Department of Finance
Working Paper Series 1998

FIN-98-041

The Dynamics of Discrete Bid and Ask Quotes

Joel Hasbrouck

July 25, 1997

This working paper series has been generously supported by a grant from

CDC Investment Management Corporation

The Dynamics of Discrete Bid and Ask Quotes

Joel Hasbrouck

First Draft: February 3, 1996
Current draft: July 25, 1997

Preliminary Draft
Not for Quotation
Comments Welcome

Professor of Finance
Stern School of Business
New York University
Suite 9-190
44 West Fourth St.
New York, NY 10012-1126

Tel: (212) 998-0310
Fax: (212) 995–4901
E-mail: jhasbrou@stern.nyu.edu
Web: http://www.stern.nyu.edu/~jhasbrou

For comments on an earlier draft, I am indebted to James Angel, Darrell Duffie, James Hamilton, Mark Rubinstein, Neil Shephard, Matt Spiegel, and seminar participants at Cornell University, Georgetown University, New York University, Stanford University, University of California at Berkeley and the UCSD Conference on High Frequency Data. All errors are my own responsibility.

The Dynamics of Discrete Bid and Ask Quotes

Abstract

This paper describes a general approach to the estimation of security price dynamics when the phenomena of interest are of the same scale or smaller than the tick size. The model views discrete bid and ask quotes as arising from three continuous random variables: the efficient price of the security, a cost of quote exposure (information and processing costs) on the bid side and a similar cost of quote exposure on the ask side. The bid quote is the efficient price less the bid cost rounded *down* to the next tick; the ask quote is the efficient price plus the ask cost rounded *up* to the next tick. To deal with situations in which the cost of quote exposure possesses both stochastic and deterministic components and the efficient price follows an EGARCH process, the paper employs a nonlinear state-space estimation method. The method is applied to intraday quotes at fifteen-minute intervals for Alcoa (a randomly chosen Dow stock). The results confirm the existence of persistent intraday volatility. More importantly they establish the existence of a persistent stochastic component of quote exposure costs that is large relative to the deterministic intraday "U" component.

1. Introduction

Although most determinants of a security price are likely to be continuous variables, institutional arrangements generally constrain the security to a discrete grid. This grid may be coarse relative to the price variation over brief intervals, and also relative to the economic costs of order submission and execution. This paper suggests that the bid and ask quotes arise from an implicit efficient price and quote-exposure costs, all of which are continuous random variables. The discrete bid quote is the implicit efficient price less the continuous bid exposure cost rounded down to the next tick; the discrete ask is the efficient price plus the continuous ask exposure cost rounded up. The paper proposes a nonlinear state-space procedure for estimation and (in real time applications) online filtering, and applies this model to fifteen-minute bid and ask quotes for a New York Stock Exchange stock.

This paper stands at a juncture of two themes in empirical financial research: the analysis of discreteness in security prices and the characterization of time-variation in microstructure data. The former concentrates on estimation of latent price moments (e.g., "true" variances) from discrete transaction price series. The latter embraces the deterministic (e.g. intraday) patterns in spreads, trading volumes and volatilities, but more importantly the time-varying volatility models (ARCH and its offspring). Beyond their obvious common concern with the dynamics of security prices, however, the two strands have evolved along methodological lines sufficiently different to impede their integration.

The classic models of discreteness in transaction prices were motivated primarily by the need for stock return volatility estimates based on daily closing prices. The dominant approach involves method of moments estimators, typically derived from a rounding transformation applied to a transaction price that lies on a continuum. The early models (Gottlieb and Kalay (1985) and Ball (1988)) assume that the latent security price follows a continuous diffusion process, and apply the rounding transformation directly to the latent price. Later studies (Harris (1990) and Dravid (1991)) incorporate more

realistic microstructure behavior, in particular the feature that transaction prices arise as incoming orders hit standing bid and offer quotes.

Although they constitute a significant methodological improvement over simply ignoring discreteness altogether, the transaction price discreteness models suffer generic economic and statistical limitations. The bid and ask quotes (when modeled) are assumed to reflect the interaction of the latent price and a spread that is fixed, and the moment estimators require stationarity assumptions. The stationarity and fixed-spread assumptions have largely prevented application of these methods in studies of intraday data or time-varying volatility.

The analysis of time-varying volatility (beginning with Engle (1982) and surveyed in Bollerslev, Engle and Nelson (1994)) has enjoyed considerable success in modeling of volatility evolution for purposes of option pricing and risk management. The literature is also marked by a trend toward applications involving shorter horizons. One motivation for this trend is statistical: volatility estimates of a pure diffusion process, for example, are enhanced by sampling at more frequent intervals. A second motivation is operational: a volatility estimate that is updated more frequently yields more timely advice concerning volatility shifts. At intraday and shorter horizons, however, market microstructure effects in general (and discreteness effects in particular) become large relative to changes in the latent security price. This suggests that satisfactory estimation of volatilities from short horizon data may be contingent on adequate modeling of these effects. The application in this paper constitutes the first estimation (to my knowledge) of a high-frequency ARCH model for an individual stock.

Related research focuses on deterministic patterns in microstructure data. Equity markets typically display elevation of volatility, spreads and volumes around the beginning and end of trading sessions (Brock and Kleidon (1992), McInish and Wood (1992), Lee Mucklow and Ready (1993) and Chan, Christie and Schultz (1995)). These intraday patterns are invariably assessed, however, from discrete transaction prices or quotes, and the intraday variation is small relative to the tick size. Lee, Mucklow and Ready, for example, report a beginning-of-day elevation in the quoted spread for NYSE stocks of

roughly 10% (relative to the daily mean). For an average spread of 1/4 (arguably on the high side for a large NYSE stock), the 10% elevation is only one fifth of the 1/8 tick size.

The approach advocated in this paper jointly models the phenomena of interest in each of these research themes. The central assumptions concern the relations between the underlying continuous market state variables (the latent efficient price and the costs of market making) and the observed discrete bid and ask quotes. The statistical model derived from these assumptions, however, allows for generality in the underlying dynamics of the continuous market state variables. These dynamics are jointly deterministic (as in the intraday pattern literature) and stochastic (as in the time-varying volatility studies).

The model is applied to bid and ask quotes collected at fifteen-minute intervals for a typical NYSE stock over a sample period when the tick size is one-eighth of a dollar. The importance of discreteness in this sample can be motivated by two observations. First, the changes in the discrete quotes overwhelmingly concentrate on zero, one or two ticks. This suggests that the tick size is large relative to short-term price movements. Second, in 45.9% of the intervals, the closing spread was one eighth. This suggests that the tick size is frequently binding on the economic spread.

Yet the NYSE recently moved to a tick size of one-sixteenth for most stocks, and plans to move to decimal trading by the year 2000. The question then arises as to whether discreteness in empirical studies is fated to be a statistical nuisance effect of fading importance.

While the diminution of the tick size obviously lessens the benefit of explicitly modeling discreteness, fundamental economic considerations suggest that the tick size will in equilibrium remain large relative to certain trading costs. Harris (1996, 1997) points out that in a market that at least partially respects time priority for orders at a given price, the tick is the cost of jumping to the head of the existing queue. An arriving buyer, for example, can step ahead of the existing limit buy orders only by bidding one tick higher than the best prevailing bid. To the extent that market institutions view time priority as a desirable feature in encouraging the supply of liquidity, they are unlikely to be indifferent to arbitrarily small tick sizes.

Although these considerations may help motivate the present approach, determination of the optimal tick size lies outside the scope of the present paper. Analyses of issues related to this important question include Ahn, Cao and Choe (1996), Angel 1994), Anshuman and Kalay (1994), Bernhardt and Hughson (1996), Brown, Laux and Schacter (1991), Chordia and Subrahmanyam (1995), Cordella and Foucault (1996), Glosten (1994), Harris (1990, 1991)). Nor does the present model account for clustering, the affinity of transaction prices and quotes for integers, halves, quarters, etc., in decreasing frequency (Niederhoffer (1965, 1966), Harris (1991, 1994), and Christie and Shultz (1995a, 1995b)). In a dynamic setting, clustering requires specification of a stochastic mapping from continuous state variables to discrete observations that is more complicated than the simple rounding functions employed here.

The paper is organized as follows. The next section surveys existing models of discreteness and time-varying volatility most closely related to the present model. The model, which describes the generating mechanism for discrete bid and ask quotes, is presented in Section 3. The paper then turns to the problem of inference: how to estimate the underlying model from the observed discrete bid and ask prices. Section 4 discusses the restrictions imposed on the underlying variables by the discrete observations. Section 5 introduces the nonlinear filtering algorithm, the associated maximum likelihood procedure and computational techniques. The full dynamic model, which incorporates stochastic and deterministic time variation in the cost and efficient price volatility, is presented in Section 6. The model is estimated for a representative NYSE stock in Section 7. A brief summary concludes the paper in section 8.

2. Background and literature review

Stock price discreteness

The present paper's primary antecedents lie in the stock price discreteness literature. Discrete transaction prices were initially modeled as random walk realizations, rounded to the nearest grid point (Gottlieb and Kalay (1985), Ball (1986)). A representative model is

$$\begin{aligned} m_t &= m_{t-1} + u_t \\ p_t &= \text{Round}[m_t] \end{aligned} \tag{1}$$

where m_t is the implicit efficient ("true") price, p_t is the reported transaction price and $\text{Round}[\cdot]$ rounds its argument to the nearest integer. (Throughout this paper all variables are assumed scaled so that the tick/grid size is unity.) In this model traders negotiate a continuous price, which is then discretized.

Harris (1990) enhances this model to incorporate bid and ask quotes that differ from the efficient price by a half-spread c. A restatement of Harris's model (in present notation) is:

$$\begin{aligned} m_t &= m_{t-1} + u_t \\ b_t &= m_t - c \\ a_t &= m_t + c \\ p_t &= \begin{cases} \text{Round}[b_t] \text{if } \pi_t = -1 \\ \text{Round}[a_t] \text{if } \pi_t = +1 \end{cases} \end{aligned} \tag{2}$$

where b_t is the bid quote, a_t is the ask quote, and π_t is a buy/sell indicator variable. The bid and ask quotes (implicit in Harris's presentation) lie on a continuum. The rounding here is symmetrically applied only to the side of the quote that is realized. Glosten and Harris (1988) generalize this to allow for a quote schedule linear in trade size.

Dravid (1991), citing a suggestion of Paul Pfleiderer, allows for asymmetric rounding of the bid and ask quotes:

$$\begin{aligned} m_t &= m_{t-1} + u_t \\ b_t &= \text{Floor}[m_t - c] \\ a_t &= \text{Ceiling}[m_t + c] \\ p_t &= \begin{cases} b_t \text{ if } \pi_t = -1 \\ a_t \text{ if } \pi_t = +1 \end{cases} \end{aligned} \tag{3}$$

where $\text{Floor}[\cdot]$ rounds its argument down to the next lowest grid point and $\text{Ceiling}[\cdot]$ rounds its argument up.

The asymmetric rounding used by Dravid is also a key feature of the model proposed in this paper. While a full discussion of this property will be deferred until section 3, some simple considerations may help motivate the present developments. Most

markets impose discreteness on quotes as well as transaction prices. By asymmetrically rounding up on the ask and down on the bid, a market maker avoids the possibility of loss on the incoming trade. If the rounding were symmetric (all prices rounded up, all prices rounded down or all prices rounded to the nearest integer), then one or both sides of the quotes might be associated with an expected loss. For example, if the efficient price is 5 and the cost is 1.1, nearest-integer rounding yields a bid of 4 and an ask of 6, both of which yield expected losses. Furthermore, symmetric rounding may imply degenerate quotes (identical bid and ask prices) if c is small.

Viewing c as a fixed parameter and assuming that the fractional part of m_t is uniformly distributed between zero and unity, it is easily shown that the expected value of the discrete spread $s_t = a_t - b_t$ is given by $Es_t = 1 + 2c$.[1] If c is stochastic and independent of the fractional part of m_t, then $Es_t = 1 + 2Ec_t$. This is important because most studies of intraday patterns in the spread estimate $2Ec_t$ by cross-day averaging of s_t taken at a given time of day. That the expectations differ by (constant) unity suggests that discreteness does not distort estimated time-of-day variations in the spread.

For variances and autocovariances, however, matters are more complicated. Even if c is constant, the discrete spread will be random and autocorrelated. For example, if c= 1/4, then the spread is one tick as long as the fractional part of m is between 1/4 and

---

[1] Assume that m is uniform on the unit interval, and for convenience, drop the time subscripts. Then $b = \begin{cases} -1, \text{if } m < c \\ 0, \text{if } m > c \end{cases}$ and $a = \begin{cases} 1, \text{if } m < 1-c \\ 2, \text{if } m > 1-c \end{cases}$.

If $0 < c < 1/2$, then $s = \begin{cases} 2, \text{if } 0 < m < c \text{ or } (1-c) < m < 1 \\ 1, \text{if } c < m < (1-c) \end{cases}$.

If $1/2 < c < 1$ then $s = \begin{cases} 2, \text{if } 0 < m < (1-c) \text{ or } c < m < 1 \\ 3, \text{if } (1-c) < m < c \end{cases}$

Integrating either of these over the distribution of m gives $Es = 1 + 2c$. This argument generalizes for c>1.

3/4; and the spread is two ticks otherwise. Furthermore, to the extent that the fractional part of m_t is persistent (e.g., if successive u_t are small relative to the tick size), the discrete spread will exhibit positive autocorrelation. Therefore, variability and autocorrelation in the discrete spread do not necessarily imply variability and autocorrelation in the underlying cost determinants of the spread.

The three models described by equations (1), (2) and (3) have different (and generally incompatible) implications for the behavior of the transaction prices. In the model of equation (1), p_t follows a rounded random walk. In (2), p_t is a rounded signal-plus-noise. In (3), p_t can be viewed as being randomly selected from one of two rounded random walks (one associated with the bid and the other with the ask).

Although the focus of each of models described by equations (1), (2) and (3) is the dynamics of transaction prices, the progression displays increasingly realistic modeling of the (implicit) quotes. In the present application, quotes are observed and are central, to the exclusion of the trades. In general, quotes are of particular interest in microstructure studies because they can be updated in the absence of trades to reflect changing information, and because they reflect perceived asymmetric information costs. Quotes furthermore constitute in many markets the primary (sometimes only) publicly available data (e.g., most bond and foreign exchange markets).

In their empirical implementations, the models described above are augmented with stationarity assumptions on u_t and π_t. In Gottlieb and Kalay (1985), Ball (1986) and Dravid (1991), the implied variances and autocovariances of the transaction price changes are used to compute moment estimators. The stationarity assumptions are crucial to the tractability of these estimators. These assumptions are, of course, violated in applications where the moments are changing in a deterministic fashion (as in studies of intraday patterns) or in a stochastic fashion (as in the time-varying volatility models).[2] Although

---

[2] These comments on the restrictiveness of the stationarity assumptions apply to most other moment-based estimators, including the GMM approach used by Madhavan, Richardson and Roomans (1997) and the vector autoregression approaches used in, for

Harris (1990) also assumes stationarity, this is not essential to his iterative maximum likelihood approach. His approach is in fact a variant of the nonlinear state space model described by Kitagawa (1987). The Kitagawa approach is used in the present paper.

Time-varying volatility

In modeling the dynamics of the latent efficient price, the present paper draws on prior studies of time-varying volatility. The aspects of this extensive literature that are most relevant for the present purpose involve general concerns in the application of ARCH-family models to short-horizon security returns and specific features of the particular process used in the present model.

As to the general concerns, the classic models of time-varying volatility in asset returns typically model the security price as a martingale, with increments that are (by definition) uncorrelated but of persistent stochastic variance. Price discreteness, as well other microstructure phenomena such as inventory control and bid-ask bounce, obviously give rise to non-martingale components. Since the original literature on stock price discreteness arose from the need for variance estimates derived from discrete data under the assumption that the variance is time-invariant, one might have expected discreteness to remain a prominent concern in the analysis of variance that is time-varying.

There are perhaps several reasons why this has not occurred. Firstly, there are methodological differences between the stationary moment estimators favored in the discreteness literature and the likelihood methods used in the volatility studies. Secondly, the neglect of discreteness in the characterization of the time-varying component of volatility might be tolerable if discreteness is viewed as introducing a bias that is approximately constant. One might further conjecture that, in many data samples the

---

example, Hasbrouck (1991) and surveyed in Hasbrouck (1996). Deterministic intraday patterns in these models are typically investigated by estimating them over subperiods (e.g., the first and last half-hour of trading) during which time the processes are assumed stationary.

significant moves in forecasted volatility are driven by price changes that dwarf the tick size.

It is also the case, however, that most applications of time-varying volatility models have been to data that are not highly discretized. Volatility models for single stocks are invariably estimated at daily or longer horizons (e.g., Cheung and Ng (1992) and Duffee (1995)). These models have been applied at shorter horizons for stock indexes, in which the discreteness problem is significantly mitigated by aggregation over securities (e.g., Nelson (1991)). They have also been used in foreign exchange markets, where the tick size is in most respects small relative to the equity markets (e.g., Andersen and Bollerslev (1997)).

An ARCH variant commonly used for security returns is the exponential generalized autoregressive conditional heteroscedasticity model (EGARCH) due to Nelson (1991):

$$\begin{aligned} m_t &= m_{t-1} + u_t \\ \sigma_t^2 &= \mathrm{Var}(u_t) \\ \ln(\sigma_t^2) &= \eta + \varphi(\ln(\sigma_{t-1}^2) - \eta) + \gamma(|\zeta_{t-1}| - \mathrm{E}|\zeta_{t-1}|) \end{aligned} \tag{4}$$

where $\zeta_t \equiv u_t/\sigma_t$ is the standardized increment. The terms on the right hand side reflect a mean (η), an autoregressive adjustment (at rate φ) toward the mean, and a disturbance component with coefficient γ driven by the prior period's shock. (The asymmetry term suggested by Nelson is omitted.) Nelson suggested that the standardized increment ζ_t be distributed in accordance with the generalized error distribution (GED), denoted $f_{GED}(\zeta_t; \nu)$ where ν is the tail-thickness parameter. When ν=2, the GED reduces to the standard normal density. The expected absolute value $\mathrm{E}|\zeta_t|$ in equation (4) is unconditional and time-invariant, depending only on the tail-thickness parameter.

Nelson applied this model to the return series computed from daily close S&P prices. The importance of discreteness should be minimal in Nelson's application due to aggregation over securities and a time horizon that is (relative to the present paper) long. An important contribution of the present paper is the extension of the ARCH framework to high-frequency highly-discretized data.

The question naturally then arises as to whether explicitly modeling discreteness leads to volatility parameter estimates that meaningfully differ from those realized by the expedient of simply estimating the continuous volatility model with discrete data. While this paper does not address this issue in general, the present data sample does illustrate a striking consequence of ignoring discreteness. In the present data set, when the conventional continuous EGARCH/GED model is applied to the discrete quote-midpoint changes, likelihood minimization with various algorithms and starting points generally fails to converge.

Simulations suggest that the problem lies in the interaction of discreteness with the GED distribution. As the tail-thickness parameter drops below two (the normal case), the GED becomes sharply peaked at zero. When zero-mean continuous data are rounded, the observations near zero are collapsed to a high peak at zero. This feature of the sample often dominates the minimization, driving the value of ν inexorably downward. In summary, the continuous EGARCH/GED model is not (by reason of computation infeasibility) a valid alternative to the present approach.

Other related studies

Econometric models of ordered discrete variables often employ ordered probit models. Probit analyses relevant to the present paper include Hausman, Lo and MacKinlay (1992) and Bollerslev and Melvin (1994). Broadly speaking, probit models are reduced-form specifications: robust and flexible in modeling data, but difficult in some respects to interpret structurally.

Effects in the standard probit model flow from observable predetermined variables to an unobservable continuous variable (via a linear specification), and thence to the observed discrete analog of the continuous variable (via a mapping function defined by a set of breakpoints). In Hausman et al, observable variables (volume, the futures return, the bid-ask spread, etc.) plus a random disturbance determine a latent continuous price change, which is then maps into the observed discrete price change.

The requirement that the latent continuous variable depend only on observable data rules out an attractive and basic feature of the discreteness models discussed above,

viz., the dependence of the latent unobservable price on its prior value. In each of the models given by equations (1), (2) and (3), for example, the latent unobservable continuous price m_t depends on m_{t-1}, not on any observable discretized transform of m_{t-1}. Viewed as a reduced form model, a probit specification of observed price changes may accord well with observed data, but there are no obvious candidates for the underlying structural model of continuous prices.

The vagueness of the underlying structural model may be tolerable in many situations. As an example, note that in a probit model the mapping from the latent continuous variable to the discrete observable may distribute or smooth out any sharp peaks in the data. It was mentioned above that the sharp peak of the GED distribution leads to computational difficulties in fitting EGARCH models to discrete fat-tailed data. A probit-like modification to the model may ameliorate these problems.

Nevertheless many interesting microstructure models may involve structural features that the probit approach cannot easily resolve. In the model given by equation (3), for example, the discrete quotes reflect a continuous variable (m_t) and a continuous parameter (c). The present paper expands this framework in numerous respects, in particular allowing c to follow a persistent stochastic process. When both m_t and c are dynamic and unobservable, structural inference from a reduced form probit approach is likely to be even more challenging. The state-space approach advocated here, in contrast, starts from specification of processes for the unobservable continuous variables. Inference concerning the parameters of these processes follows from the (numerical) computation of the sample likelihood of the observed discrete data.

Although the models for discrete transaction prices given in the first part of this section do not possess simple probit reduced forms, probit models are more general in certain respects. Most notably, the mapping from the latent continuous variable to the observable discrete variable in a probit model is described by a set of breakpoints which need only be ordered. This allows for a degree of flexibility in the implicit "rounding" function that is not available in the structural models (both the ones described above and the one proposed in the present paper).

The model proposed by Bollerslev and Melvin (1994) consists of a GARCH specification for Deutschemark/dollar quotes, the output of which (conditional volatility forecasts) feeds into a probit model for the discrete spread. They find that the spread depends positively on the forecasted volatility. While their model and the present one both allow the market-making costs (cf. c in the above models) to vary stochastically, there are several structural differences.

The Bollerslev-Melvin ARCH-type component is a continuous model applied directly to the discrete data, while the ARCH-type component of the present model is specified in terms of the continuous unobservable variable. The use of a continuous likelihood for discrete data is feasible in foreign exchange data because the tick size is relative small. As noted above, this practice is not computationally feasible in the present data set (U.S. equity data), where the tick size is larger.

A more fundamental difference concerns the way at which discreteness is imposed. In Bollerslev and Melvin, the latent cost of market making is transformed in the probit mapping into a discrete spread. In Dravid's model (equation (3)) and the one advanced here, on the other hand, the latent bid and ask quotes are discretized separately. The model proposed in this paper is in this sense more complete. It will be seen as well to incorporate different costs on the bid and ask sides of the market.

In allowing the cost of market making to reflect the volatility forecast, however, the market characterization in the Bollerslev-Melvin analysis is richer than present model. The present analysis assumes that the implicit efficient price and the costs of market making evolve independently. This simplification is mandated by present computational limitations, however, and is not fundamental to the overall approach.

Another group of analyses relevant to the present paper focuses on modeling time passage and time deformation in market processes. The connection to price discreteness arises from the view that the discrete price grid defines a set of permeable barriers in the space where the continuous latent price evolves. If transactions are assumed to occur when the latent price crosses one of these barriers, the observable crossing times can be used to infer parameters of the latent price process. This approach is advocated by Marsh

and Rosenfeld (1986) and Cho and Frees (1988). More recent studies of time passage/deformation, however, stress characterization of the underlying implicit information process rather than discreteness (Engle and Russell (1995) and Engle (1996)).

3. The model

As in the preceding section let m be the implicit efficient price of the security in the usual sense of the expectation of the security's terminal value, conditional on all public information. (For ease of exposition, the time subscripts are suppressed in this section.) The agent establishing the bid quote is assumed to be subject to a nonnegative cost of quote exposure $\beta \geq 0$ for small trades, such that in the absence of discreteness restrictions she would quote a bid price of $m-\beta$. This cost may be provisionally assumed to reflect fixed transaction costs and asymmetric information costs. As in the Dravid (1991) model, she is assumed to quote a discrete bid price that is rounded down: $b = \text{Floor}[m-\beta]$. Similarly, the agent establishing the ask quote is assumed to be driven by a quote exposure cost $\alpha \geq 0$ (also for small trades), such that in the absence of discreteness restrictions he would quote an ask price of $m+\alpha$. Constrained by discreteness, he rounds up to a discrete ask quote of $a = \text{Ceiling}[m+\alpha]$. In summary, the bid and ask prices are given by

$$\begin{aligned} b &= \text{Floor}[m-\beta] \\ a &= \text{Ceiling}[m+\alpha] \end{aligned} \tag{5}$$

This construct can be motivated by most simple models of dealer behavior. In the framework of Glosten and Milgrom (1985) quote setters face a population of informed and uninformed traders. m is the expectation of the final value of the security conditional on all public information (including the transaction price history). The quote exposure costs are defined implicitly by the conditions that $m-\beta$ and $m+\alpha$ ensure the quote-setter(s) zero expected profits and no ex post regret, an outcome supported by Bertrand competition.

Due to the asymmetric rounding, a Glosten-Milgrom dealer will achieve a profit (both ex ante and ex post) on each trade. These profits need not lead to competitive price cutting because the discreteness restriction ensures that any such action, if feasible, will

result in a loss. Nor need these profits lead to a surge of new entrants. Even markets (such as the NYSE) that allow nondealers to enter limit orders usually enforce local time priority. The probability of execution and therefore the incentives for limit order placement diminish with the length of the queue.

More generally, β is the quote setter's marginal cost on the bid side of the market at a particular time. From an economic perspective, it is useful to recognize that some of the components of this cost may be negative, *as long as the total β is nonnegative*. An example of this arises in the context of inventory control. Suppose that the cost of clearing a trade is 0.5 (ticks). A dealer who is short (relative to her desired holdings) might nevertheless bid as if β=0.2, reflecting a greater propensity to accumulate a position. There is an implicit benefit of accumulation that may be viewed as a negative cost of -0.3. If the same dealer were also offering the security, we might also expect her α to be high relative to the clearing cost, reflecting her reluctance to accommodate further sales. (Similar remarks apply, of course, to the ask exposure cost.)

Negative cost components may also arise in the case of quotes established by public limit order traders. Their principal alternative to a limit order is a market order. They are not seeking to realize a dealer's profit on average, but merely to reduce their costs of trading (Harris and Hasbrouck (1996) and Harris (1994)). Although the dealer (quote setter) in most microstructure models is assumed to be uninformed, the quote exposure costs may reflect the dealer's private information by including a displacement $m-m^*$, where m^* is the dealer's (informed) estimate of the security's value.

From a modeling viewpoint, nonnegative α and β serve to prevent the bid and ask implied by (5) from coinciding or crossing. In general practice, the bid and ask quotes prevailing at a point in time reflect entered orders that have been subjected to a matching procedure according to the rules of the market. The assumptions of a common m and nonnegativity of α and β are expedients that avoid the necessity of explicitly modeling this matching process.

Most interesting applications will involve situations where the quote exposure costs are random. This randomness can be viewed as arising from several sources. Along

the lines of the Glosten-Milgrom model, there may be random time variation in the determinants of this cost, such as perceived exposure to adverse information or holding costs. In this view all dealers and potential dealers are subject to the same cost. Alternatively, we may view the quote setter as an agent drawn from a population of traders with random cost functions. If more than one such agent is active at an instant, then the relevant costs are the minimum α's and β's in the set.

Although this model allows for randomness in α, β and m, the discreteness aspect of the model arises from a nonstochastic transformation. There is no discreteness "error" or disturbance that is required to impound the effect of discreteness.

There is no assumption that all trades take place at the posted quotes. Following Rock (1996), the posted quotes modeled here are viewed as the best available prices absent knowledge of the full size of the incoming order. A trader (such as a specialist or floor trader) who can bid or offer conditional on the incoming order size may better the posted quotes. In practice, such agents bid or offer after the order has been received, and these implicit quotes do not prevail after the transaction has occurred.

In the present model the quote setter's solution to an implicit continuous optimization problem (α or β) is subjected to a transformation to yield discrete quote placement strategies. This must be viewed as an approximation to a decision process in which discreteness is more fundamentally incorporated into the calculation, i.e., an integer programming problem. Models along these lines include Anshuman and Kalay (1994), Glosten (1994), Chordia and Subrahmanyam (1995), Bernhardt and Hughson (1996) and Cordella and Foucault (1996). These models are stylized in numerous respects (typically allowing a restricted set of traders and permissible interactions) and focus almost exclusively on information costs. In these characterizations, a continuous "pre-rounding" cost constructs (such as the present α and β) do not explicitly arise. It could nevertheless be argued that such quantities exist implicitly, and that they impound the costs of quote-setting mentioned above (although they would also incorporate discreteness effects).

4. Inference from observed bid and ask quotes.

Viewed as a transformation of continuous random inputs (m, α and β) into discrete bid and ask prices, the model described by (5) is a very simple one. From the perspective of the econometrician (and that of many market participants), however, the observed bid and ask prices are given, and inference focuses on the unobserved inputs. Viewed in this direction, the model is more complex.

As a function of the observed bid and ask quotes (b, a), the feasible region for (m, α, β) consistent with model (5) is:

$$Q(b,a) = \{(m,\alpha,\beta): \alpha > 0,\ \beta > 0,\ b \le m - \beta < b+1 \text{ and } a-1 < m+\alpha \le a\} \qquad (6)$$

The inequalities define a convex polytope (geometric solid) of up to six faces.

Although the estimations in this paper are based on (6), it is easier to visualize the special case in which the quote exposure costs are the same on both bid and offer sides. In this equal-cost model $c = \alpha = \beta$, and the feasible region is

$$Q(b,a) = \{(m,c): c > 0,\ b \le m - c < b+1 \text{ and } a-1 < m+c \le a\} \qquad (7)$$

These inequalities define a two-dimensional region. Figure 1 depicts these feasible regions for representative one-, two- and three-tick spreads. The diamond shape of the region $Q(b=0, a=2)$, for example, can be viewed as arising in the following way. When c is just slightly greater than zero or slightly less than one, the range of m consistent with b=0 and a=2 is a small neighborhood about one. When c is ½, m can range from ½ to 1½.

Given a prior probability density function $f(m,c)$, the posterior density conditional on observing bid and ask quotes b and a is:

$$f(m,c|b,a) = \begin{cases} \dfrac{f(m,c)}{\Pr(b,a)} \text{ if } (m,c) \in Q(b,a) \\ 0, \text{ otherwise} \end{cases} \qquad (8)$$

where $\Pr(b,a) = \int_{Q(b,a)} f(m,c)\,dm\,dc$ is the probability of observing discrete bid and ask quotes b and a. Since this conditioning imposes a truncation on the ranges of the variables, it might seem that the conditional densities would be simple truncated versions of the priors. The truncations defined by $Q(b,a)$, however, apply to linear combinations of

the variables, not the variables themselves. The shape of $Q(b,a)$ effectively forces a nonlinear transformation on the priors.

The nature of this distortion can be illustrated by supposing flat priors, i.e., assuming that m and c are each independently and uniformly distributed over the positive real line (or, more properly, over a finite region of the positive real line large enough to encompass any values of potential interest). It might be supposed that flat priors would suffice to generate uniform posteriors. On observing a bid of zero and an ask of two, and being unwilling or unable to model c, for example, one might first conjecture simply that the efficient price is uniformly distributed between zero and two. In fact, by integrating over the diamond-shaped feasible region defined by $Q(b=0, a=2)$ in Figure 1, the correct posterior for m is triangular between ¼ and ¾. Thus, the model imposes structure even under the assumption of flat priors.

The geometry of the feasible region also carries implications for estimation. Still assuming a flat prior for m, the posterior for m conditional on observing $b=0$, $a=2$ and in addition knowing c, the posterior for m is always symmetric triangular about one for all c in the feasible range. This suggests that knowledge about c will be less informative about the location of m, but more informative about the dispersion of m. Since the location of m is not sensitive to our prior for c, the cost of assuming a flat prior on c when we are estimating the dynamics of m may be acceptably low. It is worth reemphasizing that even in this case, however, the model imposes structure. Similar arguments can be made for the distribution of m conditional on c.

Similar remarks apply to the more general model with distinct bid- and ask-side exposure costs. By integrating over the α and β axes in the feasible region defined in equation (6), it can be shown that under flat priors on all the variables, the posterior for m conditional on observing bid b and ask a is in general trapezoidal over (b, a). The location-invariance property of the symmetric cost case does not hold here, however. That is, given $b=0$ and $a=2$, the location of the posterior for m depends on α and β.

As a numerical example of a nontrivial prior on c, consider the case where $\ln(c)$ is assumed normal with mean $\mu=-1$ and standard deviation $\sigma=0.6$. This implies

$\Pr[a-b=1] = 0.29$, $\Pr[a-b=2] = 0.58$, $\Pr[a-b=3] = 0.11$, and $\Pr[a-b>3] = 0.03$, i.e., frequencies of one-, two- and three- tick spreads that might be observed for a typical NYSE stock. Assume a flat prior for m, i.e., that m is uniformly distributed on $(0, \kappa)$ where κ is "large" (but not infinite). The choice of κ is arbitrary; it integrates out of all calculations. The cost parameter is assumed to be independent of the price level, which implies that the prior density of the latent variables may be written as $f(m, c)=f(m)\,f(c)$.

Figure 2 depicts the prior and conditional density functions conditional on observing bid and ask quotes $b=0$ and $a=2$. The prior for m is drawn as a flat line of height κ^{-1}. The conditional density for m is not uniform over the allowable range of m ($1/2<m<3/2$). If m is near an endpoint, the range of feasible c values is small, with correspondingly low probability. If m is near the center of the range, the feasible set for c is larger. Similarly the conditional density for c is not simply a truncated log normal, but slopes down gradually to the boundary defined by $c=1$. When c lies on this boundary, the set of m values consistent with the observed quotes is a single point (of probability zero, given the continuous prior assumed for m). As we move inward from this constraint, the set of feasible m becomes larger.

5. Maximum Likelihood Estimation.

Suppose now that the quote generation process occurs over time periods $t=1, \ldots, T$ with state variable realizations $z_t=\{m_t, \alpha_t, \beta_t\}$ and corresponding observed bid and ask quotes $q_t=\{b_t, a_t\}$. In most applications the state variables will not be i.i.d.. Typically m_t might follow a random walk with non-i.i.d. increments, and the latent cost variables might also exhibit serial correlation. The model in such cases is neither linear nor Gaussian. The general estimation approach follows Hamilton (1994a, 1994b) and Harvey (1991). The numerical technique is due to Kitagawa (1987), which is summarized in Hamilton (1994b). (Glosten and Harris (1988) employ another variant of this method.)

The essence of the procedure is a recursive likelihood calculation. Suppose that the probability density function of the current state variables conditional on current and past observations, $f(z_t|q_t, q_{t-1}, \ldots)$ is known for some time t. Looking ahead to $t+1$,

$$f(z_{t+1}|q_t,q_{t-1},\ldots)=\int f(z_{t+1}|z_t)f(z_t|q_t,q_{t-1},\ldots)dz_t \tag{9}$$

where $f(z_{t+1}|z_t)$ is the (possibly time-dependent) state transition density function. The conditional probability of observing q_{t+1} is

$$\Pr(q_{t+1}|q_t,q_{t-1},\ldots)=\int_{z_{t+1}\in Q_{t+1}} f(z_{t+1}|q_t,q_{t-1},\ldots)\,dz_{t+1} \tag{10}$$

where $Q_{t+1}\equiv Q(b_{t+1},a_{t+1})$ as defined in equation (6). The sequence of these probabilities may be used to construct the likelihood function.

The range of the integration in (10) is a distinctive feature of the present problem. In typical filtering applications the integration region is $\mathbf{R}^d$ where d is the dimension of the state vector. In the present application, however, the quotes serve to bound the possible values of the state variables: Q_t defines a small region of z_t space. In an online forecasting application we would be interested in computing the probabilities given by (10) for a number of possible realizations of q_{t+1}. In an estimation situation, however, we need only compute the probability for the value of q_{t+1} that actually occurs in the sample.

Next, note that the joint density of next period's state variables and quotes is $f(z_{t+1},q_{t+1}|q_t,q_{t-1},\ldots)=f(z_{t+1}|q_t,q_{t-1},\ldots)$ if $z_{t+1}\in Q_{t+1}$ and zero otherwise. This too reflects a simplification peculiar to the present problem: computation of the right-hand-side density usually involves integration over a density function of the observational errors. Here, the observations (the quotes) are a deterministic function of the state variables. Therefore $f(z_{t+1}|q_{t+1},q_t,q_{t-1},\ldots)=f(z_{t+1},q_{t+1}|q_t,q_{t-1},\ldots)/\Pr(q_{t+1}|q_t,q_{t-1},\ldots)$, if $z_{t+1}\in Q_{t+1}$ and zero otherwise. This completes the update.

In an estimation context, the state transition density is assumed to depend on a parameter vector θ: $f(z_t|z_{t-1})=f(z_t|z_{t-1};\theta)$. Maximum likelihood estimation proceeds by maximizing the sum of the log of the conditional probabilities $\Pr(q_{t+1}|q_t,q_{t-1},\ldots;\theta)$. It should be noted, however, that the procedure does not provide estimated residuals that might be used to test the specification of the transition densities.

Although straightforward in principle, this update requires the evaluation of two integrals for which closed-form solutions are not readily available. In the standard Kalman

filter, all joint, marginal and conditional densities are normal, and the results of the integrations are summarized by update formulae for the conditional means and variances. In the present case, successive updates involve computation of nested, truncated densities of increasing dimension. These are calculated using a grid approximation. Appendix A discusses the computational details. Appendix B presents some simulation results.

As an estimation approach, maximization of a grid-approximated likelihood function is flexible in accommodating a wide range of transition densities and is perhaps conceptually and computationally simplest. Other approaches may offer greater computational efficiency (Kitagawa and Gersch (1996)). Markov Chain Monte Carlo methods also hold promise (Shephard and Manrique (1997)), particularly in view of their apparent ability to accommodate state vectors of high dimension.

6. The full dynamic model

The present implementation assumes that the efficient price evolves independently of the quote exposure costs. The efficient price is modeled as the EGARCH/GED given in equation (4). In the present application, however, $|\zeta_t|$ on the r.h.s. of (4) is not observed. As a more tractable alternative to carrying σ_t as a (fourth) unobservable state variable, I assume for purposes of estimation that the variance process is driven by the conditional expectation of the absolute efficient price increment. That is, $|\zeta_t|$ in (4) is replaced by its conditional expectation $E[|\zeta_t||q_t, q_{t-1}, \ldots]$, which is easily computed in the course of the iterative update.

To allow for the intraday "U" shapes frequently exhibited by market data, the level term η in (4) is allowed to vary as a function of time-of-day. A parsimonious function that permits end-point elevation can be built from exponential decay functions. The deterministic component of the variance is modeled as:

$$\eta_t = \begin{cases} l_1 + l_2^{open} \exp\left(-l_3^{open} \tau_t^{open}\right) + l_2^{close}, & \text{if } t \text{ is an intraday interval} \\ \eta^{overnight}, & \text{if } t \text{ is an overnight interval} \end{cases} \qquad (11)$$

where τ_t^{open} is the elapsed time since the opening quote of the day (in hours) and τ_t^{close} is the time remaining before the scheduled market close (in hours).

The bid and ask quote exposure costs are assumed to evolve as:

$$\begin{aligned} \ln(\alpha_t) &= \mu_t + \phi(\ln(\alpha_{t-1}) - \mu_{t-1}) + v_t^{\alpha} \\ \ln(\beta_t) &= \mu_t + \phi(\ln(\beta_{t-1}) - \mu_{t-1}) + v_t^{\beta} \end{aligned} \tag{12}$$

where v_t^{α} and v_t^{β} are independently distributed as $N(0, \sigma_v^2)$. This specification allows for deterministic time variation in the mean, and also a persistent stochastic component. The deterministic component of the cost process is:

$$\mu_t = k_1 + k_2^{open} \exp\left(-k_3^{open} \tau_t^{open}\right) + k_2^{close} \exp\left(-k_3^{close} \tau_t^{close}\right) \tag{13}$$

where τ_t^{open} is the elapsed time since the opening quote of the day (in hours) and τ_t^{close} is the time remaining before the scheduled market close (in hours).

The assumption of independent evolution allows the state transition density $f(m_t, \alpha_t, \beta_t | m_{t-1}, \alpha_{t-1}, \beta_{t-1})$ to be factored as $f(m_t|m_{t-1}) f(\alpha_t|\alpha_{t-1}) f(\beta_t|\beta_{t-1})$. While this appreciably enhances the speed of the computations, it is less desirable from a modeling viewpoint. The assumption of independence between α and β is most appropriate to a market setting in which the bid and ask quotes reflect limit orders originating from different idiosyncratic liquidity traders. In the case where the quotes reflect the interests of a single dealer or derive from common influences, it would be more appropriate to allow for positive correlation between the two costs. One might also expect (following Bollerslev and Melvin) correlation between the quote exposure costs and volatility in m. With one exception, these generalizations are not at present computational feasible, although they certainly provide directions for future research. The exception involves the equal-cost model defined by $\alpha=\beta=c$, in which case the state space is two-dimensional (m and c).

The specification described above jointly models the bid and ask quote exposure costs and the efficient price. Given the computational complexity of the likelihood calculation, however, it is also useful to consider simpler models. For example, if only the quote exposure costs are of interest, the model might be estimated assuming at each point in time a flat prior on the efficient price. This variant, consisting solely of the cost equations (12) and (13) is termed the cost model. When m is eliminated as a state

variable, the numerical grid approach is still necessary due to the stochastic variation in the cost, but the reduction in dimension speeds computation. Alternatively, if the efficient price dynamics are the sole concern, one might estimate the EGARCH/GED component of the model under the assumption of flat priors on the quote exposure costs. This variant is termed the EGARCH/GED model. Both of the cost and EGARCH/GED submodels are in principle misspecified, but as a practical matter the cost of the misspecification may be small relative to the computational gains.

7. Estimation

Data

I estimate the specifications described in the last section to NYSE bid and ask quotes for Alcoa (ticker symbol AA) for all trading days in 1994. Alcoa is the first Dow Stock (in alphabetical ordering) and is viewed as a representative high-activity security. Bid and ask quotes are those prevailing at the close of fifteen-minute intervals. The first observation of a day generally corresponds to 9:45, the last to 16:00 (26 points). There are 6,780 observations. The fifteen-minute interval was chosen as a frequency that is high enough to be of microstructure interest, yet still yield a computationally tractable number of observations.

Table 1 reports descriptive statistics for the absolute value of the bid first-differences. (Results for ask first-differences were virtually identical.) The proportion of intervals for which the bid change is zero is 39% (intraday) and 18% (overnight). Not reported in the table is the additional finding that in 24% of the intraday intervals and in 8% of the overnight intervals, there was no change in either the bid or the ask quote. If the underlying changes in the efficient price are viewed as arising from a continuous distribution of modest leptokurtosis, these figures suggest that the efficient price changes are not large relative to the tick size. The extreme values in the sample lie roughly seven standard deviations from the mean for the intraday intervals and five standard deviations from the mean for the overnight intervals.

(For a normally-distributed variate, the probability of observing an extreme value seven standard deviations from the mean in a sample of 6,528 observations is approximately 1×10^{-10}; that of an extreme value five standard deviations from the mean in a sample of 251 observations is approximately 1×10^{-7}.)

Table 2 reports descriptive statistics for the bid-ask spread. There is clear variation in the spread. In a sense, one purpose of the present model is the allocation of this variation to deterministic and stochastic effects.

Estimates of the full model

Table 3 reports parameter estimates. For purposes of exposition, these may be grouped as cost- and EGARCH (variance)-related. The EGARCH-related parameter estimates suggest a strong persistent stochastic component of the return variance. The autoregressive variance parameter estimate of φ=0.88, however, implies a half-life of about six (15-minute) periods. The intraday persistence reflects, therefore, phenomena different from those underlying daily and longer-term volatility persistence. The GED tail-thickness parameters of $\nu^{day} = 0.86$ and $\nu^{overnight} = 1.02$. For comparison purposes, Figure 3 graphs the GED density with ν=0.86 against the standard normal.

A GED with ν=0.86 has a standardized kurtosis in excess of the normal of roughly 4.6. To place this in perspective, the intraday 15-minute bid changes for the present sample (standardized by time-of-day standard deviation) exhibit an excess kurtosis of 3.84. It is recognized, however, that both discreteness and ARCH effects acting separately can affect kurtosis (see Gottlieb and Kalay for the former; Bollerslev et al (1994) for the latter). The present analysis therefore suggests that the kurtosis in the bid price changes is similar in magnitude to that estimated for the underlying efficient price process (the driving GED). This implies that kurtosis is fundamental and not an artifact of the observation process.

The ν estimates are lower than Nelson's estimate for daily CRSP returns (about 1.6). The present estimates imply a more pronounced leptokurtosis, consistent with a "lumpy" intraday information arrival process for individual stocks. This is less pronounced in the daily index returns due to aggregation over firms and time.

Turning now to the quote-exposure cost estimates, the deterministic parameters depict the usual U-shaped intraday pattern, although the standard errors of the decay rates are large. Of more interest is the characterization of the stochastic component. Both the disturbance variance σ_v and the autoregressive parameter ϕ are strongly positive. The autoregressive parameter suggests that 37% of the excess log cost persists at the subsequent time point (fifteen minutes later).

The relative importance of the deterministic and stochastic sources of variation in the quote exposure cost can be ascertained from simulations of the model using the parameter estimates. For a simulation of 2,500 days, Figure 4 depicts the time path of the 10^{th}, 50^{th} and 90^{th} percentiles of the cost expressed in dollars per share. The 50^{th} percentile (the median) displays the "U" shape characteristically found in spreads. The median is roughly four cents per share at the open and two cents thereafter, rising slightly at the close. Most importantly, the elevation associated with the beginning and end of trading is modest compared with the stochastic variation implied by the 10^{th} and 90^{th} percentile bands. This suggests that the stochastic component is relatively large.

The finding of a mean-reverting persistent component in the cost of quote exposure is consistent with several models of economic behavior. It may reflect mean-reversion in the underlying cost determinants (such as asymmetric information exposure costs or inventory holding costs related to risk) that are common across all actual and potential quote-setters. It may also reflect, however, the arrivals and departures of individual traders with differing costs. A buyer anxious to trade, for example, might enter a limit order that betters the prevailing bid. At some point this limit order is likely to be removed, either because it has been hit or else because the trader has withdrawn it and replaced it with a market buy order.

Estimates of the cost and EGARCH/GED models.

Both the cost and EGARCH/GED models are computationally simpler subcases of the full model. The first follows from an ongoing assumption of a diffuse prior for the efficient price; the second assumes a diffuse prior on the quote exposure costs. The resulting estimates are given in the last two columns of Table 3. Not only are the

estimates virtually identical to those obtained for the full model, but so are the estimated standard errors. Although one might have suspected that the full specification would result in more precise estimates, this does not appear to be the case.

Several considerations could account for this. One possibility is simply general model misspecification. But even in a correctly specified model, the information about α and β contributed by m (and vice versa) might be small. Section 4 pointed out that the model imposes structure even under the assumption of flat priors. With flat priors in the equal-cost model defined by $c=\alpha=\beta$, it was shown that c is likely to be more informative about the dispersion of m than the location. In estimates of the equal-cost model (not reported), the parameters governing the dynamics of m were essentially similar to those of the full model reported here.

The equal-cost and full models may be thought of as resulting from polar assumptions about the correlation (one and zero, respectively) between α and β. The similarity of estimates may stem from the inability of the data to identify the correlation. Ignoring discreteness issues for the moment (which compounds the difficulties), the statistical situation is similar to one in which we are attempting to infer the joint distributional properties of two unobserved variables (α and β) on the basis of observing their sum (the spread). The decreased informativeness of the sum taxes both the data and the model specification.

8. Conclusion

This paper has presented a dynamic model of discrete bid and ask quotes. The discrete quotes are rounded transformations of a continuous efficient price and continuous quote exposure costs. The latter are presumed to capture most of the costs usually associated with market-making or limit order placement, such as fixed transaction costs and asymmetric information costs. The full statistical model is a rich one, allowing for stochastic and deterministic time variation in the efficient price volatility and the quote exposure cost. The model is estimated by maximum likelihood using a nonlinear state-space filtering approach due to Kitagawa (1987).

This specification is estimated for NYSE bid and ask quotes collected at the end of 15-minute intervals for Alcoa over 1994. The estimates confirm the existence of deterministic "U" shapes in the quote cost and efficient price volatility, and also persistence in stochastic volatility. More importantly, however, the estimates confirm the existence of a persistent stochastic component of the quote exposure cost. The magnitude of this component is roughly comparable to the variation associated with the "U" shapes.

Logical and economically desirable extensions of the model mentioned earlier include allowing for correlations between the quote exposure costs and between these costs and the efficient price variance. One would additionally wish to allow for exogenous variables (especially signed trades) in the efficient price evolution, e.g., $m_t = m_{t-1} + \lambda x_t + u_t$, where x_t is the signed trade volume (positive for buyer-initiated trades) and λ is the market-depth parameter. It is relatively easy to allow for deterministic time-variation in λ. These extensions would require modifications of the state-transition density that are conceptually simple, but (once we can no longer conveniently factor $f(z_{t+1}|z_t)$) computationally taxing.

The next level of generalization would allow for stochastic market depth and endogenous trades. This would require modeling the dynamics of these variables, and including them in the state vector. The Kitagawa approach to maximum likelihood estimation employed in this paper, however, relies on numerical integration over the state vector. The curse of dimensionality looms large with state dimensions beyond the level (three) employed in the present paper. Other methods of functional approximation or Monte Carlo approaches may offer more promising paths.

Appendix A.

Approximation Methods.

The present analysis follows Kitagawa (1987) in approximating the conditional densities required in Section 5 by step functions defined over a lattice. The three-dimensional state variable is $z_t = \{m_t, \alpha_t, \beta_t\} \in \mathbf{R} \times \mathbf{R}^+ \times \mathbf{R}^+$. The space $\mathbf{R} \times \mathbf{R}^+ \times \mathbf{R}^+$ is assumed to be partitioned into a set of lattice cells $\{C_t^1, C_t^2, \ldots\}$ where each $C_t^i \in \mathbf{R} \times \mathbf{R}^+ \times \mathbf{R}^+$ is a rectangular solid. The conditional density $f(z_t | q_t, q_{t-1}, \ldots)$ is approximated as

$$f(z_t | q_t, q_{t-1}, \ldots) \approx \sum_{i=1}^{\infty} \delta_t^i I(z_t \in C_t^i) \tag{A.1}$$

where I is an indicator variable and the δ's are parameters. The state transition densities $f(z_{t+1} | z_t)$ are replaced by the discrete transition probabilities $\Pr(C_{t+1}^j | C_t^i)$, and the integration in (9) is becomes the summation:

$$\Pr(C_{t+1}^j | q_t, q_{t-1}, \ldots) = \sum_{i=1}^{\infty} \Pr(C_{t+1}^j | C_t^i) \Pr(C_t^i | q_t, q_{t-1}, \ldots) \tag{A.2}$$

The integration in (10) becomes:

$$\Pr(q_{t+1} | q_t, q_{t-1}, \ldots) = \sum_{j=1}^{\infty} \Pr(C_{t+1}^j | q_t, q_{t-1}, \ldots) \frac{\mathrm{Vol}(C_{t+1}^j \cap Q_{t+1})}{\mathrm{Vol}(C_{t+1}^j)} \tag{A.3}$$

where $\mathrm{Vol}(\cdot)$ is the volume (in cubic ticks) of its argument. When the lattice cell lies entirely within the feasible region Q_{t+1}, the summand in (A.3) is simply $\Pr(C_{t+1}^j | q_t, q_{t-1}, \ldots)$. But when only a portion of the cell lies within the feasible region, the probability is weighted by the ratio of the feasible volume to the total. (The calculation of the intersection volume was performed using the computational geometry algorithms and software routines discussed in O'Rourke (1994). Although these algorithms have not, to my knowledge, been used in econometrics, they are standard and are implemented in a number of software packages.)

As with the integration in (10), computation of the summation in (A.3) is facilitated by the restrictions implied by Q_{t+1}. The intersection $C_{t+1}^j \cap Q_{t+1}$ is empty for

virtually all of cells in the z_t space, and it is easy to specify the small set of nonempty cells. Furthermore, if the purpose of the calculation is "off-line" estimation (rather than real-time forecasting), we can economize on the calculation of the $\Pr\left(C_{t+1}^j \middle| q_t, q_{t-1}, \ldots\right)$ in equation (A.3) by computing only the values that will be used in the subsequent probability calculation. These simplifications greatly reduce the computational burden.

In approximating the discrete transition probabilities $\Pr\left(C_{t+1}^j \middle| C_t^i\right)$, I generally took

$$\Pr\left(C_{t+1}^j \middle| C_t^i\right) \approx \mathrm{Vol}\left(C_{t+1}^j\right) f\left(z_{t+1}^j \middle| z_t^i\right) \tag{A.4}$$

where z_{t+1}^j and z_t^i were the cell midpoints. This should be a good approximation when the density function is relatively constant over C_{t+1}^j. When the transition from C_t^i to C_{t+1}^j included the point of no change in the efficient price, however, concerns over the sharp peak in the GED distribution led me to use

$$\Pr\left(C_{t+1}^j \middle| C_t^i\right) \approx \int_{z_{t+1} \in C_{t+1}^j} f\left(z_{t+1} \middle| z_t^i\right) dz_{t+1} \tag{A.5}$$

where the integration calculation used the midpoint approximation (as in (A.4)) for α_t and β_t, and Chebyshev integral approximation along the m axis.

At time t, the set of lattice cells $\left\{C_t^i\right\}$ was computed as the cross-product of three one-dimensional lattices, one for each of the state variables α, β, and m. The break-points for the α lattice were (in ticks): 0., 0.01, 0.2, 0.04, 0.07, 0.13, 0.24, 0.46, 0.88, 1.67, 3.16 and 6 (the maximum spread in the analysis). These breakpoints approximate fixed intervals in $\ln(\alpha)$. The lattice for m ranged from the lowest bid in the sample to the highest ask, in 0.2-tick increments.

The number of cells necessary to cover a given quote region $Q(a,b)$ depends only on the spread a-b. For spread sizes of one through six ticks, the corresponding cell counts are: 282, 352, 218, 218, 198, 162 (for the full model); 5, 10, 15, 25, 30 (for the restricted EGARCH-only model); 74, 95, 42, 40, 40, 25 (for the restricted cost-only model).

Appendix B.

Simulations.

The estimation technique employed here involves numerical approximations and computations of moderate complexity. To assess the adequacy of these approximations and verify performance of the software, I conducted a series of simulations. The simulated model consists of equation (4) for the EGARCH/GED evolution of the efficient price and equation (12) for the evolution of the quote exposure costs. The parameter values are $\eta = 1, \varphi = 0.9, \gamma = 0.3, \nu = 1, \quad \mu = -1.7, \phi = 0.4$ and $\sigma_v = 0.9$ (chosen to approximate the values estimated from the sample). The model is essentially identical to the one applied to the actual sample data, with the exception that I did not introduce deterministic (intraday) patterns. This was done to reduce the dimensionality of the optimizations, and should not materially affect the parameter estimates for the stochastic components of the model.

I generated 25 samples of 1,000 observations (vs. roughly 6,000 in the actual sample), and estimated parameters and standard errors (with the same approximation grids used over the actual sample). Table B-I summarizes the results. Overall, convergence to population values is good, and the estimated standard errors agree with the dispersion found in the sample estimates. As in the actual sample data, the results for the full model and two sub-models are in good agreement. On a 200 mHz Pentium Pro, the average convergence times for the likelihood optimization were 22:42 (minutes:seconds) for the full model, 00:40 for the EGARCH/GED model and 00:41 for the cost model.

Table B-I.

Simulation Results.

		Parameter						
		μ	σ_v	ϕ	η	φ	γ	ν
	True Value	-1.7	0.9	0.4	1.0	0.9	0.3	1.0
Model								
Full	Mean Estimate	-1.724	0.871	0.419	0.982	0.884	0.321	1.026
	Std. Dev. of Estimates	0.062	0.046	0.053	0.110	0.044	0.072	0.066
	Mean of Estimated SE's	0.061	0.046	0.061	0.135	0.045	0.069	0.073
EGARCH/GED	Mean Estimate				0.964	0.886	0.337	0.987
	Std. Dev. of Estimates				0.120	0.046	0.077	0.059
	Mean of Estimated SE's				0.142	0.044	0.073	0.075
Cost	Mean Estimate	-1.722	0.872	0.421				
	Std. Dev. of Estimates	0.061	0.045	0.050				
	Mean of Estimated SE's	0.061	0.046	0.062				

Notes: The table reports the summary statistics for 25 simulations and estimations of the model described in Appendix B with 1,000 observations per simulation.

References

Andersen, T. G. and T. Bollerslev, 1997, Heterogeneous information arrivals and return volatility dynamics: uncovering the long-run in high-frequency returns, Journal of Finance, 52(3), 975-1005.

Ahn, H.-J, C. Q. Cao, and H. Choe, 1996, Tick size, spread and volume, Journal of Financial Intermediation, 5.

Angel, J., 1994, Tick size, share prices and stock splits, working paper, Georgetown University.

Anshuman, V. R., and Avner Kalay, 1994, Market making rents under discrete prices: Theory and evidence, Working paper, Boston College.

Ball, C. A., 1988, Estimation bias introduced by discrete security prices, Journal of Finance, 43, 841-865.

Bernhardt, D. and E. Hughson, 1996, Discrete pricing and the design of dealership markets, Journal of Economic Theory, 71, 148-182.

Bollerslev, T., and M. Melvin, 1994, Bid-ask spreads in the foreign exchange market: an empirical analysis, Journal of International Economics.

Bollerslev, T., R. F. Engle and D. B. Nelson, 1994, ARCH models, in R. F. Engle and D. L. McFadden, eds., Handbook of Econometrics, Volume IV, (Elsevier).

Brock, W. and A. Kleidon, 1992, Periodic market closure and trading volume: A model of intraday bids and asks, Journal of Economic Dynamics and Control, 16, 451-489.

Brown, S., P. Laux and B. Schacter, 1991, On the existence of an optimal tick size, Review of Futures Markets, 10, 50-72.

Chan, K. C., W. G. Christie and P. H. Schultz, 1995, Market structure and the intraday pattern of bid-ask spreads for NASDAQ securities, Journal of Business, 68(1), 35-60.

Cheung, Y. W. and L. Ng, 1992, Stock price dynamics and firm size: An empirical investigation, Journal of Finance, 47, 1985-1997.

Cho, D. C. and E. W. Frees, 1988, Estimating the volatility of discrete stock prices, Journal of Finance, 43, 451-466.

Chordia, T. and A. Subrahmanyam, 1995, Market making, the tick size, and payment-for-order flow: theory and evidence, Journal of Business, 68, 543-575.

Christie, W. G. and P. H. Schultz, 1994a, Why do NASDAQ market makers avoid odd-eighth quotes?, Journal of Finance, 49, 1813-40.

Christie, W. G. and P. H. Schultz, 1994b, Why did NASDAQ market makers stop avoiding odd-eighth quotes?, Journal of Finance, 49, 1841-60.

Cordella, Tito and Thierry Foucault, 1996, Minimum price variations, time priority and quote dynamics, Working paper, Universitat Pompeu Fabra.

Dravid, A. R., 1991, Effects of bid-ask spreads and price discreteness on stock returns, Working Paper, Rodney L. White Center, Wharton School.

Duffee G. R., 1995, Stock returns and volatility: A firm level analysis, Journal of Financial Economics, 37, 399-420.

Engle, R. F., and J. E. Russell, 1995, Forecasting transaction rates: the autoregressive conditional duration model, UCSD Discussion Paper.

Engle, R. F., 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, Econometrica, 50, 987-1008.

Engle, R. F., 1996, The econometrics of ultra-high-frequency data, UCSD Discussion Paper.

Glosten, L. R. and L. E. Harris, 1988, Estimating the components of the bid-ask spread, Journal of Financial Economics, 21, 123-42.

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

Glosten, L. R., 1994, Is the electronic open limit order book inevitable?, Journal of Finance, 49, 1127-1161.

Gottlieb, G. and A. Kalay, 1985, Implications of the discreteness of observed stock prices, Journal of Finance, 40, 135-153.

Hamilton, J. D., 1994a, "State-Space Models," in R. F. Engle and D. L. McFadden, eds., Handbook of Econometrics, Volume IV, (Elsevier).

Hamilton, J. D., 1994b, Time Series Analysis, (Princeton University Press, Princeton).

Harris, L. E. and J. Hasbrouck, 1996, Market vs. Limit orders: the SuperDOT evidence on order submission strategy, Journal of Financial and Quantitative Analysis, 31, 213-232.

Harris, L. E., 1990, Estimation of stock variances and serial covariances from discrete observations, Journal of Financial and Quantitative Analysis, 25, 291-306.

Harris, L. E., 1990, Liquidity, trading rules and electronic trading systems, Salomon Center Monograph Series in Finance and Economics 1990-4 (Stern School of Business, New York University).

Harris, L. E., 1991, Stock price clustering and discreteness, Review of Financial Studies, 4, 389-415.

Harris, L. E., 1994, Limit Orders, working paper, University of Southern California.

Harris, L. E., 1994, Minimum price variations, discrete bid-ask spreads and quotation sizes, Review of Financial Studies, 7, 149-178.

Harris, L. E., 1996, Does a large minimum price variation encourage order exposure, working paper, School of Business, University of Southern California.

Harris, L. E., 1997, Decimalization: a review of the arguments and evidence, working paper, School of Business, University of Southern California.

Harvey, A. C., 1991, Forecasting, Structural Time Series Models and the Kalman Filter, (Cambridge).

Hasbrouck, J., 1991, Measuring the information content of stock trades, Journal of Finance, 46, 179-207.

Hasbrouck, J., 1996, Modeling microstructure time series, in *Handbook of Statistics, Volume IV*, G. S. Maddala and C. R. Rao, eds., (North Holland; Amsterdam).

Hausman, J. A., A. W. Lo and A. C. MacKinlay, 1992, "An Ordered Probit Analysis of Transaction Stock Prices," Journal of Financial Economics, 31, 319-330.

Kitagawa, G., and W. Gersch, 1996, Smoothness Priors Analysis of Time Series, Lecture Notes in Statistics 116, (Springer-Verlag: New York).

Kitagawa, G., 1987, "Non-Gaussian State-Space Modeling of Nonstationary Time Series," Journal of the American Statistical Association, 82, 1032-1041.

Lee, C. M., B. Mucklow and M. J. Ready, Spreads, depths and the impact of earnings information: an intraday analysis, Review of Financial Studies, 6, 345-374.

McInish, T. H. and R. A. Wood, 1992, An analysis of intraday patterns in bid/ask spreads for NYSE stocks, Journal of Finance, 47(2), 753-764.

Madhavan, A., M. Richardson and M. Roomans, 1997, Why do security prices change? A transaction level analysis of NYSE stocks, Review of Financial Studies, forthcoming.

Marsh, T. A. and E. R. Rosenfeld, 1986, Non-trading, market making and estimates of stock price volatility, Journal of Financial Economics, 15, 359-372.

Nelson, D. B., 1991, Conditional Heteroskedasticity in asset returns: a new approach, Econometrica, 59, 347-370.

Niederhoffer, V., 1965, Clustering of stock prices, Operations Research, 13, 258-265.

Niederhoffer, V., 1966, A new look at clustering of stock prices, Journal of Business, 39, 309-313.

O'Rourke, Joseph., 1994, Computational Geometry in C. Cambridge ; New York : Cambridge University Press.

Rock, K., 1996, The specialist's order book and price anomalies, forthcoming, Review of Financial Studies.

Shephard, N., and A. Manrique, 1997, Simulation based likelihood inference for limited dependent processes, Working paper, Nuffield College, Oxford University.

Table 1.

Descriptive statistics for 15-minute bid changes, Alcoa, 1994.

Bid changes (in 1/8 dollar ticks) were computed for Alcoa for all trading days in 1994 (plus the overnight change).

	Intraday	Overnight
N	6,528	251
Min (ticks)	-10	-15
Max (ticks)	11	19
Mean (ticks)	0.03	-0.21
Std. Dev. (ticks)	1.58	3.70
Distribution		
% with no change	39%	18%
% with 1-tick change	37%	27%
% with >1-tick change	25%	47%

Table 2.

Descriptive statistics for bid-ask spread at 15-minute intervals, Alcoa, 1994. Spreads (in 1/8 dollar ticks) were computed for Alcoa at fifteen minute intervals during the trading day, for all trading days in 1994

N	6,780
Min	1
Max	5
Mean	1.65
Std	0.67
Distribution	
1-Tick	45.9%
2-Tick	43.5%
3-Tick	10.3%
4 or more ticks	0.3%

Table 3.

The state variables in the model are $z_t = \{m_t, \alpha_t, \beta_t\}$ where m_t is the implicit efficient price, α_t is the quote exposure cost on the ask or offer side of the market, and β_t is the quote exposure cost on the bid side of the market. t indexes 15-minute intraday intervals (plus the overnight period). The dynamics of the state variables are:

$$\begin{aligned} m_t &= m_{t-1} + u_t \\ \ln(\alpha_t) &= \mu_t + \phi(\ln(\alpha_{t-1}) - \mu_{t-1}) + v_t^{\alpha} \\ \ln(\beta_t) &= \mu_t + \phi(\ln(\beta_{t-1}) - \mu_{t-1}) + v_t^{\beta} \end{aligned}$$

where v_t^{α} and v_t^{β} are independently distributed as $N(0, \sigma_v^2)$. The efficient price disturbance, u_t, has standard deviation σ_t. Its standardized value $\zeta_t = u_t/\sigma_t$ is distributed as $f_{GED}(\zeta_t; \nu)$ where ν is the tail-thickness parameter. The efficient price variance follows a modified EGARCH process:

$$\ln(\sigma_t^2) = \eta_t + \varphi(\ln(\sigma_{t-1}^2) - \eta_{t-1}) + \gamma(\mathrm{E}_{t-1}|\zeta_{t-1}| - \mathrm{E}|\zeta_{t-1}|)$$

where $E_{t-1}|\zeta_{t-1}| = E[|u_{t-1}||q_{t-1}, q_{t-2}, \ldots]/\sigma_t$ is the filtered estimate conditional on the bid and ask prices through t-1.

The deterministic component of the cost process is:

$$\mu_t = k_1 + k_2^{open} \exp(-k_3^{open} \tau_t^{open}) + k_2^{close} \exp(-k_3^{close} \tau_t^{close})$$

where τ_t^{open} is the elapsed time since the opening quote of the day (in hours) and τ_t^{close} is the time remaining before the scheduled market close (in hours). The deterministic component of the variance is:

$$\eta_t = \begin{cases} l_1 + l_2^{open} \exp(-l_3^{open} \tau_t^{open}) + l_2^{close}, & \text{if } t \text{ is an intraday interval} \\ \eta^{overnight}, & \text{if } t \text{ is an overnight interval} \end{cases}$$

The observations are the discrete bid and ask quotes: $b_t = \mathrm{Floor}[m_t - \beta_t]$ and $a_t = \mathrm{Ceiling}[m_t + \alpha_t]$.

The column corresponding to the "full" model gives parameter estimates based the Kitagawa nonlinear filtering procedure. The "cost" model is estimated assuming at each point in time a flat prior for the efficient price. The "EGARCH" model is estimated

Table 3 (Continued).

assuming at each point in time a flat prior for the quote exposure costs. The models are estimated for Alcoa over all trading days in 1994, with t indexing 15-minute intervals within the day (and the overnight interval). Standard errors are reported in parentheses.

Table 3 (Continued).

		Model		
		Full	Cost	EGARCH
Quote exposure cost parameters:	k_1	-1.67 (0.03)	-1.68 (0.03)	
	k_2^{open}	0.45 (0.06)	0.46 (0.06)	
	k_3^{open}	2.42 (0.72)	2.40 (0.69)	
	k_2^{close}	0.21 (0.07)	0.19 (0.07)	
	k_3^{close}	3.48 (2.50)	3.50 (2.68)	
	ϕ	0.37 (0.03)	0.39 (0.03)	
	σ_v	0.86 (0.03)	0.86 (0.02)	
EGARCH parameters:	l_1	0.39 (0.06)		0.35 (0.07)
	l_2^{open}	1.73 (0.23)		1.76 (0.24)
	l_3^{open}	1.11 (0.23)		1.11 (0.23)
	l_2^{close}	0.59 (0.14)		0.61 (0.15)
	$\eta^{overnight}$	2.73 (0.13)		2.73 (0.14)
	φ	0.88 (0.02)		0.90 (0.02)
	γ	0.29 (0.03)		0.28 (0.03)
	ν^{day}	0.86 (0.02)		0.81 (0.02)
	$\nu^{overnight}$	1.02 (0.12)		1.01 (0.11)

Figure 1

Feasible Regions for the Equal Cost Model.

As a function of the efficient price m and quote exposure cost c, the discrete bid and ask quotes are given by b=Floor($m-c$) and a=Ceiling($m+c$). Given bid and ask quotes a and b, the region of feasible m and c is:

$$Q(b,a) = \{(m,c): c > 0, b \le m - c < b + 1 \text{ and } a - 1 < m + c \le a\}$$

The figure depicts the regions $Q(0, 1)$, $Q(0, 2)$ and $Q(0, 3)$.

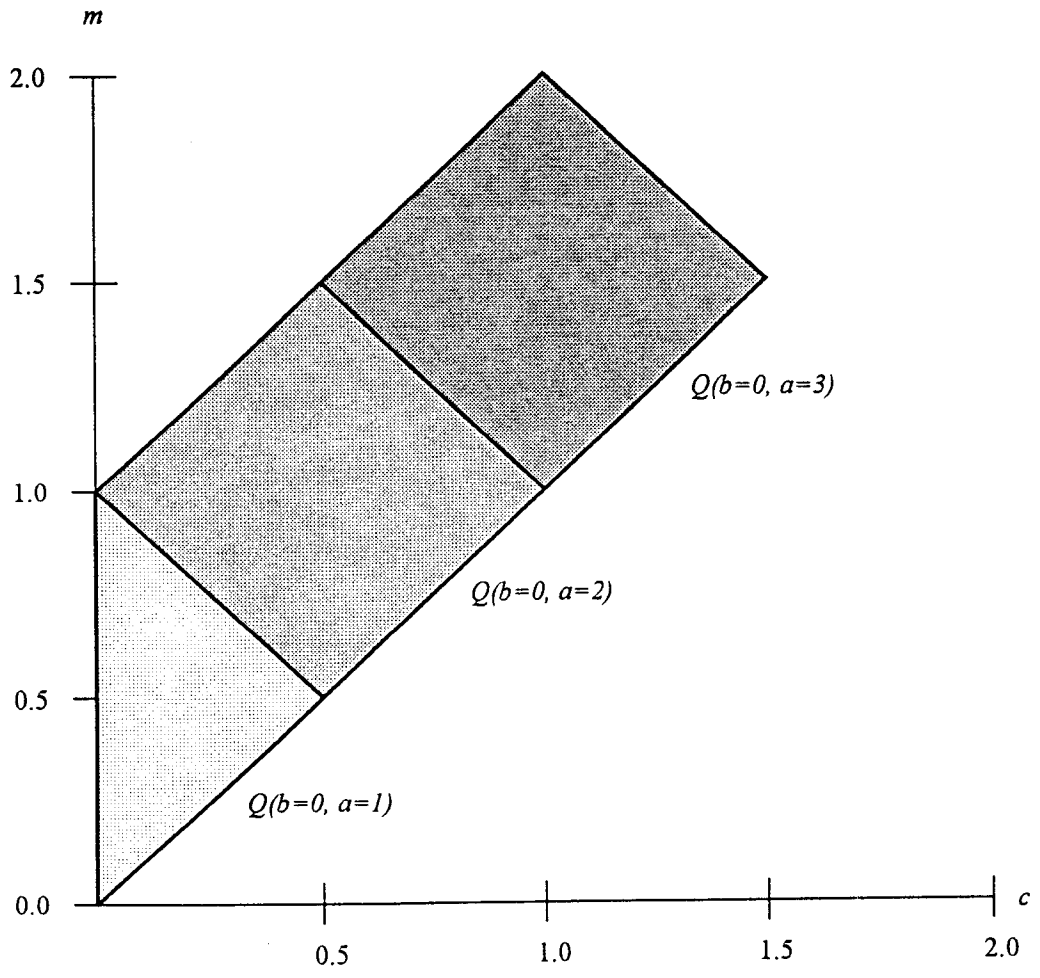
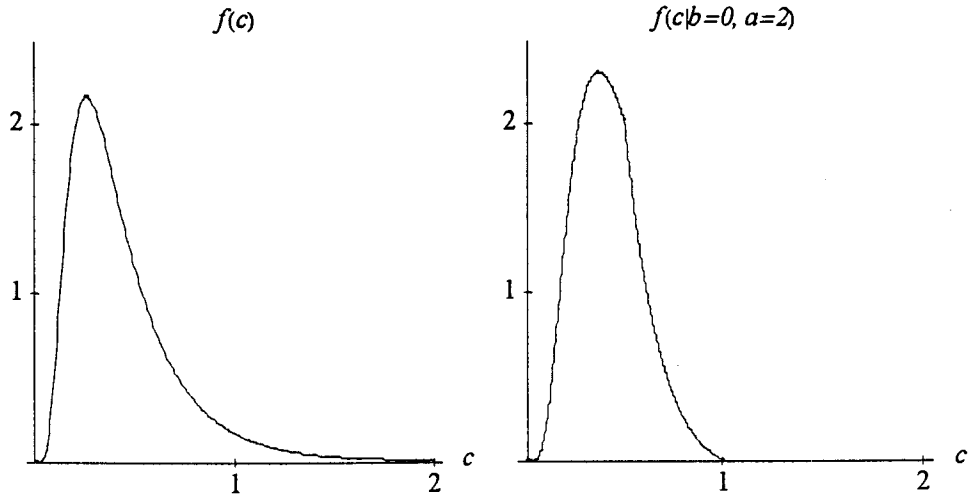


Figure 2

Figure depicts prior and poster probability densities for the efficient price m and quote exposure cost c. The prior density of c is lognormal: $\ln(c) \sim N(\mu = -1, \sigma = 0.6)$. The prior density for m is a uniform prior on the interval (0, κ), where κ is an arbitrary positive constant (and does not appear in the posterior densities). The posterior densities are conditional on observing bid and ask quotes of b=0 and a=2.

Panel A. Prior and posterior densities of the quote exposure cost c.



Panel B. Prior and posterior densities for the efficient price m.

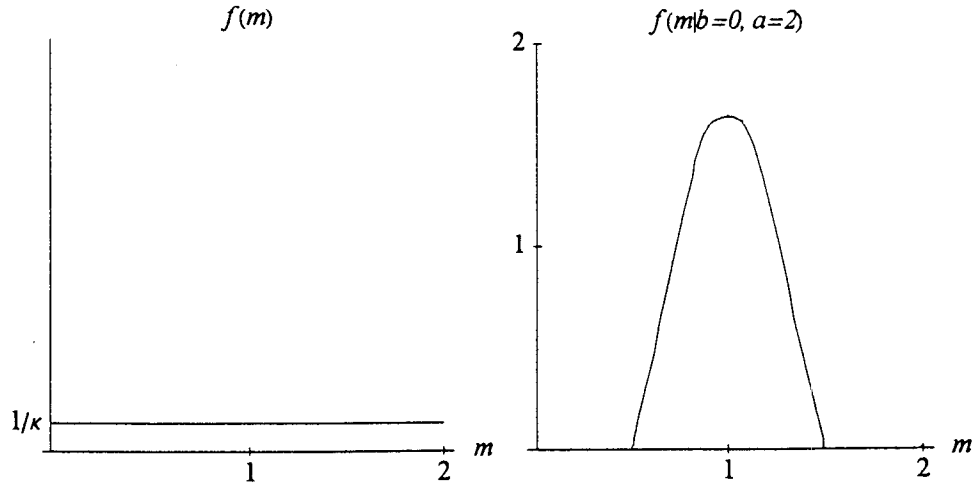


Figure 3

Figure depicts the probability density functions for the standard normal and standard GED with tail-thickness parameter ν=0.86.

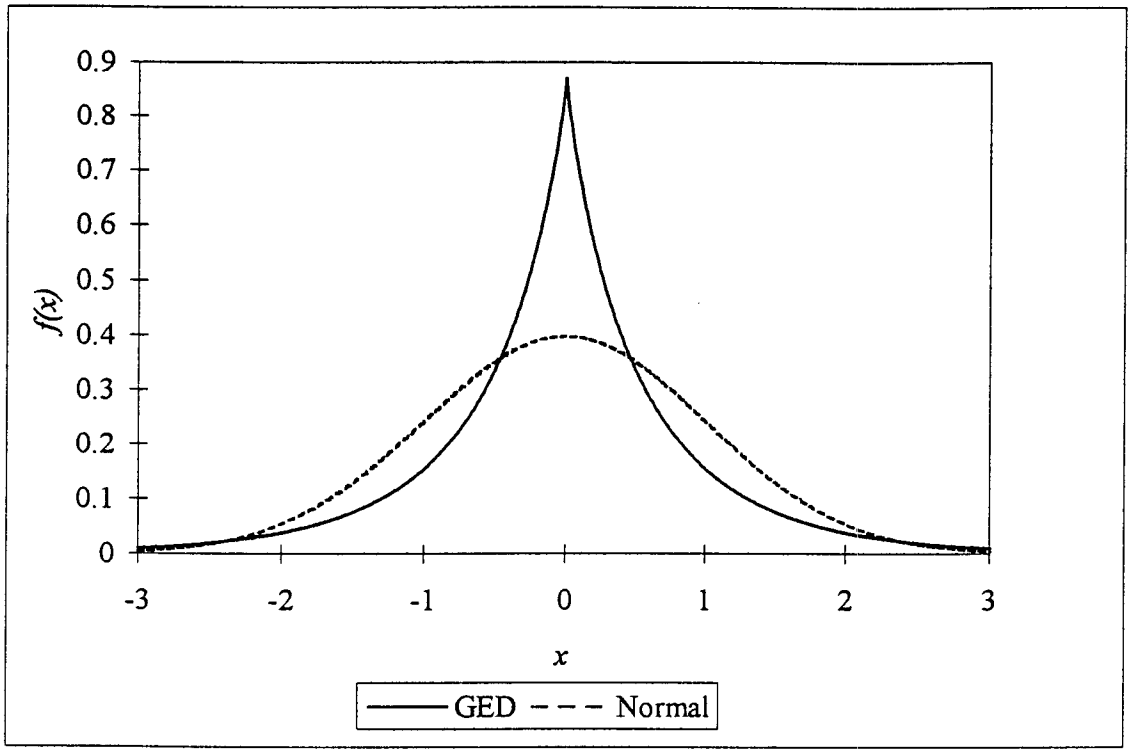


Figure 4

Figure depicts the time of day patterns in the quote exposure cost for ticker symbol AA implied by the model and estimates given in Table 3. The solid line is the 50th percentile of the cost. The upper and lower dashed lines are the 10th and 90th percentiles (respectively). (NB: these are not estimation confidence intervals.)