ERRORS IN TRADE CLASSIFICATION: CONSEQUENCES AND REMEDIES[1]

Carsten Tanggaard[2]

Working Paper D03-6
Department of Finance
Aarhus School of Business

June 2003.

[1]An earlier version of this paper has been presented to: *Euro Working Group on Financial Modeling XXXI Meeting*, November 2002, *European Meeting of the Econometric Society*, August 2002, *Annual Meeting of the French Finance Association*, June 2002. 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 help and discussions and Esben Høg and Dan Weaver for their comments on an earlier version of the paper.

[2]Department of Finance, Aarhus School of Business, Fuglesangs Alle 4, DK 8210 Aarhus V, Denmark. E-mail: carsten@tanggaard.com.

Abstract

The consequences of errors in trade classification are potentially worse than documented in existing empirical research. This is demonstrated by the use of a formal model of classification errors in a generic regression-type microstructure model. The bias is a function of the probability of trade-reversal in addition to the probability of an error. These parameters depend on stock and trade characteristics in addition to trading procedures and trade reporting standards. The bias is highly sensitive to the background variables, thus causing concern about the validity of empirical studies applying possibly erroneous classification methods without controlling for such effects. The theory, outlined in the paper, predicts that given empirical evidence on error rates, effective spreads must realistically be expected to be downward biased by more than 50%. However, the bias one can observe from using the TORQ database is less severe and has the opposite sign. This is due to special features of the NYSE trading process which may not carry over to other markets. This research also emphasizes the need for proper adjustment of classification error bias. Therefore, the paper proposes a GMM estimator for improved estimation. Simulation evidence indicates that in medium and larger sized samples the method is capable of removing virtually all the bias in market quality statistics like effective, realized, and adverse selection spreads. This is empirically verified in an application to data from TORQ.

JEL: G12, C22, C51.

Keywords: Adverse selection, classification error, effective spread, errors-in-variables, GMM, measurement errors, realized spread, 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 initiated the trade. In markets such as the NYSE auction market, the Nasdaq dealer market, and electronic exchanges with automatic match of limit orders, one can often 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 the number of shares a field in the trade record from 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 an unobserved variable. In any case improper estimation of trade direction causes noise and bias in inference.

Random errors can be due to inappropriate sampling methods and lack of data, volatile markets, manual handling of orders and trades, or from delayed trade-reports. 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, its use may be subject to errors. In particular, delays in trade reporting and non-synchronous time stamps are important sources of error as can be the price improvement efforts of the NYSE specialists.

Other errors stem from inappropriate assumptions. For example, if a brokerage service is involved on both sides of a block trade then it may be wrong to consider one side as initiating the trade, and the whole idea of trade classification becomes dubious for such trades. Another example is the assumption that trade initiators always trade at non-zero effective spreads. This assumption cannot be justified empirically. In fact, it has been shown that significant numbers of trades execute outside the opposite quotes [sells above the ask and buys below the bid]. See Ellis, Michaely and O'Hara (2000) and Finucane (2000) for evidence.

The most accurate classification methods require information about the orders participating in the trade. In an ideal situation 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. Complications arise if there is some kind of human interference in the trading process. As an example, if trades are executed in a specialist market where market orders are stopped for price improvement purposes [Lee and RadhaKrishna (2000)].

Dealer markets, where one part of the trade is a dealer and the other part is a broker or an investor, also allow for similar accurate classification in that investors and brokers [trading on behalf of investors] can be assumed to be trade-initiators [see, for example, Ellis et al. (2000)].

An alternative is to identify the initiating part with the latest arriving order. This method provides an accurate classification when applied to data from an automatic order match system. In such systems the arrival time for one of the orders - a market order or a marketable limit order - will coincide with the time of trade execution, and the latest arriving order may be regarded as initiating the trade.

Order information is generally not included in publicly available databases from the exchanges. In particular, the widely used TAQ data from the NYSE has only trade and quote information. Glosten and Harris (1988), in their study of the components of the bid-ask spread, only have a trade record. They apply a statistical estimation technique to infer the trade-sign [see Harris (1990) for a technical description].

A simpler alternative is the tick rule, which only requires a sequence of trade prices. It classifies trades according to the most recent non-zero price change. If the most recent price change was positive [negative] the trade is classified as a buy [sell]. If the most recent price change was zero, one can use the sign of the most recent classified trade. The tick rule has been applied by regulators on the NYSE and was used for classifying trades without having quote data by Holthausen, Leftwich and Mayers (1987). The tick rule can also be used in reverse order [Hasbrouck (1988), Lee and Ready (1991)].

The quote method applies information about proximity to prevailing quotes to classify trades. Assuming that trade-initiators trade at non-zero effective spreads, a trade executed above the mid-quote is a buy and a trade executed below the mid-quote is a sell. Mid-quote trades represent a tie, which can be handled by the tick rule. If one does not want to classify such trades, an alternative is to code them with a zero [effectively saying that such trades were executed at a zero effective spread].

The quote method - with the tick-rule used as a tie-breaker for mid-quote trades - is often called the Lee-Ready algorithm. Lee and Ready (1991) studied the properties of the tick rule, and examined the problem of quote changes preceding trade reports on the NYSE. They documented that comparing trades to quotes posted at least 5 seconds prior to the trade is a good heuristic solution to this problem.

The literature abounds with empirical research based on microstructure trading data and with trade classification based on the quote method and the tick rule. Finucane (2000) has a survey and an extensive list of published research using these methods on data from the NYSE, Nasdaq, and other US markets.

Due to its widespread use, it is important to know the accuracy of the Lee-Ready method and to know which factors cause it to fail.

The quote method is always wrong when a trade is executed at a negative effective spread, i.e. when a buy [sell] is executed below [above] the mid-quote. That this occurs frequently can be seen in the reported error rates for the quote rule [Odders-White (2000)] and it has been studied explicitly by Ellis et al. (2000) and Finucane (2000). Such errors can be the result of using erroneous quote data or they can be the results of price improving efforts by the NYSE specialist or a Nasdaq dealer.

The tick rule is likely to fail when markets are trending, or when volatility is high compared with the tick size [see Aitken and Frino (1996)]. Such market conditions make it more likely that a non-zero price change is due to changes in fundamentals and not a trade-reversal.

Using data from TORQ, Lee and RadhaKrishna (2000) show that under optimal conditions the expected success rate for the Lee-Ready method can approach 93%[1]. This, however, requires removal of mid-quote trades and trades which cannot be identified in publicly available databases like TAQ. Odders-White (2000) examines NYSE stocks in the TORQ database and finds an error rate of 15.03% for the Lee-Ready method applied to NYSE stocks. The reason for the discrepancy between the two studies is the exclusion by Lee and RadhaKrishna (2000) of trades that are likely to receive price improvement. Finucane (2000), Odders-White (2000) examine the factors behind misclassification of NYSE trades. Mid-quote trades, crossed trades, trades on quote revisions, and trades on zero ticks are more likely to be misclassified [Finucane (2000)] as are small trades and trades in stocks with high trading intensity and price improved trades [Odders-White (2000)]. Finucane (2000) finds that misclassification stemming from the use of the Lee-Ready method leads to an average bias of 17.29% in effective spreads. For the tick rule, the bias is only -5.30%. For the Lee-Ready method the bias is smallest for the low turnover stocks, whereas for the tick rule the bias is smallest for high volume stocks. For the Lee-Ready classification, the bias is related to the amount of price improvement as measured by sells [buys] executed above [below] the mid-quote.

Ellis et al. (2000) study the factors contributing to misclassification on Nasdaq, and find that the combined Lee-Ready method only classifies 81.05% of all Nasdaq trades correctly. They also find that trades inside the quote, trades in high volume stocks, and ECN trades have a higher than average likelihood of being misclassified.

[1]See Hasbrouck (1992) for a description of the TORQ database.

Ellis et al. (2000) propose an alternative to the Lee-Ready method: use the quote method for trades executed at the current quote, and use the tick rule for other trades. They show that this new method improves the classification success rate marginally, but improves estimates of effective spreads by 10%.

A study by Aitken and Frino (1996), using data from the Australian Stock Exchange (ASE) fully automatic order match system, documented that the tick rule classifies 74.4% of the trades correctly. This is a significantly lower success rate than the 90% reported in Lee and Ready (1991), and also somewhat worse than the results for NYSE stocks documented in Finucane (2000), Lee and RadhaKrishna (2000), Odders-White (2000) using TORQ data. In the Australian market the misclassification is related to volatility [trending market]. The problem is likely to be worsened by the relatively low tick size applied by the ASE (1 cent) compared to the 1/8$ tick size at the time TORQ was compiled.

Theissen (2001) examined data from Frankfurt Stock Exchange and found a 72.8% success rate for the Lee-Ready algorithm. Interestingly, Theissen (2001) also found that liquid stocks are more likely to be classified correctly than low volume stocks, which is in contrast with evidence from the NYSE [Odders-White (2000)] and Nasdaq [Ellis et al. (2000)]. Theissen (2001) showed that classification errors cause an upward bias of 36.6% for effective spread estimation when using the Lee-Ready method on stocks from the German market[2].

This paper presents an econometric model of classification errors in time series regression models. Apparently, the existing research [Aigner (1973), Bollinger (1996), Klepper (1988), for example] has not directly dealt with the special problems pertinent to time series data, and the existing methods are not applicable in the context of trade-classification errors. The paper presents analytical expressions for the asymptotic bias in general linear time series regressions with binary regressors subject to sign error. These results are then used within the Glosten (1987) model to illustrate the bias in market quality statistics like effective and realized spread, adverse selection spread and adverse component of the spread caused by errors in trade classification.

In addition, the paper proposes GMM estimation for proper estimation of regression-type microstructure models. The proposed method uses non-standard instruments. It can be used for bias-reduced estimation without assuming too much market specific information about factors underlying classification errors. Firstly, by direct estimation of the structural parameters together with the parameters of the contami-

[2] It must be noted that the comparisons in Theissen (2001) rely on two rather different definitions of the effective spread.

nation process. Secondly, by an explicit - and simpler - way of incorporating exogenous information about error rates. For example, one can refer to existing research and assume that the Lee-Ready method classifies 15%, say, of the trades wrongly. Even such a crude estimate may be better than ignoring the problem by assuming no errors, which appears to be a common practice today. The paper presents some evidence that even a crude exogenous estimate of the error rates results in better estimation quality than unadjusted ordinary least squares.

A simulation study of the finite-sample properties in a Glosten (1987)-type microstructure model indicates that GMM estimation reduces overall estimation error [mean-squared-error] in medium and larger sized sample [with 1000 trades or more]. The empirical application to TORQ data shows that the time series properties of trade-indicators are consistent with the model assumptions. Furthermore, it is demonstrated in the application to TORQ data that GMM estimation removes virtually all bias in the estimation of effective spreads and produces results which are as good as OLS for other market quality statistics.

One conclusion from this paper is that errors in trade-classification represent a potentially worse problem than suggested by the empirical microstructure literature[3]. This has important implications for empirical research. For example, for cross-sectional and cross-sample comparisons. This is because trading strategies and reporting standards, as well as trading intensity and security characteristics, affect the probability of error as well as the clustering of trades [see Hasbrouck and Ho (1987), Huang and Stoll (1997), and Odders-White (2000)]. Moreover, these factors are shown in this paper to have a substantial influence on the bias of estimated parameters. Thus, cross-sectional comparisons without proper control for these background variables may cause errors in inference[4].

[3] The intuition is that the trade-indicator is not only a qualitative variable. Statistics like effective spread and signed trading volume depend on the *size* of the trade-indicator even though it is a binary variable. An almost trivial example gives the intuition. If one associates buys with a 1 and sells with a -1 then a 15% probability of error leads to an average absolute error of 30% in the trade-indicator even when errors are completely random. This is not magic with numbers and the problem is further exacerbated by the well-documented serial correlation in order flow.

[4] With a few exceptions [including the papers reviewed here] there seems to be little awareness of the consequences of trade-sign errors, let alone methods for correcting bias. There seems to be a common understanding that the empirically documented error rates do not constitute a serious problem. One reason for the apparent lack of awareness is probably that the empirical research on US market document rather small empirical biases [Odders-White (2000), for example]. However, the relatively small biases observed in US market data do not necessarily carry over to non-US markets. See section 5 for a further discussion.

This paper proceeds as follows. Section 2 proceeds with a discussion of the consequences of trade classification errors for spreads and measures of the components of the bid-ask spread. Section 3, and the accompanying appendix A, presents a general framework for analyzing consequences of classification errors and for remedying the effects using GMM. Section 4 examines the finite-sample properties using a Monte Carlo simulation experiment. Section 5 presents an empirical application to NYSE trades using the TORQ database. Finally, section 6 gives some concluding remarks.

2 Consequences of classification errors: effective and realized spreads

Consider the Glosten (1987) model of transactions price changes

$$\Delta P_t = c\Delta Q_t + zQ_t + \upsilon_t, \tag{1}$$

where ΔP_t is the time difference of the traded price, Q_t is the trade-indicator [1 if the trade at time t is buyer initiated, -1 if it is seller initiated], and υ_t is a random error. The theoretical model in Glosten (1987) is similar in spirit to the model by Glosten and Milgrom (1985) of trading when some some traders have private information. A more richly parameterized version, which included size-dependent parameters, was tested in Glosten and Harris (1988) and, using a slightly different specification, in Huang and Stoll (1997).

The model decomposes trade price changes into three components: bid-ask bounce [$c\Delta Q_t$], market impact of a trade [zQ_t], and the residual term [υ_t]. The latter captures general market movements caused by public information and microstructure effects like price discreteness. More details can be found in Glosten and Harris (1988) and in the survey by Huang and Stoll (1997). The trading environment, in which the theoretical model is valid, is a rather specialized dealer market. However, similar reduced form models are routinely applied to a variety of market structures, like the NYSE specialist market and limit order driven markets [see de Jong, Nijman and Röell (1996) for an application to an order-driven market].

This model will be used throughout the paper to illustrate theoretical and empirical results. The effective spread, S_e, is defined as twice the difference between the traded price and the current average of the bid and ask. As such the effective spread is two times the market maker's gross profit in each trade. The effective spread is measured in the model (1) as $S_e = 2 \times (z + c)$.

The realized spread, S_r, is defined as the unit costs of a round trip trade: a buy followed by an immediate sale. The realized spread is measured in this model as

$S_r = 2c + z$. It can be seen as the market makers profit net of adverse selection costs. The adverse selection costs arise when the market maker trades with privately informed investors [Glosten and Milgrom (1985)], and z is called the adverse selection spread. The adverse selection component of the spread is the adverse selection spread divided by the effective spread: $\alpha = z/[2(z+c)]$. The idea of expressing the adverse selection component of the spread as a fraction of the effective spread comes from Huang and Stoll (1997). In this paper it is merely used as an example to illustrate the consequences of estimating ratios using biased estimators of nominator/denominator.

In the absence of classification errors, estimation of (1) requires no particular assumptions on the process for Q_t. However, in the presence of classification errors the properties of least squares regressors depend on $Q_{t.}$ Therefore, assume that Q_t is generated by a first order autoregression,

$$Q_t = (1-2\pi)Q_{t-1} + \zeta_t, \tag{2}$$

where π is the probability of a trade-reversal, i.e. a buy followed by a sell or vice-versa. The markov assumption is generally consistent with existing empirical research using trade-indicator data. Let the probability of a classification error in Q_t be constantly equal to δ.

Strictly speaking it is the true process, which is assumed to be markovian. When Q_t is measured subject to a sign-error the markov property may disappear and least squares regression in (2) leads to biased estimates. However, the bias of the least squares regressor, $\widehat{\pi}$, [from equation (2)] can be found from

$$\operatorname{P}\lim \widehat{\pi} = \frac{1}{2}[1 - k^2 h], \tag{3}$$

where $k = (1-2\delta)$ and $h = (1-2\pi)$. This follows from a general expression discussed in section 3.

Clearly, the estimator is consistent, $\operatorname{P}\lim \widehat{\pi} = \pi$, if $\delta = 0$. However, it is somewhat difficult to asses the magnitude of the bias when $\delta > 0$.

INSERT TABLE 1 ABOUT HERE

As table 1 shows, the bias can be quite substantial under realistic assumptions on the error rate. For example, when $\delta = 0.15$ and $\pi = 0.2$ the proportionate bias is 76.5%. This means that if the true $\pi = 0.2$ then one must expect that the least squares estimate is 76.5% higher or $\widehat{\pi} \approx 0.35$. Note that $\pi = 0.2$ is the hypothetical true value, and unbiasedness only occurs in the case of uncorrelated order flow [when

$\pi = 0.5$]. Persistence in order flow corresponding to values of π well below 0.5 is, however, well documented [see Hasbrouck and Ho (1987), Huang and Stoll (1997)]. Huang and Stoll (1997) [in their table 5] present estimates of π ranging from 0.0711 to 0.2401 with an average of π =0.1601. Thus, in general one must expect least squares estimates of π to be substantially biased, and a bias of 76.5% may even be a conservative assessment of the error.

As discussed extensively in the microstructure literature, mean-reversion in order-flow may occur when market makers stabilize inventories by adjusting quotes in response to trades [see Hasbrouck (1988), Huang and Stoll (1997)]. But negative [$\pi > 0.5$] serial correlation in Q_t causes the bias to change sign with the absolute value of the bias being of the same order of magnitude for values of π equally distant from 0.5.

Turning to model (1) the bias of all parameters can be derived from general expressions given in section 3. Details about the formulas will also be given in the tables presented below. The examples originate from a base example, in which $c = 0.3$, $z = 0.05$.

INSERT TABLE 2 ABOUT HERE

Table 2 shows the theoretical [asymptotic] proportionate bias of the effective spread for error rates ranging from 0.05 to 0.3 and for trade reversal probabilities ranging from 0.1 to 0.9. In general, the effective spread is severely biased when error rates are high and when order flow is persistent. The bias is smallest when $\pi \lessapprox 1$. For the example of $\pi = 0.2$ and $\delta = 0.15$ the effective spread is biased by -57%. The examples in table 2 correspond to a true effective spread of $S_e = 0.7$. If the theory is correct then one must expect a least squares estimate which is only $(1 - 0.57) \times 0.7 \approx 0.30$.

That the effective spread is downward biased follows from the model assumptions. The Glosten (1987) model assumes that all trades are executed at non-negative effective spreads. Thus, a misclassified trade will empirically appear as having been executed at a negative spread. Thus, the average effective spread estimated from (1) tends to be smaller than the true spread when some trades are misclassified[5].

INSERT TABLE 3 ABOUT HERE

The bias of the realized spread is exemplified in table 3. The order of magnitude is the same as for the effective spread although for $\pi < 0.5$ the realized spread is more

[5]Interestingly, the sign of the bias is the opposite of the sign documented in empirical research using US data [see Ellis et al. (2000) and Odders-White (2000) and section 5 for a further discussion].

biased than the effective spread and conversely for mean reverting order flow [$\pi > 0.5$]. Furthermore, unless some kind of trade consolidation or bunching is exercised on the data, most markets exhibit mean aversion in the reported order flow [see Hasbrouck and Ho (1987), Huang and Stoll (1997)]. Thus, the examples confirm one of the conjectures from the introduction: order flow correlation, in general, exacerbates the problems with classification error bias.

INSERT TABLE 4 ABOUT HERE

The bias of the adverse selection spread, z, decreases with increasing order flow correlation. For the example of $\pi = 0.2$ and $\delta = 0.15$, the bias is -13.4%, while it is -60.3% if the order flow correlation is $\pi = 0.8$ [see table 4]. In general, as table 4 shows, the bias can be quite substantial when order flow is strongly mean reverting.

INSERT TABLE 5 ABOUT HERE

Table 5 illustrates the dangers of constructing ratios of biased estimators. For larger error rates [$\delta \geq 0.2$] the adverse selection component, $\alpha = z/(2c + 2z)$, is remarkably more biased than any of the other statistics. For the example of $\pi = 0.2$ and $\delta = 0.15$ the theoretical bias is 101.2, which means that the estimated value is more than twice the size of the true value. Finally, the bias is smallest in the case of no serial correlation in order flow.

INSERT TABLE 6 ABOUT HERE

The numbers used in these examples [$c = 0.3, z = 0.05$] are roughly consistent with the numbers for the median firm in the TORQ database as reported in the empirical section 5 below. It is easily seen from the formulas for $\mathrm{P}\lim(\widehat{S}_r)$ and $\mathrm{P}\lim(\widehat{z})$ [see tables 3, 4] that the proportionate bias for the realized and adverse selection spread does not depend on c and z. Furthermore, as the example of table 6 shows, the bias of the effective spread, S_e, and the adverse selection component, α, are not overly sensitive to the choice of c and z. For example, if c is multiplied by a factor 1/5, the bias of S_e is reduced from -57.0% to -49.7%. A similar minor effect is observed when c is multiplied by 5. The bias of $\widehat{\alpha}$ appears slightly more sensitive, which may result from the effect of estimating a ratio using biased nominator/denominator estimates.

3 Measurement error bias in regression-type microstructure models

The classical measurement error model for regressions has a long tradition in econometrics and statistics [see Berkson (1950), Durbin (1954), Wald (1940) with even older references in Wald (1940)]. Measurement errors [errors-in-variables] occur when an estimated or proxy variable is used in place of the true unobserved observation, or simply if data is contaminated with errors. Such problems are in fact quite common in financial economics. For example, realized returns are used in the rational expectations asset pricing literature instead of unobserved expectations. Parameter estimates and averages from single-equation estimations are used routinely as data in cross-sectional comparisons. A stock market index is often used as a proxy for the return on the unobserved market portfolio, which fundamentally flaws tests of the CAPM [see Roll (1977)]. See also Maddala and Nimalendram (1996) for a further discussion of errors-in-variables problems in financial economics.

The measurement error problem for regressions can be interpreted as an identification problem, or as the result of using an inadequate estimator resulting in biased estimation.

The identification problem arises because the OLS estimator is the minimizer of the vertical distance from observation points to the regression line. Therefore, unless one variable is truly exogenous, there are two different regression lines either by regressing y on x or x on y. This causes a problem when there is a measurement error in x as x can then not serve as a truly exogenous regressor [Berkson (1950), Durbin (1954)]. The true regression line when there is only 1 regressor will have a slope, which is somewhere in-between the two opposite regressions. Wald (1940) emphasizes the identification issue when regressors are subject to measurement error: one must have exogenous information, for example, about the ratio of standard deviations of x and y in order to identify the regression. Berkson (1950) notes that the bias vanishes if the *observed* regressor is a controlled variable. This amounts to saying that the orthogonality condition in least squares regression applies to the observed data and not to the true regressors. This assumption is, of course, highly specialized and cannot be expected to apply to non-experimental financial data. Klepper and Leamer (1984) apply a method of moments technique in a least squares regression with errors in all continuous regressors.

Another interpretation of the errors-in-variables problem is that the observed residual becomes correlated with the true regressors, which violates standard assumptions of the regression model. Finally, when there are measurement errors in regressions, the average cross-product of the regressors [the $\frac{1}{T}X'X$ matrix] is not a consistent estimator of the true regressor covariance matrix. It is straightforward to see in the case of one regressor how this leads to a downward bias in estimated coefficients.

Sign errors in regressors are multiplicative in nature, and the solution methods for additive measurement errors do not carry over to binary regressors. Thus, regressions with errors in binary regressors require special treatment.

Aigner (1973) shows that in a regression with one binary regressor subject to classification error, the least squares estimator is inconsistent with a bias factor which equals the probability that a given observation is *not* subject to classification error.

A modified least squares approach conditional on knowing the covariance between classification errors and the regressor leads to a consistent estimator of all coefficients [see Aigner (1973)]. Freeman (1984) applies the results from Aigner (1973) to study the effects of measurement errors in research on union effects. Interestingly, Freeman (1984) emphasizes the severe consequences of misclassification in longitudinal labour market studies using survey data. The probability of misclassification in union studies is of the same order of magnitude as documented in trade classification research, and Freeman (1984) shows that misclassification accounts for the majority of the differences in conclusions between longitudinal and cross-sectional studies of union effects. Related research on the consequences and remedies of misclassification in labour market surveys is found in Card (1996), Mathiowetz and Duncan (1988), Poterba and Summers (1986).

Using results from Klepper and Leamer (1984), Klepper (1988) derives bounds for the true parameters when there are more than one binary regressor, and when the probability of an error does not depend on the true classification. This independence assumption was relaxed for one binary regressor in Bollinger (1996).

The existing methods for classification errors in binary regressions relate to models with one binary regressor, to cross-sectional data, or to panel data. These models do not fit with features that are pertinent to time series. The multivariate model in Klepper (1988), for example, does not apply here because the same classification error affect lagged values of the regressors.

It is assumed 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 a classification error occurs then

$$Q_t = (1 - 2I_t)Q_t^*, \tag{4}$$

where Q_t^* denotes the underlying true trade-indicator, and I_t is an error indicator taking the values 0 and 1: 1 if there is an error, 0 otherwise.

As such (4) is just a definition. The equation imposes no restriction on the possible kind of error. A proper analysis of the problem requires more structure. For that purpose assume that Q_t^* is a markov chain with time-invariant transition probabilities and with Q_t^* being independent of I_t. Unless otherwise stated, I_t is serially uncorrelated, although most results of this paper apply when I_t is serially correlated - a markov chain, for example [see A for further details].

The definition in equation (4) implies that the classification error is multiplicative rather than additive. One could replace (4) with an additive specification, $Q_t = Q_t^* + Z_t$, where $Z_t = -2I_tQ_t^*$. In fact, an additive specification was used in most of the articles cited above on binary classification errors. However, the additive error is not independent of the true classification, which makes it impossible to use IV-estimation. In addition, it is not natural to use an additive specification for a problem, which is in fact multiplicative in nature.

Full independence of Q_t^* and I_t is likely to be too strong an assumption. For example, it is well known that in NYSE market data there is serial correlation in Q_t^* due to trading features and the splitting of crossed block trades in the trade reports [see Hasbrouck and Ho (1987) and Madhavan and Cheng (1997)]. Therefore, the serial correlation in Q_t^* is probably contagious and will occur in I_t with a similar pattern. Further evidence on the similar autocorrelation patterns in Q_t^* and I_t is presented in section 5 below. Nevertheless, this paper proceeds under the assumption of independence between Q_t^* and I_t. This can be partly justified by a reference to the empirical results in section 5, which suggest that even under this simplifying assumption, the GMM estimator is a significant improvement over the least squares estimator.

3.1 Example: trade-indicator autoregression

The examples in section 2 assumed that the trade-indicator is a first order markov chain [or autoregressive process].

This markov assumption is actually taken for granted in much empirical market microstructure research. Either explicitly [Madhavan, Richardson and Roomans (1997)] or implicitly by assuming that Q_t is serially uncorrelated [Roll (1984), Glosten and Harris (1988), Harris (1990) , George, Kaul and Nimalendram (1991)] or by assuming that $\mathrm{E}_{t-1}(Q_t) = (1 - 2\pi)Q_{t-1,}$ where π is the probability of a trade-reversal [Choi, Salandro and Shastri (1988), Huang and Stoll (1997)]. Finally, it is some-times assumed that the data generating process is a vector autoregression [Hasbrouck (1991), Hasbrouck (1993)]. In addition, the assumption is empirically justified by the analysis in section 5.

Often, though, a first order autoregression does not fit empirically. For example, the results reported from VAR modeling in Hasbrouck (1991), Hasbrouck (1993) indicate that higher order lag structures are needed. In order to maintain generality and facilitate applications to VAR models, it will be assumed that Q_t^* is an mth order markov chain with transition probabilities

$$\begin{aligned} \mathrm{P}(Q_t^* &= i|Q_{t-1}^* = j_1, Q_{t-2}^* = j_2, \ldots, Q_{t-m}^* = j_m) \\ &\equiv p(i|j_1, j_2, \ldots, j_m), \; i, j_k \in \{-1, 1\}. \end{aligned} \tag{5}$$

The markov model can be given an autoregressive representation

$$Q_t^* = \sum_{k=1}^{m} \varphi_k Q_{t-k}^* + \zeta_t^*. \tag{6}$$

By construction, the $\zeta_t^{*\prime}s$ are identically distributed and serially uncorrelated, and fulfill $\mathrm{E}(\zeta_t^* Q_{t-k}^*) = 0$ for $k > 0$. However, they are clearly not *iid*.

The relationship between (5) and (6) is

$$p(1|j_1, j_2, \ldots, j_m) = \frac{1}{2}(1 + \sum_{k=1}^{m} j_k \varphi_k), \tag{7}$$

$$p(-1|j_1, j_2, \ldots, j_m) = \frac{1}{2}(1 - \sum_{k=1}^{m} j_k \varphi_k). \tag{8}$$

Thus, in the absence of classification errors, the transition probabilities can be estimated by first fitting (6) using least squares regression and then deriving $p(i|j_1, j_2, \ldots, j_m)$ from (7), (8).

Let $\mathrm{P}(I_t = 1) = \delta$, and let φ denote the vector $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_m)'$. Also, let $\mu = 1 - 2\delta$, and define $\rho_Q^*(i-j) \equiv \mathrm{Corr}(Q_{t-i,}^* Q_{t-j}^*)$.

Least squares regression is applied to the observed version of (6)

$$Q_t = \sum_{k=1}^{m} \varphi_k Q_{t-k} + \zeta_t. \tag{9}$$

It is shown in the appendix A that for the least squares regressor, $\widehat{\varphi}$

$$\mathrm{P}\lim \widehat{\varphi} = \left[P^* + \psi \Delta\right]^{-1} P^* \varphi, \tag{10}$$

where P^* is the $m \times m$ matrix with $\rho^*_Q(i-j)$ in the ijth cell, Δ is [here] the identity matrix, and $\psi = \frac{4\delta(1-\delta)}{(1-2\delta)^2}$.

It follows directly from equation (10) that $\widehat{\varphi}$ is inconsistent unless $\delta = 0$. Note, that the results in appendix A are slightly more general as they allow for serial correlation in I_t and for general multiplicative errors rather than just sign-errors.

3.2 Example: returns regressed on trade-indicator

The Glosten (1987) model [equation (1)] is a special case of the trade-indicator regression

$$\Delta P_t = \sum_{i=0}^{n} \beta_i Q^*_{t-i} + \upsilon^*_t \tag{11}$$

with $n = 1, \beta_0 = c + z, \beta_1 = -c, \beta_j = 0, j \geq 2$.

In the absence of classification errors, the term υ^*_t must fulfill standard assumptions for GMM or OLS to apply.

Let $\beta = (\beta_0, \ldots, \beta_n)'$. Then least squares estimation is applied to the observed version of (11)

$$\Delta P_t = \sum_{i=0}^{n} \beta_i Q_{t-i} + \upsilon_t \tag{12}$$

It is shown in appendix A that for the least squares estimator of β

$$\text{P}\lim \widehat{\beta} = \mu^{-1} \left[P^* + \psi\Delta\right]^{-1} P^*\beta, \tag{13}$$

where P^*, Δ - except for now being of dimension: $(n+1) \times (n+1)$ - were defined above. It follows directly that $\widehat{\beta}$ is inconsistent unless $\delta = 0$.

3.3 Bias-reduced estimation

There are several methods of circumventing errors-in-variables bias beyond simply ignoring them. A commonly applied technique is instrumental variable estimation. Instruments can be constructed, for example, by grouping of variables [Durbin (1954)]. In time series applications the serial correlation of regressors facilitates the use of lagged observations as instruments [see Greene (2000)]. The latter method is commonly applied in the time series asset pricing literature, where observed variables are used in place of unobserved rational expectations, and lagged observations are used as instruments. Unfortunately, it does not work directly as a method with binary regressors and sign-errors. This is because, by construction, the *additive* classification error is correlated with the true classification, which violates the conditions for the standard instrumental variable approach.

The GMM technique, however, provides a unifying framework, which is useful for bias correction using exogenous information, and for direct estimation of the error process under suitable model assumptions.

Suppose then that data for T trades are given. Then it is shown in appendix A that the following equations hold for the autoregression [equation (6)]

$$\frac{1}{T}\sum_{t=1}^{T}\left(\vartheta_k Q_t Q_{t-k} - \sum_{j=1}^{m}\vartheta_{k-j}\varphi_j Q_{t-j}Q_{t-k}\right) = 0, k \geq 1, \tag{14}$$

where

$$\vartheta_k = \begin{cases} 1 & k = 0 \\ 1/\mu^2 & k \neq 0 \end{cases}.$$

Turning to the time series regression [model (11)] it is shown in appendix A that the following equations hold

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mu^{-1}Y_t Q_{t-l} - \sum_{j=0}^{n}\vartheta_{l-j}\beta_j Q_{t-j}Q_{t-l}\right) = 0,\ l \geq 0, \tag{15}$$

For (14) as well as (15) it is the *observed noisy* trade-indicator which enters the equations. These equations can be applied in a GMM framework for improved estimation.

Let θ be a vector compiled of the unknown parameters φ, β, δ. Then - conditional on the given data - the right hand sides in (14), (15) are functions of θ. Collecting these functions into a vector, $d(\theta)$, the GMM estimator is

$$\widehat{\theta} = \arg\min_{\theta}\ d(\theta)'Wd(\theta), \tag{16}$$

where W is weight-matrix. This weight matrix can be fixed - the identity matrix, for example - or it can be estimated from data as described in the GMM-literature [see Hamilton (1994)].

In many cases there is exogenous information available on the magnitude of classification errors. For example, one may assume that the probability of a classification error is 0.15, say, when using the Lee-Ready method. This estimate can be justified by a reference to some of the existing studies on NYSE, Nasdaq, and other exchanges cited in the introduction. Even though the estimate is not strictly correct, it may be a more accurate assumption leading to less bias than simply assuming that there is no error. Given such an estimate, the GMM problem [equation (16)] has a linear solution. The resulting estimator can be implemented as an OLS or IV-estimator with suitably redefined product and cross-product matrices[6].

4 A simulation experiment

Standard results on asymptotic normality of GMM estimators are accurate in large samples, but may be poor descriptions in finite and small samples. The same is the case for the probability limits [the formulas (10) and (13)]. In addition, probability limits are only one dimension of estimator quality. In finite samples the variance is as important for estimation quality as the bias.

Monte Carlo simulation is a useful tool to examine the finite sample properties. An experiment was conducted within the Glosten (1987) trade-indicator model. Thus, consider equation (1) with $c = 0.30$, $z = 0.05$, and $v_t \sim \mathrm{N}(0, \sigma^2)$ with $\sigma = f \times 2(c+z) = f \times S_e$ for some fixed f. The basic experiment fixed f at $f = 0.5$. These numbers altogether are roughly consistent with the numbers for the median firm in the TORQ database.

The trade-indicators, Q_t^*, were generated from a binary markov chain with $\pi = \mathrm{P}(Q_t^* = i | Q_{t-1}^* = j), i \neq j$. π was set to 0.2 in the simulated data and subsequently estimated from (6) with $m = 1$ and $\varphi_1 = 1 - 2\pi$. The contamination process, I_t, was generated as a serially uncorrelated process with $\mathrm{P}(I_t) = \mathrm{P}(I_t = 1) = \delta$.

Each simulation run generated time series of fixed length, T. The number of simulation runs was $R = 100000$ for $T = 100$, $R = 10000$ for $T = 1000$, and $R = 1000$ for $T = 10000$. The probability of an error was fixed at three different values $\delta = 0.05, \delta = 0.10$, and $\delta = 0.15$.

The parameters of interest were estimated using GMM and OLS estimators. The GMM estimator is the solution of (16) using 6 instruments and the identity as weight-matrix. The number of instruments was chosen rather arbitrarily, but the overall conclusions are not sensitive at all to this choice.

The parameters of interest are the probability of trade-reversal [π], the effective spread [S_e], the adverse selection spread [z], and the adverse selection component of the spread [$\alpha = z/S_e$].

Mean-squared-error and bias are used as the main evaluation criteria for the simulation experiments. Let x denote some parameter with true value, x_0, and let $\widehat{x}_i$ be

[6] Details on these formulas are available on request.

the estimate from the ith simulation run. For each parameter, x, the mean-squared error, MSE, is defined as $\mathrm{E}(\widehat{x} - x_0)^2$, and estimated as

$$MSE(\widehat{x}) = \frac{1}{R}\sum_{i}^{R}(\widehat{x}_i - x_0)^2,$$

where R is the number of simulation runs, and $\widehat{x}_i$ is the estimate in simulation run $\#i$. The root mean-squared-error is $RMSE = \sqrt{MSE}$.

The bias is defined as $Bias = \mathrm{E}(\widehat{x} \text{ -} x_0)$, and it is estimated as

$$Bias(\widehat{x}) = \frac{1}{R}\sum_{i}^{R}\widehat{x}_i - x_0.$$

The following research questions are examined in the simulation study:

Q1: Is the GMM estimator better than OLS in finite samples?

Q2: How sensitive are the conclusions to changes in disturbance volatility [changes in f]?

Q3: Is a fixed δ better than GMM estimation of δ?

Q4: What are the effects of mis-specification of the contamination process, e.g. a wrong assumption on the value δ?

Q5: What is the effect of serial correlation in I_t?

Let $Variance$ denote the variance, $\mathrm{V}(\widehat{x}_i)$, of a parameter estimate. Then because $MSE = Bias^2 + Variance$, the ratio $\frac{Bias^2}{MSE}$ is a measure of the relative importance of bias compared to variance in the estimation of x. This number is reported together with MSE and $Bias$ in the tables 7, 8, and 9.

INSERT TABLE 7 ABOUT HERE
INSERT TABLE 8 ABOUT HERE
INSERT TABLE 9 ABOUT HERE

These results again demonstrate the damaging effect of trade-classification errors. Even for modest error rates, $\delta = 5\%$, and a sample size of $T = 100$, the OLS bias term [more precisely the squared bias] accounts for about 77% of the error in the estimation of the effective spread, S_e. This ratio increases, and becomes close to 1, in large samples or when the classification error rate increases [see tables 8, and 9]. The intuition is that in large samples [asymptotic] bias becomes the dominant component of the mean-squared-error.

For the GMM estimator, the bias plays a decreasing role in comparison with variance. For smaller samples [$t \leq 1000$] and $\delta = 0.05$, the estimation error is comparable to the OLS error.

In large samples [$T > 1000$], however, the GMM estimator improves relatively to ordinary least squares. But for $\delta = 0.05$, it requires a sample size of $T = 10000$ before GMM is a real improvement over OLS. The intuition is that GMM requires the estimation of an extra parameter, δ, which adds further noise, and the low probability of an error implies little bias that GMM can remove.

Turning to the results for $\delta = 0.10$ and $\delta = 0.15$ [in tables 8 and 9] the same qualitative results hold, except that GMM is relatively better than OLS for a given sample size.

The overall conclusion from this experiment is that for smaller samples [$T < 1000$] there is no benefit from using GMM in comparison with OLS, and it is only for $\delta = 0.15$ that there is a visible improvement by using GMM in comparison with OLS when $T = 1000$. Accordingly, the remaining experiments were conducted for $T = 1000$ and $T = 10000$ only.

INSERT TABLE 10 ABOUT HERE

These simulation results originated from a model with $f = 1/2$ [$\mathrm{V}(v_t) = (fS_e)^2$], i.e. with a standard deviation in the Glosten (1987) model $\approx 1/2S_e$.

Table 10 shows the results from an experiment, where f is varied from 0.1 to 0.9.

The parameters of interest are S_e, z, α. This experiment has no implication for π, and π is left out of this and the remaining experiments. The simulation results are presented as ratios of estimation error: $\frac{RMSE(\mathrm{GMM})}{RMSE(\mathrm{OLS})}$. Thus, a ratio less than 1 indicates that the GMM estimator performs relatively better than the OLS estimator. The error rate was fixed at $\delta = 0.10$.

For $T = 1000$ as well as $T = 10000$ a reduction in $\mathrm{V}(v_t)$ causes a relative improvement of GMM over OLS. This is because it causes estimation variance to contribute relatively less than bias to overall estimation error. For $T = 1000$ the worst case for the GMM estimator is for the estimation of the adverse selection spread, z, where the ratio is $\gtrapprox 1$. The reason is that variance dominates the estimation error for that parameter, and the importance of GMM in this context is bias-reduction. For $T = 10000$ the ratio is substantially smaller than 1 for almost all parameters, and for all variance regimes. The exception is again for the adverse selection spread and for large variances [$f > 0.7$], where the ratio is close to but less than 1.

INSERT TABLE 11 ABOUT HERE

One conclusion from the first experiments is that estimation of an extra parameter [δ] cannot be justified in small samples or when the bias contributes little to the overall estimation error. As an alternative to estimation of all parameters using GMM, one may assume a fixed and known value of δ and then estimate the remaining parameters conditional on the given estimate. The estimate may be an informed guess or come from external sources like in the labour market studies [see Freeman (1984)].

Table 11 presents some evidence on the performance of this procedure. The simulation experiment used a fixed value of $\delta = 0.10$ to generate data. The estimation, however, was carried out assuming three different fixed values of $\delta : \delta = 0.05, \delta = 0.10$, and $\delta = 0.15$.

The simulation results are presented as ratios of estimation error: $\frac{RMSE(\text{GMM}|\widehat{\delta})}{RMSE(\text{OLS})}$ and $\frac{RMSE(\text{GMM}|\widehat{\delta})}{RMSE(\text{GMM})}$. The former indicates the performance relative to OLS estimation, while the latter shows the results relatively to full GMM estimation. A ratio less than 1 thus indicates that the overall estimation error is smaller for the chosen fixed value, $\widehat{\delta}$, compared to the benchmark estimator.

For $T = 1000$ the GMM$|\widehat{\delta}$ method is generally better than OLS when $\widehat{\delta}$ is fixed at the true value [$\delta = 0.10$] or lower. The exception being the adverse selection spread, z, where variance dominates the estimation error and bias corrected estimation does not help a lot. GMM estimation with fixed $\widehat{\delta}$ generally performs worse than GMM with estimated $\widehat{\delta}$.

For $T = 10000$ the fixed-$\widehat{\delta}$ method outperforms OLS even when $\widehat{\delta}$ is chosen below the true value, but it does not outperform GMM with estimated $\widehat{\delta}$.

Thus, the overall conclusion from this experiment is that in order to reduce bias and estimation error [compared to OLS], one should estimate $\widehat{\delta}$ by GMM or assume an optimistic [i.e. too low] fixed value for $\widehat{\delta}$.

INSERT TABLE 12 ABOUT HERE

The analysis in the next section [section 5] indicates that the classification error process, I_t, one can observe in the TORQ database is serially correlated. More precisely I_t is well-described as an AR(1) with a 1. order serial correlation of $\rho \approx 0.4$.

The 1. order serial correlation coefficient, ρ, can be estimated jointly with the other parameters $[\delta, \varphi, \beta]$ using the GMM-estimator. Experiments - using simulated as well as trading data from TORQ - however, indicated that ρ is difficult to identify empirically.

Evidence on this problem is presented in table 12, which shows the relative estimation error when estimating $[\delta, \rho]$ jointly with the other parameters. The results are presented as ratios: $\frac{RMSE(\text{GMM})}{RMSE(\text{GMM}|\widehat{\rho}=0)}$. Thus, a relative estimation error < 1 indicates that a full GMM procedure outperforms the GMM procedure with $\widehat{\rho}$ fixed at $\widehat{\rho} = 0$. The data was simulated using two different true values for $\rho : \rho = 0.2$ and $\rho = 0.4$. The results for $T = 1000$ indicate that full GMM estimation of δ, ρ can generally not be justified in comparison with assuming a fixed value: $\widehat{\rho} = 0$. For $T = 10000$ the error ratio is > 1 for all parameters and all values of ρ. The intuition is that most of the bias is removed by introducing [estimating] the parameter δ. Introduction of an extra correlation parameter, ρ, removes little bias but instead adds extra noise. Apparently, the extra variance does not go away in large samples[7].

A summary of the simulation results is as follows. GMM and bias-corrected estimation can generally be justified in larger samples [where $T \geq 1000$], and when the probability of a classification error is substantial [$\delta \geq 0.10$]. This conclusion is generally not sensitive to the disturbance variance. Assuming a fixed value of δ is generally better than ordinary least squares unless this fixed value exceeds the true value. However, a full GMM procedure is typically even better, especially in large samples. Finally, serial correlation in the classification error process can in general be ignored by assuming that $\widehat{\rho} = 0$.

These simulation results are likely to under-estimate the need for and value of bias-corrected estimation. The simulation experiment used a single-equation approach. In empirical microstructure research one often compares cross-sectional averages between samples. This leads to further variance reduction [effectively increasing the sample size]. Thus, for empirical research, effective bias-correction by the GMM method may be even more useful than indicated by this simulation study.

[7] It is a bit surprising to see that a mis-specified model outperforms a correctly specified model also in large samples. An additional experiment [not reported here] for $T = 100000$ verified this result for a very large sample.

5 Empirical application to TORQ data

The TORQ database is the only publicly available database with detailed order and audit information from the NYSE. TORQ has trades, orders, quotes, and the audit trail for 144 stocks traded in November 90 through January 91. The sample is stratified, and it is representative for all stocks listed on the NYSE during that period [Hasbrouck (1992)]. The TORQ database provides a unique opportunity to study the properties of the true trade-indicator and the contamination process as well as the effects of bias correction methods.

The idea is to construct the true trade-indicator using information on time stamps of participating orders. This approach was used in the article by Odders-White (2000). This trade-indicator can serve as a benchmark for testing the Lee-Ready and tick rule classification methods. Clearly, the validity of the inference is conditional on the truthfulness of the benchmark.

Lee and RadhaKrishna (2000) criticize the method in Odders-White (2000) for including trades involving market orders on both sides. Such trades can be wrongly "true-classified" by the method in Odders-White (2000) because one market order may be a stopped order. Instead, Lee and RadhaKrishna (2000) characterize orders such as [non-stopped] market orders and marketable limit orders as clearly active [trade-initiators]. This information is available in TORQ, and it allows for a comparison with the Lee-Ready method. Lee and RadhaKrishna (2000) showed that 40% of the trades cannot be classified unambiguously as either buyer or seller initiated because these trades have active traders on both sides.

Despite the criticism by Lee and RadhaKrishna (2000), the analysis in this paper uses the same approach as in Odders-White (2000). Firstly, because it creates a larger sample, and secondly because the discrepancy in conclusions between the two methods is not clearly in favor of the Lee and RadhaKrishna (2000) approach. Thirdly, because the difference is not overly important for the purpose of this paper.

5.1 Constructing trade indicators

The data was pre-filtered by removing: non-NYSE trades and quotes, trades with correction code different from zero, trades executed before 9:45, audit trail records for which there were no corresponding trade record, and outlier quotes as described in Chung, Ness and Ness (2001).

Audit trail records were linked to order records by the use of the fields: `report time` in the audit trail and `buy time/sell time` in the order record. Audit records and trade records were linked using the sequence number. Audit records that could not be linked on both sides were discarded.

The hypothetical true trade indicator for each audit record was constructed by identifying the trade initiator with the latest arriving order. Trades that could not be classified were removed from the sample.

The quote method and the tick rule trade-indicators were constructed using standard methods. The combined Lee-Ready method applied the 5 seconds delay rule of thumb. The procedure in its totality seems to follow the method applied in Odders-White (2000)[8].

INSERT TABLE 13 ABOUT HERE

The total number of trades left in the sample is 316846 after filtering and removal of non-classified trades [see table 13]. This number is comparable to the 318364 trades in the sample analyzed by Odders-White (2000). The difference is 1518 trades $\approx 0.48\%$.

The error rates for the three different methods are quite similar [numbers from Odders-White (2000) in brackets]: 10.72% [10.80%] for the quote method, 18.49% [21.37%] for the tick rule, and 15.32% [15.03%] for the Lee-Ready algorithm. An exception is the tick rule, where the difference is almost 2 percentage points.

In order to mitigate effects of small samples and outliers, the trades were filtered according to the rules described in Chung et al. (2001), and stocks with fewer than 100 trades after filtering were left out. The resulting sample consisting of 122 stocks is used for the examination of time series properties.

5.2 Time series properties

The Lee-Ready method is used throughout the rest of this paper and the bias-reduction methods are applied to trades signed using this method.

INSERT TABLE 14 ABOUT HERE

The theoretical model of classification errors depends on the time series properties of the trade-indicator, Q_t^*, and the classification error process, I_t. In particular, the GMM estimator assumes that Q_t^*, is markov.

[8]Elizabeth Odders-White kindly helped by explaining details in her procedure. I am responsible for any error.

Table 14 summarizes the sample moments of the true classification process, Q_t^*, and the contamination process, I_t. The contamination process is defined as $I_t = (1 - Q_t/Q_t^*)/2$, where Q_t was estimated using the Lee-Ready method.

The statistics in table 14 show that there is substantial autocorrelation of orders 1 to 3. This is true for the trade-indicator process as well as for the classification error process. Furthermore, the partial autocorrelations are virtually zero after lag 1. This is the typical behavior for an autoregressive process of order 1. This supports the theoretical assumption about a markov chain model for the trade-indicator process as well as the contamination process. Note, the similarity in the serial correlation pattern for Q_t^* and I_t. Both processes are autoregressive, and both have a 1. order serial correlation ≈ 0.4. As discussed earlier, this can be a sign of dependence and common features in the two processes.

Hitherto, it was assumed that I_t is serially uncorrelated. The model framework of appendix A, however, allows for generally serially correlated I_t. Nevertheless, the simulation experiments indicated that the serial correlation coefficient is difficult to identify in data. Furthermore, it does not cause a lot of difference to the empirical results if it is assumed to be 0. The empirical results for TORQ reported next give further support to this view.

5.3 An application to the Glosten (1987) trade-indicator model

Table 16 presents results from an application of the OLS and GMM estimators to the Glosten (1987) model. The OLS estimator is applied to the Lee-Ready classified data as well as the benchmark data using trade-indicators obtained from order data [labelled *'benchmark'* in the table]. The difference between the two sets of estimates depends on the bias resulting from classification errors.

Based on the conclusions from the simulation experiment the sample includes only stocks with more than 1000 trades. Furthermore, the serial correlation coefficient [in I_t] was fixed at two different values, $\widehat{\rho} = 0.0$ and $\widehat{\rho} = 0.2$.

Table 16 has the results for the 3 parameters of interest the effective spread [S_e], the adverse selection spread [z], and the adverse selection component of the spread [α].

The most interesting results are for the effective spread, where the average OLS bias is 0.11 [0.10 measured at the median]. The t-statistic for the mean-difference across the sample is clearly significant.

In addition table 16 has results for the GMM estimator [equation (16)] using 6 moment equations [of the form (15)] and the identity as weight-matrix.

The GMM estimator is clearly less biased. The cross-sectional average effective spread is 0.57 for the benchmark data, while it is 0.59 for the GMM estimator using Lee-Ready classified data when $\widehat{\rho} = 0.0$. The mean-difference is only borderline significant at the 5% level. Also, the medians are very similar for $\widehat{S}^{*}$ and $\widehat{S}^{GMM}$. The results for $\widehat{\rho} = 0.2$ are similar.

For the adverse selection statistics, z and α, there does not seem to be any advantage by estimation using GMM. Although the cross-sectional average and median of z, α are both close to the benchmark values, the same is the case for the OLS estimators. This was really also one of the predictions from the simulation experiments.

Apparently, the spread-bias does not have the sign predicted by the theory, nor does it have the predicted magnitude. The effective spread estimated from the TORQ data is upward biased, whereas the theory predicts a downward bias. The order of magnitude of the empirical estimates is consistent with other empirical studies as reviewed in the introduction.

The explanation is likely to be the following. There are two sources of error when using the quote/Lee-Ready method. Firstly, the quotes used for signing trades may be erroneous. This causes a downward bias as described elsewhere in the paper.

Secondly, the quote-method assumes that all trades execute at non-negative effective spreads. That is a buy is executed at a price above the current mid-quote, and a sell is executed below. However, if a trade is price improved, and therefore effectively trade at a negative spread, then a misclassification results in an upward bias of the estimated spread. Price improvement of that size is examined and documented in the literature. See Ellis et al. (2000) for documentation from Nasdaq data, and Finucane (2000) for evidence from NYSE [TORQ]. If both types of errors occur in a given data set, then the effect on the average spread may well be a mitigating factor for the average bias across all trades. And if the latter type of "error" dominates, then the effect may be that biases are reversed as the results for TORQ indicate.

Note, however, that the opposite sign of the bias compared to predictions by theory has no effect on the GMM estimator. As expected, it appears to correct the bias in effective spreads estimation quite nicely.

6 Conclusion

This paper has demonstrated that the consequences of wrong estimation of the trade sign are potentially worse than empirical studies of the US markets have indicated. One reason is that the effect of classification errors is mitigated by two effects offsetting each other. To the extent that this mitigating effect does not exist in non-US data the consequences for inferences may be worse for such markets.

The paper presented a model of classification errors in a binary regressor model. This was used to demonstrate the damaging effect of classification errors in a regression-type microstructure model - the Glosten (1987) model. In addition, the paper showed how this model can be utilized to adjust for classification error bias using a GMM-type estimator. A simulation experiment examined the finite-sample properties. Finally, the empirical section studied the time series properties of trade-indicators and classification errors and quite successfully applied the methods to a sample of NYSE traded stocks from the TORQ database.

A Multiplicative errors in variables

The model in this appendix is slightly more general than the one discussed in the main sections of the paper. Firstly, the model allows for a general multiplicative error process - not only binary sign-errors. The specification here also allows for general sign-errors in variables such as signed trade size and - to some extent - for regressors to be functions of a binary variable. The latter allows for trade-size dependent bid-ask spread and adverse selection as in Huang and Stoll (1997). In addition, the model allows for a serially correlated measurement error process.

Let X_t denote a mean zero stationary time series process observed subject to a stationary multiplicative measurement error, U_t. Thus,

$$X_t = U_t X_t^*,$$

where U_t and X_t^* are independent processes such that $\mathrm{E}(U_t) \neq 0$.

In the classification error model of this paper: $X_t^* = Q_t^*$, $X_t = Q_t$, $U_t = (1 - 2I_t)$.

Let $\mu = \mathrm{E}(U_t), \sigma_U^2 = \mathrm{Var}(U_t)$, $\rho_U(i-j) = \mathrm{Corr}(U_{t-i}, U_{t-j})$. Furthermore, let $\psi = \sigma_U^2/\mu^2$ denote the squared coefficient of variation of U_t.

By definition $\mathrm{E}(U_{t-i}U_{t-j}) = \mathrm{V}(U_t) + \mathrm{E}(U_{t-i})\mathrm{E}(U_{t-j}) = \sigma_U^2 \rho_U(i-j) + \mu^2$. Thus, by independence

$$\begin{aligned} \mathrm{E}(X_{t-i}X_{t-j}) &= \mathrm{E}(U_{t-i}U_{t-j})\mathrm{E}(X_{t-i}^* X_{t-j}^*) \\ &= \left[\sigma_U^2 \rho_U(i-j) + \mu^2\right] \mathrm{E}(X_{t-i}^* X_{t-j}^*). \end{aligned}$$

Therefore, $\mathrm{E}(X_{t-i}^* X_{t-j}^*) = \vartheta_{i-j}\mathrm{E}(X_{t-i}X_{t-j})$, where $\vartheta_k = \left[\sigma_U^2 \rho_U(k) + \mu^2\right]^{-1}$.

Let us consider two models: an autoregression and a time series regression.

A.1 Autoregression

Let the autoregression for the true unobserved, X_t^*, be

$$X_t^* = \sum_{i=1}^{m} \varphi_i X_{t-i}^* + \zeta_t^*. \tag{17}$$

X_t is observed subject to multiplicative noise: $X_t = U_t X_t^*$, and the following autoregression is estimated instead of (17)

$$X_t = \sum_{i=1}^{m} \varphi_i X_{t-i} + \zeta_t. \tag{18}$$

Let the superscript $'$ denote the transpose of a matrix and let $\varphi = (\varphi_1, \ldots, \varphi_m)'$, $\mathbf{X}_t = (X_t, \ldots, X_{t-m+1})'$.

Then, assuming sufficient number of initial-observations, the least squares estimator [of equation (18)] is

$$\widehat{\varphi} = \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{t-1} \mathbf{X}_{t-1}' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{t-1} X_t \right).$$

Under the assumptions of this model: $\mathrm{P}\lim \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{t-1} \mathbf{X}_{t-1}' \right) = \mathrm{E}\left(\mathbf{X}_{t-1} \mathbf{X}_{t-1}' \right)$ in which the ijth element is $\mathrm{E}(X_{t-i}^* X_{t-j}^*) \left[\sigma_U^2 \rho_U(i-j) + \mu^2 \right]$. Thus,

$$\mathrm{P}\lim \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{t-1} \mathbf{X}_{t-1}' \right) = \mu^2 \left(P^* + \psi \Delta \right),$$

where P^* is the matrix with ijth element $\mathrm{E}(X_{t-i}^* X_{t-j}^*)$, and Δ is the matrix with ijth element equal to $\rho_U(i-j)\mathrm{E}(X_{t-i}^* X_{t-j}^*)$.

Furthermore, $\mathrm{P}\lim \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{t-1} X_t \right) = \mathrm{E}\left(\mathbf{X}_{t-1} X_t \right)$ with ith row equal to

$$\begin{aligned}
\mathrm{E}\left(X_{t-i} X_t \right) &= \mathrm{E}\left(X_{t-i} U_t X_t^* \right) \\
&= \mathrm{E}\left(X_{t-i} U_t \mathbf{X}_{t-1}^{*\prime} \varphi \right) + \mathrm{E}\left(X_{t-i} U_t \zeta_t \right) \\
&= \mathrm{E}\left(U_t U_{t-i} X_{t-i}^* \mathbf{X}_{t-1}^{*\prime} \varphi \right) \\
&= \mathrm{E}\left(U_t U_{t-i} \right) \mathrm{E}\left(X_{t-i}^* \mathbf{X}_{t-1}^{*\prime} \varphi \right) \\
&= \left[\mu^2 + \sigma_U^2 \rho_U(i-j) \right] \mathrm{E}\left(X_{t-i}^* \mathbf{X}_{t-1}^{*\prime} \right) \varphi, \\
i &= 1, \ldots, m.
\end{aligned}$$

Therefore,

$$\mathrm{P}\lim \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_{t-1} X_t \right) = \mu^2 \left(I + \psi \Sigma \right) P^* \varphi,$$

where $\Sigma = diag\left[\rho_U(1), \rho_U(2), \ldots, \rho_U(m) \right]$. The combination of these results yields

$$\mathrm{P}\lim \widehat{\varphi}^{OLS} = (P^* + \psi\Delta)^{-1} [I + \psi\Sigma] P^* \varphi.$$

In the special case of serially uncorrelated U_t

$$\mathrm{P}\lim \widehat{\varphi}^{OLS} = (P^* + \psi\Delta)^{-1} P^* \varphi,$$

and when $\psi = 0$ [no errors at all] the OLS estimator is clearly consistent.

A GMM estimator can be derived by exploiting the autoregressive nature of the true process, X_t^*. One has $\mathrm{E}(X_{t-k}^*\zeta_t) = 0$ for $k \geq 1$. This can be rewritten as

$$\mathrm{E}\left(X_t^* X_{t-k}^* - \sum_{j=1}^{m} \varphi_j X_{t-j}^* X_{t-k}^*\right) = 0,\ k \geq 1,$$

or in terms of observable data, as

$$\mathrm{E}\left(\vartheta_k X_t X_{t-k} - \sum_{j=1}^{m} \vartheta_{k-j}\varphi_j X_{t-j} X_{t-k}\right) = 0,\ k \geq 1, \tag{19}$$

where $\vartheta_k = \left[\sigma_U^2 \rho_U(k) + \mu^2\right]^{-1}$.

For GMM estimation one can replace the population moments in (19) with their sample equivalents.

A.2 Time series regression

Consider the time series regression

$$Y_t = \sum_{i=0}^{n} \beta_i X_{t-i}^* + \upsilon_t^*, \tag{20}$$

where Y_t is a [possibly] continuous variable, and υ_t^* is a disturbance term. The regressor, X_t, is observed subject to multiplicative noise: $X_t = U_t X_t^*$, and the following regression is estimated in place of (20)

$$Y_t = \sum_{i=0}^{n} \beta_i X_{t-i} + \upsilon_t. \tag{21}$$

The time series regression from the main section of the paper has $Y_t = \Delta P_t$.

Let $\beta = (\beta_0, \ldots, \beta_n)'$, and $\mathbf{X}_t = (X_t, \ldots, X_{t-n})'$. Then, assuming a sufficient number of initial-observations, the OLS estimator [from equation (21)] is

$$\widehat{\beta}^{OLS} = \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t Y_t\right).$$

The derivation of

$$\mathrm{P}\lim\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t'\right) = \mu^2\left(P^* + \psi\Delta\right),$$

where P^* is the matrix with ijth element $\mathrm{E}(X^*_{t-i}X^*_{t-j})$, and Δ is the matrix with ijth element equal to $\rho_U(i-j)\mathrm{E}(X^*_{t-i}X^*_{t-j})$, follows from the equivalent result for the autoregression.

Furthermore, $\mathrm{P}\lim\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_tY_t\right) = \mathrm{E}\left(\mathbf{X}_tY_t\right)$ with ith row equal to

$$\begin{aligned}
\mathrm{E}\left(X_{t-i+1}Y_t\right) &= \mathrm{E}\left(X_{t-i+1}\mathbf{X}^{*\prime}_t\beta\right) + \mathrm{E}\left(X_{t-i+1}\upsilon_t\right) \\
&= \mathrm{E}\left(U_{t-i+1}X^*_{t-i+1}\mathbf{X}^{*\prime}_t\right)\beta \\
&= \mathrm{E}\left(U_{t-i+1}\right)\mathrm{E}\left(X^*_{t-i+1}\mathbf{X}^{*\prime}_t\right)\beta \\
&= \mu\mathrm{E}\left(X^*_{t-i+1}\mathbf{X}^{*\prime}_t\right)\beta, \\
i &= 1,\ldots,n+1.
\end{aligned}$$

Therefore,

$$\mathrm{P}\lim\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_tY_t\right) = \mu P^*\beta.$$

Combining these results yields

$$\mathrm{P}\lim\widehat{\beta}^{OLS} = \mu^{-1}(P^* + \psi\Delta)^{-1}P^*\beta.$$

The GMM estimator is derived from the orthogonality condition, $\mathrm{E}(X^*_{t-i}\upsilon^*_t) = 0$. This can be rewritten as

$$\mathrm{E}\left(Y_tX^*_{t-l} - \sum_{j=0}^{n}\beta_jX^*_{t-j}X^*_{t-l}\right) = 0,\ l \geq 0,$$

or in terms of observable data

$$\mathrm{E}\left(\mu^{-1}Y_tX_{t-l} - \sum_{j=0}^{n}\beta_j\vartheta_{l-j}X_{t-j}X_{t-l}\right) = 0,\ l \geq 0. \tag{22}$$

For GMM estimation the equations (22) can be replaced by their sample equivalents.

In the sign-error example, where $U_t = (1-2I_t)$, one has $\rho_I(i-j) = \mathrm{Corr}(I_{t-i}, I_{t-j})$. U_t is a linear function of I_t, and therefore $\rho_I(k) = \rho_U(k)$ and $\sigma^2_U = 4\sigma^2_I = 4\delta(1-\delta)$. Thus, $\psi = 4\delta(1-\delta)/(1-2\delta)^2$ and $\vartheta_k = \left[4\delta(1-\delta)\rho_I(k) + \mu^2\right]^{-1}$. In the case of serially uncorrelated I_t:

$$\vartheta_k = \begin{cases} 1 & k = 0 \\ (1/\mu)^2 & k \neq 0 \end{cases}.$$

At the expense of more complicated notation these results can be extended to models with additional regressors and to vector autoregressive models of trades and transactions returns [Hasbrouck (1993)].

References

Aigner, D. J.: 1973, Regression with a binary independent variable subject to errors of observation, *Journal of Econometrics* **1**(1), 49–60.

Aitken, M. and Frino, A.: 1996, The accuracy of the tick test: Evidence from the Australian Stock Exchange, *Journal of Banking and Finance* **20**(10), 1715–1729.

Berkson, J.: 1950, Are there two regressions?, *Journal of the American Statistical Association* **45**(250), 164–180.

Bollinger, C. R.: 1996, Bounding mean regressions when a binary regressor is mismeasured, *Journal of Econometrics* **73**, 387–399.

Card, D.: 1996, The effect of unions on the structure of wages: A longitudinal analysis, *Econometrica* **64**(4), 957–979.

Choi, J. Y., Salandro, D. and Shastri, K.: 1988, On the estimation of bid-ask spreads: Theory and evidence, *Journal of Financial and Quantitative Analysis* **38**(2), 219–230.

Chung, K. H., Ness, B. F. V. and Ness, R. A. V.: 2001, Can the treatment of limit orders reconcile the differences in trading costs between NYSE and Nasdaq issues?, *Journal of Financial and Quantitative Analysis* **36**(2), 267–286.

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

Durbin, J.: 1954, Errors in variables, *Revue de l'Institut International de Statistique* **22**, 23–32.

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.

Freeman, R. B.: 1984, Longitudinal analyses of the effects of trade unions, *Journal of Labor Economics* **2**(1), 1–26.

George, T. J., Kaul, G. and Nimalendram, M.: 1991, Estimation of the bid-ask spreads and its components: A new approach, *Review of Financial Studies* **4**(4), 623–656.

Glosten, L. R.: 1987, Components of the bid-ask spread and the statistical properties of transaction prices, *Journal of Finance* **42**(5), 1293–1307.

Glosten, L. R. and Harris, L.: 1988, Estimating the components of the Bid/Ask spread, *Journal of Financial Economics* **21**(1), 123–142.

Glosten, L. R. and Milgrom, P. R.: 1985, Bid, ask and transaction prices in a specialist market with heterogenously informed traders, *Journal of Financial Economics* **14**, 71–100.

Greene, W. H.: 2000, *Econometric Analysis*, 4th edn, Prentice-Hall, Inc.

Hamilton, J. D.: 1994, *Time Series Analysis*, Princeton University Press, 41 William Street, Princeton, New Jersey 0840, USA.

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.: 1988, Trades, quotes, inventories, and information, *Journal of Financial Economics* **22**, 229–252.

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. 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.

Hasbrouck, J. and Ho, T. S. Y.: 1987, Order arrival, quote behavior, and the return-generating process, *The Journal of Finance* **42**(4), 1035–1048.

Holthausen, R. W., Leftwich, R. W. and Mayers, D.: 1987, The effects of large block transactions on security prices, *Journal of Financial Economics* **19**(2), 237–267.

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.

Klepper, S.: 1988, Bounding the effects of measurement errors in regressions involving dichotomous variables, *Journal of Econometrics* **37**(3), 343–359.

Klepper, S. and Leamer, E. E.: 1984, Consistent sets of estimates for regressions with errors in all variables, *Econometrica* **52**(1), 163–183.

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**(2), 733–746.

Maddala, G. S. and Nimalendram, M.: 1996, Errors-in-variables problems in financial models, *in* G. S. Maddala and C. R. Rao (eds), *Handbook of Statistics*, Vol. 14, Elsevier Science B.V., chapter 17, pp. 507–528.

Madhavan, A. and Cheng, M.: 1997, In search of liquidity: Block trades in the upstairs and downstairs markets, *Review of Financial Studies* **10**(1), 175–203.

Madhavan, A., Richardson, M. and Roomans, M.: 1997, Why do security prices change? A transaction-level analysis of NYSE stocks, *Review of Financial Studies* **10**(4), 1035–1064.

Mathiowetz, N. A. and Duncan, G. J.: 1988, Out of work, out of mind: Response errors in retrospective reports of unemployment, *Journal of Business and Economic Statistics* **6**(2), 221–229.

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

Poterba, J. M. and Summers, L. H.: 1986, Reporting errors and labor market dynamics, *Econometrica* **54**(6), 1319–1338.

Roll, R.: 1977, A critique of the asset pricing theory's tests: Part I, *Journal of Financial Economics* **4**, 129–176.

Roll, R.: 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* **39**(4), 1127–1139.

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.

Wald, A.: 1940, The fitting of straight lines if both variables are subject to error, *Annals of Mathematical Statistics* **11**(3), 284–300.

B Tables

Table 1: Bias of probability of trade reversal

Prob. of Trade Reversal (π)	Error rate (δ)					
	0.05	0.10	0.15	0.20	0.25	0.30
0.1	76.0	144.0	204.0	256.0	300.0	336.0
0.2	28.5	54.0	76.5	96.0	112.5	126.0
0.3	12.7	24.0	34.0	42.7	50.0	56.0
0.5	0.0	0.0	0.0	0.0	0.0	0.0
0.7	-5.4	-10.3	-14.6	-18.3	-21.4	-24.0
0.8	-7.1	-13.5	-19.1	-24.0	-28.1	-31.5
0.9	-8.4	-16.0	-22.7	-28.4	-33.3	-37.3

Bias of the least squares estimator of π [probability of trade reversal]. Derived from the autoregression, $Q_t = (1-2\pi)Q_{t-1} + \xi_t$ when Q_t is a binary trade-indicator subject to classification error. The probability of classification error is δ. The bias is measured as a percentage of the true value: $Bias(\widehat{\pi}) = 100 \times [\mathrm{P}\lim(\widehat{\pi}) - \pi]/\pi$, where

$$\mathrm{P}\lim \widehat{\pi} = \frac{1}{2}[1 - \mu^2\varphi],$$

and where $\mu = (1 - 2\delta)$, $\varphi = (1 - 2\pi)$.

Table 2: Bias of effective spread

Prob. of Trade Reversal (π)	Error rate (δ)					
	0.05	0.10	0.15	0.20	0.25	0.30
0.1	−35.5	−62.8	−72.2	−78.4	−83.0	−86.9
0.2	−28.0	−45.2	−57.0	−65.9	−73.2	−79.4
0.3	−19.0	−33.7	−45.6	−55.7	−64.5	−72.4
0.5	−10.0	−20.0	−30.0	−40.0	−50.0	−60.0
0.7	−5.9	−12.6	−20.2	−28.8	−38.5	−49.3
0.8	−5.0	−10.4	−16.8	−24.5	−33.7	−44.5
0.9	−5.1	−9.2	−14.3	−20.9	−29.5	−40.1

Bias of the least squares regression estimator of effective spread, S_e. Derived from the Glosten (1987) regression, $\Delta P_t = c\Delta Q_t + zQ_t + \upsilon_t$, when the trade-indicator, Q_t, is subject to error. The probability of classification error is δ. The effective spread is $S_e = 2 \times (c + z)$. The example parameters are $c = 0.30, z = 0.05$. The bias is measured as a percentage of the true value: $Bias(\widehat{S}_e) = 100 \times [\mathrm{P}\lim(\widehat{S}_e) - S_e]/S_e$, where

$$\mathrm{P}\lim \widehat{S}_e = 2c\left(\frac{\varphi^2\mu^3 + \varphi\mu - \mu - \varphi\mu^3}{\varphi^2\mu^4 - 1}\right) + 2z\left(\frac{\varphi^2\mu^3 - \mu}{\varphi^2\mu^4 - 1}\right),$$

with $\mu = (1 - 2\delta)$, $\varphi = (1 - 2\pi)$.

Table 3: Bias of realized spread

Prob. of Trade Reversal (π)	Error rate (δ)					
	0.05	0.10	0.15	0.20	0.25	0.30
0.1	−48.9	−67.2	−77.0	−83.1	−87.5	−90.8
0.2	−30.0	−48.1	−60.3	−69.4	−76.5	−82.3
0.3	−20.1	−35.5	−47.8	−57.9	−66.7	−74.4
0.5	−10.0	−20.0	−30.0	−40.0	−50.0	−60.0
0.7	−4.8	−10.8	−18.1	−26.6	−36.4	−47.4
0.8	−3.1	−7.5	−13.4	−21.1	−30.4	−41.6
0.9	−1.7	−4.8	−9.5	−16.1	−25.0	−36.2

Bias of the least squares regression estimator of realized spread, S_r. Derived from the Glosten (1987) regression, $\Delta P_t = c\Delta Q_t + zQ_t + \upsilon_t$, when the trade-indicator, Q_t, is subject to error. The probability of classification error is δ. The realized spread is $S_r = 2c + z$. The example parameters are $c = 0.30, z = 0.05$. The bias is measured as a percentage of the true value. $Bias(\widehat{S}_r) = 100 \times [\mathrm{P}\lim(\widehat{S}_r) - S_r]/S_r$, where

$$\mathrm{P}\lim \widehat{S}_r = (2c+z)\left(\frac{\varphi\mu - \varphi\mu^3 - \mu + \varphi^2\mu^3}{\varphi^2\mu^4 - 1}\right)$$

with $\mu = (1-2\delta)$, $\varphi = (1-2\pi)$.

Table 4: Bias of adverse selection spread

Prob. of Trade Reversal (π)	Error rate (δ)					
	0.05	0.10	0.15	0.20	0.25	0.30
0.1	−1.7	−4.8	−9.5	−16.1	−25.0	−36.2
0.2	−3.1	−7.5	−13.4	−21.1	−30.4	−41.6
0.3	−4.8	−10.8	−18.1	−26.6	−36.4	−47.4
0.5	−10.0	−20.0	−30.0	−40.0	−50.0	−60.0
0.7	−20.1	−35.5	−47.8	−57.9	−66.7	−74.4
0.8	−30.0	−48.1	−60.3	−69.4	−76.5	−82.3
0.9	−48.9	−67.2	−77.0	−83.1	−87.5	−90.8

Bias of the adverse selection spread. Derived from the Glosten (1987) regression, $\Delta P_t = c\Delta Q_t + zQ_t + v_t$, when the trade-indicator, Q_t, is subject to error. The probability of classification error is δ. The adverse selection spread is z. The example parameters are $c = 0.30, z = 0.05$. The bias is measured as a percentage of the true value: $Bias(\widehat{z}) = 100 \times [\mathrm{P}\lim(\widehat{z}) - z]/z$, where

$$\mathrm{P}\lim \widehat{z} = z \left(\frac{\varphi\mu^3 + \varphi^2\mu^3 - \varphi\mu - \mu}{\varphi^2\mu^4 - 1} \right)$$

with $\mu = (1 - 2\delta)$, $\varphi = (1 - 2\pi)$.

Table 5: Bias of adverse selection component

Prob. of Trade Reversal (π)	Error rate (δ)					
	0.05	0.10	0.15	0.20	0.25	0.30
0.1	80.4	155.7	225.1	287.5	342.1	388.1
0.2	34.7	68.6	101.2	131.8	159.4	183.4
0.3	17.5	34.5	50.7	65.8	79.3	90.9
0.5	0.0	0.0	0.0	0.0	0.0	0.0
0.7	−15.1	−26.2	−34.6	−40.9	−45.8	−49.4
0.8	−26.3	−42.0	−52.3	−59.5	−64.5	−68.1
0.9	−46.1	−63.9	−73.1	−78.7	−82.3	−84.7

Bias of the adverse selection component of the spread. Derived from the Glosten (1987) regression, $\Delta P_t = c\Delta Q_t + zQ_t + \upsilon_t$, when the trade-indicator, Q_t, is subject to error. The probability of classification error is δ. The adverse selection component of the stread is $\alpha = z/[2(z+c)]$. The example parameters are $c = 0.30, z = 0.05$. The bias is measured as a percentage of the true value: $Bias(\widehat{\alpha}) = 100 \times [\mathrm{P}\lim(\widehat{\alpha}) - \alpha]/\alpha$, where

$$\mathrm{P}\lim \widehat{\alpha} = \frac{z(\varphi\mu^3 + \varphi^2\mu^3 - \varphi\mu - \mu)}{2(c+z)\,(\varphi^2\mu^3 - \mu) + 2c\,(\varphi\mu - \varphi\mu^3)}.$$

with $\mu = (1 - 2\delta)$, $\varphi = (1 - 2\pi)$.

Table 6: Sensitivity of bias to changes in regression parameters

	c				
	0.06	0.15	0.30	0.60	1.5
$Bias(\widehat{S}_e)$	-49.7	–54.5	-57.0	-58.5	-59.6
$Bias(\widehat{\alpha})$	72.0	90.1	101.2	108.7	114.2

Sensitivity of bias to changes in model parameters. Derived from the Glosten (1987) regression, $\Delta P_t = c\Delta Q_t + zQ_t + \upsilon_t$, when the trade-indicator, Q_t, is subject to error. The probability of classification error is δ. The effective spread is $S_e = 2 \times (c + z)$, while the adverse selection component of the spread is $\alpha = z/S_e$ with $z = 0.05, \delta = 0.15, \pi = 0.2$. The bias is measured as a percentage of the true value: $Bias(\widehat{x}) = 100 \times [\mathrm{P}\lim(\widehat{x}) - x]/x$, $x = S_e, \alpha$, where.

$$
\begin{aligned}
\mathrm{P}\lim \widehat{S}_e &= 2c\left(\frac{\varphi^2\mu^3 + \varphi\mu - \mu - \varphi\mu^3}{\varphi^2\mu^4 - 1}\right) + 2z\left(\frac{\varphi^2\mu^3 - \mu}{\varphi^2\mu^4 - 1}\right), \\
\mathrm{P}\lim \widehat{\alpha} &= \frac{z(\varphi\mu^3 + \varphi^2\mu^3 - \varphi\mu - \mu)}{2(c+z)\left(\varphi^2\mu^3 - \mu\right) + 2c\left(\varphi\mu - \varphi\mu^3\right)},
\end{aligned}
$$

with $\mu = (1 - 2\delta)$, $\varphi = (1 - 2\pi)$.

Table 7: Estimation error and bias for GMM and OLS estimators when $\boldsymbol{\delta = 0.05}$.

Parameter	True	OLS			GMM		
		$RMSE$	$Bias$	$\frac{Bias^2}{MSE}$	$RMSE$	$Bias$	$\frac{Bias^2}{MSE}$
$T = 100, \delta = 0.05$							
π	0.2	0.073	0.056	0.587	0.153	0.108	0.498
S_e	0.6	0.227	-0.199	0.768	0.384	-0.300	0.611
z	0.05	0.042	-0.004	0.009	0.042	-0.016	0.135
α	1/14	0.087	0.018	0.042	0.156	0.017	0.012
$T = 1000, \delta = 0.05$							
π	0.2	0.058	0.056	0.944	0.028	0.006	0.054
S_e	0.6	0.199	-0.196	0.972	0.094	-0.029	0.096
z	0.05	0.013	-0.002	0.015	0.014	-0.002	0.021
α	1/14	0.034	0.024	0.502	0.000	0.269	0.020
$T = 10000, \delta = 0.05$							
π	0.2	0.057	0.057	0.995	0.007	0.001	0.013
S_e	0.6	0.197	-0.197	0.997	0.028	-0.005	0.037
z	0.05	0.005	-0.002	0.120	0.005	0.000	0.007
α	1/14	0.026	0.025	0.903	0.006	0.000	0.000

Estimation error and bias in the Glosten (1987) trade-indicator model: $\Delta P_t = c\Delta Q_t + zQ_t + v_t$, with $c = 0.30$ and $z = 0.05$, where P_t is the traded price, and Q_t is the trade-indicator. All numbers were obtained by simulation. Data for Q_t was generated from a markov chain with $\mathrm{P}(Q_t = i|Q_{t-1} = j) = \pi$. Classification errors were generated as a serially uncorrelated process with $\mathrm{P}(I_t = 1) = \delta$. ΔP_t was generated from the regression with $v_t \sim N\,[0, (S_e/2)^2]$. π was estimated from the autoregression $Q_t = \varphi Q_{t-1} + \xi_t$ with $\varphi = 1 - 2\pi$. The parameters of interest are the effective spread, $S_e = 2(c + z)$, the adverse selection spread, z, the adverse selection component of the spread, $\alpha = z/S_e$, and the probability of order-reversal, π. The sample size for each run is T. For a parameter, $x \in \{S, \pi, z, \alpha\}$, MSE is the simulated estimate of $\mathrm{E}(\widehat{x} - x_0)^2$, $Bias$ is the simulated estimate of $\mathrm{E}(\widehat{x} - x_0)$, and $RMSE = \sqrt{MSE}$.

Table 8: Estimation error and bias for GMM and OLS estimators when $\delta = 0.10$.

Parameter	True	OLS			GMM		
		$RMSE$	$Bias$	$\frac{Bias^2}{MSE}$	$RMSE$	$Bias$	$\frac{Bias^2}{MSE}$
$T = 100, \delta = 0.10$							
π	0.2	0.118	0.107	0.823	0.195	0.173	0.789
S_e	0.6	0.337	-0.319	0.900	0.472	-0.438	0.862
z	0.05	0.045	-0.006	0.018	0.044	-0.020	0.219
α	1/14	0.165	0.037	0.049	9.877	-0.074	0.000
$T = 1000; \delta = 0.10$							
π	0.2	0.108	0.107	0.981	0.061	0.024	0.158
S_e	0.6	0.318	-0.316	0.989	0.168	-0.085	0.257
z	0.05	0.015	-0.004	0.072	0.017	-0.005	0.092
α	1/14	0.059	0.048	0.670	0.032	0.005	0.021
$T = 10000; \delta = 0.10$							
π	0.2	0.108	0.108	0.999	0.009	0.001	0.024
S_e	0.6	0.316	-0.316	0.999	0.034	-0.008	0.055
z	0.05	0.006	-0.004	0.400	0.005	-0.001	0.013
α	1/14	0.050	0.049	0.949	0.007	0.000	0.000

Estimation error and bias in the Glosten (1987) trade-indicator model: $\Delta P_t = c\Delta Q_t + zQ_t + v_t$, with $c = 0.30$ and $z = 0.05$, where P_t is the traded price, and Q_t is the trade-indicator. All numbers were obtained by simulation. Data for Q_t was generated from a markov chain with $\mathrm{P}(Q_t = i|Q_{t-1} = j) = \pi$. Classification errors were generated as a serially uncorrelated process with $\mathrm{P}(I_t = 1) = \delta$. ΔP_t was generated from the regression with $v_t \sim N\left[0, (S_e/2)^2\right]$. π was estimated from the autoregression $Q_t = \varphi Q_{t-1} + \xi_t$ with $\varphi = 1 - 2\pi$. The parameters of interest are the effective spread, $S_e = 2(c + z)$, the adverse selection spread, z, the adverse selection component of the spread, $\alpha = z/S_e$, and the probability of order-reversal, π. The sample size for each run is T. For a parameter, $x \in \{S, \pi, z, \alpha\}$, MSE is the simulated estimate of $\mathrm{E}(\widehat{x} - x_0)^2$, $Bias$ is the simulated estimate of $\mathrm{E}(\widehat{x} - x_0)$, and $RMSE = \sqrt{MSE}$.

Table 9: Estimation error and bias for GMM and OLS estimators when $\delta = 0.15$.

Parameter	True	OLS			GMM		
		$RMSE$	$Bias$	$\frac{Bias^2}{MSE}$	$RMSE$	$Bias$	$\frac{Bias^2}{MSE}$
$T = 100, \delta = 0.15$							
π	0.2	0.161	0.153	0.895	0.226	0.218	0.932
S_e	0.6	0.415	-0.402	0.941	0.527	-0.515	0.957
z	0.05	0.049	-0.009	0.031	0.045	-0.023	0.263
α	1/14	3.573	0.065	0.000	333.324	-4.732	0.000
$T = 1000, \delta = 0.15$							
π	0.2	0.153	0.152	0.989	0.187	0.155	0.688
S_e	0.6	0.400	-0.399	0.994	0.437	-0.374	0.733
z	0.05	0.017	-0.007	0.174	0.024	-0.017	0.501
α	1/14	0.084	0.071	0.702	0.090	0.055	0.377
$T = 10000, \delta = 0.15$							
π	0.2	0.153	0.153	0.999	0.011	0.003	0.067
S_e	0.6	0.399	-0.399	1.000	0.047	-0.017	0.131
z	0.05	0.009	-0.007	0.655	0.006	-0.001	0.045
α	1/14	0.073	0.072	0.957	0.008	0.000	0.000

Estimation error and bias in the Glosten (1987) trade-indicator model: $\Delta P_t = c\Delta Q_t + zQ_t + v_t$, with $c = 0.30$ and $z = 0.05$, where P_t is the traded price, and Q_t is the trade-indicator. All numbers were obtained by simulation. Data for Q_t was generated from a markov chain with $\mathrm{P}(Q_t = i|Q_{t-1} = j) = \pi$. Classification errors were generated as a serially uncorrelated process with $\mathrm{P}(I_t = 1) = \delta$. ΔP_t was generated from the regression with $v_t \sim N\left[0, (S_e/2)^2\right]$. π was estimated from the autoregression $Q_t = \varphi Q_{t-1} + \xi_t$ with $\varphi = 1 - 2\pi$. The parameters of interest are the effective spread, $S_e = 2(c + z)$, the adverse selection spread, z, the adverse selection component of the spread, $\alpha = z/S_e$, and the probability of order-reversal, π. The sample size for each run is T. For a parameter, $x \in \{S, \pi, z, \alpha\}$, MSE is the simulated estimate of $E(\widehat{x} - x_0)^2$, $Bias$ is the simulated estimate of $E(\widehat{x} - x_0)$, and $RMSE = \sqrt{MSE}$.

Table 10: Relative estimation error for different noise variance regimes.

Parameter	Residual variance factor, f [$\text{Var}(v_t) = f(S/2)$]				
	0.1	0.3	0.5	0.7	0.9
$T = 1000, \delta = 0.10, \pi = 0.2$					
S_e	0.24	0.33	0.52	0.74	0.94
z	1.02	1.06	1.09	1.08	1.06
α	0.17	0.33	0.57	0.83	1.02
$T = 10000, \delta = 0.10, \pi = 0.2$					
S_e	0.06	0.08	0.12	0.16	0.20
z	0.50	0.70	0.85	0.94	0.99
α	0.05	0.09	0.14	0.19	0.23

The relative estimation error is $\frac{RMSE(\text{GMM})}{RMSE(\text{OLS})}$ in the Glosten (1987) trade-indicator model: $\Delta P_t = c\Delta Q_t + zQ_t + v_t$, with $c = 0.30$ and $z = 0.05$, where P_t is the traded price, and Q_t is the trade-indicator. All numbers were obtained by simulation. Data for Q_t was generated from a markov chain with $P(Q_t = i|Q_{t-1} = j) = \pi$. ΔP_t was generated from the regression with $v_t \sim N\left[0, (fS_e)^2\right]$. The parameters of interest are the effective spread, $S_e = 2(c + z)$, the adverse selection spread, z, and the adverse selection component of the spread, $\alpha = z/S_e$. The sample size for each run is T, and δ is the true probability of a classification error. For a parameter, $x \in \{S_e, z, \alpha\}$, MSE is the simulated estimate of $E(\widehat{x} - x_0)^2$ and $RMSE = \sqrt{MSE}$.

Table 11: Relative estimation error for fixed $\widehat{\boldsymbol{\delta}}$.

Parameter	$\frac{RMSE(\text{GMM}\mid\widehat{\delta})}{RMSE(\text{OLS})}$			$\frac{RMSE(\text{GMM}\mid\widehat{\delta})}{RMSE(\text{GMM})}$		
	Probability of classification error, $\widehat{\delta}$					
	0.05	0.10	0.15	0.05	0.10	0.15
$T = 1000, \delta = 0.10, \pi = 0.2$						
S_e	0.68	0.26	3.52	1.13	0.43	5.84
z	1.01	1.04	1.07	0.91	0.94	0.97
α	0.67	0.37	0.66	0.98	0.55	0.98
$T = 10000, \delta = 0.10, \pi = 0.2$						
S_e	0.67	0.06	2.36	5.74	0.55	20.17
z	0.87	0.84	0.88	1.02	0.98	1.03
α	0.58	0.14	0.72	4.17	1.04	5.23

The relative estimation errors are $\frac{RMSE(\text{GMM}\mid\widehat{\delta})}{RMSE(\text{OLS})}$ and $\frac{RMSE(\text{GMM}\mid\widehat{\delta})}{RMSE(\text{GMM})}$ in the Glosten (1987) trade-indicator model: $\Delta P_t = c\Delta Q_t + zQ_t + v_t$, with $c = 0.30$ and $z = 0.05$, where P_t is the traded price, and Q_t is the trade-indicator. All numbers were obtained by simulation. Data for Q_t was generated from a markov chain with $P(Q_t = i|Q_{t-1} = j) = \pi$. ΔP_t was generated from the regression with $v_t \sim N\left[0, (S_e/2)^2\right]$. The parameters of interest are the effective spread, $S_e = 2(c + z)$, the adverse selection spread, z, and the adverse selection component of the spread, $\alpha = z/S_e$. The sample size for each run is T, and δ is the true probability of a classification error. For a parameter, $x \in \{S_e, z, \alpha\}$, MSE is the simulated estimate of $E(\widehat{x} - x_0)^2$ and $RMSE = \sqrt{MSE}$. OLS is the ordinary least squares estimator, GMM is the GMM estimator with estimated $\widehat{\delta}$ and GMM$|\widehat{\delta}$ is the GMM estimator with fixed and known probability of classification error, $\widehat{\delta}$.

Table 12: Relative estimation error for fixed $\widehat{\rho}$= 0.

Parameter	Probability of classification error (δ)					
	0.05	0.10	0.15	0.05	0.10	0.15
	$\rho = 0.2$			$\rho = 0.4$		
$T = 1000, \pi = 0.2$						
S_e	1.27	1.04	0.87	1.12	0.92	0.84
z	1.03	0.99	0.93	1.03	0.96	0.90
α	0.97	0.87	0.78	0.93	0.83	0.73
$T = 10000, \pi = 0.2$						
S_e	1.71	1.70	1.70	1.99	1.89	1.87
z	1.06	1.08	1.09	1.23	1.26	1.33
α	1.08	1.10	1.15	1.12	1.19	1.29

The relative estimation error is $\frac{RMSE(\text{GMM})}{RMSE(\text{GMM}|\widehat{\rho}=0)}$ in the Glosten (1987) trade-indicator model: $\Delta P_t = c\Delta Q_t + zQ_t + v_t$, with $c = 0.30$ and $z = 0.05$, where P_t is the traded price, and Q_t is the trade-indicator. All numbers were obtained by simulation. Data for Q_t was generated from a markov chain with $P(Q_t = i|Q_{t-1} = j) = \pi$. ΔP_t was generated from the regression with $v_t \sim N[0, (S_e/2)^2]$. Classification errors were generated from a 1. order autoregression with parameter, ρ. The parameters of interest are the effective spread, $S_e = 2(c + z)$, the adverse selection spread, z, and the adverse selection component of the spread, $\alpha = z/S_e$. The sample size for each run is T, and δ is the true probability of a classification error. For a parameter, $x \in \{S_e, z, \alpha\}$, MSE is the simulated estimate of $E(\widehat{x} - x_0)^2$ and $RMSE = \sqrt{MSE}$. GMM is the GMM estimator with $\widehat{\delta}$,$\widehat{\rho}$ estimated from data. GMM$|\widehat{\rho} = 0$ is the GMM estimator of $\widehat{\delta}$ and the structural parameters, $\widehat{\rho} = 0$.

Table 13: Summary of classification statistics

Method	Classification				Succes
	Buys	Sells	Undetermned	Errors	Rate
True classification	168475	148371	0	–	–
Quote method	127073	105611	56232	27930	89.28%
Tick rule	141602	116606	67	58571	81, 51%
Lee-Ready	145622	122667	6	48551	84.68%

Breakdown of trades according to trade sign using different methods. The data source is 144 NYSE traded stocks in the TORQ database. The total number of trades is 316, 846, which is the number of trades that can be classified according to time of order arrival using the same method as in Odders-White (2000). Each trade was classified according to the quote method, the tick rule as well as the combined Lee-Ready algorithm using a 5 seconds delay. The success rate is the number of buys plus sells divided by the sum of the number of buys, sells, and unclassified trades.

Table 14: Sample moments of trade-indicator and contamination processes.

	Cross-sectional				
	Average	Std. Error	25%	50% Quantile	75%
Trade-indicator, Q_t^*					
$Mean$	-0.018	0.019	-0.136	0.039	0.122
ACF_1	0.460	0.010	0.381	0.453	0.538
ACF_2	0.240	0.011	0.154	0.240	0.338
$ACF3$	0.159	0.009	0.087	0.152	0.234
$PACF_1$	0.461	0.010	0.381	0.453	0.539
$PACF_2$	0.025	0.006	-0.007	0.033	0.060
$PACF_3$	0.036	0.005	0.006	0.034	0.060
Contamination process, I_t					
$Mean$	0.106	0.006	0.056	0.106	0.160
ACF_1	0.418	0.015	0.316	0.432	0.541
ACF_2	0.196	0.014	0.053	0.194	0.311
$ACF3$	0.122	0.011	0.022	0.120	0.204
$PACF_1$	0.419	0.015	0.316	0.433	0.542
$PACF_2$	-0.003	0.007	-0.040	0.014	0.054
$PACF_3$	0.024	0.005	-0.011	0.025	0.050

Summary of time series moments for Q_t^* and I_t. The cross-sectional distribution is based on 122 NYSE traded stocks from the TORQ database with more than 100 trades during the sample period, November 90 through January 91. ACF_i is the autocorrelation at lag i, and $PACF_i$ is the partial autocorrelation at lag i.

Table 15: Univariate time series models

	Cross-sectional				
	Average	Std. Error	25%	50% Quantile	75%
Autoregression for Q_t^*					
const	-0.018	0.011	-0.069	0.015	0.054
Q_{t-1}^*	0.446	0.009	0.375	0.443	0.525
Q_{t-2}^*	0.010	0.006	-0.016	0.022	0.053
Q_{t-3}^*	0.036	0.005	0.006	0.034	0.060
Autoregression for I_t					
const	0.058	0.003	0.035	0.056	0.079
I_{t-1}	0.417	0.014	0.326	0.439	0.529
I_{t-2}	-0.014	0.008	-0.058	0.003	0.040
I_{t-3}	0.024	0.005	-0.011	0.025	0.050

Summary of autoregressions for Q_t^* and I_t. The cross-sectional distribution is based on 122 NYSE traded stocks from the TORQ database with more than 100 trades during the sample period, November 90 through January 91.

Table 16: OLS and GMM estimators of the Glosten (1987) model using data from TORQ.

	Cross-sectional		
	Average	t-statistic	Median
Effective spread, S_e			
$\widehat{S}_e^{benchmar}$	0.57	–	0.38
$\widehat{S}_e^{OLS}$	0.68	10.94	0.48*
$\widehat{S}_e^{GMM}\vert\widehat{\rho}=0$	0.59	2.04	0.39
$\widehat{S}_e^{GMM}\vert\widehat{\rho}=0.2$	0.56	-0.52	0.37
Adverse selection spread, z			
$\widehat{z}^{benchmark}$	0.03		0.02
$\widehat{z}^{OLz}$	0.04	4.54	0.03
$\widehat{z}^{GMM}\vert\widehat{\rho}=0$	0.04	4.35	0.03
$\widehat{z}^{GMM}\vert\widehat{\rho}=0.2$	0.04	3.36	0.03
Adverse selection component			
$\widehat{\alpha}^{benchmark}$	0.15	–	0.13
$\widehat{\alpha}^{OLS}$	0.13	-2.52	0.12
$\widehat{\alpha}^{GMM}\vert\widehat{\rho}=0.0$	0.16	1.78	0.13
$\widehat{\alpha}^{GMM}\vert\widehat{\rho}=0.2$	0.18	2.59	0.14

Cross-sectional distribution of the effective spread, $S_e = 2(c+z)$, the adverse selection component of the spread, z, and the adverse selection component, $\alpha = z/S_e$, in the Glosten (1987) trade-indicator model: $\Delta P_t = c\Delta Q_t + zQ_t + v_t$. Estimates labelled with a *'benchmark'* use the trade classification using time of arrival for orders [hypothetical true classification]. The other estimators use the Lee-Ready method. ρ is the first order serial correlation of I_t [the classification error process], and δ is the probability of a classification error. OLS is the least squares estimator, and GMM is a GMM estimator with estimated $\widehat{\delta}$ and $\widehat{\rho}$ fixed at $\widehat{\rho} = 0.0$ and $\widehat{\rho} = 0.2$ respectively. A'$*$' label on a median value indicates a significant meadian difference [Wilcoxon test] to the '*benchmark'* least squares estimator. The data is 51 NYSE traded stocks from the TORQ database with more than 1000 trades during the sample period, November 90 through January 91.