-Ready method). When  $\pi$  increases to 0.20 the bias reduces to -58% and to -52% for  $\pi = 0.25$ .

<sup>2</sup>Tables 2 and 5 of Huang and Stoll [1997].

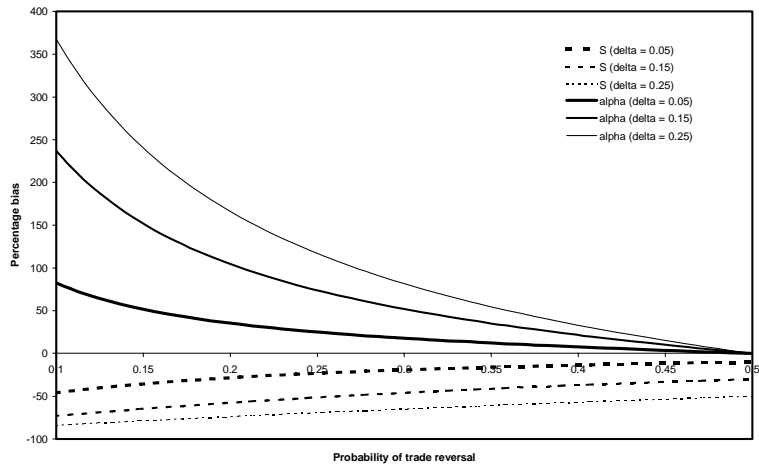


Figure 2: Bias in effective spread and relative market impact

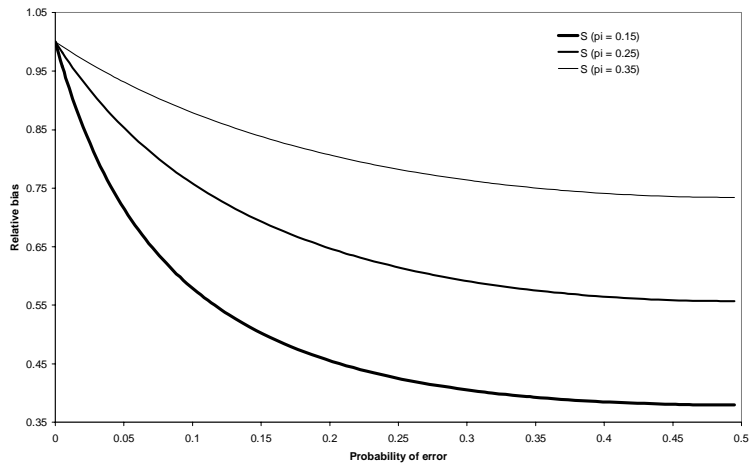


Figure 3: Relative bias of average and regression based effective spread

The problem is even more pronounced for the adverse selection component,  $\alpha$ , as figure 1 illustrates. Note, however, that the bias of  $\hat{c}$  (market impact) is much smaller as it is the product of an upward biased estimate,  $\alpha$ , and a downward biased component,  $\hat{S}/2$ . As an example with  $\pi = 0.15$  and  $\delta = 15\%$  the downward bias in  $\hat{c}$  is only about -11%. Thus, the parameterization used in Huang and Stoll [1997] is extremely sensitive to classification error bias. Furthermore, the bias depends on  $\pi$ . For example when  $\pi \approx 1/2$  the bias in  $\hat{\alpha}$  is negligible as is illustrated in figure 2 and the classification error bias in  $\hat{S}$  is also somewhat (but not as much) reduced.

These calculations point towards a drawback of regression based measures of bid-ask spread: the classification error bias depends on the error rate as well as  $\pi$ .

Above we saw that the bias on the average effective spread,  $\bar{S}$ , does not depend on  $\pi$ , whereas  $\hat{S}$  does. Figure 3 shows the relative bias,  $[\text{P lim } \bar{S} / \text{P lim } \hat{S}]$  plotted for three values of  $\pi$  and for a continuum of different error rates. Even modest error probabilities make the regression based estimator much more biased than the average effective spread. In the next section, I show how to use instrumental variables to obtain an estimator with reduced bias, and one for which the bias does not depend on  $\pi$ .

### 3.3 An instrumental variable estimator

It is well known that if errors are additive and independent of the regressors then instrumental variable estimation will remove the errors-in-variables bias. Furthermore, in the context of time series regression lagged regressors can serve as instruments. In the trade-indicator model classification errors are multiplicative and one can not hope to get consistent estimators from a standard IV-estimation. Nevertheless, let us consider the instrumental variable estimator using  $(Q_{t-2}, Q_{t-3})$  as instruments. It can be shown that

$$\text{P lim} \begin{bmatrix} \widehat{S/2} \\ c \end{bmatrix}_{IV} = \frac{1}{1-2\delta} \begin{bmatrix} S/2 \\ c \end{bmatrix}. \quad (3)$$

Thus, the IV estimator of  $\widehat{S/2}$  is upward biased by a factor  $1/(1-2\delta)$ . The same is the case for the estimator of absolute market impact,  $\hat{c}$ , while the adverse selection component,  $\hat{\alpha}$ , is unbiased.

This is important as the bias is now of a much simpler nature. In addition the bias is of a small magnitude (compare with the graphical examples) and even a crude

estimate of  $\delta$  can further reduce the bias. Finally, the instrumental variable approach can even provide an estimate of the error rate,  $\delta$ .

This follows from consideration of the autoregression,  $Q_t = \theta Q_{t-1} + e_t$ . The first order serial correlation,  $\theta$ , is related to the trade-reversal probability by  $\theta = 1 - 2\pi$ . The OLS estimator is, however, biased in the case of classification errors. This follows from  $\text{P lim } \hat{\theta}_{OLS} = (1 - 2\delta)^2\theta$ . Now we are in a position to completely remove the bias by using IV estimation (using lagged  $Q$ 's as instruments).

The IV estimator is asymptotically unbiased:  $\text{P lim } \hat{\theta}_{IV} = \theta$ . This gives us a simple estimator of  $\delta$ , namely  $\hat{\delta} = \left[1 - \left(\hat{\theta}_{OLS}/\hat{\theta}_{iv}\right)^{1/2}\right]/2$ , which can then be used to obtain an estimate of the multiplicative bias of  $(\hat{S}, \hat{\alpha})$  and  $\hat{c}$  and remove the bias term in (3).

#### 4 Empirical examination

I examined classification errors using the TORQ database, which is the only publicly available database with detailed order and audit information from the NYSE [Hasbrouck 1992]. TORQ has trades, orders, quotes, and the audit trail for 144 stocks traded in November 90 through January 91. The sample is stratified and is representative for all stocks listed on the NYSE during that period. The same data was used in recent studies of trade classification methods by Lee and RadhaKrishna [2000], Finucane [2000], and Odders-White [2000].

The TORQ database allows the identification of the true indicator using one of several methods. I use the same approach as in Odders-White [2000] by using time of order arrival to identify order arrival, whereas Lee and RadhaKrishna [2000] used trade participation (market/limit order) to identify the trade initiator.

In constructing the data, I did as follows. In the first round I filtered away the following records: a) non-NYSE quotes and trades, b) those with correction code different from zero, c) trades executed before 9:45.

Next, I link the audit trail records to the order records. In that process I removed audit trail records for which there was no corresponding trade record. Link fields are `report time` in the audit trail and `buy time/sell time` in the order record. Audit records that could not be linked on both sides were subsequently discarded. Finally, I construct the true trade indicator for each audit record by identifying the trade initiator with the latest arriving order. The quote method and the tick rule were constructed using standard methods. The combined Lee-Ready method applied

	Buys	Sells	Unclassified	Errors	Succes rate
True classification	168475	148376		-	-
Classified by quote rule	127141	105848	55950	27912	89.30%
Classified by tick rule	141602	116606	72	58571	81,51%
Not classified by LR	145640	122728	8	48475	84.70%

The table includes all trades for which the true classification exists and for which it was possible to identify the trade indicator according to one of the three methods: the tick rule, the quote match rule or the Lee-ready method. The success rate is the number of buys plus sells divided by the sum of the number of buys, sells, and unclassified trades.

Table 1: Summary of classification statistics

the 5 seconds delay rule of thumb. The procedure in it's totality seems to follow the method applied in Odders-White [2000]<sup>3</sup>.

The total number of included trades is 316,851 (see table 1). This is comparable to 318,364 included trades in Odders-White [2000]. I have no explanation for the difference (1513 trades). The error rates for the three different methods are, however, quite similar (numbers from Odders-White [2000] in parentheses): 10.70% (10.80%) for the quote rule, 18.49% (21.37%) for the tick rule, and 15.30% (15.03%) for the Lee-Ready method. An exception is the tick rule, where the difference is almost 2 percentage points.

After having documented consistency with Odders-White (2000), I excluded stocks with fewer than 1000 trades. This leaves us with 51 stocks for the rest of the analysis.

Table 2 shows details about the stochastic properties of the true trade-indicator,  $Q_t^*$ , and the error process,  $U_t$ . I estimated univariate and bivariate markov chains of order 1 to 4 for each stock [Billingsley 1961]. The Schwartz criterion was constructed as  $-2 \ln Lik + \ln(T)d$ , where  $T$  is the number of trades, and  $d$  is the number of estimated parameters. I also tested for serial independence under the more general hypothesis of a first order Markov chain and for independence (contingency) between  $Q_t^*$  and  $U_t$  [Billingsley 1961]. In general, there is no need for higher order terms as the cross-sectional average of the markov order is between 1 and 2 and only in very

<sup>3</sup>I thank Elizabeth Odders-White for helping me with explaining details in her procedure.

Markov order (MBICE)	Average order	# [1]	# [2]	# [>2]
True trade indicator, $Q_t^*$	1.45	70.60%	17.64%	11.76%
Error process, $U_t$	1.24	82.36%	11.76%	5.88%
Joint process, $Q_t^*, U_t$	1.08	92.16%	7.84%	0.00%
Test for serial independency	Average $-2 \ln LR$	Minimum $-2 \ln LR$		
True trade indicator, $Q_t^*$	1993.99*	79.34*		
Error process, $U_t$	2006.02*	84.39*		
Joint process, $Q_t^*, U_t$	6505.85*	308.09*		
Test for contingency	Average $-2 \ln LR$	Minimum $-2 \ln LR$		
Joint process, $Q_t^*, U_t$	1004.73*	17.14*		
Moments	Average	Std. dev.	Median	
Error rate ( $\bar{U}$ )	14.23%	7.31%	14.22%	
Number of buys, [ $\#(Q_t^* = 1)$ ]	53.83%	5.42%	54.14%	
Timeseries-regression	Average	Std.error		
<i>const</i>	0.064	0.004		
$U_{t-1}$	0.537	0.013		
$Q_t^*$	-0.013	0.002		
$\Delta Q_{t-1}^*$	-0.006	0.002		
<i>PACF</i> , $U_t$	Average	Std.error		
<i>Lag</i> = 1	0.517	0.012		
<i>Lag</i> = 2	0.039	0.006		
Correlations	Average	Minimum	Maximum	
$\text{Corr}(Q_t^*, U_t)$	-0.068	-0.198	0.077	

The sample consists of all stocks from the TORQ with more than 1000 trade records. All numbers in the table were constructed from the cross-sectional distribution. The markov order were estimated from markov chain models (of  $u, Q_t^*$ ) using the Schwartz (*BIC*) criterion. The #[] columns describe the cross-sectional distribution of markov orders estimated by *BIC*. The likelihood ratio tests for serial correlation and contingency are  $\chi^2$  tests. An '\*' indicates that the statistic is significant at the 5% level using the appropriate  $\chi^2$  distribution [Billingsley 1961, see ]. The error rate is the fraction of wrongly classified trades. The buy rate is the fraction of trades with  $Q_t^* = 1$ . The timeseries regression is  $U_t = \hat{a}_0 + \hat{a}_1 U_{t-1} + \hat{a}_2 Q_t^* + \hat{a}_3 \Delta Q_{t-1}^* + v_t$ . The *PACF* is the partial autocorrelation of the error process,  $U_t$ .

Table 2: Properties of trade indicator and error process.

few cases was the selected order greater than 2. The test for contingency (under the alternative of a first order chain) was rejected for all stocks.

This evidence points to a rejection of some of the assumptions from the classification error model. What does this mean for the analysis?

Firstly, there is only weak indication of higher (i.e.  $\geq 2$ ) order correlations in  $Q_t^*$  and even if there were the instrumental variable estimator could be adapted to this. Next, the serial independence of  $U_t$  is not critical. As long as the 2. order correlation vanishes, one can still - by lagging the instruments - use the instrumental variable estimator. This assumption can be checked by inspection of the partial autocorrelations in table 2. One can see that the average first order *PACF* is 0.517 and highly significant (cross-sectionally). The average *PACF* at lag 2 is also significant although less so, and the average is only 0.039. Thus, we will continue under the assumption of serial correlation of order 1 and no higher order partial correlations.

The final assumption about independence of  $U_t$  and  $Q_t^*$  was rejected by the contingency test. However, the timeseries regression (of  $U_t$  on  $U_{t-1}$ ,  $Q_t^*$ ,  $\Delta Q_t^*$ ) indicates that the relevant coefficients are very small in magnitude and in fact statistically insignificant. Also the correlation coefficient  $\text{Corr}(U_t, Q_t^*)$  is very small, and insignificant. Thus, there is conflicting evidence between the contingency test and the timeseries based test. Because the average correlation between  $U_t$  and  $Q_t^*$  is in fact very small, we will ignore this problem in the following.

I then applied an IV estimator (2 instruments) to the regression  $Q_t = \theta Q_{t-1} + e_t$  to estimate the classification error,  $\hat{\delta}_{IV}$ . Using more than 2 instruments did not change the results in any significant way. The estimate indeed showed up to be quite crude. In general I found that  $0 \leq \hat{\delta}_{IV} \approx \frac{1}{2} \hat{\delta}^*$ , where  $\hat{\delta}^*$  is the estimate of  $\delta$  using data for the true indicator. Nevertheless, even a crude estimate of the error rate may be better than relying on OLS-based regressions, which ignore the classification errors.

To investigate this, I constructed several estimators of the effective spread,  $S$ , the adverse selection component,  $\alpha$ , and  $c = \alpha \frac{S}{2}$  in the regression model (see the notes to table 3 for an explanation of the different estimators. I used 6 lagged  $Q_t$ 's as instruments for the regression.

The results are summarized in table 3. The first observation is that the effective spread,  $S^*$ , obtained from the true indicators, is smaller than  $\hat{S}_{OLS}$ . This is not in accordance with the predictions of the theoretical model and is in fact counter-intuitive. It is, however, consistent with other empirical studies using the same data

SPREAD AND IMPACT	Average	Std. error	Median
$\widehat{S}^*$	0.580*	0.088	0.379
$\widehat{\alpha}^*$	0.128*	0.015	0.112
$\widehat{c}^*$	0.026*	0.004	0.018
$\widehat{S}_{OLS}$	0.697*	0.091	0.505
$\widehat{\alpha}_{OLS}$	0.111*	0.011	0.102
$\widehat{c}_{OLS}$	0.031*	0.004	0.025
$\widehat{S}_{IV,ADJ}$	0.560*	0.080	0.381
$\widehat{\alpha}_{IV,ADJ}$	0.157*	0.022	0.136
$\widehat{c}_{IV,ADJ}$	0.044*	0.009	0.020
$\widehat{S}_{IV}$	0.657*	0.096	0.406
$\widehat{\alpha}_{IV}$	0.157*	0.022	0.136
$\widehat{c}_{IV}$	0.051*	0.011	0.023
DIFFERENCES	Average	Std. error	Median
$\widehat{S}_{OLS} - \widehat{S}^*$	0.117*	0.011	0.100
$\widehat{\alpha}_{OLS} - \widehat{\alpha}^*$	-0.017*	0.007	-0.006
$\widehat{c}_{OLS} - \widehat{c}^*$	0.005*	0.001	0.004
$\widehat{S}_{IV,ADJ} - \widehat{S}^*$	-0.020	0.069	-0.014
$\widehat{\alpha}_{IV,ADJ} - \widehat{\alpha}^*$	0.030	0.017	0.029
$\widehat{c}_{IV,ADJ} - \widehat{c}^*$	0.018*	0.008	0.004
$\widehat{S}_{IV,ADJ} - \widehat{S}_{OLS}$	-0.137*	0.068	-0.098
$\widehat{\alpha}_{IV,ADJ} - \widehat{\alpha}_{OLS}$	0.046*	0.015	0.044
$\widehat{c}_{IV,ADJ} - \widehat{c}_{OLS}$	0.013	0.008	0.003

Effective spreads,  $S$ , and market impacts,  $c$ , are estimated from the regression  $\Delta P_t = \frac{\delta}{2}\Delta Q_t + cQ_{t-1} + \varepsilon_t$ , where  $c = \alpha S/2$ , and the transactions price,  $P$ , was logged and multiplied by 100 (thus making all estimates expressed as percentages of the traded price). Estimators labelled with an '\*' were obtained from true indicators. Estimators labelled with  $IV$  are instrumental variable estimators (using 6 lags of  $Q_t$ ). Estimators labelled with  $ADJ$  were adjusted (by multiplying by  $1 - 2\delta$ ). Estimators labelled  $OLS$  are ordinary least squares estimators. OLS and instrument estimators were obtained by using the erroneous trade-indicators,  $Q_t$ .

Table 3: Effective spread and market impact using OLS and IV estimators.

	AVERAGES			MEDIANS		
	$\hat{S}^*$	$\hat{S}_{OLS}$	$\hat{S}_{OLS} - \hat{S}^*$	$\hat{S}^*$	$\hat{S}_{OLS}$	$\hat{S}_{OLS} - \hat{S}^*$
	0.581	0.699	0.117*	0.382	0.511	0.099
	AVERAGES			MEDIANS		
Instruments	$\hat{S}_{IV,ADJ}$	$\hat{S}_{IV,ADJ} - \hat{S}^*$		$\hat{S}_{IV,ADJ}$	$\hat{S}_{IV,ADJ} - \hat{S}^*$	
3	0.525	-0.057		0.361	0.099	
4	0.538	-0.043		0.398	-0.005	
5	0.567	-0.013		0.362	0.026	
6	0.562	-0.020		0.381	-0.014	
7	0.573	-0.006		0.424	0.017	

This table illustrates the robustness of the spread estimators to the number of instruments. The median columns for  $S_{OLS} - \hat{S}^*$ ,  $\hat{S}_{IV,ADJ} - \hat{S}^*$  are the medians of cross-sectional pairwise differences.

Table 4: Robustness with respect to choice of instruments.

[Finucane 2000, Odders-White 2000]. One explanation is that some buys (sells) are reported as taking place below (above) the bid (ask). In this case buys mistaken for sells will tend to increase the spread instead of narrowing the spread as discussed above.

Nevertheless, one sees that the effective spread estimated using either of the IV estimators is much close to  $\hat{S}^*$  than the OLS estimator. Furthermore, the difference,  $\hat{S}_{IV,ADJ} - \hat{S}^*$ , is not significant while the other differences are. This confirms my conjecture: OLS estimates are biased, while the bias is significantly reduced by using any of the IV estimators. Finally, there is no sign that the IV estimators have higher variance than the OLS estimators.

Similar comments apply to the estimators of  $\alpha$  and  $c = \alpha \frac{S}{2}$ . The OLS estimators are biased, while the differences,  $\hat{\alpha}_{IV,ADJ} - \hat{\alpha}^*$ , and  $\hat{c}_{IV,ADJ} - \hat{c}^*$ , do not deviate significantly from 0. Furthermore,  $\hat{\alpha}_{IV,ADJ}$  is significantly different from  $\hat{\alpha}_{OLS}$ , while the difference  $\hat{c}_{IV,ADJ} - \hat{c}_{OLS}$  is not significant (outside 2 standard errors).

Overall, I conclude that OLS estimation is likely to provide biased estimates of regression type market microstructure models using erroneous trade classifications. On the other hand the example shows that IV estimation reduces bias in estimation of effective spreads and adverse selection parameters.

These results used 6 instruments in the spread regression (2). This was based

on a somewhat informed guess. Nevertheless, the results do not appear sensitive to this choice as table 4 shows. The estimate,  $\hat{S}_{IBV,ADJ}$ , appears to stabilize when the number of instruments increase.

In applied research one would like to have a data driven method for the number of instruments. There are tests in the literature for the adequacy of instrumental variables estimators, which could be applied. However, things become a little bit complicated by the double application of IV estimation. An information criterion like BIC would also require a formal ML estimation of the model.

## 5 Conclusion and comments on future research

It is well-known that errors in trade classification will cause bias in OLS estimators for regression type microstructure models.

The bad news of the paper is that this bias can be quite substantial under reasonable assumptions on the probability of errors and trade reversals. Thus, cross-sectional comparisons may be incorrect alone for the fact that the probability of classification error and trade-reversal varies between stocks. For example, it is known from other studies that classification errors depend on factors such as proximity of prices to quotes, on firm size, and trading intensity. Furthermore, it is well-known that reporting standards may or may not split trades into small chunks, and therefore severely affect the serial correlation properties of trade-indicators. This fact combined with the findings of this paper is reason for concern for applied research.

The good news in this paper is that a simple IV estimation, which can be done using any econometric package, completely removes the dependency on trade reversal probabilities. In fact the bias depends only - and in a simple way - on the error probability. It was demonstrated that even a crude estimate of the error rate yields a better (that is less biased) estimate than the raw OLS estimator.

Research, however, needs to be done with respect to methods for determining the optimal number of instruments. A computer simulation study may be useful for understanding the consequences of classification errors under more general model assumptions. Also research is needed for understanding the problems related to other regression-type microstructure models. As an example, VAR-models [Hasbrouck 1991, Hasbrouck 1993], are typically estimated by applying OLS to regression equations of exactly the same type. Therefore, inference in such models will also be biased. The precise nature of this bias is not clear.

## References

- Aitken, M. and Frino, A.: 1996, The accuracy of the tick test: Evidence from the Australian Stock Exchange, *Journal of Banking and Finance* **20**, 1715–1729.
- Billingsley, P.: 1961, *Statistical Inferences for Markov Processes*, Vol. II of *Statistical Research Monographs*, The University of Chicago Press, Chicago and London.
- Ellis, K., Michaely, R., and O’Hara, M.: 2000, The accuracy of trade classification rules: Evidence from Nasdaq, *Journal of Financial and Quantitative Analysis* **35**(4), 529–551.
- Finucane, T. J.: 2000, A direct test of methods for inferring trade direction from intra-day data, *Journal of Financial and Quantitative Analysis* **35**(4), 553–576.
- Glosten, L. R. and Harris, L.: 1988, Estimating the components of the Bid/Ask spread, *Journal of Financial Economics* **21**, 123–142.
- Harris, L.: 1990, Estimation of stock price variances and serial covariances from discrete observations, *Journal of Financial and Quantitative Analysis* **25**(3), 291–308.
- Hasbrouck, J.: 1991, The summary informativeness of stock trades: An econometric analysis, *The Review of Financial Studies* **4**(3), 571–595.
- Hasbrouck, J.: 1992, Using the TORQ database, *Technical report*, New York Stock Exchange, 11 Wall Street, New York, NY 10005.
- Hasbrouck, J.: 1993, Assessing the quality of a security market: A new approach to transaction-cost measurement, *Review of Financial Studies* **6**(1), 191–212.
- Huang, R. D. and Stoll, H. R.: 1997, The components of the bid-ask spread: A general approach, *The Review of Financial Studies* **10**(4), 995–1034.
- Lee, C. M. C. and RadhaKrishna, B.: 2000, Inferring investor behavior: Evidence from TORQ data, *Journal of Financial Markets* **3**(2), 83–111.
- Lee, C. M. C. and Ready, M.: 1991, Inferring trade directions from intraday data, *Journal of Finance* **46**, 733–746.

Odders-White, E. R.: 2000, On the occurrence and consequences of inaccurate trade classification, *Journal of Financial Markets* **3**(3), 259–286.

Theissen, E.: 2001, A test of the accuracy of the Lee/Ready trade classification algorithm, *Journal Of International Financial Markets, Institutions & Money* **11**(2), 147–165.