Empirical Market Microstructure

Economic and Statistical Perspectives on the Dynamics of Trade in Securities Markets

Teaching notes for B40.3392, Fall, 2003

Prof. Joel Hasbrouck

Department of Finance
Stern School of Business
New York University
44 West 4th St.
New York NY 10012

email: jhasbrou@stern.nyu.edu
web: http://www.stern.nyu.edu/~jhasbrou

Lastest versions and supporting material (programs, datasets, etc.) document are contained in the Empirical Market Microstructure link off of my web page.

Draft 1.1
13:32 on Thursday, January 8, 2004

Preface

This document is a collection of teaching notes from a one-semester PhD course given in the Fall of 2003. My intent was to cover some of the empirical approaches to market microstructure, the theory that motivated them, and the results from time series analysis necessary to understand them. I assume that the reader has some prior exposure to or a working knowledge of basic financial economics and statistics, but beyond this the presentation is self-contained.

Part I discusses the economic structure underlying the martingale property of security prices, and discusses some preliminary features of actual security price data. I then turn to consideration of fixed transaction costs and the Roll (1984) model of the bid-ask spread, which then becomes the central construct going forward. In particular, the Roll model is used to introduce moving-average and autoregressive representations of time series. The next two sections cover the basic asymmetric information models: the sequential trade and continuous auction approaches. I then return to the Roll model and discuss generalizations that incorporate asymmetric information. These generalizations all feature a transaction price that behaves as random walk (the efficient price) plus noise. The last section of Part I turns to general methods for characterizing random-walk and noise components from statistical evidence. All of the statistical specifications discussed in Part I are univariate representations of price changes.

Part II discusses trades, i.e., quantities that can be signed "buy" or "sell", usually from the viewpoint of a customer demanding liquidity. Trades constitute an essential component of the asymmetric information models described in Part I. They also give rise to what have historically been called "inventory control effects". Part II discusses basic inventory control models. The discussion then shifts to multivariate time series models, specifically those that involve prices and trades. I examine purely statistical models (vector autoregressions), and discuss characterizations of random-walk and noise components in these models. These results are generalizations of the univariate results. I discuss a number of structural economic models that fit into this framework. It is logical at this point to consider estimates of information asymmetry based solely on trades (the "probability of informed trading", PIN). Another useful generalization involves multiple prices on the same security.

Electronic limit order books have emerged as the preeminent security market structure. Part III discusses the economics of limit orders and markets organized around them. Part IV describes links between market microstructure and asset pricing. These last two areas are especially active fields of research.

It is sometimes useful to have a sense of the actual trading institutions. A descriptive piece on US equity markets (originally written as a separate working paper) is included in the appendix to this document.

The bibliography to this ms. has live web links. Some of the links are to working paper sites. Others are directly to journals, JSTOR or Econbase. You (or your institution) may need a subscription to follow these.

The scope of this manuscript is limited, and the selection of material is idiosyncratic. It is most certainly not a comprehensive treatment of the field of market microstructure. A partial list of omitted topics would

include: transaction cost measurement; comparative market design; the "industrial organization" aspects of market structure (fragmentation, consolidation, etc.); behavioral aspects of trading; experimental evidence; the role of time (duration models, asynchronous trading, etc.); price/volume analyses. In addition, the paper is primarily concerned with equity markets. The microstructures of bond, foreign exchange, futures and options markets are different.

Nor is the book a full treatment of time series analysis. In fact, there are many excellent books on time series analysis. Why attempt the awkward task of bringing this material into a microstructure treatise at all? There are several reasons. In the first place, time series analysis concepts are useful (perhaps essential) to critically evaluating the empirical work in the field. Second, the interplay between economic and statistical microstructure models often helps to clarify both. As a final and perhaps more subtle point, exposition in most statistics texts (coverage, sequencing, balance) is usually driven, implicitly at least, by the nature of the data to be modeled. It is a fact that most applications and illustrations in the extant literature of time series econometrics are drawn from macroeconomics. Now a theorem is a theorem irrespective of the sampling frequency. But microstructure data and models are distinctive: normality is often an untenable assumption; sample sizes are usually enormous; measurement of "time" itself is open to various interpretations. Moreover, topics such as random-walk decompositions and cointegration, which might walk on in Act IV of a macroeconomic analysis, merit starring roles in microstructure dramas. It is my hope that seeing this material organized from a microstructure perspective will help readers to apply it to microstructure problems.

The notes contain a few assigned problems and empirical "cases". Problems look like this:

Problem 0.1 Information asymmetries in the gold market

In the following model, what is the implied price impact of a $1M gold purchase? ...

Where I've worked out the answer, it is indicated as:

Answer

The value of $0.02 per ounce is obtained as follows ...

■ **Note:**

The answers are not distributed with the pdf version of this document.

Although this document is text of lecture notes that can be printed or viewed on a screen, it is also a computer program. It was composed in *Mathematica*, a software package for working with symbolic mathematics. The "code" for many of the derivations, solutions, graphs, etc. is embedded in the text. For the sake of expositional clarity, display of this code is suppressed in the printed and pdf versions of the document. (Large sections of code are identified by "***Mathematica***" in the right-hand margin.) If you're curious, though, you can download the *Mathematica* notebook and examine and/or run the code. To view the code, you'll need the (free) *MathReader*, available at www.wolfram.com. To run the code, you'll need the full *Mathematica* system.

Contents

Part I: Univariate models of security prices 1

1. Market microstructure: an overview 2

• Sources of value and reasons for trade • Mechanisms in economic settings • Multiple characterizations of prices • "Liquidity" • Econometric issues • The questions • Readings • *Mathematica* initializations

2. The long-term dynamics of security prices 7

2.a Macroeconomic models of asset prices 7

• A sample of market prices

2.b Martingales in microstructure analyses 11

3. A dealer market with fixed transaction costs: the Roll model 12

3.a Model structure 12

3.b Inference 13

4. Moving average and autoregressive representations of price changes 15

4.a Stationarity and ergodicity 15

4.b Moving average models 16

4.c Autoregressive models 18

4.d The lag operator and representations . 18

4.e Forecasting …. 19

4.f Problems 20

5. Sequential trade models of asymmetric information 21

5.a Overview 21

5.b A simple sequential trade model 22

• A numerical example • Market dynamics over time • Numerical example, continued

5.c Extensions ….. 28

• Fixed transaction costs • Price-sensitive liquidity traders and market failures • Event uncertainty • Orders of different sizes • Orders of different types

5.d Empirical implications 31

5.e Problems 32

6. Strategic trade models of asymmetric information 33

6.a The single-period model 33

• The informed trader's problem • The market maker's problem • Properties of the solution

6.b The multiperiod model 37

• Setup • Solution • Analysis of solution • Numerical example • Autocorrelation in trades • Increasing the number of auctions (when total noise trading remains unchanged)

6.c Problems based on the single-period model ... 42

7. The generalized Roll model 44

7.a Overview 44

7.b Model description 44

• Alternative representations and special cases • The autocovariance structure of Δp_t

7.c Identification of σ_w^2 … 46

7.d The moving average (MA) representation 47

• Forecasting and filtering • Proof

7.e How closely does p_t track m_t? 50

• Overview • σ_s^2 in the generalized Roll model

8. Univariate random-walk decompositions 53

8.a Overview 53

8.b The autocovariance generating function 54

8.c The random-walk variance …. 56

8.d Further identification in special cases 56

• The special case of $\theta_\eta(L)\,\eta_t = 0$: Additional results • The special case of $\theta_w(L) = 0$

8.e Smoothing (optional) . 58

• General setup • Exclusively private information • Exclusively public information

8.f Filtering 60

8.g Variance of the pricing error: σ_s^2 60

• Other approaches

8.h Problems 62

9. Estimation of time series models 64

9.a Estimating the MA model. 64

• Maximum likelihood • Direct moment estimates • Estimation based on autoregression

9.b Structural estimates and their distributional properties 67

• The "delta" method • Subsampling • Starting values

9.c Case study I … 69

• Accessing WRDS • Using SAS • Analyzing the output

Part II: Multivariate models of trades and prices 71

10. The trade process and inventory control 72

10.a The dealer as a smoother of intertemporal order imbalances. 72

• Background: the exponential/Poisson arrival model • The Garman model

10.b Active inventory control 74

10.c How do dealer inventories actually behave? .. 75

• Is the visible quote the control variable for inventory control?

10.d The properties of the trade direction series …. 78

11. Random walks, etc. 79

11.a Is it a random walk? . . 79

11.b Invertibility 81

11.c The Wold theorem revisited . . 81

• Summary

12. Multivariate time series 86

12.a Vector moving average and autoregressive models … 86

12.b Impulse response functions: their use and interpretation 88

12.c Cholesky factorizations 89

12.d Attributing explanatory power 90

12.e Forecast variance decompositions 91

13. Prices and trades: statistical models 93

13.a Trade direction variables: constructing q_t 93

13.b Simple trade/price models 93

• Model 1 (Generalized Roll model, with both p_t and q_t observed) • Model 2: Autocorrelated trades • Model 3: Endogenous trades • Model 4: Contemporaneous trade and public information effects

13.c General VAR specifications .. 99

13.d Summary of asymmetric information measures 101

• The trade impact coefficient, λ • Variance decomposition measures

13.e Case Study II . 101

14. Prices and trades: structural models 103

14.a Glosten & Harris (1988) 103

14.b Madhavan, Richardson and Roomans (1997) 104

14.c Huang and Stoll (1997) 104

14.d The components of the spread 105

15. The probability of informed trading (*PIN*) 107

15.a Model structure 107

15.b A mixture of two Normal Poisson approximations 109

15.c Mixture aspects of EHKOP .. 112

15.d Summary 114

16. What do measures of information asymmetry tell us? 116

17. Linked prices: cointegration and price discovery 117

17.a Two securities 117

17.b One security, two markets 118

17.c The general case of multiple prices 120

• Price discovery

17.d Sources of cointegration 121

• Linear arbitrage conditions • Nonlinear arbitrage conditions

17.e Case Study III 122

Part III: Limit orders 123

18. Limit orders and dealer quotes 124

18.a Overview 124

18.b Limit order placement when faced with incoming orders of varying size 125

18.c Empirical evidence 131

18.d Introduction of a dealer/specialist 133

19. Bidding and offering with uncertain execution 135

19.a Expected utility 135

19.b Setting the bid for a single risky security. 135

• Extension: Bid as a function of quantity
19.c Setting the bid with correlated risky assets 137
• The bid for asset 1: • Bids for portfolios
20. Limit order submission strategies 140
• Broader models of choice and strategy
21. Dynamic equilibrium models 146
• Foucault (1999) • Parlour (1998)

Part IV: Microstructure and asset pricing 149
22. Trading and asset pricing with fixed transaction costs 150
22.a Theory 150
• Amihud and Mendelson (1986): The model • Constantinides (1986) • Heaton and Lucas (1996)
22.b Empirical Analyses 158
• Amihud and Mendelson (1986) • Brennan and Subrahmanyam (1996)
22.c Alternative measures of "liquidity" 161
• Liquidity ratio • Illiquidity ratio • Reversal measures
22.d Stochastic liquidity 164

Appendix: US equity markets: overview and recent history 165

Bibliography 188

Part I: Univariate models of security prices

Chapter 1. Market microstructure: an overview

Market microstructure is the study of the trading mechanisms used for financial securities.

There is no “microstructure manifesto," and historical antecedents to the field can probably be found going back to the beginning of written language. But at some point, the field acquired a distinct identity. As good a starting point as any is the coinage of the term “market microstructure” in the paper of the same title by Garman (1976):

> “[W]e depart from the usual approaches of the theory of exchange by (1) making the assumption of asynchronous, temporally discrete market activities on the part of market agents and (2) adopting a viewpoint which treats the temporal microstructure, i.e., moment-to-moment aggregate exchange behavior, as an important descriptive aspect of such markets.”

Analysis from this perspective typically draws on one or more of the following themes.

- **Sources of value and reasons for trade**

In many economic settings, the value of something is often thought to possess private and common components. Private values are idiosyncratic to the agent and are usually known by the agent when the trading strategy is decided. Common values are the same for everyone in the market and are often known or realized only after trade has occurred.

In security markets, the common value component reflects the cash flows from the security, as summarized in the present value of the flows or the security’s resale value. Private value components arise from differences in investment horizon, risk-exposure, endowments, tax situations, etc. Generally, common value effects dominate private value effects.

A necessary condition for gains from trade within a set of agents is contingent on some sort of differentiation. In modeling, this is often introduced as heterogeneous private values.

■ Mechanisms in economic settings

Once motives for trade are established, microstructure analyses generally focus on the mechanism, or protocol, used to effect trade.

Most economists first encounter the Walrasian auction. An auctioneer calls out a hypothetical price, and agents specify their excess demands. The process iterates until the total excess demand is zero. This mechanism is rarely encountered in practice (the London gold "fixing" being the most important example). It is nevertheless a useful point of departure for modeling, and is frequently used as a basis for computing the efficiency of a set of trades.

Here are some of the more common mechanisms:

- When there are two agents, trade is accomplished by bargaining. Ultimatum situations arise when one side can (credibly) make a “take it or leave it” offer. When there is the possibility of counter-offers, we have sequential bargaining.
- When there is one seller and many potential buyers, we often encounter an auction.
- When there have many buyers and many sellers convening at a single time, we have a call market. (On securities exchanges organized as floor markets, the convening is often coordinated by having an exchange representative “call” the security.)
- In continuous security markets, trades can potentially occur at any time. Continuous security markets are frequently categorized as dealership (quote driven) or double auction (order driven) markets.

Most real-world security markets are hybrids. Continuous markets dominate, but there are a fair number of periodic call markets as well. Furthermore, although security markets viewed from afar usually involve many agents, some interactions viewed closely resemble bargaining situations. As a result, economic perspectives from bargaining and auction literatures (which predate financial market microstructure) are often useful.

■ Multiple characterizations of prices

There is rarely a “single” price in microstructure analyses. Prices are sometimes actual trade prices; sometimes they are bids or offers (proposed prices). Typically, the price depends on agent’s identity, whether she’s buying or selling, the market venue, etc.

■ “Liquidity”

Liquidity is a summary quality or attribute of a security or asset market. There are no formal definitions, except those that are very context-specific. The underlying qualities are sufficiently widely accepted to make the term useful in practical and academic discourse.

Here are some of the component attributes of “liquidity”. Liquidity is like the static concept of elasticity (“How much will an order (incremental demand or supply) move the price?”. Liquidity, however, also has time and cost dimensions. (How much will it cost me to trade? How long will it take me to trade?)

“In a liquid market, you can trade a large amount without moving the price very much. Any price perturbations caused by the trade quickly die out.”

A common definition of liquidity is: “Depth, breadth, resilience”

- Depth. If we look a little above the “current” market price, there is a large incremental quantity available for sale. If we look a little below the current price, there is a large incremental quantity that is sought (by a buyer or buyers).
- Breadth. The market has many participants.
- Resilience. Price impacts caused by the trading are small and quickly die out.

Where does liquidity come from? Here is one thought-provoking viewpoint:

> Liquidity is created through a give and take process in which multiple counterparties selectively reveal information in exchange for information ultimately leading to a trade.

The excerpt is taken from the offering materials for the Icor Brokerage (an electronic swaps platform).

One sometimes encounters the term "liquidity externality. This is a network externality. As more agents participate in a market, the market clearing price will become more stable (less noisy). This benefits the individual participants.

■ Econometric issues

Microstructure time series are distinctive. Market data are typically:

- Discrete events realized in continuous time (“point processes”)
- Well-ordered.

  Most macroeconomic data are time-aggregated. This gives rise to simultaneity, and findings that must be qualified accordingly. For example, quarterly labor income and quarterly consumption expenditure are postively correlated. We can estimate a linear least-squares relation between the two, but we won't be able to say much about causality. Market events, however, are typically time-stamped to the

second. This supports stronger conclusions about causality (at least in the *post hoc ergo propter hoc sense*).

- Driven by unspecified (and unobserved) information processes with time-varying characteristics
- Detailed (e.g., the state of a single limit order book is specified by numbers of orders and quantities at all price points)

Microstructure data samples are typically:

- Large: there are many observations (10,000 would not be unusual)
- Small: the covered intervals of calendar time are usually short, on the order of days or months.
- New: we don't have much long-term historical data.
- Old: market institutions are changing so rapidly that even samples a few years previous may be seriously out of date.

The range of econometric techniques applied to market data is extremely broad. Always remember that economic significance is very different from (and much more difficult to achieve) than statistical significance.

■ The questions

Here is a partial list of "big questions" in market microstructure:

- What are optimal trading strategies for typical trading problems?
- Exactly how is information impounded in prices?
- How do we enhance the information aggregation process?
- How do we avoid market failures?
- What sort of trading arrangements maximize efficiency?
- What is the trade-off between "fairness" and efficiency?
- How is market structure related to the valuation of securities?
- What can market/trading data tell us about the informational environment of the firm?
- What can market/trading data tell us about long-term risk?

Although they might have been worded differently, most of these problems have been outstanding as long as the field has been in existence.

■ Readings

- Background readings in financial economics include Ingersoll (1987(, Huang and Litzenberger (1998), Duffie (2001)
- For econometric background, see Greene (2002).
- O'Hara (1995) is the standard reference for the economics of market microstructure. Surveys include: Hasbrouck (1996); Madhavan (2000); Biais, Glosten, and Spatt (2002); Harris (2003).
- The paper's discussion of time series analysis emphasizes concepts rather than proofs. Hamilton (1994) is a deeper, though still accessible, treatment. Gourieroux and Jasiak (2001) and Tsay (2002) also provide useful developments.
- The institutional details about trading arrangements are rapidly changing. Some places to start include the appendix to this document: Hasbrouck, Sofianos, and Sosebee (1993) (for the NYSE); Smith, Selway, and McCormick (1998) (for Nasdaq); Euronext (2003) (for Euronext).

■ *Mathematica* initializations

If you are reading the pdf or printed version of this document,the code associated with the *Mathematica* sections (like the one immediately following) will not be visible.

Comments and initializations ***Mathematica***

Chapter 2. The long-term dynamics of security prices

It is often useful in economic analysis to separate, conceptually at least, long-run and short-run effects. When we apply this perspective to security markets, we view long-run price dynamics as driven by "fundamental" considerations of security value: expected cash flows, long-term risk and required returns. The effects of liquidity and trading mechanism are short-run. In a sense, then, microstructure phenomena can be viewed as an "overlay" on a long-term valuation process.

This is, of course, a simplification. In most economic analysis, and certainly here, "long-term" and "short-term" are linked. The long-term characteristics of a security will determine in part who holds it, who trades it, and how it will be traded. Conversely, the features of the trading environment may affect the long-term return on the security. In extreme circumstances, the limitations of the trading mechanism may preclude a security's existence.

The overlay view of market mechanisms is nevertheless a useful place to start. The first question is then, what are the long-term dynamics of security prices? Or, in a world with perfectly frictionless (costless and infinitely liquid) markets, how would we expect security prices to behave?

2.a Macroeconomic models of asset prices

The basic result from classical asset pricing theory is that a security price should behave as a martingale. A martingale is a time series with unforecastable increments: we can't predict where it will go. Slightly more formally, a time series $\ldots x_{t-1},\ x_t,\ x_{t+1}$ can be considered a martingale if $E[x_{t+1} \mid x_t, x_{t-1}, \ldots] = x_t$. This implies that the changes (increments) are in expectation zero: $E[x_{t+1} - x_t \mid x_t, x_{t-1}, \ldots] = 0$.

Cochrane (2001), Ch. 1: illustrates this with a simple two-period consumption/investment model. Consider an agent whose utility depends on current and future consumption:

$$U(c_t, c_{t+1}) = u(c_t) + \beta\, u(c_{t+1}) \tag{2.a.1}$$

The agent has consumption endowments e_t and e_{t+1}. There is a risky security with current share price p_t and payoff x_{t+1}. The agent's choice variable is the number of shares purchased, ξ. Negative ξ correspond to short sales. It is assumed that the agent can buy or sell any amount of the asset at price p_t. Given ξ, the levels of consumption are

$$\begin{aligned} c_t &= e_t - p_t\,\xi \\ c_{t+1} &= e_{t+1} + x_{t+1}\,\xi \end{aligned} \tag{2.a.2}$$

The agent maximizes expected utility $E_t\, U(c_t, c_{t+1})$ over ξ subject to these consumption dynamics. The first-order condition is

$$-p_t\, u'(c_t) + E_t[\beta\, u'(c_{t+1})\, x_{t+1}] = 0 \tag{2.a.3}$$

The asset payoff consists of time $t+1$ market value plus dividends:

$$x_{t+1} = p_{t+1} + d_{t+1} \tag{2.a.4}$$

Microstructure analyses are typically short-term, i.e., over horizons sufficiently brief that:

- $d_{t+1} = 0$ (The stock does not go ex dividend during the analysis.)
- $\beta \approx 1$ (There is negligible time preference.)

Then:

$$p_t = E_t\left[\beta \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1}\right] \approx E_t\, m_t\, p_{t+1} \text{ where } m_{t+1} = \frac{u'(c_{t+1})}{u'(c_t)} \tag{2.a.5}$$

Under risk-neutrality, $u'(c)$ is constant, so

$$p_t = E_t\, p_{t+1} \tag{2.a.6}$$

Thus, p_t is a martingale. The expectation here is said to be taken with respect to the natural (actual) probability measure. More generally, if we drop the assumption of risk-neutrality, the martingale property holds with respect to the risk-neutral probability measure.

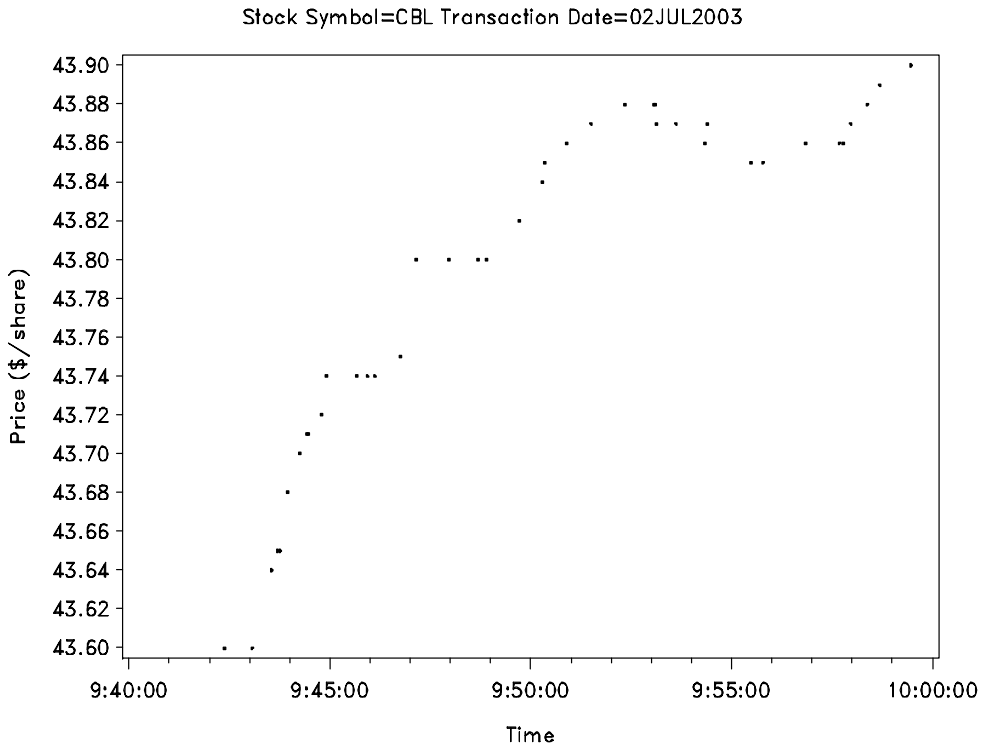
This development follows from the analysis of a single investor's problem. Different investors might have different ms, different probability assessments, and therefore different risk-neutral probabilities. Under more structure (complete markets, absence of arbitrage) there exists one risk-neutral probability measure that is common across all agents (see Cochrane).

In much empirical microstructure work, risk neutrality is (implicitly at least) assumed.

A particularly important variant of martingale is the random walk. For example, suppose that the logarithm of the security price, m_t, follows $m_t = m_{t-1} + u_t$ where $Eu_t = 0$. We might also include a drift term: $m_t = m_{t-1} + \mu + u_t$, where μ represents the expected price change due to a positive expected return.

■ A sample of market prices

Here is a graph of NYSE trades in CBL on July 2, 2003, between 9:30 and 10:00. (Although the NYSE formally opens at 9:30, the first trade did not occur until 9:42. The program that produced this graph is AnalyzeCBL01.sas.)



The stock was selected as one that was, on average, traded reasonably frequently (but not as often as, say, IBM). It is representative of many NYSE-listed stocks. The date, however, was chosen as one on which the volume was (for CBL) unusually high. High volume is often associated with the arrival or announcement of significant new information relevant for the stock's value, and often this information is associated with a large price change as well.

How might we characterize this sample? Assume that the data were generated by a log random walk: $p_t = p_{t-1} + \mu + u_t$ where the u_t are i.i.d. with $Eu_t = 0$ and $Eu_t^2 = \sigma_u^2$. Supposing that we have a sample $\{p_0,\ p_1,\ ...,\ p_T\}$, a natural estimate of μ is $\hat{\mu} = \sum_{t=1}^{T} \Delta p_t / T$ where $\Delta p_t = p_t - p_{t-1}$. A natural estimate of $\mathrm{Var}(u_t) = \sigma_u^2$ is $\hat{\sigma}_u^2 = \sum_{t=1}^{T} (\Delta p_t - \hat{\mu})^2 / T$.

For the CBL data above, there are 40 prices. The estimates are:
$\hat{\mu} = 0.000176;\ \mathrm{SE}(\hat{\mu}) = 0.000047;\ \hat{\sigma}_u = 0.0029$. These numbers are presented for the sake of completeness only. The sample is not a random one and the estimates therefore possess little validity. But sample paths from random walks often appear to exhibit trends and other regularities.

But in samples that *are* random, similar estimates are often used. In their computation and interpretation, these issues typically arise.

- What is t?
- Are the moments we're trying to estimate finite?

- How should we estimate the mean μ?

Each of these concerns requires some explanation.

In most time series analysis, the time subscript t is conventional wall-clock or calendar time. This is customary in dealing with most economic or physical variables, where the mechanism that generates the data is fundamentally cast or anchored in natural time. In securities markets, though, trade occurrences and price changes are often viewed as arising from information that can arrive with intensity that is varying (in wall-clock time). Therefore, "event time", i.e., letting t index trades, is often a sensible alternative to natural time.

Turning to the second issue, recall that the nth order moment of a random variable x is defined as Ex^n. The centered moment of order n is $E(x - Ex)^n$. The variance is therefore the second-order centered moment. A moment may be infinite because as x increases or decreases toward $\pm\infty$ the quantity x^n or $E(x - Ex)^n$ increases faster than the (tail) probability density declines. In general, if an uncentered moment of order n is finite, the sample estimate $\sum_{t=1}^{T} x_t^n / T$, where T is the sample size, is an asymptotically consistent estimate (using a Law of Large Numbers). Hypothesis testing, however, often relies on the asymptotic distribution of the sample estimate. Constructed using a Central Limit Theorem. The essential properties of this distribution require existence of moments of order $2n$.

Classical and generalized moment estimates are used in many settings where the existence of the required moments is taken for granted. In many market microstructure applications, however, some skepticism is warranted. Recent evidence from extreme-value analyses suggests that finite moments for returns exist only up to order 3, and for volume only up to order 1.5. (Gabaix, Gopikrishnan, Plerou, and Stanley (2003)). If this is indeed the case, conventional return variance estimates are consistent, but the distribution of these estimates is not well-defined. For volume (an essential component of many analyses), the variance is infinite, and the quantities that depend on the variance (like the standard error of the mean) are undefined.

Finally, we turn to estimation of the mean. The surprising result here is that in microstructure data, we are usually better off setting the estimate of the unconditional return mean to zero. There are two reasons for this. First, the cost of borrowing or lending within trading sessions is often literally zero. In US equity markets, for example, a trade on day T is settled on day $T + 3$ irrespective of when during day T the trade actually occurred. The second reason is the expected returns are usually small relative to their estimation errors.

To see this, suppose that we have a year's worth of daily data for a typical US stock. Assume an annual return of $\mu_{\text{Annual}} = 0.10$ ("10%") and a volatility of $\sigma_{\text{Annual}} = 0.25$. The implied daily expected return is $\mu_{\text{Day}} = 0.10 / 365 = 0.000274$. The implied daily volatility is $\sigma_{\text{Day}} = 0.25 / \sqrt{365} = 0.0131$. With 365 observations, the standard error of estimate for the sample mean is $\text{SE}(\hat{\mu}_{\text{Day}}) = \sigma_{\text{Day}} / \sqrt{365} = \sigma_{\text{Annual}} / 365 = 0.000685$. This is about two-and-a-half times the true mean.

Let's consider another estimate of μ_{Day}: zero. Clearly this is biased downward, but its standard error of estimate is only 0.000274. At the cost of a little bias, we can greatly reduce the estimation error. The point extends to estimates of centered moments, such as variance, skewness, etc. In most cases, the uncentered

(that is, not “de-meaned”) estimates will have substantially lower measurement error than the unbiased estimates. Are the numbers here realistic? Microstructure data samples are typically shorter than one year, and the problem would actually be worse than indicated.

In a sense, the transition for macro-finance to microstructure can be thought of as a refinement of the interval of observation. In a given annual sample, say, we progress from annual observations to daily, from daily to hourly, etc. This progression clearly increases the number of observations. More numerous observations usually enhance the precision of our estimates. Here, though, the increase in observations is not accompanied by any increase in the calendar span of the sample. So do we gain or not? It depends. Merton (1980) shows that estimates of second moments (variances, covariances) are helped by more frequent sampling. Estimates of mean returns are not.

2.b Martingales in microstructure analyses

When we drop the assumption that the agent can buy or sell any amount ξ of the asset at a single price p_t, the formal argument in support of the martingale property of prices falls apart.

Suppose that the agent can only buy at a dealer’s ask price p_t^a and sell at a dealer’s bid price p_t^b (with, of course, $p_t^a > p_t^b$). The first order condition resulting from the agent’s optimization then becomes $p_t^b \le E_t\, m_{t+1}\, x_{t+1} \le p_t^a$. This establishes bounds, but certainly does not imply that either the bid or the ask follows a martingale.

The martingale continues to possess a prominent role, however. Suppose that we have a random variable X and a sequence of sets of conditioning information $\Phi_1, \Phi_2, \ldots$. For example, suppose that there is a set of variables $\{z_1, z_2, \ldots\}$ that are useful in predicting X, and we let $\Phi_1 = \{z_1\},\ \Phi_2 = \{z_1, z_2\},\ \ldots,\ \Phi_k = \{z_1, z_2, \ldots, z_k\}$. Then the sequence of conditional expectations $E[X \mid \Phi_k]$ for $k = 1, 2, \ldots$ is a martingale.

It is common in microstructure analyses for an agent’s objective function to depend on the terminal payoff of the security. The conditional expectation of this payoff will be important in formulating strategy. Over time, the set of conditioning information expands (or, at least, does not contract), and therefore this conditional expectation evolves as a martingale. When the conditioning information is “all public information”, this is sometimes called (with a nod to the asset pricing literature) “efficient” price of the security.

One of the basic goals of microstructure analysis is a detailed and realistic view of how informational efficiency arises, that is, the process by which new information comes to be impounded or reflected in prices. In microstructure analyses, observed prices are usually not martingales. By imposing economic or statistical structure, though, it is often possible to identify a martingale component of the prices. This allows the information attribution to proceed.

Chapter 3. A dealer market with fixed transaction costs: the Roll model

The model described in this section is due to Roll (1984). The Roll construct is the basic black dress of microstructure models: it's appropriate in many different situations, and it's easy to accessorize. Furthermore, the model offers an excellent pedagogical framework. By virtue of the fact that it that maps cleanly into a statistical model, it is useful for motivating and illustrating the basics of time series analysis.

3.a Model structure

The evolution of the (log) efficient price is given by:

$$m_t = m_{t-1} + u_t \tag{3.a.1}$$

The market has the following features:

- All trading is conducted through specialized intermediaries ("dealers"). A dealer posts bid and ask (offer) prices, b_t and a_t. If a customer wants to buy (any quantity), he must pay the dealer's ask price. If a customer wants to sell, she receives the dealer's bid price.
- Dealers are competitive and bear a per-trade cost c.

Then the bid and ask are given by:

$$\begin{aligned} b_t &= m_t - c \\ a_t &= m_t + c \end{aligned} \tag{3.a.2}$$

That is, the dealers set their quotes to recover their costs. At time t, we observe a transaction price p_t:

The actual trade price is:

$$p_t = m_t + c\, q_t \tag{3.a.3}$$

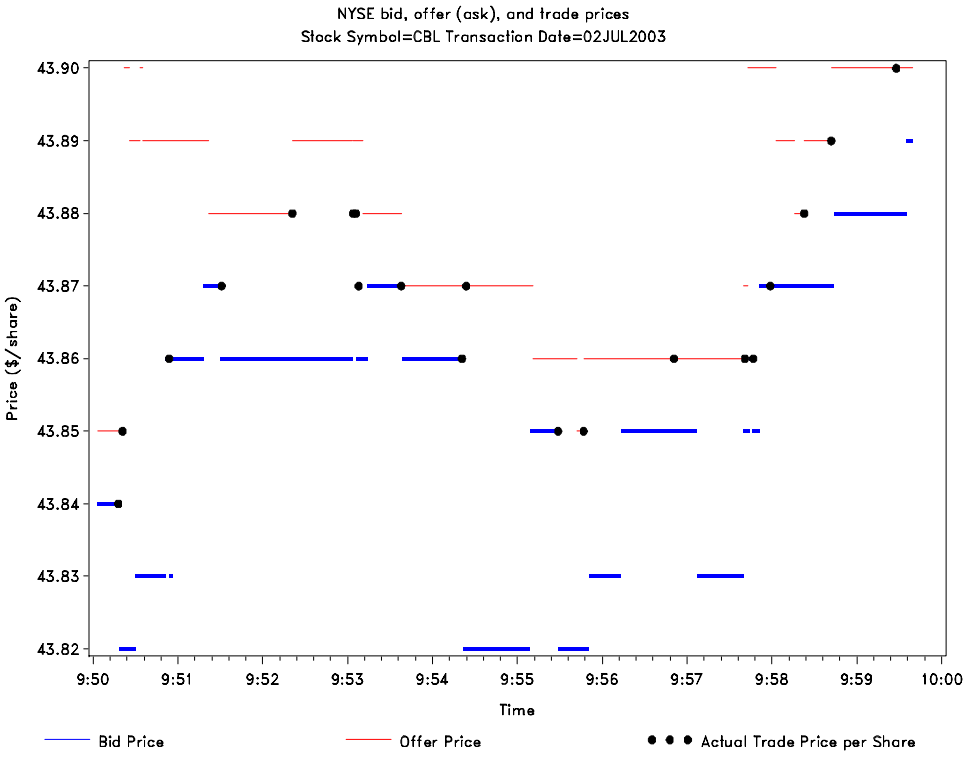
where q_t is the trade direction indicator (+1 if the customer is buying at the ask; –1 if the customer is selling at the bid).

Suppose for the moment that m_t follows a homoscedastic random walk. What are some reasonable assumptions about q_t?

- Buys and sells are equally likely.
- q_t are serially independent

- q_t are independent of the u_t.

Before considering estimation and inference, it might be helpful to look at some actual bid/ask/trade data. Here is a record of trades and quotes for CBL for a portion of July 2, 2003:



This graph and the statistics discussed in this section are produced by the SAS program AnalyzeCBL01.

The most obvious feature of the data is that the spread between the bid and ask, assumed to be a constant $2c$ in the Roll model is actually varying, approximately between one and five cents in this sample. Furthermore, trades at the bid tend to cause a downward revision in the bid, and trades at the ask cause an upward revision in the ask. This calls into question the assumed independence of q_t and u_t. Finally, although it is not obvious in this particular sample, the q_t tend to be positively autocorrelated: buys tend to follow buys and sells tend to follow sells.

Nevertheless, the Roll model often achieves a characterization of price dynamics that is adequate for many purposes.

3.b Inference

The Roll model has two parameters, c and σ_u^2. These are most conveniently estimated from the variance and first-order autocovariance of the price changes.

Inference in this model is based on the price changes Δp_t:

$$\Delta p_t = p_t - p_{t-1} = -c\, q_{t-1} + c\, q_t + u_t \tag{3.b.4}$$

To obtain $\mathrm{Var}(\Delta p_t) \equiv \gamma_0$, note that

$$\Delta p_t^2 = q_{t-1}^2\, c^2 + q_t^2\, c^2 - 2\, q_{t-1}\, q_t\, c^2 - 2\, q_{t-1}\, u_t\, c + 2\, q_t\, u_t\, c + u_t^2 \tag{3.b.5}$$

In expectation, all of the cross-products vanish except for those involving q_t^2, q_{t-1}^2 and u_t^2. So:

$$\gamma_0 = 2\, c^2 + \sigma_u^2 \tag{3.b.6}$$

To obtain $\mathrm{Cov}(\Delta p_t, \Delta p_{t-1}) = \gamma_1$, we examine:

$$\begin{aligned}\Delta p_t \Delta p_{t-1} = {} & -q_{t-1}^2\, c^2 + q_{t-2}\, q_{t-1}\, c^2 - q_{t-2}\, q_t\, c^2 + \\ & q_{t-1}\, q_t\, c^2 - q_{t-1}\, u_{t-1}\, c + q_t\, u_{t-1}\, c - q_{t-2}\, u_t\, c + q_{t-1}\, u_t\, c + u_{t-1}\, u_t\end{aligned} \tag{3.b.7}$$

In expectation, all of the cross-products vanish except for the first, so:

$$\gamma_1 = -c^2 \tag{3.b.8}$$

It is easily verified that all autocovariances of order two or higher are zero. From the above, it is clear that $c = \sqrt{-\gamma_1}$ and $\sigma_u^2 = \gamma_0 + 2\,\gamma_1$. Faced with a sample of data, it is sensible to estimate γ_0 and γ_1, and apply these transformations to obtain estimates of the model parameters. Harris (1990) reports distributional results.

For CBL on July 2, 2003, there were 821 NYSE trades. The estimated first-order autocovariance of the price changes is $\hat{\gamma}_1 = -0.0000251$. This implies $c = 0.005$ ($/share) and a spread of $2\,c = 0.01$ ($/share).

The Roll model is often used in situations where we don't possess bid and ask data. Here, we do. The (time-weighted) average NYSE spread in the sample is 0.022 ($/share), so the Roll estimate appears to be substantially on the low side. There are several possible explanations for this. One obvious possibility is sampling error. Also, as noted above, some of the assumptions underlying the Roll model are unrealistic. There are also institutional considerations. When there is variation in the spread, agents may wait until the spread is small before trading. In addition, NYSE brokers on the floor will sometimes take the other side of an incoming order at a price better than the opposing quote. In this sample, for example, trade prices are on average 0.0079 ($/share) away from the quote midpoint. This implies an *effective* spread of 0.0158 $/share, which is somewhat closer to the Roll spread estimate.

We can obtain further results on the Roll model. But these results are best developed in a time series analysis framework. This will lay the ground for generalization of the model.

Chapter 4. Moving average and autoregressive representations of price changes

The Roll model described in the last section is a simple structural model, with a clear mapping to parameters (the covariance and autocovariance of price changes) that are easily estimated.

There are many interesting questions, though, that go beyond parameter estimation. For example, we might want to forecast prices beyond the end of our data sample. Alternatively, we might wish to identify the series of m_t (the unobserved efficient prices) underlying our data. Finally, in situations where the structural model is possibly misspecified, we might prefer to make assumptions about the data, rather than about the model.

To answer these questions, we'll begin with the structural model, and then construct a statistical model. Then, we'll pretend that we don't know the structural model, and investigate the properties of the data that might enable us to identify the statistical model. Finally, we'll work from the statistical model back to the structural model. In the process of working from the structural model to the statistical one and thence to the data, and back again, we will illustrate econometric techniques that are very useful in more general situations. Starting from a known structural model helps to clarify matters.

4.a Stationarity and ergodicity

Whereas most statistical analysis is based on observations that are independently distributed, time series observations are typically dependent. When realizations are serially dependent, we effectively have only one observation: a single sample path. To fill in for the independence assumption when invoking a law of large numbers or central limit theorem, we often rely on properties of stationarity and ergodicity.

A time series $\{x_t\}$ where the mean and covariances don't depend on t ($Ex_t = \mu$, $\text{Cov}(x_t, x_{t-k}) = \text{Cov}(x_s, x_{s-k})$ for all s, t and k) with this property is said to be *covariance stationary*. If all joint density functions of the form $f(x_t)$, $f(x_t, x_{t+1})$, ..., $f(x_t, x_{t+1}, x_{t+2})$, ... don't depend on t, then the series is (strictly) stationary. Strict stationarity, of course, implies covariance stationarity.

The price changes implied by the Roll model, Δp_t, are covariance stationary: $E\Delta p_t = 0$ and $\text{Cov}(\Delta p_t, \Delta p_{t-k}) = \gamma_k$. The price levels, p_t, are not covariance stationary. Among other things, $\text{Var}(p_t)$ increases with t. Covariance stationarity for the Δp_t would also fail if we replaced the homoscedasticity assumption $Eu_t^2 = \sigma_u^2$ with something like $Eu_t^2 = 5 + \text{Cos}(t)$, or similar time-dependent feature. $\text{Cos}(t)$ here is a *deterministic* component of the series. Such components can also arise from time trends (linear or otherwise). When the deterministic component is periodic (like $\text{Cos}(t)$), it is said to be *seasonal (*a term that says much about the frequency of observation traditionally assumed for time series data). Market data typically exhibit intra-day seasonalities (*sic*): trading volumes and return volatilities tend to be elevated at the start and end of trading sessions.

A time series is *ergodic* if its local stochastic behavior is (possibly in the limit) independent of the starting point, i.e. initial conditions. Essentially, the process eventually "forgets" where it started. The price *level* in the Roll model is not ergodic: the randomness in the level is cumulative over time. But the price changes are ergodic: Δp_t is independent of Δp_{t-k} for $k \geq 2$. Non-ergodicity could be introduced by positing $m_t = m_{t-1} + u_t + z$, where z is a zero-mean random variable drawn at time zero.

The economic models discussed in later chapters (particularly the asymmetric information models) are often placed in settings where there is a single random draw of the security's terminal payoff and the price converges toward this value. The price changes in these models are not ergodic because everything is conditional on the value draw. Nor are they covariance stationary (due to the convergence). Empirical analyses of these models use various approaches. We might assume that reality consists of a string of these models placed end-to-end (for example, a sequence of "trading days"). In this case, we view the sample as an *ensemble*, a collection of independent sample path realizations. Alternatively, we might view the models as stylized descriptions of effects that in reality overlap in some fashion that yields time invariance. For example, in each time period, we might have a new draw of some component of firm value.

4.b Moving average models

A *white noise* process is a time series $\{\epsilon_t\}$ where $E\epsilon_t = 0$, $\text{Var}(\epsilon_t) = \sigma_\epsilon^2$ and $\text{Cov}(\epsilon_t, \epsilon_s) = 0$ for $s \neq t$. This is obviously covariance stationary. In many economic settings, it is convenient and plausible to assume that $\{\epsilon_t\}$ are strictly stationary and even normally distributed, but these assumptions will be avoided here.

White noise processes are convenient building blocks for constructing dependent time series. One such construction is the *moving average* ("MA") model. The moving average model of order one (the "MA(1) process") is:

$$x_t = \epsilon_t + \theta\epsilon_{t-1} \tag{4.b.1}$$

The white noise series in a time series model is variously termed the disturbance, error or innovation series. From a statistical viewpoint, they all amount to the same thing. The economic interpretations and connotations, however, vary. When randomness is being added to a non-stochastic dynamic structural model, the term "disturbance" suggests a shock to which the system subsequently adjusts. When estimation is the main concern, "error" conveys a sense of discrepancy between the observed value and the model

prediction. "Innovation" is the term that is most loaded with economic connotations. The innovation is what the econometrician learns about the process at time t (beyond what's known from prior observations). Moving forward in time, it is the update to the econometrician's information set. In multivariate models, when x_t comprises a particularly varied, comprehensive and economically meaningful collection of variables, the innovation series is often held to proxy the update to the *agents'* common information set as well.

The Δp_t in the Roll model have the property that the autocovariances are zero beyond lag one. The MA(1) model also has this property. For this process, $\gamma_0 = (1 + \theta^2)\,\sigma_\epsilon^2$, $\gamma_1 = \theta\sigma_\epsilon^2$ and $\gamma_k = 0$ for $k > 1$.

More generally, the moving average model of order K ("MA(K)") is

$$x_t = \epsilon_t + \theta_1\,\epsilon_{t-1} + \ldots + \theta_K\,\epsilon_{t-K} \tag{4.b.2}$$

the MA(K) process is covariance stationary and has the property that $\gamma_j = 0$ for $j > K$. If we let $K = \infty$, we arrive at the infinite-order moving average process.

Now comes a point of some subtlety. If we believe that the data are generated by the Roll model (a *structural* model), can we assert that a corresponding moving average model (a *statistical* model) exists? By playing around with the θ and σ_ϵ^2 parameters in the MA(1) model, we can obviously match the variance and first-order autocovariance of the structural Δp_t process. But this is not quite the same thing as claiming that the full joint distribution of the Δp_t realizations generated by the structural model could also be generated by an MA(1) model. Moreover, there's at least one good reason for suspecting this shouldn't be possible. The structural model has two sources of randomness, u_t (the efficient price innovations) and q_t (the trade direction indicators). The MA(1) model has only one source of randomness, ϵ_t.

Why do we care? Why can't we just limit our analysis to the structural model and be done with it? The answer to these questions lies in the fact that the econometrician does not observe the u_t and q_t, nor, therefore does the econometrician know the efficient price. The moving average representation is a useful tool for constructing an estimate of the efficient price, as well as for forecasting.

Fortunately, an MA(1) representation does exist. The basic result here is the Wold (not Wald) Theorem:

Any zero-mean covariance stationary process $\{x_t\}$ can be represented in the form

$$x_t = \sum_{j=0}^{\infty} \theta_j\,\epsilon_{t-j} + \kappa_t \tag{4.b.3}$$

where $\{x_t\}$ is a zero-mean white noise process, $\theta_0 = 1$ (a normalization), and $\sum_{j=0}^{\infty} \theta_j < \infty$.

κ_t is a linearly-deterministic process, which in this context means that it can be predicted arbitrarily well by a linear projection (possibly of infinite order) on past observations of x_t.

For proofs, see Hamilton (1984) or Sargent (1979).

For a purely stochastic series, $\kappa_t = 0$ and we are left with a moving average representation.

A related result due to Ansley, Spivey, and Wrobleski (1977) establishes that if a covariance stationary process has zero autocovariances at all orders higher than *K*, then it possesses a moving average representation of order *K*. This allows us to assert that an MA(1) representation exists for the Roll model.

Empirical market microstructure analyses often push the Wold Theorem very hard. The structural models are often stylized and underidentified (we can't estimate all the parameters). The data are frequently non-Normal (like the trade indicator variable in the Roll model). Covariance stationarity of the observations (possibly after a transformation) is often a tenable working assumption. For many purposes, as we'll see, it is enough. (Chapter 11 presents an illustration of the Wold Theorem applied to discretely-valued data.)

4.c Autoregressive models

Although the moving average model has many convenient properties, it is difficult in that the driving disturbances are generally unobserved. Moreover, direct estimation of the moving average model is difficult unless we're willing to make distributional assumptions on the errors. Most of the time, it's more convenient to work with an alternative representation of the model -- the autoregressive form.

To develop this, note that we can rearrange $\Delta p_t = \epsilon_t + \theta\epsilon_{t-1}$ as

$$\epsilon_t = \Delta p_t - \theta \epsilon_{t-1} \tag{4.c.4}$$

This gives us a backward recursion for ϵ_t: $\epsilon_{t-1} = \Delta p_{t-1} - \theta\epsilon_{t-2}$, $\epsilon_{t-2} = \Delta p_{t-2} - \theta\epsilon_{t-3}$, and so forth. Using this backward recursion in $\Delta p_t = \epsilon_t + \theta\epsilon_{t-1}$ gives

$$\begin{aligned} \Delta p_t &= \theta(\Delta p_{t-1} - \theta(\Delta p_{t-2} - \theta(\Delta p_{t-3} - \theta\epsilon_{t-4}))) + \epsilon_t \\ &= -\epsilon_{t-4}\theta^4 + \Delta p_{t-3}\theta^3 - \Delta p_{t-2}\theta^2 + \Delta p_{t-1}\theta + \epsilon_t \end{aligned} \tag{4.c.5}$$

If $|\theta| < 1$, then in the limit, the coefficient of the lagged ϵ_t converges to zero. Then:

$$\Delta p_t = \theta\Delta p_{t-1} - \theta^2 \Delta p_{t-2} + \theta^3\Delta p_{t-3} + \ldots + \epsilon_t \tag{4.c.6}$$

This is the autogressive form: Δp_t is expressed as a convergent linear function of its own lagged values and the current disturbance.

4.d The lag operator and representations

To go move between various representations, it is convenient to use the lag operator, L (sometimes written as the backshift operator, B). It works in a straightforward fashion, and can generate leads as well as lags:

$$Lx_t = x_{t-1};\ L^2 x_t = x_{t-2};\ L^{-3} x_t = x_{t+3},\ \text{etc.} \tag{4.d.7}$$

Using the lag operator, the moving average representation for Δp_t is:

$$\Delta p_t = \epsilon_t + \theta L\epsilon_t = (1 + \theta L)\epsilon_t \tag{4.d.8}$$

The autoregressive representation is:

$$\Delta p_t = \epsilon_t + \theta L \Delta p_t - \theta^2 L^2 \Delta p_t + \theta^3 L^3 \Delta p_t + ... = \epsilon_t + (\theta L - \theta^2 L^2 + \theta^3 L^3 + ...) \Delta p_t \quad (4.d.9)$$

In the previous section we derived this by recursive substitution. But there is an alternative construction that's particularly useful when the model is complicated. Starting from the moving average representation, $\Delta p_t = (1 + \theta L) \epsilon_t$, we may write

$$(1 + \theta L)^{-1} \Delta p_t = \epsilon_t \quad (4.d.10)$$

where we've essentially treated the lag operator term as an algebraic quantity. If L were a variable and $|\theta| < 1$, we could construct a series expansion of the left hand side. This expansion, through the third order is:

$$[1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + O(L^4)] \Delta p_t = \epsilon_t \quad (4.d.11)$$

where $O(L^4)$ represents the higher order terms. This can be rearranged to get the autoregressive representation.

4.e Forecasting

A martingale has differences that are uncorrelated with the history of the series, and therefore can't be forecast. The unobservable efficient price in the Roll model is a martingale, but the observed trade price is not. If we know θ and have a full (infinite) price history up the time t, $\{p_t, p_{t-1}, p_{t-2}, ...\}$, then using the autoregressive representation we can recover the innovation series $\{\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}\}$. Then:

$$E[\Delta p_{t+1} \mid p_t, p_{t-1}, ...] = E[\epsilon_{t+1} + \theta \epsilon_t \mid p_t, p_{t-1}, ...] = \theta \epsilon_t \quad (4.e.12)$$

Therefore, the forecast of next period's price is:

$$p_t^* \equiv E[p_{t+1} \mid p_t, p_{t-1}, ...] = p_t + \theta \epsilon_t \quad (4.e.13)$$

How does p_t^* evolve?

$$p_t^* - p_{t-1}^* = p_t + \theta \epsilon_t - (p_{t-1} + \theta \epsilon_{t-1}) = (\epsilon_t + \theta \epsilon_{t-1}) + \theta \epsilon_t - \theta \epsilon_{t-1} = (1 + \theta) \epsilon_t \quad (4.e.14)$$

The increment to the conditional expectation is a scaled version of the innovation in the process. This is not surprising. Recall that martingales often arise as a sequence of conditional expectations. Since the ϵ_t are serially uncorrelated, p_t^* is a martingale.

Now for a more difficult question. Is it true that $p_t^* = m_t$? That is, have we identified the implicit efficient price?

If $p_t^* = m_t$, then $p_t = p_t^* + c q_t$ and $\Delta p_t = \Delta p_t^* + c \Delta q_t$. But this implies

$$\epsilon_t + \theta \epsilon_{t-1} = (1 + \theta) \epsilon_t + c \Delta q_t \Leftrightarrow -\theta(\epsilon_t - \epsilon_{t-1}) = c \Delta q_t. \quad (4.e.15)$$

In other words, all of the randomness in the model is attributable to the q_t. But this is structurally incorrect: we know that changes in the efficient price, u_t, also contribute to the ϵ_t.

Thus the random-walk property assumed for m_t *does not* suffice to identify it from the observed data. We will see later that there are an infinite number of candidates for m_t that are compatible with the data.

4.f Problems

These problems investigate modifications to the Roll model.

Problem 4.1 Autocorrelation in trades

The Roll model assumes that trade directions are serially uncorrelated: $\text{Corr}(q_t, q_s) = 0$ for $t \neq s$. In practice, one often finds positive autocorrelation (buys tend to follow buys; sells tend to follow sells). See Hasbrouck and Ho (1987) and Choi, Salandro and Shastri (1988).

Suppose that $\text{Corr}(q_t, q_{t-1}) = \rho > 0$ and $\text{Corr}(q_t, q_{t-k}) = 0$ for $k > 1$. Suppose that ρ is known. What are the autocovariances of the Δp_t process? What is the moving average structure? What is the estimate of c?

Problem 4.2 Trade directions correlated with changes in the efficient price.

In the basic Roll model, $\text{Corr}(q_t, u_t) = 0$. Now suppose that $\text{Corr}(q_t, u_t) = \rho$, where ρ is known, $0 < \rho < 1$. The idea here is that a buy order is associated with an increase in the security value, a connection that will be developed in the models of asymmetric information. Suppose that ρ is known. What are the autocovariances of the Δp_t process? What is the moving average structure? What is the estimate of c?

Chapter 5. Sequential trade models of asymmetric information

5.a Overview

Much current work in market microstructure concentrates on the role that trading and markets play in aggregating information. That is, the essential outputs of the trading process are signals (most importantly the trade price) that summarize diverse private information of market participants.

This role of markets is emphasized in Grossman (1976) and Grossman and Stiglitz (1980). The title of the latter piece, "On the *impossibility* of informationally efficient markets" (italics mine) is not intended as an ironclad universal law, but rather as an invitation for us to reflect on the economic forces and mechanisms that facilitate or discourage informational efficiency. The asymmetric information models in microstructure are very much in this spirit, and are often important for their negative predictions as well as their positive ones.

The general features of the microstructure asymmetric information models might be described as follows.

- They are generally dominated by common value considerations. The primary benefit derived from ownership of the security is the resale value or terminal liquidating dividend that is the same for all holders. But in order for trade to exist, we also need private value components, e.g., diversification or risk exposure needs that are idiosyncratic to each agent. The private values are often modeled in an ad hoc fashion. Sometimes we simply assert the existence of unspecified private values that generate the assumed behavior.
- Generally, public information initially consists of common knowledge concerning the probability structure of the economy, in particular the unconditional distribution of terminal security value and the distribution of types of agents. As trading unfolds, the most important updates to the public information set are market data, such as bids, asks, and the prices and volumes of trades. Many of the models make no provision for the arrival of nontrade public information (e.g., "news announcements") during trading.
- Private information may consist of a signal about terminal security value, or more commonly, perfect knowledge of the terminal security value.

When all agents are ex ante identical, they are said to be symmetric. This does not rule out private values or private information. It simply means that all individual-specific variables (e.g., the coefficient of risk aversion, a value signal) are identically distributed across all participants. In an asymmetric information model, some subset of the agents has superior private information.

The majority of the asymmetric information models in microstructure examine market dynamics subject to a single source of uncertainty, i.e., a single information event. At the end of trading, the security payoff

(terminal value) is realized and known.

Thus, the trading process is an adjustment from one well-defined information set to another. From a statistical perspective, the dynamics of this adjustment are not stationary. These are not models of ongoing trading, although they can be "stacked" one after another to provide a semblance of ongoing trading.

Theoretical market microstructure has two main sorts of asymmetric information models.

- In the sequential trade models, randomly-selected traders arrive at the market singly, sequentially, and independently. This line of inquiry begins with Glosten and Milgrom (1985).
- The other class of models usually features a single informed agent who can trade at multiple times. Following O'Hara (1995), we'll describe these as strategic trader models. When an individual trader only participates in the market once (as in the sequential trade models), there is no need for her to take into account the effect her actions might have on subsequent decisions of others. A trader who revisits the market, however, must make such calculations, and they involve considerations of strategy. This second class of models is also sometimes described as "continuous auction," but the continuity of the market is not really an essential feature. This line of thought begins with Kyle (1985). (Note: "Albert S." is pronounced "Pete".)

The essential feature of both models is that a trade reveals something about the agent's private information. A "buy" from the dealer might result from a trader who has private positive information, but it won't originate from a trader who has private negative information. Rational, competitive market makers will set their bid and ask quotes accordingly. All else equal, more extreme information asymmetries lead to wider quotes. Trades will also engender a "permanent" impact on subsequent prices. The spread and trade-impact effects are the principal empirical implications of these models.

We begin with the sequential trade models.

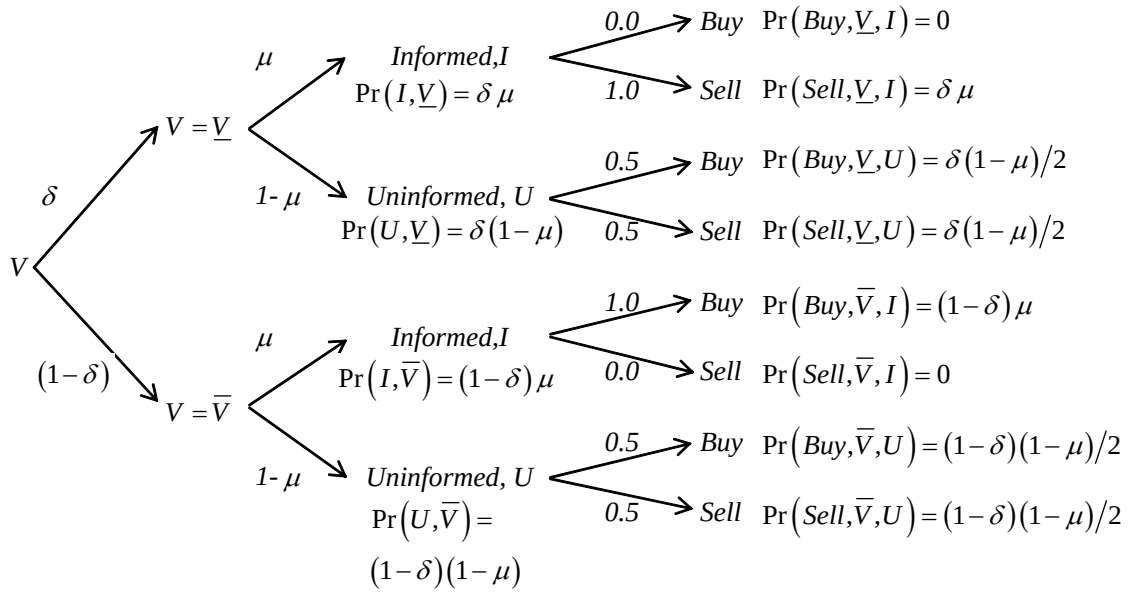
5.b A simple sequential trade model

The essential sequential trade model is a simple construct. The model presented here is a special case of Glosten and Milgrom (1985). It is also contained in many other analyses.

The terminal security value is V, which at the end of the day will be either high or low, $\overline{V}$ or $\underline{V}$. The probability of a high outcome is $P_{\overline{V}} = \delta$. The trading population consists of informed and uninformed traders. Informed traders ("insiders") know the realization of V. The proportion of informed traders in the population is μ.

A dealer posts bid and ask quotes, B and A. A trader is drawn at random from the population. If the trader is informed, she buys if $V = \overline{V}$ and sells if $V = \underline{V}$. If the trader is uninformed, he buys or sells randomly and with equal probability.

The event tree for the first trade looks like this:



In the probability notation, "$\overline{V}$ " is shorthand for the event that $V = \overline{V}$, etc. Note that in this model there is always a trade. (This is not always the case for these models.)

Mathematica

The unconditional buy and sell probabilities are:

$$\begin{aligned} \Pr(Buy) &= \tfrac{1}{2}\,(-2\,\delta\,\mu + \mu + 1) \\ \Pr(Sell) &= \left(\delta - \tfrac{1}{2}\right)\mu + \tfrac{1}{2} \end{aligned} \qquad (5.b.1)$$

In the case where $\delta = \frac{1}{2}$ (equal probabilities of good and bad outcomes), the buy and sell probabilities are also equal.

The unconditional expectation of terminal value is:

$$EV = \overline{V}\,(1-\delta) + \underline{V}\,\delta \qquad (5.b.2)$$

The various conditional expectations are:

$$\begin{aligned} &E[V \mid U, \text{Buy}] = EV \\ &E[V \mid U, \text{Sell}] = EV \\ &E[V \mid I, \text{Buy}] = \overline{V} \\ &E[V \mid I, \text{Sell}] = \underline{V} \end{aligned} \tag{5.b.3}$$

Now consider the dealer's situation. The demands of the uninformed traders are inelastic. So if the dealer is a monopolist, expected profits are maximized by setting the bid infinitely low and the ask infinitely high. Obviously, at these prices, only the uninformed trade.

In practice, the dealer's market power is constrained by competition and regulation. Competition arises from other dealers, but also and more generally from anyone who is setting a visible quote, such as a public customer using a limit order. In some venues, regulation limits the dealers' power. For example, NASD's Rules of Fair Practice (Article III, Section IV) generally prohibit markups (sale price over purchase price) in excess of 5%.

To proceed, we'll assume that dealers are competitive, driving all expected profits to zero. Furthermore, for the usual reasons, the dealer can't cross-subsidize "buys" with "sells" or vice versa. (If he were making a profit on the "sells", for example, another dealer would undercut his ask.) It thus suffices to consider buys and sells separately.

We'll look at customer buys (trades at the dealer's ask price). The dealer's realized profit on the trade is $\pi = A - V$, or in expectation, conditional on the customer's purchase,

$$E[\pi \mid \text{Buy}] = A - E[V \mid \text{Buy}] \tag{5.b.4}$$

Under the zero-expected profit condition, a customer buy at the ask price occasions no ex post regret. The revenue received by the dealer (A) is equal to the value of the security surrendered.

Continuing, we may write the dealer's expected profit as:

$$E[\pi \mid Buy] = A - (E[V \mid U, Buy]\, P(U \mid Buy) + E[V \mid I, Buy] \Pr(I \mid Buy)) \tag{5.b.5}$$

Setting this to zero it establishes the ask price:

$$A = E[V \mid U, Buy]\, P(U \mid Buy) + E[V \mid I, Buy] \Pr(I \mid Buy) \tag{5.b.6}$$

Alternatively, it can be rearranged as:

$$(A - E[V \mid U, Buy])\, P(U \mid Buy) + (A - E[V \mid I, Buy]) \Pr(I \mid Buy) = 0 \tag{5.b.7}$$

The first term on the l.h.s. is the expected profits from uninformed buyers; the second term is the expected losses to informed buyers. Essentially, the dealer's losses to informed traders are passed on to uninformed traders.

If the uninformed traders lose on average, why do they play? Are they stupid? It can't be ruled out, but there are also considerations outside of the stylized model that are consistent with rational uninformed trading.

There may be gains to trade from risk-sharing and long-run returns of security ownership (see O'Hara (2003)).

Now to complete the calculation, $E[V \mid U, Buy] = EV$ where $EV = \delta \overline{V} + (1 - \delta)\, \underline{V}$, the unconditional expectation. The conditional probability of an uninformed buyer is

$$\Pr(U|Buy) = \frac{1 - \mu}{-2\,\delta\,\mu + \mu + 1} \tag{5.b.8}$$

For a purchase originating from an informed trader, $E[V \mid I, Buy] = \overline{V}$. The probability of this event is

$$\Pr(I|Buy) = -\frac{2\,(\delta - 1)\,\mu}{-2\,\delta\,\mu + \mu + 1} \tag{5.b.9}$$

Therefore the ask price is $A =$

$$A = \frac{\underline{V}\,\delta\,(\mu - 1) + \overline{V}\,(\delta - 1)\,(\mu + 1)}{(2\,\delta - 1)\,\mu - 1} \tag{5.b.10}$$

Similarly the bid is:

$$B = \frac{\overline{V}\,(\delta - 1)\,(\mu - 1) + \underline{V}\,\delta\,(\mu + 1)}{(2\,\delta - 1)\,\mu + 1} \tag{5.b.11}$$

The bid-ask spread is

$$A - B = \frac{4\,(\overline{V} - \underline{V})\,(\delta - 1)\,\delta\,\mu}{(1 - 2\,\delta)^2\,\mu^2 - 1} \tag{5.b.12}$$

In the symmetric case of $\delta = \frac{1}{2}$,

$$A - B = (\overline{V} - \underline{V})\,\mu \tag{5.b.13}$$

■ A numerical example

This example is programmed on an Excel spreadsheet (SimpleSequentialTradeModel.xls) available on my web site.

SimpleSequentialTradeModel.xls. (c) Joel Hasbrouck, 2004, All rights reserved.

This spreadsheet describes a sequential trade model with informed and uninformed traders. It is adapted from Glosten and Milgrom, "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," J. Financial Economics, 1985, v. 14, 71-100.

It is now morning. At the end of the day, the stock value, V, will be either:

$\underline{V}$ = **$100** with prob. δ = **0.400**

$\overline{V}$ = **$150** with prob 1 – δ = **0.600**

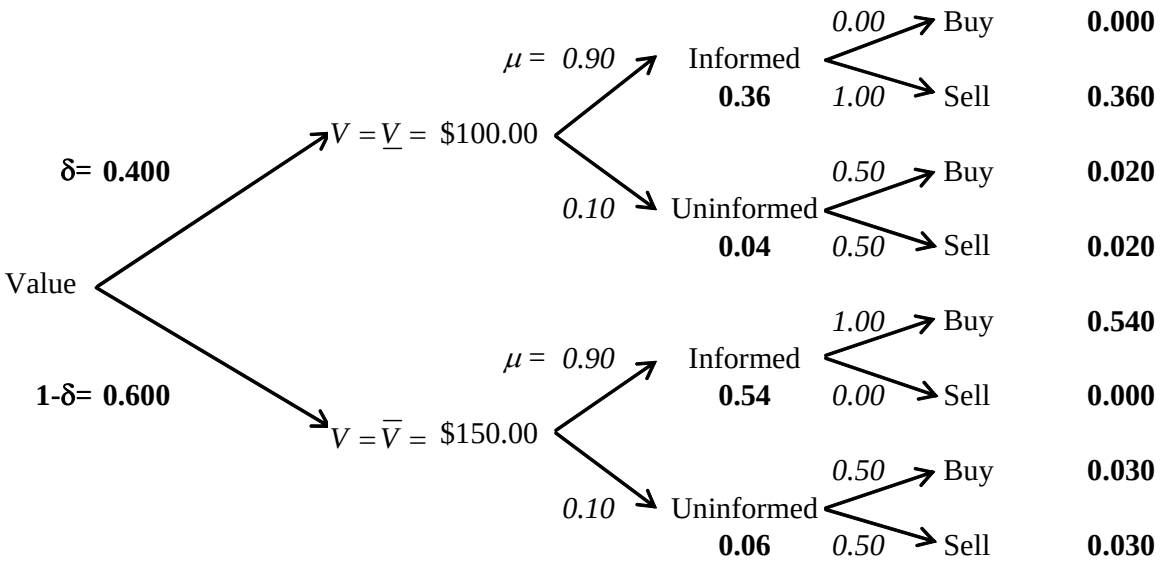
A dealer is trying to set the bid and ask quote, against which incoming market orders will trade. There are two kinds of traders. "Informed" traders know what the final value of V will be. "Uninformed" traders are trading for idiosyncratic reasons having nothing to do with V, and buy or sell with equal probability. The dealer doesn't know the type of the incoming traders, but he does know the probability of an informed trader:

μ = **0.900**

Result summary: Bid = $103.66; Ask = $148.31; Spread = $44.65

The following tree describes the sequence of events.
Numbers in italics are conditional probabilities. For example, the probability of a buy, given that an uninformed trader has arrived is 0.50

Numbers in bold are total probabilities. For example, the probabality of a low value, followed by the arrival of an uninformed trader, followed by a 'buy' is 0.000



■ Market dynamics over time

After the initial trade, the dealer updates his conditional estimate of δ and his quotes. The next trader arrives, etc.

Denote by δ_k the probability of $\underline{V}$ conditional on observing the sign (buy or sell) of the kth trade, i.e., $\delta_{k-1} = \delta$ as defined above. If the kth trade is a buy, then by reference to the event tree:

$$\delta_k(\text{Buy}_k) = \frac{\delta_{k-1} - \mu\,\delta_{k-1}}{-2\,\delta_{k-1}\,\mu + \mu + 1} \tag{5.b.14}$$

A similar expression exists for $\delta_k(\text{Sell}_k)$. The updating expression can be expressed in general form because all probabilities in the event tree except δ are constant over time.

Market dynamics have the following features:

- The trade price series is a martingale.

  Recall from the above analysis that $B_k = E[V \mid \text{Sell}_k]$ and $A_k = E[V \mid \text{Buy}_k]$. Since the trade occurs at one or the other of these prices, the sequence of trade prices $\{p_k\}$ is a sequence of conditional expectations $E[V \mid \Phi_k]$ where Φ_k is the information set consisting of the history (including the kth trade) of the buy/sell directions. A sequence of expectations conditioned on expanding information sets is a martingale.

- The order flow is not symmetric.

  Using q_k to denote the trade direction as we did in the Roll model (+1 for a buy, –1 for a sell), $E[q_k]$ is in general nonzero.

- The orders are serially correlated.

  Although the agents are drawn independently, one subset of the population (the informed traders) always trades in the same direction.

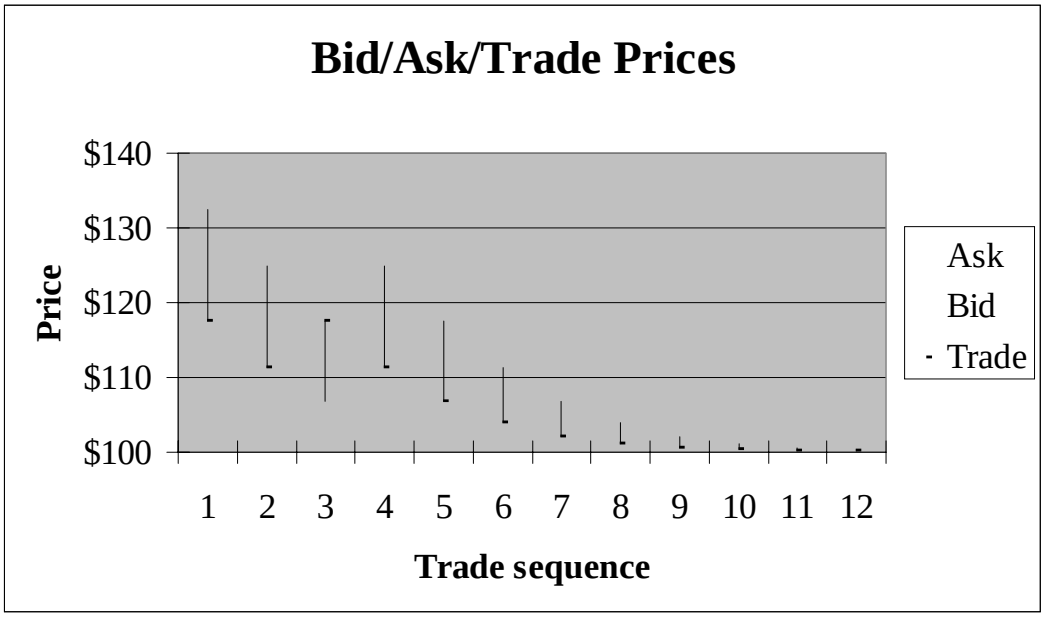
- There is a price impact of trades. For any given pattern of buys and sells through trade k, a buy on the k+1st trade causes a downward revision in the conditional probability of a low outcome, and a consequent increase in the bid and ask.

  The trade price impact is a particularly useful empirical implication of the model. It can be estimated from market data, and is plausibly a useful proxy for information asymmetries. This point will be examined subsequently in greater depth.

- The spread declines over time. Knowing the long-run proportion of buys and sells in the order flow is tantamount to knowing the outcome. With each trade, the dealer can estimate this proportion more precisely, and hence his uncertainty is reduced.

■ **Numerical example, continued**

The second worksheet in the spreadsheet shows what the path of bid, ask and trade prices is for any given sequence of buys or sells. Here is the path when the third trade is a buy, but all the others are sells:



5.c Extensions

The sequential trade framework accommodates a range of interesting generalizations.

■ **Fixed transaction costs**

Suppose that in addition to asymmetric information considerations, the dealer must pay a transaction cost c on each trade (as in the Roll model). The modification is straightforward. The ask and bid now are set to recover c as well as the information costs:

$$\begin{aligned} A &= E(V \mid \text{Buy}) + c \\ B &= E(V \mid \text{Sell}) - c \end{aligned} \tag{5.c.15}$$

The ask quote sequence may still be expressed as a sequence of conditional expectations:

$$A_k = E(V \mid \Phi_k) + c \tag{5.c.16}$$

where Φ_k is the information set that includes the direction of the kth trade. Therefore the ask sequence is a martingale. So too, is the bid sequence. Since trades can occur at either the bid or the ask, however, the sequence of trade prices is not a martingale (due to the $\pm c$ asymmetry in the problem). In terms of the

original Roll model, the effect of asymmetric information is to break the independence between trade direction q_t and the innovation to the efficient price u_t. Developments along these lines are discussed in Glosten and Milgrom (1985), p. 83.

■ Price-sensitive liquidity traders and market failures

The uninformed traders in the basic model are, although not necessarily stupid, rather simple. They aren't price sensitive: their trading demands are inelastic. If they have to buy, for example, they'll pay whatever price is necessary to get the trade done. Such desperate traders do exist, and they are a market-maker's dream customers, but they are not the rule. Most traders, even if driven by private value considerations, are somewhat price sensitive.

The traders (both informed and uninformed) in GM are actually modeled as agents subject to a random utility, $U = \rho x V + c$. ρ is the rate of substitution between current and future consumption, where "future" is the terminal payoff date; x is the number of shares held at the payoff date, and c is current consumption (*not*, in this context, transaction cost). ρ is random across traders, and its distribution is common knowledge. High ρ implies a strong preference for future consumption, and therefore (other things equal), a tendency to buy the security. The dealer's ρ is normalized to unity. The price of current consumption may also be normalized to unity.

Initially for an uninformed trader $EU = \rho x EV + c$. He will buy (paying the dealer's ask price A) if $\rho EV > A$. He will sell (at the dealer's bid price B) if $\rho EV < B$. If $B < \rho EV < A$, the agent won't trade. (In the present model, a non-trade event is uninformative. When there is event uncertainty, a non-trade is informative. This point is developed below.)

With inelastic uninformed trading demands, the dealer can set the bid and ask as wide as necessary to cover her losses to the informed traders. With elastic demands, though, there will generally be fewer uninformed agents willing to trade as these prices. The zero-expected-profit equilibrium will generally therefore exhibit a wider spread than in the inelastic case.

It is also possible that there exist no bid and ask values (other than $B = \underline{V}$ and $A = \overline{V}$) at which the dealer's expected profit is non-negative. That is, the uninformed traders are so price-sensitive that they are unwilling to participate in sufficient number to cover the dealer's losses to the informed traders (GM, p. 84). Agents trying to access the market bid and ask quotes see a blank screen. This is a market failure.

The market failure can be repaired by information dissemination that removes the asymmetry, or requiring the dealer to trade at a loss (presumably to be offset by some other benefit or concession). Both do in fact occur. Trading often stops (or is officially halted) pending a major news announcement. Exchanges, dealer associations, and simple considerations of reputation often effectively force a dealer to maintain a market presence when he would prefer to withdraw.

This is a point of considerable social and regulatory importance. While coverage and enforcement varies widely, most countries now have laws that prohibit "insider" trading. These prohibitions are grounded in considerations of fairness and economic efficiency. The economic efficiency argument holds that market

failures are extremely costly for the uninformed traders, who are denied the gains from trade (such as improved risk-sharing, etc.).

■ Event uncertainty

In the basic model an information asymmetry exists, and this fact is common knowledge. In real markets, however, significant information often arrives in a lumpy fashion. Long periods with no new information and steady or sluggish trading are punctuated by periods of extremely active trading before, during, and after major news announcements. The latter are sometimes referred to as "fast markets." Often the dealer's first inkling that an information asymmetry has arisen is a change in the pattern of incoming orders. A trading halt may be declared on the NYSE, for example, solely as a consequence of an "order flow imbalance."

This gives rise to what Easley and O'Hara (1992) model as event uncertainty. I'll discuss this model in detail in Chapter 15, but some general observations are useful at this point. Unlike the simple model, "nature's" first draw determines whether or not an information event occurs. The events of information occurrence and nonoccurrence will be denoted I and $\sim I$, respectively. Only the set of branches stemming from the I-node has a signal realization and the possibility of informed traders. If $\sim I$, then all traders are uninformed.

An informed trader always trades (in the direction of her knowledge). An uninformed trader might not trade. The no-trade probabilities for uninformed agents are the same whether I or $\sim I$, but the proportion of uninformed in the customer mix is higher with I. To the dealer, therefore, non-trade suggests an increased likelihood of $\sim I$.

■ Orders of different sizes

The basic sequential trade model has one trade quantity. Trades in real markets, of course, occur in varying quantities. Easley and O'Hara (1987) present a framework similar to that utilized in the last section. Their model features event uncertainty and two possible order sizes. The market-maker posts one set of bid and ask quotes for small trades and another set for large trades.

The most challenging thing about the model construction is the requirement that the zero-expected profit condition must hold for all quantities and directions. Expected losses on large buy orders, for example, can't be cross-subsidized by expected profits on small sell orders.

In the models considered to this point, all trades in which the market-maker might participate have some non-zero probability of involving an uninformed trader. This is a "pooling" feature of the trade mix. Were some class of trades to involve only informed traders (and therefore certain losses), no bid and ask prices (except the extrema of the value distribution would be possible. Such outcomes are "separating".

Informed traders maximize their profits by trading in the largest possible size. For a pooling equilibrium to exist, large orders must have some chance of originating from uninformed traders. A pooling equilibrium is also contingent on the existence of event uncertainty.

■ Orders of different types

The only orders permissible to this point have been marketable ones – orders that would result in an immediate execution. Real world security markets admit a much wider range. Many of the variations arise when a customer has a trading strategy that can be codified in a simple rule, that when communicated with the order avoids the necessity for further monitoring or modification on the customer's part.

One common variant is the price-contingent order. On the sell side, these are called stop-loss orders. When the trade price hits or drops through a preset barrier, the order becomes marketable. For example, consider a stop-loss order to sell triggered (“elected”) at a price of 50. When the trade price reaches 50, this is converted into a market order. Note that actual execution price for this order may well be below 50 if the market is moving quickly. There are also buy stop orders, which become marketable when the price rises through a preset barrier.

Easley and O'Hara (1991) analyze a sequential trade model where the market accepts stop orders. The main implications of the model are:

- Informed traders will never use stop orders.
- The information content of prices declines (the market becomes “less informationally efficient”)
- There is a greater probability of large price changes. In the model (and in real markets), a trade can trigger a wave of elections.

5.d Empirical implications

The sequential trade models convey two useful empirical predictions.

- Spread: At a given point in time, more extreme information asymmetry implies a larger spread.
- Price impact: For any given trade, more extreme information asymmetry implies a larger quote revision (price impact).

What sort of statistical approach should we follow?

Observations in the sequential trade models (whether of spreads, quotes, trade prices or first-differences of these variables) are not i.i.id. nor are they covariance stationary. Furthermore, because the process described by these models is a adjustment in response to non-reccurent initial conditions, the sequence of observations is non-ergodic. Therefore, standard time series analysis is not directly applicable. To proceed, we can assume that our sample consists of multiple paths of adjustment processes, stacked end-to-end, and a given or known mapping to our sample. We might assume, for example, that the model describes what happens between the beginning and end of the calendar/wall-clock trading day, and that our sample consists of independent days. Then we can treat each day as a separate observation;. This approach will be discussed in detail in a later chapter.

Alternatively, we can assume that our data are generated by a structural model that incorporates the asymmetric information effects in some unspecified fashion. This approach suggests "reduced-form" time-series models that have a much more statistical flavor. This approach is used, implicitly if not explicitly, in the many studies that rely on time-averages of spreads.

These two approaches lie on a continuum: structural economic models at one end and reduced-form statistical models at the other. The trade-off is the usual one in econometrics. Structural models offer stronger economic content and predictions, but they are more subject to misspecification. Reduced-form models are more robust to misspecification, but are more limited in the economic insights they can afford.

5.e Problems

Problem 5.1 A modified model

As in the basic model, there are two possible values for V. $V = \underline{V}$ with probability δ; $V = \overline{V}$ with probability $1 - \delta$. There are two types of traders. A type-X agent receives a signal (H or L) that is correct with probability π^X: $\Pr(L \mid \underline{V}) = \Pr(H \mid \overline{V}) = \pi^X$. Similarly, a type-Y traders receives a signal with accuracy $\pi^Y > \pi^X$. Traders always trade in the direction of their signal. If they get a low signal, they sell; if they get a high signal they buy. The fraction of type-Y traders in the population is μ. In a competitive dealer market, what are the initial bid and ask?

Chapter 6. Strategic trade models of asymmetric information

In the sequential trade framework, there are many informed agents, but each can trade only once, and only if he/she is "drawn" as the arriving trader. Furthermore, if order size is a choice variable, the informed agent will always trade the largest quantity. The Kyle (1985) model, discussed in this chapter, differs in both respects.

In the Kyle model, there is a single informed trader who behaves strategically. She sets her trade size taking into account the adverse price concession associated with larger quantities. She can furthermore, in the multiple-period version of the model, return to the market, spreading out her trades over time.

The practice of distributing orders over time so as to minimize trade impact is perhaps one of the most common strategies used in practice. With decimalization and increased fragmentation of trading activity, market participants have fewer opportunities to easily trade large quantities. In the present environment, therefore, order splitting strategies are widely used by all sorts of traders (uninformed as well as informed).

Although the Kyle model allows for strategic trade, while the sequential trade models don't, it is more stylized in some other respects. There is no bid and ask, for example; all trades clear at an informationally-efficient price.

Useful extensions of the Kyle model include: Admati and Pfleiderer (1988); Foster and Viswanathan (1990); Subrahmanyam (1991); Subrahmanyam (1991); Holden and Subrahmanyam (1994); Foster and Viswanathan (1995); Back (1992). Back and Baruch (2003) suggest a synthesis of the sequential and strategic trade models.

Initializations for the analysis of the Kyle model. ***Mathematica***

6.a The single-period model

The elements of the model are:

- The terminal security value is $v \sim N(p_0, \Sigma_0)$.
- There is one informed trader who knows v and enters a demand x (buying if $x > 0$, selling if $x < 0$).
- Liquidity traders submit a net order flow $u \sim N(0, \sigma_u^2)$, independent of v.
- The market-maker (MM) observes the total demand $y = x + u$ and then sets a price, p.

- All of the trades are cleared at p. If there is an imbalance between buyers and sellers, the MM makes up the difference.

Note that nobody knows the market clearing price when they submit their orders.

Since the liquidity trader order flow is exogenous, there are really only two players we need to concentrate on: the informed trader and the market maker. The informed trader wants to trade aggressively, e.g., buying a large quantity if her information is positive. But the MM knows that if he sells into a large net customer "buy", he his likely to be on the wrong side of the trade. He protects himself by setting a price that is increasing in the net order flow. This acts as a brake on the informed trader's desires: if she wishes to buy a lot, she'll have to pay a high price. The solution to the model is a formal expression of this trade-off.

We first consider the informed trader's problem (given a conjectured MM price function), and then show that the conjectured price function is consistent with informed trader's optimal strategy.

■ The informed trader's problem

The informed trader conjectures that the MM uses a linear price adjustment rule:

$$p = y\,\lambda + \mu \tag{6.a.1}$$

where y is the total order flow: $y = u + x$.

λ in the price conjecture is an inverse measure of liquidity. The informed trader's profits are:

$$\pi = (v - p)\,x \tag{6.a.2}$$

Substituting in for the price conjecture and y:

$$\pi = x\,(v - (u + x)\,\lambda - \mu) \tag{6.a.3}$$

In the sequential trade models, an informed trader always makes money. This is not true here. For example, if the informed trader is buying ($x > 0$), it is possible that a large surge of uninformed buying ($u >> 0$) drives the $\lambda(u + x) + \mu$ above v.

The expected profits are $\mathrm{E}\pi$:

$$\mathrm{E}\pi = x\,(v - x\,\lambda - \mu) \tag{6.a.4}$$

The informed trader maximizes expected profits by trading x:

$$x = \frac{v - \mu}{2\,\lambda} \tag{6.a.5}$$

The second-order condition for the max is

$$-2\lambda < 0 \tag{6.a.6}$$

■ The market maker's problem

The MM conjectures that the informed trader's demand is linear in v:

$$x = \alpha + v\beta \tag{6.a.7}$$

Knowing the optimization process that the informed trader followed, the MM can solve for α and β:

$$\alpha + v\beta = \frac{v - \mu}{2\lambda} \tag{6.a.8}$$

for all v. This implies:

$$\begin{aligned} \alpha &= -\frac{\mu}{2\lambda} \\ \beta &= \frac{1}{2\lambda} \end{aligned} \tag{6.a.9}$$

The relation between β and λ is particularly important. As the liquidity drops (i.e., as λ rises), the informed trader trades less.

Now the MM must figure out $E[v \mid y]$. In computing this, it is useful to recall that if $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N(\mu, \Sigma)$, then the conditional mean of Y given X is:

$$E_{Y|X}(\mathrm{x}) = \frac{\mathrm{Cov}(X, Y)\,(x - \mathrm{mean}(X))}{\mathrm{Var}(X)} + \mathrm{mean}(Y) \tag{6.a.10}$$

where an upper case letter like "X" denotes a random variable and the corresponding lower case "x" denotes a realization of that variable.

Given the definition of the order flow variable and the MM's conjecture about the informed traders behavior,

$$y = u + \alpha + v\beta \tag{6.a.11}$$

Thus:

$$E_{v|y}(y) = p_0 + \frac{\beta\,(y - \alpha - \beta\,p_0)\,\Sigma_0}{\Sigma_0\,\beta^2 + \sigma_u^2} \tag{6.a.12}$$

Market efficiency requires $E_{v|y} = p$:

$$p_0 + \frac{\beta\,(y - \alpha - \beta\,p_0)\,\Sigma_0}{\Sigma_0\,\beta^2 + \sigma_u^2} = y\lambda + \mu \tag{6.a.13}$$

This must hold for all values of y, so:

$$\mu = -\frac{\alpha\,\beta\,\Sigma_0 - \sigma_u^2\,p_0}{\Sigma_0\,\beta^2 + \sigma_u^2} \qquad \lambda = \frac{\beta\,\Sigma_0}{\Sigma_0\,\beta^2 + \sigma_u^2} \tag{6.a.14}$$

Now both the informed trader's problem and the MM's problem have been solved (given their respective conjectures). Collecting these results:

$$\mu = -\frac{\alpha\,\beta\,\Sigma_0 - \sigma_u^2\,p_0}{\Sigma_0\,\beta^2 + \sigma_u^2} \qquad \lambda = \frac{\beta\,\Sigma_0}{\Sigma_0\,\beta^2 + \sigma_u^2} \qquad \alpha = -\frac{\mu}{2\,\lambda} \qquad \beta = \frac{1}{2\,\lambda} \tag{6.a.15}$$

It just remains to solve for the parameters of the conjectures in terms of the problem inputs.

$$\alpha = -\frac{\sqrt{\sigma_u^2}\,p_0}{\sqrt{\Sigma_0}} \qquad \mu = p_0 \qquad \lambda = \frac{\sqrt{\Sigma_0}}{2\sqrt{\sigma_u^2}} \qquad \beta = \frac{\sqrt{\sigma_u^2}}{\sqrt{\Sigma_0}} \tag{6.a.16}$$

■ Properties of the solution

Both the liquidity parameter λ and the informed trader's order coefficient β depend only on the value uncertainty Σ_0 relative to the intensity of noise trading σ_u^2.

The informed trader's expected profits are:

$$\mathrm{E}\pi = \frac{\sqrt{\sigma_u^2}\,(v - p_0)^2}{2\sqrt{\Sigma_0}} \tag{6.a.17}$$

These are increasing in the divergence of the value (known by the informed trader) from the expectation of the uninformed agents (p_0). They're also increasing in the variance of noise trading. We can think of the noise trading as providing camouflage for the informed trader. This is of practical importance. All else equal, an agent trading on inside information will be able to make more money in a widely held and frequently traded stock (at least, prior to apprehension).

The informed trader's demand is:

$$x = \frac{\sqrt{\sigma_u^2}\,(v - p_0)}{\sqrt{\Sigma_0}} \tag{6.a.18}$$

How much of the private information is impounded in the price? If $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N(\mu, \Sigma)$, then the conditional variance of Y given X is:

$$\mathrm{Var}_{Y|X} = \mathrm{Var}(Y) - \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(X)} \tag{6.a.19}$$

Note that this does not depend on the realization of X. Thus,

$$\mathrm{Var}_{v|y} = \mathrm{Var}(v) - \frac{\mathrm{Cov}(y, v)^2}{\mathrm{Var}(y)} = \Sigma_0 - \frac{\beta^2 \Sigma_0^2}{\Sigma_0 \beta^2 + \sigma_u^2} \tag{6.a.20}$$

Or, in terms of the input parameters:

$$\mathrm{Var}_{v|p} = \mathrm{Var}_{v|y} = \frac{\Sigma_0}{2} \tag{6.a.21}$$

That is, half of the insider's information gets into the price. This does not depend on the intensity of noise trading.

The problems to this chapter discuss modifications to the single-period model.

6.b The multiperiod model

■ Setup

There are $k = 1, \ldots, N$ auctions. These are equally-spaced on a unit time interval. In real time, the kth auction occurs at time $\frac{k}{T}$, so the increment between auctions is $\Delta t = \frac{1}{T}$. At the kth auction, noise traders submit an order flow $u_k \sim N(0, \sigma_u^2 \Delta t)$. The informed trader submits an order flow Δx_t.

The informed traders profits are given recursively as $\pi_k = (v - p_k)\,\Delta x_k + \pi_{k+1}$ for $k = 1, \ldots, N$ and $\pi_{N+1} \equiv 0$.

■ Solution

Kyle's Theorem 2 gives the solution as follows

The informed trader's demand in auction n is linear in the difference between the true value v and the price on the preceding auction, p_{n-1}:

$$\Delta x_n = \Delta t\,(v - p_{n-1})\,\beta_n \tag{6.b.22}$$

The MM's price adjustment rule is linear in the total order flow:

$$\Delta p_n = (\Delta u_n + \Delta x_n)\,\lambda_n \tag{6.b.23}$$

Expected profits are quadratic:

$$E\pi_n = \alpha_{n-1}\,(v - p_{n-1})^2 + \delta_n \tag{6.b.24}$$

The constants in the above are given by the solutions to the difference equation system:

$$\begin{aligned}\alpha_k &= \frac{1}{4\,\lambda_{k+1}\,(1-\alpha_{k+1}\,\lambda_{k+1})}\\ \delta_k &= \Delta t\,\alpha_{k+1}\,\lambda_{k+1}^2\,\sigma_u^2+\delta_{k+1}\\ \beta_n &= \frac{1-2\,\alpha_n\,\lambda_n}{\Delta t\,(2\,\lambda_n\,(1-\alpha_n\,\lambda_n))}\\ \lambda_n &= \frac{\beta_n\,\Sigma_n}{\sigma_u^2}\end{aligned} \tag{6.b.25}$$

subject to the terminal conditions $\alpha_N = \delta_N = 0$.

The above recursions are backwards. Σ_n is the variance of v conditional on all order flow and prices through auction n. It is given by the *forward* recursion:

$$\Sigma_n=(1-\Delta t\,\beta_n\,\lambda_n)\,\Sigma_{n-1} \tag{6.b.26}$$

The solutions for $\{\alpha_k,\,\delta_k,\,\beta_k,\,\lambda_k,\,\Sigma_k\}$ don't depend on the realization of v. That is, given $\{\Sigma_0,\,p_0,\,{\sigma_u}^2\}$, agents can perfectly forecast the depth and demand coefficients.

■ Analysis of solution

To compute a solution given N and the model parameters $\{\Sigma_0,\,p_0,\,{\sigma_u}^2\}$, start at the *nth* auction. Taking the solution for λ_n and plugging in from the solution for β_n yields a cubic polynomial equation for λ_n:

$$\lambda_n = \frac{(1-2\,\alpha_n\,\lambda_n)\,\Sigma_n}{2\,\Delta t\,\sigma_u^2\,\lambda_n\,(1-\alpha_n\,\lambda_n)} \tag{6.b.27}$$

The equation has three roots. They are not pretty ones. If you really want to see them, run the following *Mathematica* line (which is not visible in the pdf/printout versions of this document).

The full solution procedure is as follows. The model parameters are $\Sigma_0,\ v,$ and σ_u^2.

1. Pick a trial value of Σ_N. By the terminal conditions, $\alpha_N = \delta_N = 0$. Solve the polynomial equation for λ_N. In general, this is a cubic, but at step N, it is quadratic. Take λ_N as the positive root. Compute β_N and Σ_{N-1}.
2. At step $N-1$, compute α_{N-1} and δ_{N-1} using the above formulas. Solve for λ_{N-1}, taking the middle root. Compute β_{N-1} and Σ_{N-1}.
3. Iterate over step 2, backwards in time until we arrive at the first auction ($k = 1$). Compute the value of Σ_0 implied by this backward recursion, given our initial guess at Σ_N. Compare this to the desired Σ_0.

Using numerical optimization, repeat steps 2 and 3 until we've found a value of Σ_N, which implies (via the backward recursions) the desired value of Σ_0.

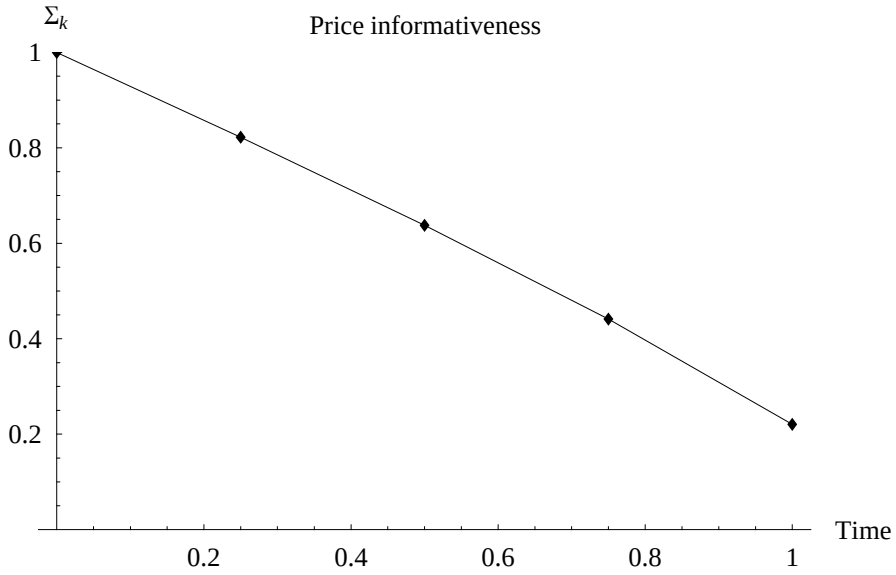
Program to implement numerical solution ***Mathematica***

■ Numerical example

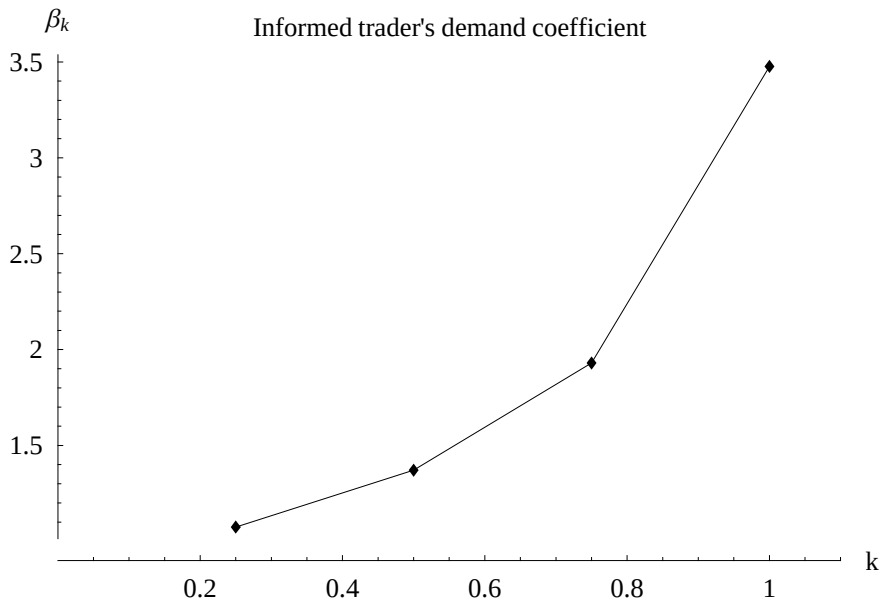
Consider the case with $T = 4$ clearings, $\sigma_u^2 = \Sigma_0 = 1$.

	α	δ	λ	β	Σ_k	Σ_{k-1}
1	0.591587	0.102483	0.662334	1.07417	0.822136	1.
2	0.541962	0.0442885	0.655372	1.37071	0.637499	0.822136
3	0.43462	0	0.63844	1.92957	0.441163	0.637499
4	0	0	0.575215	3.47696	0.220582	0.441163

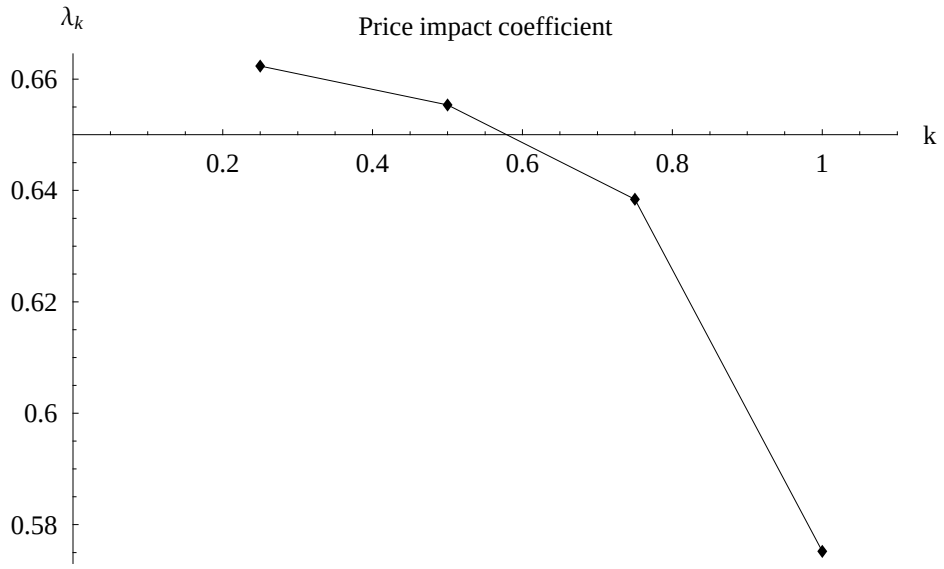
The amount of information in the price over time is given by Σ_k:



The price becomes more informative over time. The informed traders demand coefficient is β_k:



The informed trader trades more aggressively over time. The price impact parameter is given by λ_k:



The price impact coefficient declines over time: an early trade has more impact than a later trade of the same size.

■ Autocorrelation in trades

We have seen that in the sequential trade models, orders are positively autocorrelated (buys tend to follow buys). Does a similar result hold here?

Since the informed trader splits her orders over time, and tends to trade on the same side of the market, her order flow is positively autocorrelated. This should induce positive autocorrelation in the total order flow.

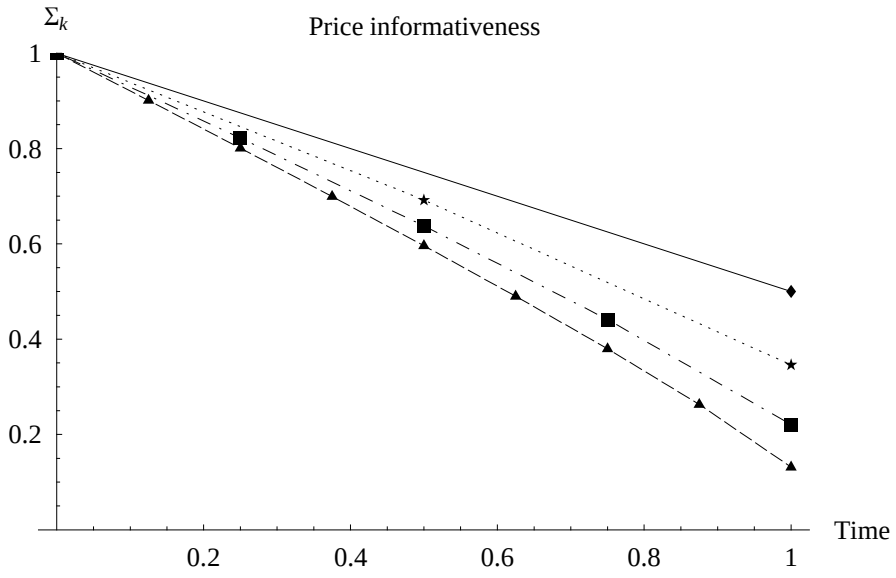
This seems simple, sensible and obvious. It's also completely wrong. Remember that market efficiency requires that the price follow a Martingale. The increments to a Martingale aren't autocorrelated. Furthermore, the price change is proportional to the net order flow. If the price change isn't autocorrelated, the net order flow can't be either.

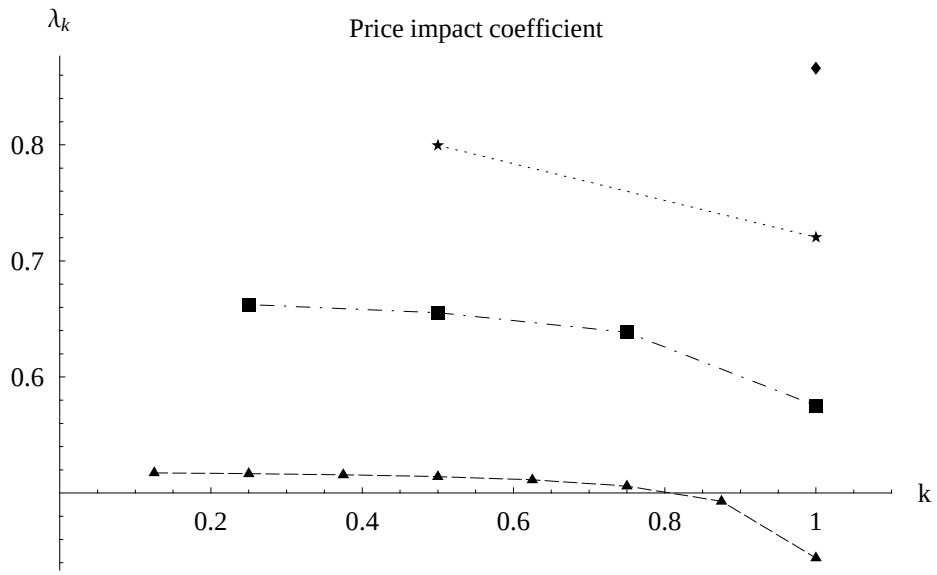
From a strategic viewpoint, the informed trader is sometimes said to "hide" behind the uninformed order flow. This means that she trades so that the MM can't predict (on the basis of the net order flow) what she will do next.

We now examine what happens when the number of auctions increases.

■ Increasing the number of auctions (when total noise trading remains unchanged)

In this example, we let consider the case with $\sigma_u^2 = 1$ and $\Sigma_0 = 4$. We examine $T = 1,\ 2,\ 4,\ 8$. Recall that as T increases, the noise trading *per auction* decreases.





6.c Problems based on the single-period model

The essential properties of the model that make it tractable arise from the multivariate normality (which gives linear conditional expectations) and a quadratic objective function (which has a linear first-order condition). The multivariate normality can accommodate a range of modifications. The following problems explore some.

Problem 6.1 Informative noise traders

The noise traders in the basic model are pure noise traders: u is independent of v. Consider the case where the u order flow is positively related to the value: $\text{Cov}(u, v) = \sigma_{uv} > 0$. Proceed as above. Solve the informed trader's problem; solve the MM's problem; solve for the model parameters $(\alpha, \beta, \mu, \lambda)$ in terms of the inputs, σ_u^2, Σ_0 and σ_{uv}. Interpret your results. Show that when $\text{Corr}(u, v) = 1$, the price becomes perfectly informative.

Answer

Problem 6.2 Informed trader gets a signal

The informed trader in the basic model has perfect information about v. Consider the case where she only gets a signal s about v. That is, $s = v + \epsilon$ where $\epsilon \sim N(0, \sigma_\epsilon^2)$, independent of v. Solve the model by proceeding as in the basic case. Solve the informed trader's problem; solve the MM's problem; solve for the model parameters $(\alpha, \beta, \mu, \lambda)$ in terms of the inputs, σ_u^2, Σ_0 and σ_ϵ^2. Interpret your results. Verify that when $\sigma_\epsilon^2 = 0$, you get the original model solutions.

Answer

Problem 6.3 Frontrunning by the informed trader's broker

In a securities market, "front-running" generally refers to a broker holding a customer order and trading before his customer. An example in the narrow sense arises when a broker holds a customer market buy order in a security and buys before executing the customer order. This is a clear violation of the broker's fiduciary duty.

Other examples are less clear. Suppose a customer puts in a limit order to buy *XYZ* at $100. The broker then puts in a limit order to buy *XYZ* at $100.01. Or, suppose a customer puts in a market order to buy *XYZ*. The broker immediately puts in a market order to buy *ABC* in the same industry, or an index security in which *XYZ* is a component. In both of these examples, the broker's actions might disadvantage the customer. Under present standards, though, it is unlikely that the customer would have a sustainable case.

Suppose that when the informed trader in the basic model puts in an order x, her broker simultaneously puts in an order γx, with $\gamma > 0$. That is, the broker piggy-backs on the informed traders information. (Improbable? See *Den of Thieves*, by James B. Stewart.)

Solve the model by proceeding as in the basic case. Solve the informed trader's problem; solve the MM's problem; solve for the model parameters $(\alpha, \beta, \mu, \lambda)$ in terms of the inputs, σ_u^2, Σ_0 and γ.

Answer

Chapter 7. The generalized Roll model

7.a Overview

Following the economic perspectives developed in the last two sections, we now turn to the problem of generalizing the Roll model to take into account asymmetric information. One sensible first step is to allow the efficient price to be partially driven by the trade direction indicator variables.

A number of models along these lines have been proposed. See, for example, Glosten (1987); Glosten and Harris (1988); Stoll (1989; George, Kaul and Nimalendran (1991); Lin, Sanger and Booth (1995); Huang and Stoll (1997).

The present development is compatible with (i.e., a special case of) most of these models. In connecting the present section to these papers, however, there are a few special considerations.

- Most of the models in the literature were estimated with observations on both prices and trades (the q_t). In contrast, this section will examine representations solely in terms of the prices. The reason for this is that there are some features of these models that are best initially encountered in a univariate setting. A second consideration is that, although we have good recent data on US equity markets that allow us to infer q_t, this is not universally the case. In many data samples and markets, only trade prices are recorded.

- A second point is that some of these models adopt the perspective of explaining "components of the spread," i.e., what proportion of the spread is due to fixed costs, what to asymmetric information and so forth. This is nothing more or less than a parameter normalization, convenient for some applications, less so for others. The underlying dynamics, however, are essentially the same as in the present development.

The term "generalized Roll model" is not in common use. It is used here to emphasize the roots of the Roll model in the present development.

7.b Model description

The evolution of the efficient price is given by:

$$m_t = m_{t-1} + w_t \tag{7.b.1}$$

The increments to the efficient prices are driven by trades and public information.

$$w_t = \lambda\, q_t + u_t \tag{7.b.2}$$

This reduces to the usual Roll model when $\lambda = 0$. The actual trade price is:

$$p_t = m_t + c\, q_t \tag{7.b.3}$$

A buy order lifts the ask, so the ask is the trade price when $q_t = +1$:

$$A_t = c + \lambda + m_{t-1} + u_t \tag{7.b.4}$$

Similarly, the bid is the trade price when $q_t = -1$:

$$B_t = -c - \lambda + m_{t-1} + u_t \tag{7.b.5}$$

Thus, the bid and ask are set symmetrically about $m_{t-1} + u_t$. The spread is $2\,(c + \lambda)$, where c reflects the fixed costs of the trade (clearing costs, clerical costs, etc.) and λ reflects the adverse selection.

This implies the following timing. Immediately after the time $t - 1$ trade, the efficient price is m_{t-1}. Then public information arrives as the realization of u_t. The market maker sets the bid and ask symmetrically about $m_{t-1} + u_t$. Then a trade arrives as the realization of q_t, and the efficient price is updated to m_t.

■ Alternative representations and special cases

For the original Roll model, we developed moving average and autoregressive representations that were useful in parameter estimation and forecasting. Here, we examine the time series structure of the generalized Roll model.

Consider the price changes $\Delta p_t = p_t - p_{t-1}$. Substituting in for p_t, m_t and w_t gives:

$$\Delta p_t = -c\, q_{t-1} + (c + \lambda)\, q_t + u_t \tag{7.b.6}$$

The model has three parameters $\{\lambda, c, \sigma_u^2\}$ and two sources of randomness: u_t and q_t. We'll consider the general case, but it will also sometimes be useful to look at the two special cases:

- *Exclusively public information* ($\lambda = 0$, the original Roll model)
- *Exclusively private information* ($u_t = 0$ for all t, or equivalently $\sigma_u^2 = 0$).

■ The autocovariance structure of Δp_t

To obtain $\mathrm{Var}(\Delta p_t) = \gamma_0$, consider:

$$\begin{aligned}\Delta p_t^2 = {} & q_{t-1}^2\, c^2 + q_t^2\, c^2 - 2\, q_{t-1}\, q_t\, c^2 + 2\,\lambda\, q_t^2\, c - \\ & 2\,\lambda\, q_{t-1}\, q_t\, c - 2\, q_{t-1}\, u_t\, c + 2\, q_t\, u_t\, c + \lambda^2\, q_t^2 + u_t^2 + 2\,\lambda\, q_t\, u_t\end{aligned} \tag{7.b.7}$$

In expectation, all of the cross-products vanish except for those involving q_t^2, q_{t-1}^2 and u_t^2. So:

$$\gamma_0 = c^2 + (c+\lambda)^2 + \sigma_u^2 \tag{7.b.8}$$

To obtain $\mathrm{Cov}(\Delta p_t, \Delta p_{t-1}) = \gamma_1$, we examine:

$$\begin{aligned}\Delta p_t \, \Delta p_{t-1} = &-q_{t-1}^2 \, c^2 + q_{t-2} \, q_{t-1} \, c^2 - q_{t-2} \, q_t \, c^2 + q_{t-1} \, q_t \, c^2 - \lambda \, q_{t-1}^2 \, c - \lambda \, q_{t-2} \, q_t \, c + 2 \, \lambda \, q_{t-1} \, q_t \, c - \\ &q_{t-1} \, u_{t-1} \, c + q_t \, u_{t-1} \, c - q_{t-2} \, u_t \, c + q_{t-1} \, u_t \, c + \lambda^2 \, q_{t-1} \, q_t + \lambda \, q_t \, u_{t-1} + \lambda \, q_{t-1} \, u_t + u_{t-1} \, u_t\end{aligned} \tag{7.b.9}$$

In expectation, all of the cross-products vanish except for the second and third terms, so:

$$\gamma_1 = -c \, (c+\lambda) \tag{7.b.10}$$

The second-order cross-product involves no contemporaneous products:

$$\begin{aligned}&\Delta p_t \, \Delta p_{t-2} = \\ &q_{t-3} \, q_{t-1} \, c^2 - q_{t-2} \, q_{t-1} \, c^2 - q_{t-3} \, q_t \, c^2 + q_{t-2} \, q_t \, c^2 - \lambda \, q_{t-2} \, q_{t-1} \, c - \lambda \, q_{t-3} \, q_t \, c + 2 \, \lambda \, q_{t-2} \, q_t \, c - \\ &q_{t-1} \, u_{t-2} \, c + q_t \, u_{t-2} \, c - q_{t-3} \, u_t \, c + q_{t-2} \, u_t \, c + \lambda^2 \, q_{t-2} \, q_t + \lambda \, q_t \, u_{t-2} + \lambda \, q_{t-2} \, u_t + u_{t-2} \, u_t\end{aligned} \tag{7.b.11}$$

So it vanishes, as do higher order autocovariances.

7.c Identification of σ_w^2

The two estimates of $\{\gamma_0, \gamma_1\}$ are not sufficient to identify the three parameters of the model $\{\lambda, c, \sigma_u{}^2\}$. Each of the special cases drops a model parameter, so these cases are identified. But the restrictions they impose (exclusively public information, or alternatively, exclusively private information) are not attractive ones.

Interestingly, though, one derived parameter from the general model can be identified without further restrictions. This is $\mathrm{Var}(w_t) = \sigma_w^2$, the variance of the efficient-price increments. To see this, first note:

$$w_t^2 = \lambda^2 \, q_t^2 + 2 \, \lambda \, u_t \, q_t + u_t^2 \tag{7.c.12}$$

Since u_t and q_t are uncorrelated, and $E q_t^2 = 1$,

$$\sigma_w^2 = \lambda^2 + \sigma_u^2.$$

Now consider the expression $\gamma_0 + 2\,\gamma_1$. With the autocovariance calculations we derived above,

$$\gamma_0 + 2\gamma_1 = \lambda^2 + \sigma_u^2 = \sigma_w^2 \tag{7.c.13}$$

It will later be shown that the identifiability of σ_w^2 is a general result, extending to multiple lags and multivariate and/or multiple price models.

Intuitively, σ_w^2 is the variance per unit time of the random-walk component of the security price. This variance is time-scaled, in the sense that if we use a longer interval to compute the change, the variance is simply multiplied by the length of the interval:

$$\text{Var}(m_t - m_{t-k}) = k\sigma_w^2.$$

But over long periods, microstructure effects become relatively less important. Most of the long-term dynamics in p_t are attributable to m_t. More precisely, as k gets large,

$$\sigma_w{}^2 = \frac{\text{Var}(m_t - m_{t-k})}{k} \approx \frac{\text{Var}(p_t - p_{t-k})}{k}$$

(How large does k have to be? Is one day good enough? A week? A month?)

To identify the other parameters in the model, we need more data or more structure.

7.d The moving average (MA) representation

Since the autocovariances vanish above the first-order, using the Wold theorem, the price changes can be represented as $\Delta p_t = \epsilon_t + \theta\epsilon_{t-1}$. In terms of this representation, the autocovariances are:

$$\{\gamma_0 = (\theta^2 + 1)\,\sigma_\epsilon^2,\ \gamma_1 = \theta\,\sigma_\epsilon^2\} \tag{7.d.14}$$

Given sample autocovariances, we can solve for the MA parameters. There are two solutions:

$$\begin{array}{lll} \text{Solution 1:} & \sigma_\epsilon^2 = \frac{1}{2}\left(\gamma_0 - \sqrt{\gamma_0^2 - 4\,\gamma_1^2}\right) & \theta = \frac{\gamma_0 + \sqrt{\gamma_0^2 - 4\,\gamma_1^2}}{2\,\gamma_1} \\ \text{Solution 2:} & \sigma_\epsilon^2 = \frac{\gamma_0}{2} + \frac{1}{2}\sqrt{\gamma_0^2 - 4\,\gamma_1^2} & \theta = \frac{\gamma_0 - \sqrt{\gamma_0^2 - 4\,\gamma_1^2}}{2\,\gamma_1} \end{array} \tag{7.d.15}$$

With some hypothetical values ($\gamma_0 = 1,\ \gamma_1 = -.2$), the MA parameters are

$$\begin{array}{lll} \text{Solution 1:} & \sigma_\epsilon^2 = 0.0417424 & \theta = -4.79129 \\ \text{Solution 2:} & \sigma_\epsilon^2 = 0.958258 & \theta = -0.208712 \end{array} \tag{7.d.16}$$

Remember that, for the basic Roll model, we were able to recursively construct the ϵ_t from the p_t:

$$\epsilon_t = \Delta p_t - \theta\Delta p_{t-1} + \theta^2\,\Delta p_{t-2} + \theta^3\,\Delta p_{t-3} - \ldots \tag{7.d.17}$$

From this we see that the two solutions for the moving average parameters are not equally attractive. In the first solution $|\theta| > 1$, and the above expression does not converge. Formally, it is not invertible.

There's an interesting relationship between the two solutions. Suppose that, rather than pressing for a full solution, we simply eliminate σ_ϵ^2. Then

$$\gamma_1\,(\theta^2 + 1) = \gamma_0\,\theta \tag{7.d.18}$$

So, θ is the solution to $\gamma_1 \theta^2 - \gamma_0 \theta + \gamma_1 = 0$. From this, it's easy to see that if θ^* is a solution, then so is $1/\theta^*$.

Therefore the invertible and noninvertible solutions must be related as $\theta^{\text{Invertible}} = 1/\theta^{\text{Noninvertible}}$.

■ Forecasting and filtering

In the basic Roll model the price forecast has been shown to be:

$$f_t = \lim_{k\to\infty} E[p_{t+k} \mid p_t,\ p_{t-1},\ ...] = E[p_{t+1} \mid p_t,\ p_{t-1},\ ...] = p_t + \theta\epsilon_t \tag{7.d.19}$$

Recall that, although this forecast is a martingale, it does *not* equal the efficient price m_t from the structural model. But if it isn't m_t, what exactly is it?

It turns out that $f_t = E[m_t \mid p_t,\ p_{t-1},\ ...]$. This is sometimes called a *filtered* estimate: the expectation of an unobserved state variable conditional on current and past observations.

If you want to see why, read the following section. (You might want to skip it on a first reading.)

■ Proof

We'll now proceed to construct the linear filters for $m_t = p_t - cq_t$. Since we know p_t and c, the trick is forming an expectation of q_t.

We'll be working a linear projection, essentially a linear regression of the form

$$q_t = \alpha_0\ p_t + \alpha_1\ p_{t-1} + ... + v_t \tag{7.d.20}$$

(for the filtered estimate) where the αs are linear projection coefficients and v_t is the projection error. Now while we could compute the αs directly, it's a messy calculation because the p_t are correlated with each other.

Think of a regression $y_t = x_t \beta + u_t$. The linear projection coefficients are given by

$$\beta = (E x_t\, x_t')^{-1}\, E x_t'\, y_t. \tag{7.d.21}$$

The calculation is a lot easier if the x_t are not mutually correlated. Then $(E x_t\, x_t')$ is diagonal and each coefficient may be computed as

$$\beta_i = \frac{\text{Cov}(x_{i,t}, y_t)}{\text{Var}(x_{i,t})}. \tag{7.d.22}$$

In the present case, it's much easier to work with the projection

$$q_t = \beta_0\, \epsilon_t + \beta_1\, \epsilon_{t-1} + \beta_2\, \epsilon_{t-2} + ... + v_t \tag{7.d.23}$$

Since the ϵ_t are uncorrelated, $\beta_i = \frac{\text{Cov}(q_t, \epsilon_{t-i})}{\sigma_\epsilon{}^2}$. So how do we compute $\text{Cov}(q_t, \epsilon_{t-i})$? We have two ways of representing Δp_t: the statistical and the structural. They obviously must agree:

$$\theta \epsilon_{t-1} + \epsilon_t = -c\, q_{t-1} + c\, q_t + \lambda\, q_t + u_t \tag{7.d.24}$$

Rearranging this to isolate ϵ_t:

$$\epsilon_t = -c\, q_{t-1} + (c + \lambda)\, q_t + u_t - \theta \epsilon_{t-1} \tag{7.d.25}$$

From which it is clear that $\text{Cov}(q_t, \epsilon_t) = c + \lambda$. Recursively substituting in again gives:

$$\epsilon_{t-1} = -c\, q_{t-2} + (c + \lambda)\, q_{t-1} + u_{t-1} - \theta \epsilon_{t-2} \tag{7.d.26}$$

Thus, $\text{Cov}(q_t, \epsilon_{t-1}) = 0$, and in fact $\text{Cov}(q_t, \epsilon_{t-k}) = 0$ for $k \geq 1$. So the projection $q_t = \beta_0 \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + ... + v_t$ becomes

$$q_t = \beta_0 \epsilon_t + v_t \text{ where } \beta_0 = \frac{c+\lambda}{\sigma_\epsilon{}^2}. \tag{7.d.27}$$

Next, recall that $E[m_t \mid p_t, \ ...] = p_t - cE[q_t \mid p_t, \epsilon_t, \epsilon_{t-1,} ...] = p_t - c\, \beta_0 \epsilon_t$.

Is there a more intuitive way of expressing this? Substituting in for β_0 gives:

$$-\text{c}\ \beta_0 = -\frac{c\,(c+\lambda)}{\sigma_\epsilon^2} \tag{7.d.28}$$

Recall next that:

$$\{\gamma_0 = c^2 + (c+\lambda)^2 + \sigma_u^2,\ \gamma_1 = -c\,(c+\lambda)\} \tag{7.d.29}$$

From which it is clear that $-c\beta_0 = \gamma_1 / \sigma_\epsilon{}^2$. Analyzing the latter expression using the invertible solution for the moving average parameters gives:

$$-\text{c}\beta_0 = \frac{2\,\gamma_1}{\gamma_0 + \sqrt{\gamma_0^2 - 4\,\gamma_1^2}} \tag{7.d.30}$$

Now the solution set for the MA parameters was

$$\begin{aligned} \sigma_\epsilon^2 &= \frac{1}{2}\left(\gamma_0 - \sqrt{\gamma_0^2 - 4\,\gamma_1^2}\right) & \theta &= \frac{\gamma_0 + \sqrt{\gamma_0^2 - 4\,\gamma_1^2}}{2\,\gamma_1} \\ \sigma_\epsilon^2 &= \frac{\gamma_0}{2} + \frac{1}{2}\sqrt{\gamma_0^2 - 4\,\gamma_1^2} & \theta &= \frac{\gamma_0 - \sqrt{\gamma_0^2 - 4\,\gamma_1^2}}{2\,\gamma_1} \end{aligned} \tag{7.d.31}$$

where the first solution is noninvertible. By inspection, it is clear that $\gamma_1 / \sigma_\epsilon{}^2 = 1 / \theta^{\text{Noninvertible}}$. But we earlier showed that $1 / \theta^{\text{Noninvertible}} = \theta^{\text{Invertible}}$. Thus

$$E[m_t \mid p_t,\ p_{t-1},\ ...] = p_t - c\beta_0 \epsilon_t = p_t + \theta^{\text{Invertible}} \epsilon_t = f_t$$

So while f_t is not in general equal to the efficient price, it can be interpreted as the expectation of the efficient price conditional on current and past information. That is, f_t is the *filtered* estimate of m_t.

Suppose that we also have at our disposal the future realizations:

$E[m_t \mid ..., p_{t+1}, p_t, p_{t-1}, ...]$ is the *smoothed* estimate of m_t.

For example, given a full data sample, we might be interested in estimating the implicit efficient price at some point in the middle of the sample. As in the filtering case, we could start with $p_t - cq_t$ and form a linear expectation of q_t:

$$q_t = ... + \beta_{-1}\,\epsilon_{t+1} + \beta_0\,\epsilon_t + \beta_1\,\epsilon_{t-1} + ... + \nu_t$$

Rather than pursue this line of development, though, we'll defer discussion of the smoother until a later section.

7.e How closely does p_t track m_t?

■ Overview

We've motivated the c parameter in the model as a cost variable. If "customers" come in and trade against "dealer" bids and asks, then c is the amount by which a customer buyer overpays relative to the efficient price (and similarly for a customer seller). This does not imply that terms of trade or unfair, or that dealers make profits after their costs, but it does imply a clear distinction between those who supply liquidity and those who demand it.

Many markets, though, don't have such a clean dichotomy between "dealer" and "customer". In limit-order-book markets, bids and asks are set by other customers. Sometimes we consider the customers who supply liquidity as quasi-dealers, i.e., dealers in all but name. More generally, though, a customer in such a market has a choice between using a market or a limit order, and (if a limit order) how it is to be priced. In such markets, the dealer/customer or liquidity supplier/demand roles become blurry.

Even when we can't directly impute a cost to either side in trade, though, it is still of interest to know how closely the trade prices track the efficient price. This is measured by $\mathrm{Var}(s_t) \equiv \sigma_s^2$ where $s_t = p_t - m_t$.

■ σ_s^2 in the generalized Roll model

The structural model implies $s_t = q_t\,c$, so $\sigma_s^2 = c^2$. Unfortunately, since c is not identified by the data, σ_s^2 isn't either. It does possess, however, a lower bound.

To see this, note first that

$$s_t = p_t - m_t = (p_t - f_t) - (m_t - f_t) \tag{7.e.32}$$

Now since f_t is constructed from $\{p_t,\ p_{t-1},\ ...\}$, the filtering error $m_t - f_t$ is uncorrelated with $p_t - f_t$. Therefore

$$\sigma_s^2 = \mathrm{Var}(p_t - f_t) + \mathrm{Var}(m_t - f_t) \tag{7.e.33}$$

Next we use the property that $f_t = p_t + \theta\epsilon_t$ is not dependent on the structural model parameters. This means that the first term on the r.h.s. is invariant. Furthermore, under one parameterization (that of exclusively private information, $u_t = 0$), $m_t - f_t = 0$. This parameterization defines the lower bound.

Specifically, if $u_t = 0$, we've seen that $m_t = p_t + \theta\epsilon_t$, so $\sigma_s^2 = \theta^2\,\sigma_\epsilon^2 = c^2$. To establish the last equality, recall that we have a mapping from the structural parameters to the autocovariances, and from the autocovariances to the moving average parameters. Using the earlier results, $\theta^2\,\sigma_\epsilon^2$ is:

$$\theta^2\sigma_\epsilon^2 = \frac{1}{2}\left(\gamma_0 - \sqrt{\gamma_0^2 - 4\,\gamma_1^2}\right) \tag{7.e.34}$$

The lower bound is:

$$\underline{\sigma_s^2} = \frac{1}{2}\left(c^2 + (c+\lambda)^2 + \sigma_u^2 - \sqrt{(\lambda^2 + \sigma_u^2)\,(4\,c^2 + 4\,\lambda\,c + \lambda^2 + \sigma_u^2)}\right) = c^2 \tag{7.e.35}$$

So in the case of exclusively private information, the lower bound is correct.

In the case of exclusively public information ($\sigma_u^2 \neq 0,\ \lambda = 0$), though, the lower bound is (in terms of the structural parameters):

$$\frac{1}{2}\left(2\,c^2 + \sigma_u^2 - \sqrt{\sigma_u^2\,(4\,c^2 + \sigma_u^2)}\right) \tag{7.e.36}$$

This is not equal to c^2, the structurally-correct answer.

Does there exist an upper bound?

In general, no. The problem is that there are many alternative structural models that are observationally equivalent (have the same θ and σ_ϵ^2). For example, consider $p_t = m_{t-2} + cq_t$. Here, trade price is driven by an efficient price that is two periods "stale". The difference $s_t = p_t - m_t = -w_t - w_{t-1} + cq_t$, and its variance is inflated by $2\,\sigma_w^2$.

This does not affect the lower bound result. In the present case, we can write

$$s_t = p_t - m_t = p_t - (m_{t-2} + w_{t-1} + w_t) = (p_t - f_t) + (f_t - m_{t-2}) - (w_t + w_{t-1}) \tag{7.e.37}$$

Here, given the lagged dependence, neither p_t nor f_t depend on $\{w_t,\ w_{t-1}\}$. The lower bound will understate the true σ_s^2 by $2\,\sigma_w^2$.

Now one can make economic arguments that it is unlikely that the price is extremely lagged relative to beliefs. Were quotes set relative to yesterday's efficient price, customers would be unwilling to trade on one side of the market. Arguments like this might justify at least a provisional assumption about how stale the price is likely to be. The point here is that the arguments must be based on economics, not statistics. Statistical analysis does not provide an upper bound.

Chapter 8. Univariate random-walk decompositions

The previous section generalized the Roll model to incorporate asymmetric information effects, and then examined the implications of the more general structural model for the reduced-form ("statistical") time-series representation of the price changes.

The present section generalizes these results. Rather than start with a structural model, though, we take a more empirical perspective. That is, we start without knowing the structural model. We begin with a moving-average representation for the price changes. This is not as restrictive as it might appear. If the price changes are covariance-stationary, then we know by the Wold theorem that such a representation exists. It may also be identified and estimated in a straightforward fashion. From the MA representation, then, we'll attempt to draw economically meaningful inferences.

This is an important approach because our existing structural models are not comprehensive and realistic. Trading processes are so complex as to make definitive structural models unattainable. This is not to say that the pursuit of such models is pointless, only to suggest that the statistical models implied by them are likely to be misspecified. Statistical time series models impose less structure on the data, and may therefore be more robust.

The key results are that MA representation for the price changes suffices to identify the variance of the implicit efficient price (σ_w^2), the projection of the efficient price on past price changes, and a lower bound on the variance of the difference between the transaction price and the efficient price. It is important that these quantities can be constructed without further economic assumptions about the model.

8.a Overview

In empirical microstructure studies, we often need to construct a proxy for an unobservable "efficient" price and examine the joint dynamics of this proxy and (often) some information set. Random-walk decompositions are especially useful here. The present development is based on Watson (1986).

The framework is one in which an observed integrated time series contains both random-walk and stationary components. Watson's perspective is a macroeconomic one: the random-walk component represents the long-term trend and the stationary component reflects the business cycle. The macroeconomic orientation accounts for the trend/cycle terminology, and the illustration is an application to GNP.

In our setting, the random-walk component is m_t, economically interpreted as the "efficient" price in the sense of market beliefs conditional on all public information.

$$m_t = m_{t-1} + w_t, \tag{8.a.1}$$

where the w_t reflect new information. The observed series is the price:

$$p_t = m_t + s_t \tag{8.a.2}$$

where s_t is a zero-mean covariance stationary process. In the basic Roll model, $s_t = cq_t$, independent of w_t. More generally, though, s_t can be serially-correlated and correlated with w_t. We represent this as:

$$s_t = \theta_w(L)\, w_t + \theta_\eta(L)\, \eta_t \ , \tag{8.a.3}$$

where η_t and w_t are uncorrelated at all leads and lags. $\theta_w(L)$ and $\theta_\eta(L)$ are lag polynomials. Note that since w_t is already fixed as the random-walk innovation, we can't generally normalize so that the leading term in $\theta_w(L)$ is unity: $\theta_w(L) = \theta_{w,0} + \theta_{w,1}\, L + \theta_{w,2}\, L^2 + \ldots$. In the second term, though, we can scale η_t so that the first term in $\theta_\eta(L)$ is unity.

In economic terms, s_t impounds all microstructure effects of a transient nature that might cause observed prices to deviate from optimal beliefs. s_t will impound, for example, fixed transaction costs, price effects stemming from inventory control, lagged adjustment, etc.

This is a structural model: we can observe p_t, but not s_t and m_t. In terms of the structural model,

$$\Delta p_t = w_t + (1 - L)\, s_t = (1 + (1 - L)\, \theta_w(L))\, w_t + (1 - L)\, \theta_\eta(L)\, \eta_t \tag{8.a.4}$$

The statistical model for the Δp_t is a moving-average process:

$$\Delta p_t = \theta(L)\, \epsilon_t \tag{8.a.5}$$

If we'd started with an autoregressive model for the price-change series, $\phi(L)\, \Delta p_t = \epsilon_t$, then we'd set $\theta(L) = \phi(L)^{-1}$ and continue.

The challenge is to make inferences about the structural model from the statistical one.

8.b The autocovariance generating function

The autocovariance generating function is a tool that will be used frequently in developing the general properties of random-walk decompositions. The following summarizes material in Hamilton, pp. 61-67.

The autocovariances of a time series $\{x_t\}$ are $\gamma_i \equiv \mathrm{Cov}(x_t,\, x_{t-i})$ for $i = \ldots, -1, 0, 1, \ldots$. We're implicitly assuming that the series is covariance-stationary, so γ_i does not depend on t. Furthermore, for a real-valued time series, $\gamma_i = \gamma_{-i}$.

The autocovariance generating function of x is defined as the polynomial:

$$g_x(z) = \ \ldots + \gamma_{-2}\, z^{-2} + \gamma_{-1}\, z^{-1} + \gamma_0 + \gamma_1\, z + \gamma_2\, z^2 + \ldots \tag{8.b.6}$$

The autocovariance generating function is a concise and useful way of representing the dynamic structure of the series.

Sometimes we can compute the γs by analysis of the structural model that generated the time series. Often, though, we just have a statistical representation for the series. We analyze these cases as follows.

Suppose that the series can be represented as a moving average model, $x_t = \theta(L)\,\epsilon_t$ where L is the lag operator and $\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \ldots$. Then $g_x(z) = \theta(z^{-1})\,\theta(z)\,{\sigma_\epsilon}^2$. For example, the first-order moving average arises in connection with the Roll model: $\Delta p_t = \epsilon_t + \theta\epsilon_{t-1} = (1 + \theta L)\,\epsilon_t$. So plugging in (and collecting powers of z) gives:

$$g_x(z) = z\,\theta\,\sigma_\epsilon^2 + \frac{\theta\,\sigma_\epsilon^2}{z} + (\theta^2 + 1)\,\sigma_\epsilon^2 \tag{8.b.7}$$

From which it is clear that $\gamma_0 = (1 + \theta^2)\,\sigma_\epsilon^2$ and $\gamma_1 = \gamma_{-1} = \theta\sigma_\epsilon^2$.

Alternatively, suppose that x_t can be represented by an autoregressive model, $\phi(L)\,x_t = \epsilon_t$, where $\phi(L) = 1 + \phi_1 L + \phi_2 L^2 + \ldots$. Then

$$g_x(z) = \phi(z^{-1})^{-1}\,\phi(z)^{-1}\,{\sigma_\epsilon}^2$$

Intuitively, we can convert the autoregressive representation to a moving average one, $x_t = \phi(L)^{-1}\,\epsilon_t$, and then use the previous result for moving average processes. The autocovariance generating function for an autoregressive model is slightly more involved than that of a moving average model. Because of the inversion, we usually have to construct an infinite-order expansion for $\phi(z)^{-1}$.

For example, consider the first-order autoregressive process $x_t = -\phi\,x_{t-1} + \epsilon_t$, or $\phi(L)\,x_t = \epsilon_t$ where $\phi(L) = (1 + \phi L)$. The series expansion for $\phi(z)^{-1}$ (around zero, through the fifth order is) is:

$$1 - \phi\,z + \phi^2\,z^2 - \phi^3\,z^3 + \phi^4\,z^4 - \phi^5\,z^5 + O(z^6) \tag{8.b.8}$$

The expansion of $\phi(z^{-1})^{-1}$ is:

$$1 - \frac{\phi}{z} + \phi^2\left(\frac{1}{z}\right)^2 - \phi^3\left(\frac{1}{z}\right)^3 + \phi^4\left(\frac{1}{z}\right)^4 - \phi^5\left(\frac{1}{z}\right)^5 + O\left(\left(\frac{1}{z}\right)^6\right) \tag{8.b.9}$$

In computing the autocovariance generating function we take the product of the two expansions:

$$\begin{aligned} g_x(z) = {} & \sigma_\epsilon^2\,\phi^{10} + \sigma_\epsilon^2\,\phi^8 + \sigma_\epsilon^2\,\phi^6 - \frac{\sigma_\epsilon^2\,\phi^5}{z^5} + \sigma_\epsilon^2\,\phi^4 + \\ & \sigma_\epsilon^2\,\phi^2 + \sigma_\epsilon^2 + \frac{\sigma_\epsilon^2\,\phi^6 + \sigma_\epsilon^2\,\phi^4}{z^4} + \frac{-\sigma_\epsilon^2\,\phi^7 - \sigma_\epsilon^2\,\phi^5 - \sigma_\epsilon^2\,\phi^3}{z^3} + \\ & \frac{\sigma_\epsilon^2\,\phi^8 + \sigma_\epsilon^2\,\phi^6 + \sigma_\epsilon^2\,\phi^4 + \sigma_\epsilon^2\,\phi^2}{z^2} + \frac{-\sigma_\epsilon^2\,\phi^9 - \sigma_\epsilon^2\,\phi^7 - \sigma_\epsilon^2\,\phi^5 - \sigma_\epsilon^2\,\phi^3 - \sigma_\epsilon^2\,\phi}{z} + \\ & z\,(-\sigma_\epsilon^2\,\phi^{11} - \sigma_\epsilon^2\,\phi^9 - \sigma_\epsilon^2\,\phi^7 - \sigma_\epsilon^2\,\phi^5 - \sigma_\epsilon^2\,\phi^3 - \sigma_\epsilon^2\,\phi) + \\ & z^2\,(\sigma_\epsilon^2\,\phi^{12} + \sigma_\epsilon^2\,\phi^{10} + \sigma_\epsilon^2\,\phi^8 + \sigma_\epsilon^2\,\phi^6 + \sigma_\epsilon^2\,\phi^4 + \sigma_\epsilon^2\,\phi^2) + \\ & z^3\,(-\sigma_\epsilon^2\,\phi^{13} - \sigma_\epsilon^2\,\phi^{11} - \sigma_\epsilon^2\,\phi^9 - \sigma_\epsilon^2\,\phi^7 - \sigma_\epsilon^2\,\phi^5 - \sigma_\epsilon^2\,\phi^3) + \\ & z^4\,(\sigma_\epsilon^2\,\phi^{14} + \sigma_\epsilon^2\,\phi^{12} + \sigma_\epsilon^2\,\phi^{10} + \sigma_\epsilon^2\,\phi^8 + \sigma_\epsilon^2\,\phi^6 + \sigma_\epsilon^2\,\phi^4) + \\ & z^5\,(-\sigma_\epsilon^2\,\phi^{15} - \sigma_\epsilon^2\,\phi^{13} - \sigma_\epsilon^2\,\phi^{11} - \sigma_\epsilon^2\,\phi^9 - \sigma_\epsilon^2\,\phi^7 - \sigma_\epsilon^2\,\phi^5) \end{aligned} \tag{8.b.10}$$

This expression neglects the higher-order terms. In fact, each coefficient of z is an infinite order sum. For example, the coefficient of $z^0 (= 1)$ is:

$$\gamma_0 = \sigma_\epsilon^2 (1 + \phi^2 + \phi^4 + ...) = \sigma_\epsilon^2 \frac{1}{1-\phi^2}.$$

The coefficient of z (which is equal to the coefficient of z^{-1}) is

$$\gamma_1 = \sigma_\epsilon^2 (-\phi - \phi^3 - \phi^5 - ...) = -\sigma_\epsilon^2 \frac{\phi}{1-\phi^2} = -\phi\, \gamma_0.$$

The coefficient of z^2 is $\gamma_2 = \sigma_\epsilon^2 \frac{\phi^2}{1-\phi^2} = -\phi\, \gamma_1$. There's a general recurrence relation: $\gamma_k = -\phi \gamma_{k-1}$.

8.c The random-walk variance

The first result comes from considering the autocovariance generating function for Δp_t in both the statistical and structural representations. From the statistical representation:

$$g_{\Delta p}(z) = \theta(z)\, \theta(z^{-1})\, \sigma_\epsilon^2 \tag{8.c.11}$$

From the structural representation:

$$g_{\Delta p}(z) = (1 + (1 - z)\, \theta_w(z))\, (1 + (1 - z^{-1})\, \theta_w(z^{-1}))\, \sigma_w^2 + (1 - z)\, \theta_\eta(z)\, (1 - z^{-1})\, \theta_\eta(z^{-1})\, \sigma_\eta^2 \tag{8.c.12}$$

In general, the autocovariance generating function for a series (like Δp_t) that is the sum of two component series will involve cross-terms between the components. These cross-terms vanish here because w_t and η_t are uncorrelated processes.

We equate the two representations and set $z = 1$, yielding:

$$\sigma_w{}^2 = \theta(1)^2\, \sigma_\epsilon^2 \tag{8.c.13}$$

The polynomial $\theta(z)$ evaluated at $z = 1$ is simply the sum of the coefficients: $\theta(1) = 1 + \theta_1 + \theta_2 + ...$

We have seen a special case of this result. In the Roll model (with or without trade impacts) it was demonstrated that $\sigma_w^2 = (1 + \theta)^2\, \sigma_\epsilon^2$.

8.d Further identification in special cases

■ The special case of $\theta_\eta(L)\, \eta_t = 0$: Additional results

When the stationary component is driven entirely by w_t, the correspondence between the structural and statistical models is:

$$(1 + (1 - L)\, \theta_w(L))\, w_t = \theta(L)\, \epsilon_t \tag{8.d.14}$$

There is only one source of randomness in the observed series, so w_t and ϵ_t are perfectly correlated. Given the variance result above,

$$w_t = \theta(1)\,\epsilon_t \tag{8.d.15}$$

Using this and expanding both sides of the prior relation yields

$$\begin{aligned} &1 + (1-L)\,\theta_w(L) = 1 + (1-L)\,(\theta_{w,0} + \theta_{w,1}\,L + \theta_{w,2}\,L^2 + ...) = \\ &\quad 1 + \theta_{w,0} + (\theta_{w,1} - \theta_{w,0})\,L - (\theta_{w,2} - \theta_{w,1})\,L^2 - ... = \frac{1}{\theta(1)}\,(1 + \theta_1\,L + \theta_2\,L^2 + \cdots) \end{aligned} \tag{8.d.16}$$

Collecting powers of L on both sides:

$$\begin{aligned} &(1 + \theta_{w,0}) = 1/\theta(1) \\ &(\theta_{w,1} - \theta_{w,0}) = \theta_1/\theta(1) \\ &... \\ &(\theta_{w,k} - \theta_{w,k-1}) = \theta_k/\theta(1) \\ &... \end{aligned} \tag{8.d.17}$$

The solution to this set of equations is:

$$\theta_{w,k} = -\textstyle\sum_{j=k+1}^{\infty} \theta_j/\theta(1) \text{ for } k = 0,\ ... \tag{8.d.18}$$

It's also sometimes convenient to write s_t in terms of the ϵs as

$$s_t = \theta_\epsilon(L)\,\epsilon_t \text{ where } \theta_{\epsilon,k} = -\textstyle\sum_{j=k+1}^{\infty} \theta_j\,. \tag{8.d.19}$$

This development was first presented by Beveridge and Nelson (1981)

In the Roll framework, this special case corresponds to the special case of exclusively private information. Recall that in this case, $\epsilon_t = -\frac{c}{\theta}\,q_t$, so $w_t = -\frac{(1+\theta)\,c}{\theta}\,q_t$. We have $\theta < 0$, so the coefficient of q_t is positive. In the representation of the stationary component, $\theta_w(L) = \frac{-\theta}{1+\theta}$. The stationary component is $s_t = p_t - m_t = p_t - (p_t + \theta\epsilon_t) = -\theta\epsilon_t = cq_t$.

Alternatively, we can obtain the stationary component as $\theta_w(L)\,w_t = -\frac{-\theta}{(1+\theta)}\,\frac{(1+\theta)\,c}{\theta}\,q_t = cq_t$.

■ The special case of $\theta_w(L) = 0$

Here, the stationary component is uncorrelated with w_t. The correspondence between structural and statistical models is

$$w_t + (1-L)\,\theta_\eta(L)\,\eta_t = \theta(L)\,\epsilon_t \tag{8.d.20}$$

The autocovariance generating functions of both sides must be equal:

$$\sigma_w^2 + (1-z)\,\theta_\eta(z)\,\theta_\eta(z^{-1})\,(1-z^{-1}) = \theta(z)\,\theta(z^{-1})\,\sigma_\epsilon^2 \tag{8.d.21}$$

The $\theta_\eta(L)$ coefficients are determined by solving this equation. In the Roll framework, this corresponds to the special case of exclusively public information.

8.e Smoothing (optional)

■ General setup

Watson's equation (3.1) states that the linear smoothed state estimate here is

$$E[m_t \mid \ldots p_{t+1},\ p_t,\ p_{t-1},\ \ldots] = \sum_{k=-\infty}^{\infty} v_k\ p_{t+k} \tag{8.e.22}$$

where the v_i are the coefficients in the polynomial

$$V(z) = \sigma_w^2[1 + (1 - z^{-1})\,\theta_w(z^{-1})][\theta(z)\,\theta(z^{-1})\,\sigma_\epsilon^2]^{-1} \tag{8.e.23}$$

(using present notation).

We'll use this formula for smoothing and filtering in the generalized Roll model, where $\theta(L) = 1 + \theta L$. We'll construct the smoother for two special cases. The derivations are informal ones. In particular, we'll be asserting the behavior of infinite series based on examination of the leading terms.

■ Exclusively private information

For this model, $\theta_w(z) = \frac{-\theta}{(1+\theta)}$, i.e., there is no dependence on z here. Furthermore, $\frac{\sigma_w^2}{\sigma_\epsilon^2} = (1+\theta)^2$.

Here's a low-order expansion of $V(z)$:

$$\begin{aligned}&\frac{(\theta+1)\,\theta^{11}}{z} - \frac{(\theta+1)\,\theta^{10}}{z^2} + \frac{(\theta+1)\,\theta^9}{z^3} - \frac{(\theta+1)\,\theta^8}{z^4} + \frac{(\theta+1)\,\theta^7}{z^5} - \\ &\frac{(\theta+1)\,\theta^6}{z^6} - z^5\,(\theta+1)\,(\theta^{10}+\theta^8+\theta^6+\theta^4+\theta^2+1)\,\theta^5 + z^3\,(\theta+1)\,(\theta^{12}-1)\,\theta^3 + \\ &z\,(\theta+1)\,(\theta^{12}-1)\,\theta + (\theta+1)\,(1-\theta^{12}) + z^2\,(\theta+1)\,(\theta^2-\theta^{14}) + z^4\,(\theta+1)\,(\theta^4-\theta^{16})\end{aligned} \tag{8.e.24}$$

In the development, we'll be using a higher-order expansion of $V(z)$ where the output is (mercifully) suppressed. (The nuts and bolts are visible in the *Mathematica* version of this document.)

The "center" term in $V(z)$ is the coefficient of p_t in the smoother, and is equal to $1 + \theta$. The coefficient of z^{-1} in $V(z)$, the coefficient of p_{t-1} in the smoother, is $-\theta(1+\theta)$. The coefficient of z in $V(z)$ is the coefficient of p_{t+1} (a lead term) in the smoother, and is equal to zero. In fact, all of the coefficients of p_{t+k} for $k > 0$ are zero. The coefficient of z^{-2} in $V(z)$ is the coefficient of p_{t-2} in the smoother: $\theta^2(1+\theta)$.

The pattern of coefficients appears to be:

$$E[m_t \mid \ldots,\ p_{t+1},\ p_t,\ p_{t-1},\ \ldots] = (1+\theta)\ p_t - \theta(1+\theta)\ p_{t-1} + \theta^2(1+\theta)\ p_{t-2} - \ldots \tag{8.e.25}$$

The coefficients of lagged prices decline exponentially. Furthermore, the sum of the coefficients is equal to unity. Thus, we have a one-sided exponentially weighted average. Since the smoother is one-sided, the filter and smoother are identical.

Mathematica

Another way of viewing the filter/smoother here is:

$$\begin{aligned} &E[m_t \mid \ldots,\ p_{t+1},\ p_t,\ p_{t-1},\ \ldots] = (1+\theta)\ p_t - \theta(1+\theta)\ p_{t-1} + \theta^2(1+\theta)\ p_{t-2} - \ldots \\ &= p_t + \theta((p_t - p_{t-1}) - \theta(p_{t-1} - p_{t-2}) + \theta^2(p_{t-2} - p_{t-3}) - \ldots) \\ &= p_t + \theta(\Delta p_t - \theta \Delta p_{t-1} + \theta^2 \Delta p_{t-2} - \ldots) \\ &= p_t + \theta\epsilon_t \\ &= m_t \end{aligned} \tag{8.e.31}$$

So the filter agrees with what we've previously derived. It is exact.

■ Exclusively public information

In this case $w_t = u_t$, and $\theta_w(L) = 0$. It turns out that the smoother has a particularly simple form. (Again, the nuts and bolts are visible in the *Mathematica* version.)

The zeroth order term (coefficient of p_t) is:

$$(\theta+1)^2\ (\theta^{20} + \theta^{18} + \theta^{16} + \theta^{14} + \theta^{12} + \theta^{10} + \theta^8 + \theta^6 + \theta^4 + \theta^2 + 1) \tag{8.e.32}$$

Assuming that the series is infinite, this simplifies to:

$$\frac{(\theta+1)^2}{1-\theta^2} \tag{8.e.33}$$

The coefficient of z (coefficient of p_{t+1} in the smoother) is:

$$(\theta+1)^2\ (-\theta^{21} - \theta^{19} - \theta^{17} - \theta^{15} - \theta^{13} - \theta^{11} - \theta^9 - \theta^7 - \theta^5 - \theta^3 - \theta) \tag{8.e.34}$$

$$-\frac{\theta\,(\theta+1)^2}{1-\theta^2} \tag{8.e.35}$$

The coefficient of z^{-1} (the coefficient of p_{t-1} in the smoothed average) is identical.

The coefficient of z^2 (coefficient of p_{t+2}) is:

$$(\theta+1)^2\ (\theta^{22} + \theta^{20} + \theta^{18} + \theta^{16} + \theta^{14} + \theta^{12} + \theta^{10} + \theta^8 + \theta^6 + \theta^4 + \theta^2) \tag{8.e.36}$$

$$-\frac{\theta^2 (\theta+1)^2}{\theta^2 - 1} \tag{8.e.37}$$

The established pattern suggests that the smoother is:

$$E[m_t \mid \ldots,\ p_{t+1},\ p_t,\ p_{t-1},\ \ldots] =$$

$$\cdots + \theta^2 \frac{(1+\theta)^2}{1-\theta^2}\ p_{t+2} - \theta \frac{(1+\theta)^2}{1-\theta^2}\ p_{t+1} + \frac{(1+\theta)^2}{1-\theta^2}\ p_t - \theta \frac{(1+\theta)^2}{1-\theta^2}\ p_{t-1} + \theta^2 \frac{(1+\theta)^2}{1-\theta^2}\ p_{t-2} + \ldots \tag{8.e.38}$$

The smoothed estimate of m_t has exponentially declining weights. The sum of the coefficients is unity.

So the smoothed estimate of m_t is a *double-sided* exponentially-weighted average of the prices.

8.f Filtering

Watson shows that given the statistical model, all compatible structural models have the same filter. That is, the coefficients of the current and lagged prices in the projection $E[m_t \mid p_t,\ p_{t-1},\ ...]$ do not depend on knowing $\theta_w(L)$, $\theta_\eta(L)$ and σ_η^2 in eq. (3). In the case where $\theta_\eta(L)\,\eta_t = 0$, the filter is without error $m_t = E[m_t \mid p_t,\ p_{t-1},\ ...]$.

In the generalized Roll model, we defined $f_t = E[p_{t+1} \mid p_t,\ p_{t-1},\ ...] = p_t + \theta\epsilon_t$. In the subcase where all information was trade related ($u_t = 0$), we showed that $f_t = m_t$. In the subcase where all information was public ($\lambda = 0$), we showed that $f_t = E[m_t \mid p_t,\ p_{t-1},\ ...]$. The Watson result is a generalization of this.

We defined f_t as the expectation of next period's price. More generally,

$$f_t = \lim_{k\to\infty} E[p_{t+k} \mid p_t,\ p_{t-1},\ ...] = E[m_t \mid p_t,\ p_{t-1},\ ...] \tag{8.f.39}$$

That is, the method of construction we used in the Roll model is generally applicable, and gives us the optimal linear filter. (See Beveridge and Nelson).

This is an important result. Countless empirical studies examine the impact of some informational datum on a security price. The Watson result (and its multivariate generalization) assert that we can identify a component of the price, f_t, that behaves as a martingale. We can't claim that this is the true efficient price, i.e., the expectation formed in agents' minds. The Watson result tells us, though, that we can at least identify the projection of this price on a given information set. This is often enough to support a compelling economic story.

8.g Variance of the pricing error: σ_s^2

As in the generalized Roll model,

$$s_t = p_t - m_t = (p_t - f_t) - (m_t - f_t) \tag{8.g.40}$$

The two r.h.s. terms are orthogonal; f_t is identified, and $m_t - f_t = 0$ for one special case ($\theta_\eta(L)\,\eta_t = 0$). The value of σ_s^2 computed in this special case thus establishes a lower bound. There is, for the same reason as in the generalized Roll analysis, no upper bound.

■ Other approaches

There is a long tradition in empirical finance of measuring market efficiency (informational and operational) by measuring or assessing how closely security prices follow a random walk. Statistical measures commonly focus on autocovariances, autocorrelations or variance ratios.

The autocovariances and autocorrelations of a random-walk are zero at all non-zero leads and lags. This makes for a clean null hypothesis, and there exist a large number of tests to evaluate this null. But if a random-walk is rejected (and in microstructure data it usually is), how should we proceed. Statistical significance (rejecting the null) does not imply economic significance. It is difficult to reduce a set of autocovariances and autocorrelations to a single meaningful number.

One approach is to compare the variances of returns computed at different intervals or endpoints. It was noted above that transaction price returns computed over long horizons are dominated by the random-walk component. A variance ratio compares the variance per unit time implied by a long horizon with a variance per unit time computed from a short horizon:

$$V_{M,N} = \frac{\frac{\mathrm{Var}(p_t - p_{t-M})}{M}}{\frac{\mathrm{Var}(p_t - p_{t-N})}{N}} \tag{8.g.41}$$

where $M,\ N > 0$. If p_t follows a random-walk, $V_{M,N} = 1$ for all M and N. Usually, though if microstructure effects dominate short-horizon returns, then typically, with $M < N,\ V_{M,N} > 1$. That is, microstructure effects inflate the variance per unit time in the short run. If we set N large and examine how $V_{M,N}$ changes as M goes from 1 to N, $V_{M,N}$ generally declines. In a sense, then, this can summarize how quickly (in terms of return interval) the prices come to resemble a random walk. As a single summary statistic, though, $V_{M,N}$ is problematic. There are few principles to apply in choosing M and N. Furthermore, negative autocorrelation at some lags can be offset by positive correlation at others, resulting in a $V_{M,N}$ near unity, even though the process exhibits complicated dependent behavior.

Variance ratios are also computed when the horizons are the same, but endpoints differ. In some markets, for example, the first and last trades of the day occur using different mechanisms, Typically, the opening price (first trade) is determined using a single-price call, and the closing price is that last trade in a continuous session. The relative efficiencies of the two mechanisms are sometimes assessed by variance ratios like

$$\frac{\mathrm{Var}(p_t^{\mathrm{Open}} - p_{t-1}^{\mathrm{Open}})}{\mathrm{Var}\,(p_t^{\mathrm{Close}} - p_{t-1}^{\mathrm{Close}})} \tag{8.g.42}$$

Studies along these lines include Amihud and Mendelson (1987, 1990, 1991) and Ronen (1998).

8.h Problems

Problem 8.1 Stale prices

The beliefs of market participants at time t are given by $m_t = m_{t-1} + w_t$. But due to slow operational systems, trades actually occur relative to a stale price: $p_t = m_{t-1} + c\,q_t$. Assume that w_t and q_t are uncorrelated at all leads and lags. What is the moving average representation of Δp_t. From this representation, determine σ_w^2.

Answer

Problem 8.2 Lagged adjustment

The beliefs of market participants at time t are given by $m_t = m_{t-1} + w_t$. But due to slow operational systems, trade prices adjust to beliefs gradually:

$$p_t = p_{t-1} + \alpha(m_t - p_{t-1}).$$

There's no bid-ask spread (see the next problem). What is the autoregressive representation for Δp_t? What is σ_w^2 (in terms of the parameters of the AR representation)?

Answer

Problem 8.3 Lagged adjustment with a bid-ask spread

The beliefs of market participants at time t are given by $m_t = m_{t-1} + w_t$. But due to slow operational systems, prices adjust gradually. The adjustment process is as follows. There is a notional price level, π_t, that adjusts toward m_t:

$$\pi_t = \pi_{t-1} + \alpha(m_t - \pi_{t-1}) \tag{8.h.43}$$

Intuitively, π_t may be thought of as the quote midpoint. Actual transaction prices occur as:

$$p_t = \pi_t + cq_t \tag{8.h.44}$$

where q_t and w_t are uncorrelated. What is the process for Δp_t? (It will have both autoregressive and moving average terms.) What is σ_w^2?

This is a special case of Hasbrouck and Ho (1987), which is in turn based on Beja and Goldman (1980). HH also allow for autocorrelated trades, in which case Δp_t is ARMA(2,2).

Answer

Chapter 9. Estimation of time series models

The material to this point has mostly dealt with the correspondence between structural and statistical representations of $\{\Delta p_t\}$. Given a statistical MA(1) model for the $\{\Delta p_t\}$, we could compute structural parameters σ_w^2, $\underline{\sigma_s^2}$, and (if we make the appropriate restrictive assumptions), $\{\lambda,\ c,\ \sigma_u^2 = 0\}$ or $\{c,\ \sigma_u^2,\ \lambda = 0\}$.

We now turn to estimation of structural parameters, based on a sample of prices $\{p_0,\ p_1,\ \ldots\ p_T\}$. We'd at least like consistent estimates of the structural parameters. In addition, for hypothesis testing, we'd like distributional results as well.

The overall estimation strategy will involve first estimating the MA model, and then transforming the MA estimates into estimates of the structural parameters.

9.a Estimating the MA model.

■ Maximum likelihood

Standard discussions of estimation in time series models usually focus on maximum likelihood methods for Gaussian processes. (Hamilton, Ch. 5.) This is generally appropriate for macroeconomic applications, where

- Normality is, if not a proven property, at least a tenable assumption.
- We can compute a likelihood function that is exact in small samples. Macroeconomic applications often have relatively few observations.

In microstructure price data, though, normality is not a plausible assumption. The price grid is coarse relative to the observations. U.S. equity prices, for example, are quoted in $0.01 increments (ticks), and successive price changes are mostly zero, one or two ticks in magnitude.

Furthermore, having an exact likelihood function is less important here. Observations are typically so numerous that asymptotic ("large sample") properties of estimators are more closely attained.

Therefore, in microstructure applications, we usually work with moment estimates. Within this class of estimators, there are two common approaches.

■ Direct moment estimates

MA parameters may be estimated directly using generalized method of moments (GMM, Hamilton, Ch. 14). Consider the MA(1) process $x_t = \epsilon_t + \theta\epsilon_{t-1}$. The natural moment conditions are those that define the autocovariances:

$$\begin{aligned} \gamma_0 &= E x_t^2 = (1+\theta^2)\,\sigma_\epsilon^2 \\ \gamma_1 &= E x_t\, x_{t-1} = \theta\sigma_\epsilon^2 \end{aligned} \tag{9.a.1}$$

Essentially, GMM picks θ and σ_ϵ^2 values that minimize

$$\begin{aligned} &\frac{1}{T}\,\Sigma x_t^2 - (1+\theta^2)\,\sigma_\epsilon^2 \\ &\text{and} \\ &\frac{1}{T}\,\Sigma x_t\, x_{t-1} - \theta\sigma_\epsilon^2 \end{aligned} \tag{9.a.2}$$

GMM also provides distributional results.

This is a sensible and practical way to estimate an MA(1) model.

The approach becomes less attractive for more complex models. For a moving average process of order q, denoted MA(q), there are $q+1$ parameters. There are also $q+1$ nonzero autocovariances. There are 2^q sets of parameters that will generate these autocovariances, only one of which is invertible. Even when q is modest, this a numerically-challenging exercise. When we extend the framework to model multivariate (vector) processes, the dimensionality of the problem increases further.

■ Estimation based on autoregression

We've seen that an MA(1) model possesses an equivalent autoregressive representation:

$$x_t = \epsilon_t + \theta\epsilon_{t-1} \Leftrightarrow\ x_t = -\theta x_{t-1} + \theta^2\, x_{t-2} - \theta^3\, x_{t-3} + \cdots + \epsilon_t \tag{9.a.3}$$

The autoregressive representation can be used as a basis for estimation. Generally, if an MA representation is of finite order, then the AR representation is of infinite order (and vice versa). This is the case here. We nevertheless note that the AR coefficients are declining geometrically, and that we might obtain a good approximation by truncating the representation at some point K.

Such a specification looks like this:

$$x_t = \phi_1\, x_{t-1} + \phi_2\, x_{t-2} + \cdots + \phi_K\, x_{t-K} + \epsilon_t{}^a \tag{9.a.4}$$

where the a superscript attached to ϵ_t denotes "approximate". This specification may be consistently estimated by ordinary least squares.

The Wold theorem ensures that if the AR model is correctly specified, the disturbances are serially uncorrelated and homoscedastic. That is, in computing Var(($\hat{\phi}_1$ $\hat{\phi}_2$ ... $\hat{\phi}_K$)) there is no reason to use anything more complicated than the usual OLS estimates of the coefficient covariance matrix. The possibility of misspecification, though, might militate in favor of a more general approach. Specifically, if our choice of K is lower than the true value or if σ_ϵ^2 has deterministic variation, then the ϵ_t^a might be serially correlated and/or heteroscedastic. A White or Newey-West estimate might be used instead.

There's one other small problem. If we know that the true statistical model is MA(1), then in estimating the equivalent AR specification, we should constrain the AR coefficients to follow the geometrically-declining pattern implied by the moving average specification.

In practice, though, the AR approach is generally used in less structured situations, when we don't know the order of the MA specification. In this case, we try to set K large enough to ensure that the ϵ_t in the AR specification are not serially correlated. We then invert the estimated AR representation to obtain the MA parameters.

There are two ways of performing this inversion. Both have their uses.

First, we can invert the AR lag polynomial. The compact form of the AR representation is $\phi(L)\, x_t = \epsilon_t$ where $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_K L^K$. Inverting gives $x_t = \phi(L)^{-1}\, \epsilon_t$. Thus, $\theta(L) = \phi(L)^{-1}$. The inversion is carried out by series expansion. This approach is useful when we don't need the full MA representation. Recall that in $\Delta p_t = \theta(L)\, \epsilon_t$, the variance of the random-walk component is $\sigma_w^2 = \theta(1)^2\, \sigma_\epsilon^2$. Since we only need the sum of the MA coefficients, we may compute $\sigma_w^2 = \sigma_\epsilon^2 / \phi(1)$, where $\phi(1)$ is the sum of the autoregressive coefficients. That is, we only need to invert the *sum* of the AR coefficients, not the full polynomial.

The second approach is based on forecasting. Given an MA(q) representation $x_t = \theta(L)\, \epsilon_t$, the forecast, in the sense of the linear expectation (projection) conditional on a given disturbance is:

$$E[x_{t+k} \mid \epsilon_t] = \theta_k\, \epsilon_t \tag{9.a.5}$$

This follows from simply taking the expectation of the MA representation, noting that the ϵ_t are uncorrelated.

This forecast may also be computed from the AR representation. Suppose that at time t, we set all lagged xs to their unconditional mean of zero ($x_{t-1} = x_{t-2} = \ldots = 0$). The current observation is then simply $x_t = \epsilon_t$. Noting that $E[\epsilon_{t+1} \mid \epsilon_t] = 0$,

$$E[x_{t+1} \mid \epsilon_t] \;=\; \phi_1\, x_t = \phi_1\, \epsilon_t \tag{9.a.6}$$

Iterating one more step ahead,

$$E[x_{t+2} \mid \epsilon_t] = \phi_1\, E[x_{t+1} \mid \epsilon_t] + \phi_2\, \epsilon_t = (\phi_1^2 + \phi_2)\, \epsilon_t \tag{9.a.7}$$

Etc.

Thus, the coefficients in the MA representation are $\theta_0 = 1$, $\theta_1 = \phi_1$, $\theta_2 = (\phi_1^2 + \phi_2)$, …

The MA coefficients developed in this fashion are also called the impact multipliers. A plot of $E[x_{t+k} \mid \epsilon_t]$ conditional on some ϵ_t (usually $\epsilon_t = 1$) describes the impulse response function of the process.

When the variable is a difference of an integrated series, like Δp_t, it is more natural to compute and plot the cumulative impulse response function $E[\sum_{j=0}^{k} \Delta p_{t+k} \mid \epsilon_t]$. Plotted over time, this quantity depicts the dynamic response of prices to ϵ_t.

In the present context, impact multipliers and moving average coefficients are the same thing. This is not always the case. Impact multipliers can also be computed from a nonstochastic version of an AR model, i.e., one in which the disturbances are suppressed, a linear difference equation.

9.b Structural estimates and their distributional properties

Given estimates $\{\hat{\theta}(L), \hat{\sigma}_\epsilon^2\}$ for the MA parameters, we may form estimates of the structural parameters (e.g., σ_w^2) by solving for these parameters using the estimates in lieu of the true $\{\theta(L), \sigma_\epsilon^2\}$.

There are two approaches to characterizing the distributions of these estimates.

■ The "delta" method

We can construct an asymptotic covariance matrix by the "delta" method (Greene, section 5.2.4 or Cochrane p. 207). The intuition is as follows.

Suppose that we have a random vector distributed as a multivariate normal: $x \sim N(\mu, \Omega)$. Linear transformations of x are also multivariate normal. If A is some $m \times n$ matrix of coefficients, then $Ax \sim N(A\mu, A\Omega A')$.

Now consider the situation where we have a parameter vector θ and we're interested in a (possibly nonlinear) continuous function $f(\theta)$ where f is $m \times 1$.

Suppose that we possess an estimate of θ that is asymptotically normal: $\sqrt{T}\left(\hat{\theta} - \theta\right) \sim N(0, \Omega)$. Then

$$\sqrt{T}\left(f(\hat{\theta}) - f(\theta)\right) \sim N(0,\ J\Omega J') \quad \text{where } J = \left(\frac{\partial f_i}{\partial \theta_j}\right)_{i,j}. \tag{9.b.8}$$

In the present case, for example, suppose that we seek estimates of $\{\sigma_w^2, \underline{\sigma_s^2}\}$. We start by estimating an AR model of order K. The $\phi(L)$ coefficients can be estimated by least squares. Denote the coefficient vector by $\phi = (\ \phi_1 \quad \phi_2 \quad \dots \quad \phi_K\ )$, with corresponding estimate $\hat{\phi}$. We can also form estimates $\hat{\sigma}_\epsilon{}^2$ and Var($\hat{\phi}$) by the usual methods.

Since $\{\sigma_w^2, \underline{\sigma_s^2}\}$ both depend on σ_ϵ^2 as well as the ϕ coefficients, however, we'll need to know the joint distribution of $\hat{\Psi} = (\hat{\phi} \quad \hat{\sigma}_\epsilon^2)$. In the normal (Gaussian) case, $\hat{\phi}$ and $\hat{\sigma}_\epsilon^2$ are asymptotically independent (Hamilton, pp. 300-301). We'll also need the function mapping ϕ and σ_ϵ^2 to $\{\sigma_w{}^2, \underline{\sigma_s^2}\}$. We compute the Jacobian of this function (possibly numerically) and apply it to the $\hat{\Psi}$ variance matrix.

This approach can work well if the mapping function is approximately linear. Most of those we work with in microstructure, unfortunately, are not. Random-walk decomposition parameters, impulse response functions, etc., are usually highly nonlinear.

■ Subsampling

An alternative approach involves partitioning the full sample into subsamples, computing an estimate for each subsample, and examining the distributional properties of the subsample estimates. For example, if the T observations span D days, it is natural to form subsamples for each day. We estimate our model (MA, VAR, whatever) for each day and compute any estimates of interest for the day. In the case of the random-walk variance, for example, we would then have a series $\hat{\sigma}_{w,d}^2$ for $d = 1, \ ..., \ D$. We would then compute the mean across days, and the standard error of this mean by the usual methods.

This is formally correct if different days are statistically independent. If we're modeling short-run microstructure effects, this is roughly accurate.

This approach for estimating the properties of time series data was originally advocated by Bartlett (for spectral estimates). In finance, inference based on subsamples in this fashion is generally called the "Fama-McBeth" approach.

■ Starting values

Suppose we're modeling price changes. What should we do with the overnight return?

The naive approach is to simply treat the price sample $\{p_0, \ p_1, \ ..., \ p_T\}$ as an undifferentiated sequence, and make no special provision for cases where, in computing $\Delta p_t = p_t - p_{t-1}$, p_t is the first trade of the day and p_{t-1} is the last trade of the previous day.

Although this usually simplifies the data analysis and programming, it is highly problematic. For one thing, opening and closing prices are often determined by different market mechanisms (e.g., single price call vs. continuous trading). Another consideration is that the overnight dynamics of the efficient price are almost certainly different from those of the trading day. As a general rule, it is better to treat each day as a separate sample, and to discard the first price change.

If we're estimating a VAR of order K, though, we'll need K lagged price changes. Here, one may either set lagged unobserved price changes to zero, or else begin the estimation sample with Δp_K.

Standard approaches to this problem advocate a formal modeling of the initial observations, essentially creating marginal distributions in which any dependence on unobserved values has been integrated out (e.g., Hamilton, Ch. 5). In principle, these approaches assume that the true process has been evolving all along and our sample starts when we begin collecting data. Although this view may be appropriate in macroeconomic data, it is usually far less so in microstructure analyses. At the NYSE, the curtain goes up at 9:30 in the morning. There may have been some prior trading activity, but if so, the dynamics were almost certainly different.

9.c Case study I

Here, we'll download and analyze the TAQ data record for a single ticker symbol on a single data. Each class participant will receive a different symbol and day.

You will need to access WRDS using the supplied account. You'll then extract and download a SAS dataset using the WRDS web interface. You'll then analyze the data using SAS. To do this, you'll need access to SAS on a PC or mainframe. Most of the class will probably be using NYU's Eureka machine. If you're using Columbia machines, please see me. I'll supply a SAS shell for the program you'll need to run. You might be able to run it as is, but it might need a little modification. You'll then take the output of the program and proceed to compute Roll spread estimates and other parameters.

Your write-up should look like the "results" section of an article. That is, there should be (at most) a few pages of summary. The summary should report the key statistics, of course, but should also go a little beyond this. The study calls for you to estimate some simple models. Are these models appropriate *for your stock*? Do they fit the data well? In some cases, the same value (e.g., σ_w^2) is estimated by various approaches. Are there big differences? Why? Etc.

In an article, you'd present the numbers in tables. Here it suffices to attach you SAS output to the back of the summary.

■ Accessing WRDS

Go to the WRDS website at http://wrds.wharton.upenn.edu/. Go to the 'members login' page and log in.

Then ⟶ NYSE TAQ ⟶ 'Consolidated Trades'. In this menu, specify your ticker symbol and your date. Select all datafields. Select as the output format 'SAS dataset'. Submit your request. When the request is processed, download the file.

Next, go to the 'Consolidated Quotes' menu. Again, specify your symbol and date. Select all datafields. Output and download as a SAS dataset.

■ Using SAS

On the course website, there is a SAS shell program (named 'AnalyzeTaq01.sas') for you to work with. There is also a sample listing and log file. You can view and edit these files with any text editor (like notepad).

Download this program and the CT and CQ datasets to the machine where you'll be running SAS. Note: the shell program assumes that the CT dataset is named 'ct2' and the CQ dataset is named 'cq2', both in your home directory. To run the program, at the machine prompt, you'll enter something like 'sas AnalyzeTaq01.sas'. SAS should put its log output in 'AnalyzeTaq01.log' and its listing output in 'AnalyzeTaq01.lst'. Download both of these files to your PC and print them out.

SAS has good online documentation at its website (www.sas.com). You need to register to use it, giving your email address, etc.

■ Analyzing the output

The listing output first contains summary statistics from the CQ file, including means, mins and maxes of absolute spreads ($ per share), log spreads (log(ofr/bid)) and the bid-ask midpoint ('BAM'). These summary spread statistics will be the point of reference for comparing some of the other estimates.

You should compute:

1. The Roll estimate of the spread. The output from 'proc arima' contains the autocovariances you need (based on first differences of log transaction prices). Compare the Roll spread estimate to the primary market (NYSE) average log spread.
   NOTE: Proc arima appears to automatically center autocovariance and autocorrelation estimates around the mean. This is generally not the best choice for microstructure price data, but I don't know of any easy way to turn it off.
2. σ_w^2 and $\underline{\sigma_s^2}$ for the MA(1) model (estimated in 'proc arima'). Report these as standard deviations for ease of interpretation.
3. σ_w^2 and $\underline{\sigma_s^2}$ for the MA(3) model (estimated in 'proc arima'). Report these as standard deviations for ease of interpretation.
4. Finally, find out the name of your company. What was the market return on that day? Was there any news on the company? (Search the Dow-Jones index.)

The assignment is due on Wednesday, November 5. Let me know early on if you're encountering difficulties.

Part II: Multivariate models of trades and prices

To this point, although trade variables have been used in the models, they've entered in fairly simple ways. Furthermore, inference has been based on univariate analyses of price changes.

In this section, we focus more closely on trades and how they are incorporated into specification and estimation.

Chapter 10. The trade process and inventory control

The asymmetric information models address one aspect of trade/price dynamics. The probability that the trade arose from the order of an informed trader gives rise to an immediate and permanent price impact.

In this section we investigate another mechanism, generally termed the inventory control effect. The inventory control models actually predate the asymmetric information models. I discuss in some detail Garman (1976) and Amihud and Mendelson (1980). Related papers include Ho and Macris (1984); Stoll (1976); Stoll (1978); O'Hara and Oldfield (1986); Madhavan and Smidt (1991); Madhavan and Smidt (1993); Hasbrouck and Sofianos (1993); Reiss and Werner (1998).

10.a The dealer as a smoother of intertemporal order imbalances.

Garman (1976) suggests that a dealer is needed because buyers and sellers do not arrive synchronously. In this model, buy and sell orders arrive randomly in continuous time. The arrival processes are Poisson.

■ Background: the exponential/Poisson arrival model

Suppose that an event of some sort (e.g., a buy order arrival) has just occurred. Let τ be the random waiting time until the next occurrence. Suppose that τ is exponentially distributed with parameter λ:

$$f(\tau) = e^{-\tau\lambda}\,\lambda \tag{10.a.1}$$

The exponential distribution has the property that $E[\tau] = 1/\lambda$ and $\mathrm{Var}(\tau) = 1/\lambda^2$. Thus, λ has units of time^{-1}, e.g. "events per second".

A Poisson random variable n with parameter μ has a distribution defined over $n \in \{0, 1, \ldots\}$ as

$$f(n) = \frac{e^{-\mu}\,\mu^n}{n!} \tag{10.a.2}$$

The mean and variance are $E[n] = \mu$ and $\mathrm{Var}[n] = \mu$. If inter-event arrival times are exponentially distributed with parameter λ, then the number of events occurring within a time interval of duration Δ is a Poisson variable with parameter $\mu = \Delta/\lambda$. This framework is often called the Poisson arrival model. λ is the arrival intensity. If λ is measured in seconds^{-1}, and we let $\Delta = 1$, then λ^{-1} is the expected number of events per second.

■ The Garman model

The arrival intensities for buyers and sellers are λ_B and λ_S. These are functions of the prices faced by customers. Suppose that the dealer posts a single price p. Then $\lambda_S(p)$ is monotone increasing and $\lambda_B(p)$ is monotone increasing.

These functions describe supply and demand curves. Demand and supply are not static. Let I_t denote the number of shares held by a dealer, i.e., the dealer's inventory of stock. If there is to be no net drive in I_t, then we must have $\lambda_S = \lambda_B$.

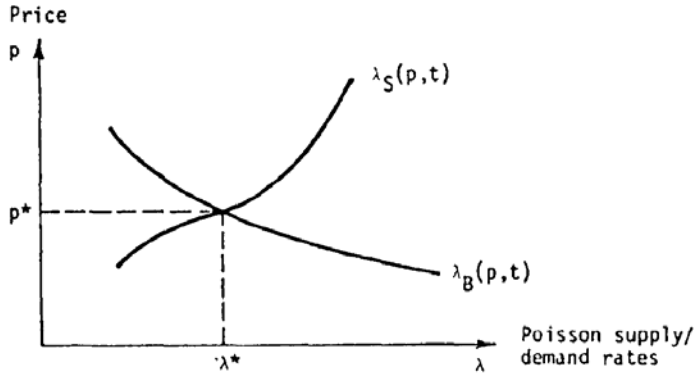


Fig. 1. Stochastic aggregate supply and demand rates as a function of price, at time t.

The sense in which the market "clears" at p^* is that average supply and demand per unit time are equal.

p^* the only *single* equilibrium price, but we're not in a single-price world. Suppose that the dealer can post an ask price, P_B, a price at which buyers trade, and a bid price P_S , at which sellers trade. The condition of equal arrival rates is now $\lambda_S(P_S) = \lambda_B(P_B)$. For the moment, we'll treat this as a constraint on the dealer's pricing strategy.

The dealer earns the spread $P_S - P_B$ on each buyer-seller pair ("the dealer's turn"). From the dealer's perspective, suppose that, subject to equal arrival rates, we set $P_B < p^* < P_S$. By setting a wide spread

- We increase the revenue per buyer-seller pair.
- We decrease the number of traders per unit time.

Revenue per unit time is given by the shaded area:

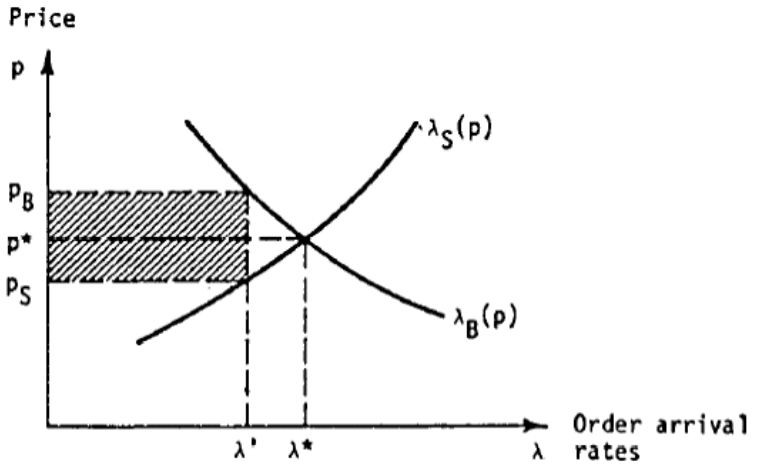


Fig. 2. Market-maker's profit rate under the assumption of no stock inventory drift.

The dealer's inventory of stock is

$$I_s(t) = I_s(0) + N_S(t) - N_B(t) \qquad (10.a.3)$$

where $N_B(t)$ is the cumulative number of trades at the ask (customer buys, dealer sells) through time t; $N_S(t)$ is the cumulative number of trades at the bid (customer sells, dealer buys). $I_s(0)$ is the dealer's starting position. There is a similar expression for the dealer's holding of cash. The key constraint is that dealer holdings of stock and cash cannot drop below zero ("ruin").

Clearly, if $\lambda_S(P_S) = \lambda_B(P_B)$, holdings of stock follow a zero-drift random walk. Cash holdings follow a positive-drift random walk (remember the turn).

Garman points out that if $\lambda_S(P_S) = \lambda_B(P_B)$, the dealer is eventually ruined with probability one. (A zero-drift random-walk will eventually hit any finite barrier with probability one.) Furthermore, with realistic parameter values, the expected time to ruin is a matter of days. The view of equilibrium as a balance of stochastic arrival rates is utilized by Saar (1998).

The practice of modeling buyer and seller arrivals as Poisson event processes is a very active area of empirical research. Modern approaches allow the arrival rate to be time-varying, with the intuition that arrival rate corresponds to informational intensity (Engle and Russell (1998)). Domowitz and Wang (1994) examine the properties of a limit order book where order arrivals at each price are Poisson.

10.b Active inventory control

The dealer in the above variant of the Garman model sets the bid and ask prices once and for all. As he sees an inventory barrier approaching, he simply watches and prays that the barrier isn't hit. Commenting on the short expected failure times implied by this strategy under realistic parameter values, Garman notes, "[T]he order of magnitude makes it clear that the specialists [dealers] must pursue a policy of relating their prices to their inventories in order to avoid failure."

This statement lays out the intuition behind an important aspect of microstructure analysis called the inventory control principle. The essential mechanism is that dealers change their bid and ask quotes in order to elicit an expected imbalance of buy and sell orders, in the direction of restoring their inventories to a preferred position.

In Amihud and Mendelson (1980), the dealer maximizes expected profits per unit time (given risk neutrality). The bid and ask prices as a function of the inventory level are depicted as follows:

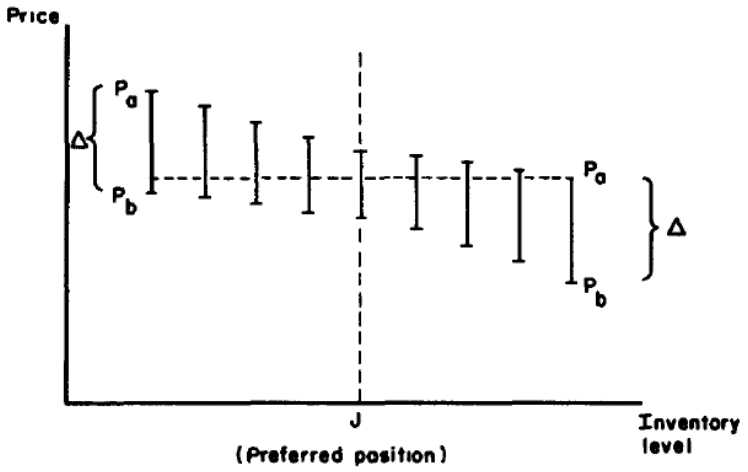


Fig 3 The bid–ask prices and the corresponding spread, Δ, as a function of the market-maker's inventory level

The key results are:

- Bid and ask are monotone decreasing functions of the inventory level.
- Dealer has a preferred position.
- There is a positive spread.
- The spread is increasing in distance from preferred position.
- The bid-ask midpoint is not always where the true value lies.
- Price fluctuations associated with inventory control are transient.
- There are no manipulative strategies.

In both Garman and AM, the spread results from market power.

10.c How do dealer inventories actually behave?

Here are some sweeping generalizations:

Ruins do occur, but infrequently.

Furthermore, in practice ruins aren't usually caused by trades that drive the dealer's inventory into the barrier. Ruin generally arises because security inventories are levered (partially financed with debt). A sudden price movement triggers a default in the dealer's borrowing arrangements. A holding of 200,000 shares may be perfectly okay when the price of the security is $50 per share, but not when the price is $10.

In a sense, ruin is caused not by moving inventory hitting a fixed barrier, but by a moving barrier hitting the existing level of inventory.

- Inventories are mean reverting. They do not follow random-walk-type processes.

  "Mean-reverting" simply means that the process seems to return over time to some long-run average value. A mean-reverting process does not diverge over time (like a random walk). Mean-reversion does not necessarily imply stationarity: the dynamics of the reversion process might change over time.

- Inventory data are difficult to obtain. They reveal market-makers' trading strategies and profit mechanisms.

Here are representative data (Hasbrouck and Sofianos (1993)).

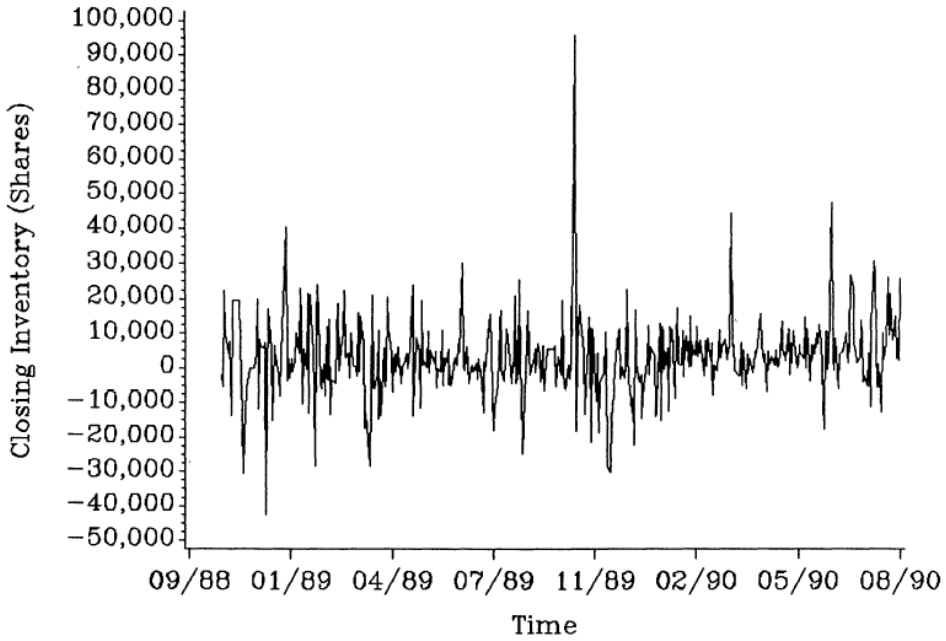


Figure 1. Daily closing specialist inventory in shares for "stock A." Source: NYSE Specialist Performance Evaluation Trade Summary file.

Some salient features:

- Inventory sometimes takes on a negative value (short positions).
- There is no obvious drift or divergence.

- The mean inventory is near zero. A closing inventory
- There is a sharp spike in late 1989. This corresponds to the “mini-crash”.

This inventory graph is well-behaved (in the sense that it corresponds to our economic intuition).

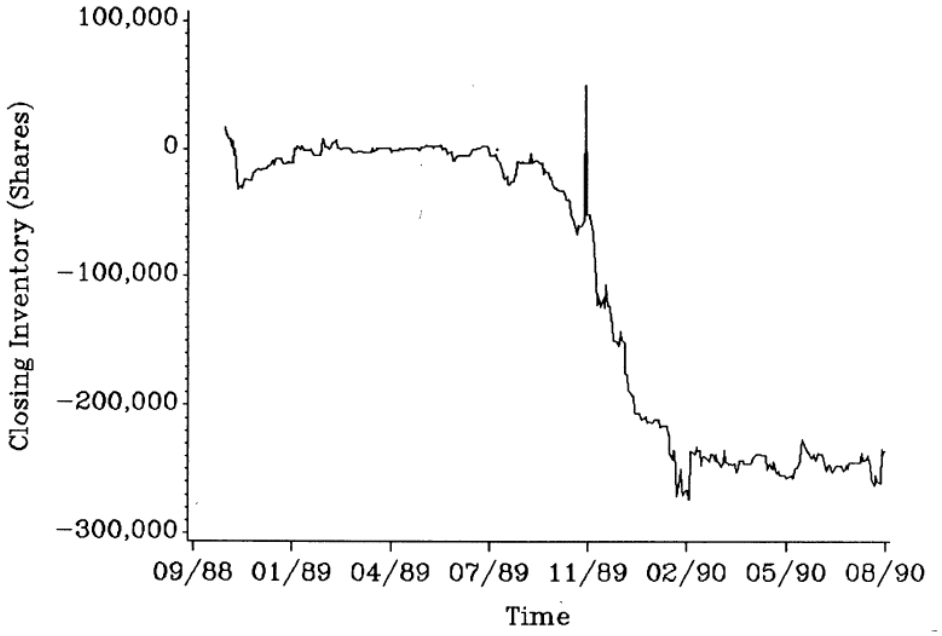


Figure 2. Daily closing specialist inventory in shares for “stock B.” Source: NYSE Specialist Performance Evaluation Trade Summary file.

The long-term component is much larger than the typical daily variation.

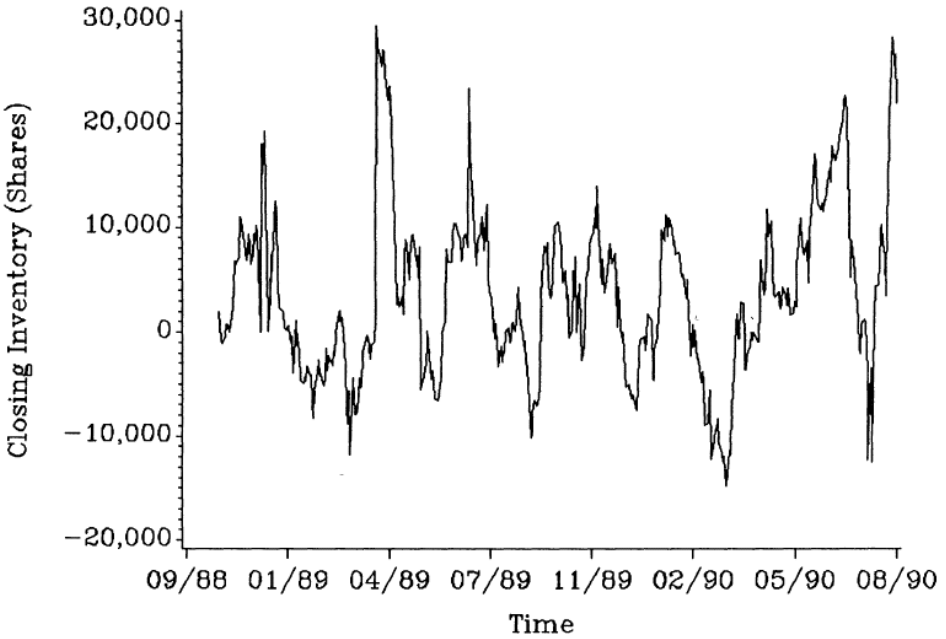


Figure 3. Daily closing specialist inventory in shares for “stock C.” Source: NYSE Specialist Performance Evaluation Trade Summary file.

The inventory appears to be mean-reverting, but has protracted departures from the mean.

■ Is the visible quote the control variable for inventory control?

In both Garman and AM models, the dealer changes his bid and ask to induce an imbalance of incoming orders.

I'll discuss some exceptions below, but as a general rule most empirical analysis of inventory control refutes the basic mechanism. In my experience, when I've sought confirmation of the practice from real-world dealers, my inquiries were met with tolerant amusement. A dealer who would pursue the hypothesized mechanism would be signaling to the world at large his desire to buy or sell. This puts him at a competitive disadvantage.

This doesn't settle matters. Some sort of inventory control must be used because inventories aren't divergent. If the adjustment mechanism isn't quote-based, then what else could it be? Here are some possibilities:

- In many markets, dealer quotes are not publicly available. They are given only in response to an inquiry by a customer or another dealer. It is safer here to reveal a quote that indicates an adjustment desire. The inquiries are not anonymous. If the counterparty (customer or dealer) uses the information against the dealer, he will find that the next time he inquires, the dealer will make a poor market. The implicit (sometimes explicit) message is: "You bagged me on our last deal. I'm quoting wide to you in order to protect myself. And punish you."
- Interdealer brokers (see Reiss and Werner (1998))
- Selectively "going along"
- Eighthing/pennying
- Other anonymous venues.

Nevertheless, although the price-based inventory control mechanism has not proven relevant to dealers, the basic lines of thought have emerged as mainstays of the order strategy literature.

10.d The properties of the trade direction series

It was assumed in the basic Roll model that trade directions were not serially correlated ($\text{Corr}(q_t, q_{t-k}) = 0 \text{ for } k \neq 0$). In practice, however, this variable tends to exhibit striking positive autocorrelation (Hasbrouck and Ho (1987)). Etc.

Chapter 11. Random walks, etc.

The last section noted the connection between the trade direction indicator variable q_t and the dealer's inventory, I_t. Assuming that all buys and sells are for one unit, $I_t = I_{t-1} - q_t$ or $q_t = -\Delta I_t$. (If we wanted to allow for various quantities, we'd just use the signed order volume in lieu of q_t.)

Now if q_t are independent (as assumed by the basic Roll model), then I_t will behave like a random walk. It will tend to diverge over time (as suggested by Garman). But if I_t is covariance stationary, what does that imply about q_t?

The resolution of these questions turns on the concepts of unit roots and invertibility. We develop these, and then revisit the Wold theorem.

11.a Is it a random walk?

How do we know whether a time series is a random walk or stationary?

The question is actually ill-phrased. In the Roll model, for example, the price is neither a random-walk nor stationary. It's a mixture of both sorts of components. A somewhat better question is, how do we know if a time-series contains a random-walk component?

From a statistical viewpoint, however, even this is too vague. For reasons that will become clear in a moment, it's more precise to ask "Does the process contain a unit root?" Formally, the material in this section applies to processes that might have a unit root and are covariance-stationary after first-differencing.

When the seminar speaker says, "the price of stock *XYZ* is nonstationary, so we take first differences before computing our statistics," this is verbal shorthand, and drops some additional assumptions (with which the audience and speaker are presumed to be familiar). In general, you don't make a nonstationary time series stationary simply by first-differencing it.

The the term "unit root" arises in connection with the autoregressive representation for a time series. Consider the autoregressive form of a time series x_t in terms of the lag polynomial:

$$\phi(L)\, x_t = \epsilon_t \text{ where } \phi(L) = 1 + \phi_1 L + \phi_2 L^2 + ... + \phi_K L^K \qquad (11.a.1)$$

The stationarity of x_t depends critically on the form of $\phi(L)$. The criterion is based on the solutions to the polynomial equation $\phi(z) = 0$, i.e., the roots of the lag polynomial with the L operator replaced by a complex variable z.

If any of the solutions are equal to one, then x_t has a random-walk component. In the long run, this component dominates the behavior of the series, causing it to diverge. A solution to $\phi(z)=0$ is called a root.

Hence, we say in this situation that "x_t has a unit root."

Suppose that we factor the polynomial as:

$$\phi(z) = (1 - a_1 z)(1 - a_2 z) \ldots (1 - a_K z) \tag{11.a.2}$$

If $z = 1/a_i$ for $i = 1, \ldots, K$ then $z = 1/a_i$ are the roots of the equation. The criterion is this: if $|a_i| > 1$ for $i = 1, \ldots, K$, i.e., if the roots lie outside of the unit circle, then the process is stationary.

For example, $x_t = 2 x_{t-1} + \epsilon_t$ is autoregressive (that is, linear in past values). It is, however, explosive: we double the last value and add a disturbance. From the polynomial perspective: $\phi(z) = (1 - 2z)$, which is zero when $z = 1/2$. This is inside the unit circle.

The Roll model also provides a nice illustration.

The structural model has the MA representation

$$\Delta p_t = \epsilon_t + \theta \epsilon_{t-1} \tag{11.a.3}$$

or, using the lag operator:

$$(1 - L)\, p_t = \theta(L)\, \epsilon_t, \text{ where } \theta(L) = 1 + \theta L \tag{11.a.4}$$

The autoregressive representation for the price *level* is:

$$\phi(L)\, p_t = \epsilon_t \text{ where } \phi(L) = \theta(L)^{-1} (1 - L) \tag{11.a.5}$$

We can identify at least one root here, and its value is unity. This is not surprising because we built a random walk into the structural model.

But if we didn't know the structural model, we'd have to make an inference based on a sample of data. There are various tests available. In practice we use:

- Economic logic.
- The eyeball test. Does a plot of the series look like it's diverging?
- Statistical unit root tests.

The eyeball and statistical tests are good ones, but it is too easy in microstructure data to conjure up situations in which they would give the wrong answer. In the Roll model, for example, a large trading cost coupled with a small random-walk volatility can generate a sample in which the dominant feature is bid-ask bounce and the sample path is apparently stationary.

11.b Invertibility

Suppose we have a series like a dealer's inventory series, I_t, that can be presumed covariance-stationary. By the Wold Theorem, it possesses a moving average representation: $I_t = \theta(L)\,\epsilon_t$. The first-difference of I_t will be stationary as well. It also possesses a moving average representation: $\Delta I_t = (1-L)\,\theta(L)\,\epsilon_t$

When we encountered in the analysis of the Roll model a series (like p_t or m_t) that was (or contained) a random-walk component, we arrived at stationarity by taking the first difference. Suppose with a dealer's inventory series, we aren't sure if it possesses a unit root or not. To be on the safe side, shouldn't we take the first difference anyway? If it was stationary to begin with, the first difference is still stationary, so what's the harm?

The problem with "over differencing" is that it ruins the recursion that underlies the autoregressive representation for the series. To see this, consider the simple case where $I_t = \epsilon_t$. The recursion then becomes

$$\Delta I_t = \epsilon_t - \epsilon_{t-1} = \epsilon_t - (\Delta I_{t-1} + \epsilon_{t-2}) = \ldots = \epsilon_t - \Delta I_{t-1} - \Delta I_{t-2} - \ldots \qquad (11.b.6)$$

The coefficients on the lagged values of ΔI_t never converge.

Despite the fact that an autoregressive representation does not exist, it is always possible to compute least-squares estimates for autoregressive models in finite samples. Often these estimated models will appear quite reasonable, with apparently well-behaved residuals, respectable goodness-of-fit tests, etc.

One additional caveat. Suppose that in lieu of dealer inventories, the data identify dealer trades: "100 shares purchased from the dealer, 200 shares sold by the dealer, etc." The trade series is (minus) the first difference of the inventory series. So if inventories are stationary, the trade series is noninvertible.

Can you estimate a non-invertible moving average model? Yes, but not by forcing it into an autoregressive straitjacket. Hamilton discusses a maximum likelihood approach.

11.c The Wold theorem revisited

The Wold theorem assures us that if q_t is covariance stationary, then it possesses a moving average representation. If this representation is invertible, then q_t possesses an autoregressive representation as well. Suppose, for example, we have a low order AR representation

$$q_t = \phi q_{t-1} + \epsilon_t \qquad (11.c.7)$$

Autoregressions are particularly useful because we can estimate them using ordinary least squares.

But wait a minute. The variable we're trying to model here takes on discrete values: $q_t = \pm 1$. This means that if we try to estimate the autoregression, our dependent variable will be a *limited* dependent variable.

Wasn't there something in Econometrics 101 that explicitly warned against these sorts of estimates? Don't we have to use probit or logit instead?

The concern is an important one. Virtually all microstructure series except time are discretely valued. Prices, for example, live on a grid that was \$1/8 for a long time, and is presently \$0.01. The vast majority of trades occur in round-lot multiples (units of 100 shares). Probit models are occasionally used (Lo, MacKinlay and Hausmann), but they are not the norm. If we couldn't assert some sort of validity for specifications like (c.1), empirical market microstructure would be quite difficult.

The purpose of this discussion, then, is to establish the force and the limitations of the Wold theorem.

We start with some reassurances. The Wold theorem is not contingent on the time series being continuously-valued, Gaussian, etc. Discrete time series are fine. Given covariance stationarity, we can confidently write $q_t = \theta(L)\,\epsilon_t$ where $\mathrm{E}\epsilon_t = 0$, $\mathrm{E}\epsilon_t^2 = \sigma_\epsilon^2$ and $\mathrm{E}\epsilon_t\,\epsilon_s = 0$ for $t \neq s$. Furthermore, since ϵ_t is uncorrelated with ϵ_{t-1}, ϵ_{t-2}, ..., then it is also uncorrelated with q_{t-1}. This means that the ϵ_t in (c.1) satisfy the main requirements for consistency of OLS estimators: they are zero-mean, homoscedastic and uncorrelated with the explanatory variables. So (c.1) is a sensible specification for estimation and forecasting.

Now for the limitations. The Econometrics 101 cautionary note points out that in a linear probability model, the disturbances might have to have weird distributions in order to generate discrete values for the dependent variable. In a specification that is based on a behavioral model, this is a disturbing point. Suppose I'm estimating $y_i = \beta_0 + \beta_1\,x_i + u_i$ where $y_i = 1$ if individual i buys an ice-cream cone and x_i is the temperature at the time of decision. It is pretty clear that no standard distribution is likely to generate zero/one values for y_i. While we might assert, therefore, that u_i is uncorrelated with x_i, it is virtually impossible for u_i to be *independent* of x_i. The same argument applies to the ϵ_t in (c.1). Even though they are not serially correlated, they are almost certainly not serially independent.

An example might clarify matters. Instead of working with q_t, though, we'll construct a simpler indicator variable. Suppose that t indexes minutes in the trading session. In terms of observed activity b_t is an indicator variable, equal to one if there is at least one trade in minute t and zero otherwise.

Suppose that the b_t are generated in the following way. There is an unobserved i.i.d. series $\{a_t\}$: $a_t = 1$ with probability η; $a_t = 0$ with probability $1 - \eta$. Then, $b_t = 1$ if $a_t + a_{t-1} = 2$ and zero otherwise.

Then the outcomes and their associated probabilities are:

	a_t	a_{t-1}	a_{t-2}	Prob	b_t	b_{t-1}
1	1	1	1	η^3	1	1
2	1	1	0	$(1-\eta)\,\eta^2$	1	0
3	1	0	1	$(1-\eta)\,\eta^2$	0	0
4	1	0	0	$(1-\eta)^2\,\eta$	0	0
5	0	1	1	$(1-\eta)\,\eta^2$	0	1
6	0	1	0	$(1-\eta)^2\,\eta$	0	0
7	0	0	1	$(1-\eta)^2\,\eta$	0	0
8	0	0	0	$(1-\eta)^3$	0	0

(11.c.8)

The mean is

$$\mathrm{E}b_t = \mathrm{E}b_{t-1} = \eta^2 \tag{11.c.9}$$

The variance is:

$$\gamma_{b,0} = \mathrm{E}[b_t - \mathrm{E}b_t]^2 = \eta^2 - \eta^4 \tag{11.c.10}$$

The first-order autocovariance is:

$$\gamma_{b,1} = \mathrm{E}[b_t - \mathrm{E}b_t][b_{t-1} - \mathrm{E}b_{t-1}] = -(\eta - 1)\,\eta^3 \tag{11.c.11}$$

Thus, $\gamma_{b,1} > 0$. Trades will appear to cluster in time.

By way of explanation, the "standard" model of random event occurrence is the Poisson/exponential model, where waiting times between events are exponentially distributed and (in consequence) the number of trades in any interval is a Poisson variate. In this model events occur “evenly” (that is, with constant intensity) in time.

In most real securities market, trading activity is more clustered than would be predicted by the exponential/Poisson model. That is, if the current trade occurred quickly after the last trade, it is likely to be quickly followed by another trade. When the event occurrences are plotted over time, they visually “cluster”.

Thus, although the latent mechanism in the problem is fanciful, the behavior of the observed series is not.

Autocovariances at all orders higher than one vanish, so we can write $b_t = \epsilon_t + \theta\epsilon_{t-1}$. We earlier saw that the invertible solution for the MA(1) parameters in terms of the autocovariances is:

$$\sigma_\epsilon^2 = \frac{\gamma_0}{2} + \frac{1}{2}\sqrt{\gamma_0^2 - 4\,\gamma_1^2} \qquad \theta = \frac{\gamma_0 - \sqrt{\gamma_0^2 - 4\,\gamma_1^2}}{2\,\gamma_1} \tag{11.c.12}$$

So:

$$\sigma_\epsilon^2 = \frac{1}{2}\left(-\eta^4 + \eta^2 + \sqrt{-(\eta-1)^3\,\eta^4\,(3\,\eta+1)}\right) \qquad \theta = \frac{\eta^4 - \eta^2 + \sqrt{-(\eta-1)^3\,\eta^4\,(3\,\eta+1)}}{2\,(\eta-1)\,\eta^3} \tag{11.c.13}$$

For example, with $\eta = 0.8$,

$$\gamma_{b,0} = 0.2304, \ \gamma_{b,1} = 0.1024 \tag{11.c.14}$$

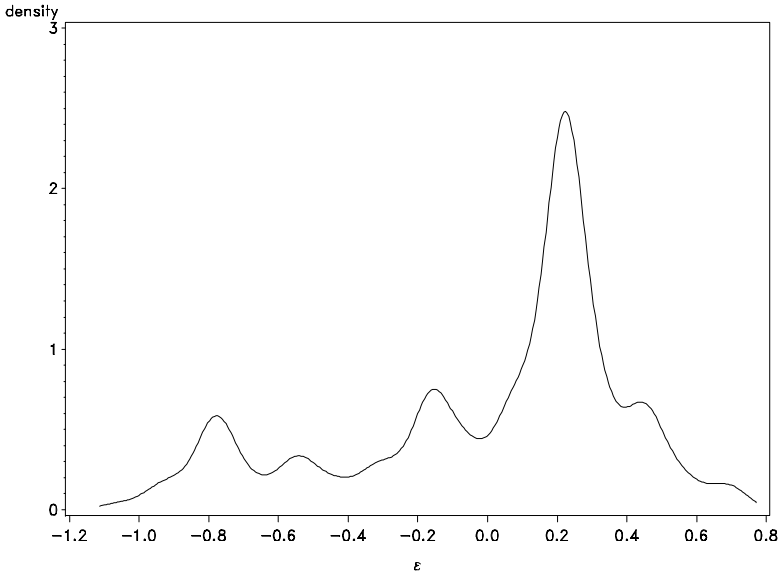
and

$$\sigma_\epsilon^2 = 0.167976 \qquad \theta = 0.609612 \tag{11.c.15}$$

We know from the Wold theorem that the ϵ_t are uncorrelated, but not much more.

At this point, it's easier to simulate the process and look at the properties of the estimated ϵ_t. (See MA1Problem.sas.) In a generated random sample of 100,000 observations:
$\overline{b_t} = 0.6449$, $\hat{\gamma}_{b,0} = 0.229$, $\hat{\gamma}_{b,1} = 0.1018$, $\hat{\sigma}_\epsilon^2 = 0.1670$, and $\hat{\theta} = 0.6101$. All are reasonably close to the population values.

Now what do the ϵ_t look like? In the first place, their autocorrelations are very close to zero, as the Wold theorem would predict. A kernel density (smoothed) histogram, though, reveals a very irregular distribution:



Furthermore, the higher-order autocorrelations are quite different from zero. For example, $\text{Corr}(\epsilon_t, \epsilon_{t-1}^2) = -0.18$. Thus, the ϵ_t are certainly not serially independent.

■ Summary

Assuming covariance stationarity, we're on firm econometric ground when we estimate linear autoregressions (and, later, vector autoregressions). Interpreting them, though, calls for a little caution. We'd generally like to interpret ϵ_t as an innovation, i.e., "new information". This interpretation must be qualified as "conditional on a linear model". There might be nonlinear models that would offer better forecasting performance and different innovations.

Chapter 12. Multivariate time series

In the univariate models, p_t actually serves in two roles. On the left hand side of an autoregression, p_t is the quantity of interest, the variable containing the martingale component that we identify with the efficient price. On the right hand side, the lagged p_t constitute the information set, the variables on which the martingale is (implicitly, in our procedures) projected. Both of these roles are open to extension and generalization. Initially, we will consider information sets expanded to include trade-direction variables, and anything else deemed relevant. Later, we will analyze techniques for drawing inferences about the martingale component of multiple price series.

This section summarizes relevant terminology and results. The material here is covered in Hamilton, Ch. 11 and 12.

12.a Vector moving average and autoregressive models

Consider a vector time series $\{y_t\}$ where y_t is an $(n \times 1)$ vector. For example, we might have $y_t = (\Delta p_t \quad q_t)$ where q_t is a trade direction variable.

The analysis broadly follows the univariate case. The multivariate autocovariances are matrices:

$$\Gamma_k = E(y_t - \mu)(y_{t-k} - \mu)' \tag{12.a.1}$$

In suppressing any dependence on t, we're implicitly assuming that $\{y_t\}$ is covariance stationary. Note that $\Gamma_k = \Gamma'_{-k}$.

The univariate autocovariance generating function generalizes to:

$$g(z) = \cdots + \Gamma_{-2} z^{-2} + \Gamma_{-1} z^{-1} + \Gamma_0 + \Gamma_1 z + \Gamma_2 z^2 \cdots \tag{12.a.2}$$

The multivariate Wold theorem ensures that we can write y_t as a (possibly infinite order) moving average:

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots = \theta(L)\epsilon_t \tag{12.a.3}$$

where the ϵ_t is a vector zero-mean white noise process: $E\epsilon_t = 0$, $E\epsilon_t \epsilon_t' = \Omega$, $E\epsilon_t \epsilon_{t-k}' = 0$ for $k \neq 0$. $\theta(L) = I + \theta_1 L + \theta_2 L^2 + \cdots$ is a matrix lag polynomial: each of the θ_i is $(n \times n)$. This is a vector moving average (VMA). The autocovariance generating function may be computed as

$$g(z) = \theta(z^{-1})\Omega\theta(z) \tag{12.a.4}$$

If the VMA is invertible, it can be written as vector autoregression (VAR):

$$\phi(L) y_t = y_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots = \epsilon_t \tag{12.a.5}$$

where $\phi(L) = I + \phi_1 L + \phi_2 L^2 + \cdots$, with each ϕ_i an $(n \times n)$ matrix.

As with univariate processes, it's useful to be able to go back and forth between AR and MA representations. Recall that for univariate processes, computing the correspondence between AR and MA representations generally required computing series expansions of the lag polynomials. This is also true for the vector processes.

For example, suppose that y_t is a vector moving average of order 1, VMA(1): $y_t = \epsilon_t + \theta\epsilon_{t-1}$. The matrix lag polynomial is $(I + \theta L)$. The autoregressive parameters may be computed from the matrix series expansion $\phi(L) = I - \theta L + \theta^2 L - \theta^3 L + \cdots$. Formally, this is identical to the univariate expansion, but the sums and products here are sums and products of matrices. Hamilton gives further results.

In microstructure applications, one usually estimates the VAR and then (if necessary) transforms the VAR into a VMA. (Although it is important that we can go in the other direction if need be, the need arises far less frequently.) In the discussion of the univariate case, we went from autoregressive to moving average representations by forecasting the process subsequent to a one-unit shock. The same approach works here.

Suppose that we possess (or have estimated) a VAR of the form given above. Suppose that all lagged values are set to their unconditional mean (zero): $y_{t-1} = y_{t-2} = \ ... \ = 0$. Consider the forecasts subsequent to a shock at time t of ϵ_t:

$$\begin{aligned} &y_t = \epsilon_t \\ &E[y_{t+1} \mid \epsilon_t] = \phi_1 \, y_t = \phi_1 \, \epsilon_t \\ &E[y_{t+2} \mid \epsilon_t] = \phi_1 \, E[y_{t+1} \mid \epsilon_t] + \phi_2 \, y_t = (\phi_1^2 + \phi_2) \, \epsilon_t \\ &... \end{aligned} \tag{12.a.6}$$

This implies that the leading terms in the VMA are:

$$y_t = \epsilon_t + \theta_1 \, \epsilon_{t-1} + \theta_2 \, \epsilon_{t-2} + ... \ = \epsilon_t + \phi_1 \, \epsilon_{t-1} + (\phi_1^2 + \phi_2) \, \epsilon_{t-2} + ... \tag{12.a.7}$$

Note: Some statistical software packages (like SAS's PROC MODEL) don't directly compute the VMA representation. In these cases, you can obtain the VMA by forecasting the estimated model subsequent to unit shocks in each of the disturbances.

Briefly, a truncated version of the VAR in eq. (12.5) can be estimated by least squares. By inversion (or forecasting), one obtains the VMA representation. From this, one computes estimates of impulse response functions, σ_w^2, $\underline{\sigma_s^2}$, etc. Distributional properties of these estimates may be inferred using the delta or subsampling methods described in the univariate case.

12.b Impulse response functions: their use and interpretation

In many empirical settings, an economic hypothesis makes a clear prediction about the sign and/or size of a regression coefficient. VAR's, though, are used in situations where interest centers on joint dynamics of the variables. The individual VAR coefficient estimates are not usually very illuminating in this respect: it is a rare hypothesis that confidently asserts sign and size of a particular entry of $\phi(L)$. Usually, the content of a VAR is assessed by summary transformations of the coefficients. The impulse response functions are among the most important of these transformations because they enable us to map out (in a form that is easily graphed) the time path of the system variables. The time path depicted, though, will depend on the starting point chosen. What starting point is most meaningful?

Suppose that we have a bivariate vector process $y_t = (y_{1,t} \quad y_{2,t})'$ with VMA representation $y_t = \theta(L)\,\epsilon_t$ and $\text{Var}(\epsilon_t) = \Omega$. For a given innovation ϵ_t, the conditional forecast k periods ahead is $E[y_{t+k} \mid \epsilon_t] = \theta_k\,\epsilon_t$. If (as is usually the case) θ_0 is normalized to I, then $E[y_t \mid \epsilon_t] = y_t = \epsilon_t$. The impulse response function is the mapping (over time) of the effect of variable j on variable i: the series of $(i,\ j)$ entries in the $\theta(L)$ lag polynomial.

Having computed the VMA coefficients, we'd like to make statements like "a one-unit shock to $y_{2,t}$ causes $y_{1,\,t+k}$ to be $\theta_{k,1,2}$, on average," (where $\theta_{k,1,2}$ is the (1,2) entry of the θ_k, the matrix coefficient of L^k). This statement is supposed to convey the intuition of what would happen if one initial variable were changed, while all others were held constant.

The problem is that if the two variables have a contemporaneous relationship, this sort of shock might be an extremely unrepresentative occurrence. Suppose, for example, that the two variables are the daily returns on overlapping indexes (like the Dow and the S&P 100). There certainly exist days when the former rises and latter stays the same or even falls, but these are relatively infrequent events. So when the two innovations are contemporaneously correlated, how should we construct hypothetical innovations, to use as starting points for impulse response functions, that are more representative?

The situation is similar to what happens in an ordinary linear regression (projection). Suppose (for concreteness and simplicity) that $(y_t \quad x_{1\,t} \quad x_{2\,t})$ are multivariate normal with zero mean and covariance matrix $\Sigma = \left(\begin{array}{c|c} \sigma_y^2 & \Sigma_{yx} \\ \hline \Sigma_{xy} & \Sigma_{xx} \end{array} \right)$ where Σ_{xx} is the 2×2 covariance matrix of the xs, Σ_{yx} is 1×2 and $\Sigma_{yx} = \Sigma_{xy}'$.
Consider the linear projection

$$y_t = x_t\,\beta + u_t \text{ where } x_t = (x_{1\,t} \quad x_{2\,t}) \text{ and } \beta = (\beta_1 \quad \beta_2)' = \Sigma_{xx}^{-1}\,\Sigma_{xy}. \tag{12.b.8}$$

A coefficient, like β_1 is normally interpreted as the effect of a unit change in $x_{1\,t}$, holding $x_{2\,t}$ fixed. But it is certainly not the case that $E[y_t \mid x_{1\,t} = 1] = \beta_1$. This latter conjecture ignores the information contained in $x_{1\,t}$ that is relevant for predicting $x_{2\,t}$ (reflected in the $\text{Cov}(x_{1\,t},\, x_{2\,t})$ on the off-diagonal of Σ_{xx}. There are several ways of computing $E[y_t \mid x_{1\,t} = 1]$. Perhaps the most straightforward is to consider a new projection, one in which y_t is solely projected onto $x_{1\,t}$: $y_t = \beta_1^*\,x_{1\,t} + u_t$ where $\beta_1^* = \text{Cov}(y_t,\, x_{1\,t}) / \sigma_y^2$. Then
$E[y_t \mid x_{1\,t} = 1] = \beta_1^*$.

Alternatively, we could first project x_{2t} onto x_{1t}: $x_{2t} = \alpha_1 x_{1t}$, where $\alpha_1 = \text{Cov}(x_{2t}, x_{1t}) / \text{Var}(x_{1t})$. If we were to set $x_{1t} = 1$, we'd expect $x_{2t} = \alpha_1$. The predicted value of y would then be $E[y_t \mid x_{1t} = 1] = (1 \mid \alpha_1) \beta$. The ordering in which we did things here was arbitrary. We'd get the same prediction if we conditioned on $x_{2t} = \alpha_1$. That is, $E[y_t \mid x_{1t} = 1] = E[y_t \mid x_{2t} = \alpha_1]$.

It is more difficult to make causal effects. If x_{1t} is a control variable, for example, we can't assert that if we dialed x_{1t} to unity, we'd expect the realization of y_t to be $E[y_t \mid x_{1t} = 1]$ (as computed by either of the above methods), or β_1 for that matter. To proceed, we need to assume or impose a causal ordering.

If we assume that causality (in the familiar sense) flows from x_{1t} to x_{2t}, then $E[y_t \mid x_{1t} = 1]$ would be the value computed above, which took into account the effect of x_{1t} on x_{2t} and hence y_t. If causality were to flow entirely in the other direction, $E[y_t \mid x_{1t} = 1] = \beta_1$, (and a computation of $E[y_t \mid x_{2t} = 1]$ would involve the indirect effects of x_{2t} on x_{1t}).

Assertion of a causal direction is tantamount to asserting a recursive structure for the variables. A convenient tool for computing this structure is the Cholesky factorization.

12.c Cholesky factorizations

Sometimes, for a given covariance matrix, we seek to construct a factor representation in which the factors, considered sequentially, capture the variation in a variable not explained by factors that were included earlier. This could be done by performing successive linear projections. The Cholesky factorization is an alternative. In a Cholesky decomposition, a symmetric positive definite matrix is factored into a lower triangular matrix and its transpose: $\Omega = F' F$, where F' is lower triangular. The lower triangular matrix can be interpreted as a transformation matrix for recursively generating the original variables from a set of underlying uncorrelated zero-mean unit-variance factors.

Consider the 2×2 covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \tag{12.c.9}$$

The Cholesky factor

$$F = \begin{pmatrix} \sigma_1 & \rho \sigma_2 \\ 0 & \sqrt{1 - \rho^2}\, \sigma_2 \end{pmatrix} \tag{12.c.10}$$

$F' F$ recreates the original covariance matrix

Now consider the lower triangular matrix

$$F' = \begin{pmatrix} \sigma_1 & 0 \\ \rho \sigma_2 & \sqrt{1 - \rho^2}\, \sigma_2 \end{pmatrix} \tag{12.c.11}$$

Suppose we posit a factor structure for x:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\,\sigma_2 & \sqrt{1-\rho^2}\;\sigma_2 \end{pmatrix} \begin{pmatrix} z_{1t} \\ z_{2t} \end{pmatrix} \text{ where } \begin{pmatrix} z_{1t} \\ z_{2t} \end{pmatrix} \sim N(0,\, I)$$

z_1 explains all of x_{1t}, so it is natural to view this as the "x_1 factor". z_2 reflects the information contained in x_2 that is not in x_1. This corresponds to a causal ordering that places primacy on x_1. This factor structure is purely a consequence of the ordering. If we'd arranged the variables as $\begin{pmatrix} x_{2t} \\ x_{1t} \end{pmatrix}$, x_{2t} would have been the principal driver.

Orthogonalized impulse response functions

Suppose that we have an innovations representation (VMA) for a multivariate time series y_t:

$$y_t = \epsilon_t + \theta_1\,\epsilon_{t-1} + \theta_2\,\epsilon_{t-2} + \ldots \tag{12.c.12}$$

With the Cholesky factorization $F'\,F = \Omega = \text{Var}(\epsilon_t)$, we may then write $\epsilon_t = F'\,z_t$ where $z_t \sim N(0,\, I)$. This expresses the model innovations in terms of underlying unobserved uncorrelated factors. The VMA written in this fashion is:

$$y_t = F'\,z_t + \theta_1\,F'\,z_{t-1} + \theta_2\,F'\,z_{t-2} + \ldots \tag{12.c.13}$$

where the $\theta_i\,F'$ coefficient matrices represent the orthogonalized impulse response coefficients. For example, if $z_t = (\,1 \quad 0 \quad \ldots \quad 0\,)'$, $F'\,z_t$ will be an $n \times 1$ vector of the contemporaneous effects of a one-standard-deviation shock to ϵ_{1t}, assuming that this shock affects all other variables; $\theta_1\,F'\,z_t$ will be the effect in period $t+1$ and so on.

When we wish to investigate behavior of the system under alternative causal orderings, it is often easiest to re-order the variables in the original analysis, letting the statistical software do the work. This will usually result in a fresh estimation of the model, however. If compositional efficiency is a consideration, an alternative procedure is to simply permute the variables in the coefficient and covariance matrices, and recompute the Cholesky factorization.

12.d Attributing explanatory power

An important related issue involves the attribution of explanatory power. The explained variance in the regression is $\beta'\,\text{Var}\left(\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix}\right)\beta = \beta_1^2\,\sigma_1^2 + 2\,\beta_1\,\beta_2\,\sigma_{12} + \beta_2^2\,\sigma_2^2$. If $\sigma_{12} = 0$, there is a clean decomposition of how much is explained by the two variables. If $\sigma_{12} \neq 0$, ambiguity arises. We may nevertheless identify two extremes. We may associate the covariance term entirely with x_{1t}, or alternatively, entirely with x_{2t}. The first case corresponds to placing x_{1t} first in the causal ordering; the second, to placing x_{2t} first. Notice that since the covariance term can be negative, it is not possible to say a priori which ordering maximizes the explanatory power.

The situation can also be viewed as one in which we sequentially add explanatory variables to a regression. The incremental explanatory power of a variable depends on what variables were included earlier, and we can't assume that the incremental explanatory power of a variable is maximized by placing it first. For example, consider a signal extraction problem that arises frequently in microstructure models. The true value is $v \sim N(0, \sigma_v^2)$; the observed signal is $s = v + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon^2)$, independent of v. In models, we usual need to project v on s. Here, consider the (perfect) projection of v on s and ϵ. Suppose we put ϵ "first". The projection of v solely on ϵ has no explanatory power (due to the independence of v and ϵ). But if we add ϵ to a projection that already includes s, the explanatory power jumps (from some $R^2 < 0$ to $R^2 = 1$). So the incremental explanatory power of ϵ is actually maximized by including it last.

The lesson seems to be that if we seek the maximum or minimum possible incremental explanatory power for a variable, we must consider its incremental explanatory power under all possible orderings (permutations) of the variables. Actually, we need not investigate all permutations. The incremental R^2 associated with adding x^* to the variable set depends only on the preceding explanatory variables, not their ordering. If we have a total of $n+1$ explanatory variables, and wish to examine the incremental R^2 associated with adding x^* as the $k+1^{\text{st}}$ variable, there are $\binom{k}{n} = \frac{n!}{(n-k)!\,k!}$ possible combinations. So the total number of combinations associated with adding x^* first, second, etc. is $\sum_{k=0}^{n} \frac{n!}{(n-k)!\,k!} = 2^n$. This is smaller than the number of permutations, $n!$. With $n = 9$, we need to investigate 512 cases, while $9! = 362,880$.

12.e Forecast variance decompositions

In constructing the impulse response functions (moving average representation), we forecast the system conditional only on knowing ϵ_t. (Recall that all lagged values were set to zero.) In a real-time forecasting situation, we'd possess the history of the process. The forecast at lead k in this situation is $E[y_{t+k} \mid \epsilon_t, \epsilon_{t-1}, \ ...]$. In the table below, we present the actual value of y_{t+k}, the forecast of y_{t+k} and the difference between the two (the forecast error):

Actual	y_{t+k}	$=$	ϵ_{t+k}	$+\theta_1 \epsilon_{t+k-1}$	$+ \ldots$	$\theta_{k-1} \epsilon_{t+1}$	$+\theta_k \epsilon_t$	$+\theta_{k-1} \epsilon_{t-1}$	$+ \ldots$
Forecast	$E[y_{t+k} \mid \epsilon_t, \epsilon_{t-1}, \ ...]$	$=$					$\theta_k \epsilon_t$	$+\theta_{k-1} \epsilon_{t-1}$	$+ \ldots$
Forecast error	$y_{t+k} - E[y_{t+k} \mid \epsilon_t, \epsilon_{t-1}, \ ...]$	$=$	ϵ_{t+k}	$+\theta_1 \epsilon_{t+k-1}$	$+ \ldots$	$\theta_{k-1} \epsilon_{t+1}$			

The forecast error covariance at lead k is therefore: $\sum_{j=0}^{k-1} \theta_j \, \Omega \, \theta_j'$. In the case of diagonal Ω, the forecast error variance can be cleanly dichotomized into contributions from each of the system innovations. If there are off-diagonal elements, we can bound these contributions using different Cholesky factorizations and permutations as described above. A particularly important special case of this technique arises in the limit as $k \to \infty$. In this case, the forecast error variance is equal to the total variance of the system variables, $\text{Var}(y_t)$.

Note: $\sum_j \theta_j \, \Omega \, \theta_j' \neq [\sum_j \theta_j] \, \Omega \, [\sum_j \theta_j]'$. Confusion on this point is especially problematic when one of the variables is a price change, Δp_t. In this case, the corresponding term of limiting $\sum_j \theta_j \, \Omega \, \theta_j'$ is $\text{Var}(\Delta p_t)$,

while the corresponding term of the limiting $[\sum_j \theta_j] \Omega [\sum_j \theta_j]'$ is Var(w_t), the variance of the random-walk component of the price.

Chapter 13. Prices and trades: statistical models

13.a Trade direction variables: constructing q_t

Most markets disseminate and record bid and ask quotes. When these are merged with the trade reports, it is often possible to judge the trade price relative to the bid and ask quotes. A trade at the ask (or more commonly, above the bid-ask midpoint) is signed as a "buy" ($q_t = +1$); a trade at the bid (or below the bid-ask midpoint) is a "sell" ($q_t = 1$).

Although simple in principle, the procedure has its limitations. Two of the more commonly-encountered difficulties are:

- In some markets, trades may occur at prices other than the posted bid and ask. In US equity markets, for example, trades occurring exactly at the bid-ask midpoint frequently occur.
- Reporting practices may induce incorrect sequencing of trades and quotes.

Despite these limitations, however, q_t constructed in this way often have substantial power in explaining price dynamics.

13.b Simple trade/price models

This section describes four models of increasing complexity.

■ Model 1 (Generalized Roll model, with both p_t and q_t observed)

When we observe both p_t and q_t, the generalized Roll model can be estimated via single-equation least squares.

Recall that the models is:

$$\begin{aligned} m_t &= m_{t-1} + w_t \\ w_t &= \lambda\, q_t + u_t \\ p_t &= m_t + c\, q_t \\ \Delta p_t &= -c\, q_{t-1} + c\, q_t + \lambda\, q_t + u_t \end{aligned} \tag{13.b.1}$$

Previously, we assumed that only the p_t were observed. In many applications, though, we possess the q_t as well. If this is the case, we can easily estimate the $\{c,\ \lambda,\ \sigma_u^2\}$ parameters via OLS regression applied to the last equation.

OLS suffices here because the q_t in this equation are both known and predetermined (with respect to the Δp_t). The residual, u_t, is uncorrelated with the explanatory variables. The long-run price forecast is:

$$f_t = E[p_{t+1} \mid p_t,\ p_{t-1},\ ...,\ q_t,\ q_{t-1},\ ...] = p_t + E[\Delta p_{t+1} \mid p_t,\ p_{t-1},\ ...,\ q_t,\ q_{t-1},\ ...] = p_t - cq_t \qquad (13.b.2)$$

By inspection, it is clear that $f_t = m_t$. The variance of the random walk component is:

$$\sigma_w^2 = \underbrace{\lambda^2 \sigma_q^2}_{\text{Trade-related / Private Information}} + \underbrace{\sigma_u^2}_{\text{Non-trade-related / Public information}} \qquad (13.b.3)$$

Given the structure of q_t, $\sigma_q^2 = 1$, a result that we've used earlier. Here we leave it in symbolic form.

This expression implies a clear decomposition of random-walk variance into one component that is attributable to trades and another that is uncorrelated with trades. Given the economic rationale for the specification, the trade-related component is due to the market's assessment of the private-information component of the trade, while the non-trade component is due to public information.

The $\lambda^2 \sigma_q^2$ quantity is in a sense an absolute measures of private information. Sometimes its useful to have a relative measure as well. An natural candidate is

$$\lambda^2 \sigma_q^2 / \sigma_w^2. \qquad (13.b.4)$$

This can be viewed as the coefficient of determination (R^2) in a regression of w_t on q_t.

■ Model 2: Autocorrelated trades

When the q_t are serially correlated, the Δp_t regression must include lags of q_t. As long as the q_t are exogenous, though, we don't need to estimate a joint specification.

Suppose, for example that q_t is MA(1):

$$q_t = \beta \epsilon_{q,t-1} + \epsilon_{q,t} \qquad (13.b.5)$$

In most markets, the autocorrelation in trade directions is positive: the order following a "buy" also tends to be a "buy". Thus, it's realistic to expect $\theta > 0$.

In the generalized Roll model, q_t appears in two contexts. First, it simply determines whether the trade price is at the bid or ask. Second, it drives the revision in the efficient price due to inferred private information. In this latter context, it is important to note that in the present case, $E[q_t \mid q_{t-1},\ q_{t-2},\ ...] \neq 0$. Therefore, the information content of q_t, i.e., the informational innovation, what we learn that we didn't know before, is $q_t - E[q_t \mid q_{t-1},\ q_{t-2},\ ...] = \epsilon_t^q$. The increment to the efficient price is therefore

$$w_t = u_t + \lambda \epsilon_{q,t} \qquad (13.b.6)$$

Now, Δp_t becomes:

$$\Delta p_t = u_t - c\,(\beta\,\epsilon_{q,t-2} + \epsilon_{q,t-1}) + \lambda\,\epsilon_{q,t} + c\,(\beta\,\epsilon_{q,t-1} + \epsilon_{q,t}) \tag{13.b.7}$$

Due to the presence of ϵ^q_{t-2}, Δp_t will have a nonzero autocovariance at lag two (it is now a second-order moving average). The price at time $t+2$ is:

$$p_{t+2} = p_t + \Delta p_{t+1} + \Delta p_{t+2} = p_t + u_{t+1} + u_{t+2} - c\,\beta\,\epsilon_{q,t-1} - c\,\epsilon_{q,t} + (c\,\beta + \lambda)\,\epsilon_{q,t+1} + (c + \lambda)\,\epsilon_{q,t+2} \tag{13.b.8}$$

Taking the expectation of this, conditional on what we know at time t gives

$$f_t = p_t - c\beta\epsilon_{q_{t-1}} - c\epsilon_{q,t} \tag{13.b.9}$$

As above, we can verify that $f_t = m_t$. The random-walk decomposition is now:

$$\sigma^2_w = \lambda^2\,\mathrm{Var}(\epsilon_{q,t}) + \sigma^2_u \tag{13.b.10}$$

If the q_t are still to be unconditionally distributed as equally-probable realizations of ± 1, then (assuming $\theta > 0$), $\mathrm{Var}(\epsilon_{q,t}) < 1$. So the trade-related contribution to the efficient price variance is lower than in the uncorrelated case.

Estimation in this model is slightly more complicated because the expression for Δp_t, which we were using as a regression specification, involves the unobserved $\epsilon_{q,t}$ innovations.

One approach would be to estimate the q_t process, compute $\hat{\epsilon}_{q,t}$, the estimated innovations and use them in the Δp_t regression:

$$\Delta p_t = u_t - c\,(\beta\,\hat{\epsilon}_{q,t-2} + \hat{\epsilon}_{q,t-1}) + \lambda\,\hat{\epsilon}_{q,t} + c\,(\beta\,\hat{\epsilon}_{q,t-1} + \hat{\epsilon}_{q,t}) \tag{13.b.11}$$

The $\hat{\epsilon}_t$, though, are generated regressors. OLS coefficient estimates will be consistent here. (This is true only because the q_t are exogenous.) The asymptotic distribution of the OLS estimates, though is complicated.

An easier and more general approach is to estimate the model by regressing the Δp_t onto the q_t. To see what this implies, first rewrite the regression in terms of the lag operator:

$$\Delta p_t = u_t + (\lambda + c - c(1-\beta)\,L - c\beta L^2)\,\epsilon_{q,t} \tag{13.b.12}$$

Then recall that since $q_t = (1 + \beta L)\,\epsilon_{q,t}$,

$$\epsilon_{q,t} = (1 + \beta L)^{-1}\,q_t \tag{13.b.13}$$

Substituting into the Δp_t equation gives:

$$\Delta p_t = u_t + (\lambda + c - c(1-\beta)\,L - c\beta L^2)\,(1 + \beta L)^{-1}\,q_t \tag{13.b.14}$$

The expansion of $(1 + \beta L)^{-1}\,q_t$ is of infinite order. In practice, we'd get approximate results by estimating a truncated specification. (Specification would be simpler if we'd started with an autoregressive representation for q_t, like $q_t = \beta q_{t-1} + \epsilon_{q,t}$, in the first place.)

In this case, we don't need to estimate the joint dynamics of q_t and Δp_t. It might be more efficient to do so, though, since β appears in both processes. We could stack the last two equations as a vector autoregression.

We might also want to specify the VAR in a way that explicitly models the contemporaneous causality (the impact of q_t on Δp_t). The generic VAR is usually specified as $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ \dots \ + \epsilon_t$, where the first term on the r.h.s. is y_{t-1}

In the present model, we have a recursive relationship at time t. It could be specified as:

$$y_t = \phi_0 y_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ \dots \ + \epsilon_t \text{ where } \phi_0 = \begin{pmatrix} 0 & \lambda + c \\ 0 & 0 \end{pmatrix} \tag{13.b.15}$$

Essentially, we have a recursive contemporaneous structure: q_t affects Δp_t, but not vice versa. When this model is estimated, the residual covariance matrix $\text{Var}\begin{pmatrix} u_t \\ \epsilon_{q,t} \end{pmatrix}$ will be diagonal by construction.

How should the VAR be estimated? VARs are conventionally (and conveniently) estimated using ordinary least-squares. Here, though, we have a preferred structural model, and GMM is a reasonable alternative. Applying GMM here, we'd have five model parameters $\{c, \lambda, \beta, \sigma_u^2, \sigma_\epsilon^2\}$. These parameters determine the (vector) autocovariances of the process, and it would be logical to use these as the moment conditions.

■ Model 3: Endogenous trades

When the q_t are not exogenous, it is necessary to model the joint dynamics. In this variation of the model, price changes can affect subsequent q_t.

To this point, q_t have been assumed exogenous to the public information process u_t. This simplifies the analysis because there is a clear causal direction of the effects in the model. It is not, however, particularly realistic.

Returns might affect subsequent trades for several reasons. Recall that the dealer inventory control hypothesis suggests that dealers respond to inventory imbalances by changing their quotes to elicit an imbalance in the subsequent incoming order flow. More broadly, we suspect that some agents in the economy follow price-sensitive strategies. If the price goes up purely by reason of public information, momentum traders may leap in and buy. Alternatively, an options trader who is hedging a short call position will buy when the price rises. Either of these effects (and probably many others) break the assumption that q_t is exogenous.

When we are modeling multiple time series and can't assert a priori a one-way causal structure, the model must allow for joint dynamics. The models that we can test and interpret are fairly general and flexible ones. But to illustrate the approach, we'll consider a simple modification to our structural model.

The new trade direction process is:

$$q_t = \alpha u_{t-1} + \beta \epsilon_{q,t-1} + \epsilon_{q,t} \tag{13.b.16}$$

Note that $\epsilon_{q,t}$ is still the innovation in the trade.

$$\Delta p_t = u_t - c\,(\alpha\, u_{t-2} + \beta\, \epsilon_{q,t-2} + \epsilon_{q,t-1}) + \lambda\, \epsilon_{q,t} + c\,(\alpha\, u_{t-1} + \beta\, \epsilon_{q,t-1} + \epsilon_{q,t}) \tag{13.b.17}$$

The equations may be stacked to form a vector moving average:

$$\begin{pmatrix} \Delta p_t \\ q_t \end{pmatrix} = \begin{pmatrix} 1 & \lambda + c \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_t \\ \epsilon_{q,t} \end{pmatrix} + \begin{pmatrix} c\alpha & -c(1-\beta) \\ \alpha & \beta \end{pmatrix} \begin{pmatrix} u_{t-1} \\ \epsilon_{q,t-1} \end{pmatrix} + \begin{pmatrix} -c\alpha & -c\beta \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_{t-2} \\ \epsilon_{q,t-2} \end{pmatrix} \tag{13.b.18}$$

This can be written more concisely in vector/matrix notation

$$y_t = \theta_0.\epsilon_t + \theta_1.\epsilon_{t-1} + \theta_2.\epsilon_{t-2} \tag{13.b.19}$$

where "." denotes matrix multiplication and

$$y_t = \begin{pmatrix} \Delta p_t \\ q_t \end{pmatrix};\ \epsilon_t = \begin{pmatrix} u_t \\ \epsilon_{q,t} \end{pmatrix};\ \theta_0 = \begin{pmatrix} 1 & c+\lambda \\ 0 & 1 \end{pmatrix};\ \theta_1 = \begin{pmatrix} c\,\alpha & -c\,(1-\beta) \\ \alpha & \beta \end{pmatrix};\ \theta_2 = \begin{pmatrix} -c\,\alpha & -c\,\beta \\ 0 & 0 \end{pmatrix} \tag{13.b.20}$$

Let $\theta_{k,1}$ denote the first row of θ_k, i.e., the row corresponding to Δp_t. Then

$$\Delta p_t = \theta_{0,1}\, \epsilon_t + \theta_{1,1}\, \epsilon_{t-1} + \theta_{2,1}\, \epsilon_{t-2} \tag{13.b.21}$$

where

$$\theta_{0,1} = \{1,\ c+\lambda\};\ \theta_{1,1} = \{c\,\alpha,\ -c\,(1-\beta)\};\ \theta_{2,1} = \{-c\,\alpha,\ -c\,\beta\} \tag{13.b.22}$$

Recall that in the univariate case, $\Delta p_t = \theta(L)\, \epsilon_t$, we could compute the random-walk variance as $\sigma_w^2 = \theta(1)^2\, \sigma_\epsilon^2$. The corresponding result here, derived from the multivariate autocovariance generating function is

$$\sigma_w^2 = (\theta_{0,1} + \theta_{1,1} + \theta_{2,1})\, \Omega\, (\theta_{0,1} + \theta_{1,1} + \theta_{2,1})' \tag{13.b.23}$$

where $\Omega \equiv \mathrm{Var}(\epsilon_t) = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \mathrm{Var}(\epsilon_{q,t}) \end{pmatrix}$.

The sum of the $\theta_{k,1}$'s is:

$$\Sigma\theta = \{1,\ \lambda\} \tag{13.b.24}$$

So

$$\sigma_w^2 = \Sigma\theta\ \Omega\ \Sigma\theta' = \mathrm{Var}(\epsilon_{q,t})\, \lambda^2 + \sigma_u^2 \tag{13.b.25}$$

This is the same σ_w^2 as we obtained for the simpler case when there was no feedback from u_{t-1} to q_t. Why?

Although the model is more complex dynamically, the informational dynamics are identical. That is, w_t is generated the same way in both models. In the summation of the $\theta_{k,1}$, the transient effects drop out and we're left with the variance of the random-walk component.

As in the previous case, we could estimate this model with a VAR like $y_t = \phi_0\, y_t + \phi_1\, y_{t-1} + \phi_2\, y_{t-2} + \ \ldots\ + \epsilon_t$ where $\phi_0 = \begin{pmatrix} 0 & \lambda + c \\ 0 & 0 \end{pmatrix}$. The estimated residual covariance matrix is diagonal by construction.

To this point, we've concentrated on the first row of the system. The full sum of the θ_k is:

$$\begin{pmatrix} 1 & \lambda \\ \alpha & \beta + 1 \end{pmatrix} \tag{13.b.26}$$

The second row corresponds to q_t, which is stationary (and therefore doesn't contain a random walk component). The coefficient sum can nevertheless be interpreted as summarizing the effect of a given innovation on long-run cumulative trades.

■ Model 4: Contemporaneous trade and public information effects

This model allows public information to affect trades:

$$q_t = \alpha\, u_t + \beta\, \epsilon_{q,t-1} + \epsilon_{q,t} \tag{13.b.27}$$

Interpreting u_t as public information, $\alpha > 0$ might arise as a consequence of buying by uninformed traders on positive news. Market-makers observe the public information prior to setting their quotes, so from their perspective, $\epsilon_{q,t}$ is still the informational innovation in the trade.

$$\Delta\, p_t = u_t - c\,(\alpha\, u_{t-1} + \beta\, \epsilon_{q,t-2} + \epsilon_{q,t-1}) + \lambda\, \epsilon_{q,t} + c\,(\alpha\, u_t + \beta\, \epsilon_{q,t-1} + \epsilon_{q,t}) \tag{13.b.28}$$

The VMA coefficient matrices are now:

$$\theta_0 = \begin{pmatrix} 1 & c + \lambda \\ 0 & 1 \end{pmatrix};\ \theta_1 = \begin{pmatrix} c\,\alpha & -c\,(1-\beta) \\ \alpha & \beta \end{pmatrix};\ \theta_2 = \begin{pmatrix} -c\,\alpha & -c\,\beta \\ 0 & 0 \end{pmatrix} \tag{13.b.29}$$

Unlike the previous case, there is no clear contemporaneous recursive structure. Therefore, we could not estimate (as we did in the previous cases) a VAR like:

$$y_t = \phi_0\, y_t + \phi_1\, y_{t-1} + \phi_2\, y_{t-2} + \ \ldots\ + \epsilon_t \tag{13.b.30}$$

where ϕ_0 has all entries on the main diagonal and below equal to zero.

A specification like:

$$\begin{pmatrix} y_{1\,t} \\ y_{2\,t} \end{pmatrix} = \begin{pmatrix} 0 & \gamma_{12} \\ \gamma_{21} & 0 \end{pmatrix} \begin{pmatrix} y_{1\,t} \\ y_{2\,t} \end{pmatrix} + \begin{pmatrix} \epsilon_{1\,t} \\ \epsilon_{2\,t} \end{pmatrix} \tag{13.b.31}$$

can be estimated by single-equation least-squares. The problem is one of interpretation. We can rewrite this as $(I - \gamma)\, y_t = \epsilon_t$, where $\gamma = \begin{pmatrix} 0 & \gamma_{12} \\ \gamma_{21} & 0 \end{pmatrix}$, from which it is clear that ϵ_t is a (particular) linear transformation of y_t.

The structural model here actually possesses additional identifying restrictions on the VMA and VAR coefficients that we could in principle exploit. More generally, if we can't identify a contemporaneous recursive structure, it is better to estimate a VAR like:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ \dots \ + \epsilon_t \tag{13.b.32}$$

where contemporaneous effects will show up in the off-diagonal elements of $\text{Var}(\epsilon_t) = \Omega$. Estimation proceeds as follows. We estimate a truncated VAR for $y_t = (\ \Delta p_t \quad q_t\ )'$:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ \dots \ + \phi_K y_{t-K} + \epsilon_t \tag{13.b.33}$$

which we write more compactly as

$$(I - \phi(L))\, y_t = \epsilon_t \text{ where } \phi(L) = \phi_1 L + \phi_2 L^2 + \ \dots \ + \phi_K L^K \tag{13.b.34}$$

Consider the expression $\theta(1)\,\Omega\,\theta(1)'$ where $\theta(L)$ is the VMA lag polynomial. Since $\theta(L) = (I - \phi(L))^{-1}$,

$$\theta(1)\,\Omega\,\theta(1)' = (I - \phi(1))^{-1}\,\Omega\,(I - \phi(1))^{-1\,\prime} \tag{13.b.35}$$

The first element of $\theta(1)\,\Omega\,\theta(1)'$ is σ_w^2. I.e.,

$$\sigma_w^2 = c\,\Omega\,c\prime \text{ where } c \text{ is the first row of } (I - \phi(1))^{-1}. \tag{13.b.36}$$

Since Ω is not generally diagonal, the decomposition of σ_w^2 into trade- and non-trade-related components is not identified. Using the Cholesky factorization approach described above, though, we can determine an upper and lower bound for each contribution.

13.c General VAR specifications

The structural models described above are intended to illustrate the various sorts of joint dynamics that can arise between trades and prices. We can compute many of the derived statistics from these models, though, without knowing the precise structure and identifying the structural parameters. This is fortunate because most economic microstructure models are stylized constructs, intended primarily to illustrate the broad features of an economic mechanism. We have no plausible theory, for example, that might predict that q_t is MA(1), as opposed to, say, AR(1).

For a particular stock, we might attempt a precise identification of the orders of the VAR and VMA components of a model, but in practice we usually seek robust specifications that might be estimated across different stocks and different time samples. These considerations militate in favor of general specifications. The approaches discussed here are covered in Hasbrouck (1988, 1991a, 1991b, 1993).

Both the Kyle and Easley-O'Hara models suggest that larger order flows convey more information. It therefore makes sense to expand the set of signed trade variables to include signed volume, i.e., a quantity like $q_t V_t$ where V_t is the volume (usually the dollar volume) of the trade. It is also common to include signed nonlinear transformations of the volume to allow for more flexibility in the trade-impact function. Commonly used variables include $q_t V_t^2$, $q_t \sqrt{V_t}$ and $q_t \log(V_t)$.

Let Q_t denote the collection of signed-trade variables employed in a specification, for example, $Q_t = \begin{pmatrix} q_t \\ q_t V_t \end{pmatrix}$. The complete set of variables in the VAR is then $y_t = \begin{pmatrix} \Delta p_t \\ Q_t \end{pmatrix}$.

We estimate a general VAR of the form:

$$y_t = \phi(L)\,\epsilon_t \text{ where } \phi(L) = \phi_1 L + \phi_2 L^2 + \ldots + \phi_K L^K. \tag{13.c.37}$$

The covariance matrix of the disturbances is $\text{Var}(\epsilon_t) = \Omega$. It will be useful to partition this as $\text{Var}(\epsilon_t) = \begin{pmatrix} \sigma_1^2 & \sigma_{1Q} \\ \sigma_{Q1} & \Omega_Q \end{pmatrix}$ where $\sigma_1^2 = \text{Var}(\epsilon_{1t})$, the variance of the error associated with the Δp_t equation, and Ω is the covariance matrix of the trade variables.

Assuming the joint process to be covariance stationary and invertible, the y_t possess a VMA representation $y_t = \theta(L)\,\epsilon_t$.

In the univariate case, with $y_t = \Delta p_t$, the autoregression could be expressed in the form $y_t = \phi(L)\,y_t + \epsilon_t$ and the moving average representation in the form $y_t = \theta(L)\,\epsilon_t$, with the correspondence given by $\theta(L) = (1 - \phi(L))^{-1}$. The variance of the random-walk component of p_t was $\sigma_w^2 = |\,\theta(1)\,|^2\,\sigma_\epsilon^2 = |\,1 - \phi(1)\,|^{-2}\,\sigma_\epsilon^2$.

The corresponding development in the present multivariate case is $\theta(1)\,\Omega\,\theta(1)'$. This is not a scalar, but rather an $n \times n$ matrix, in which σ_w^2 is the first-row, first-column entry. That is:

$$\sigma_w^2 = a\,\Omega\,a' \text{ where } a \text{ is the first row of } \theta(1), \text{ or equivalently, the first row of } (I - \phi(1))^{-1}. \tag{13.c.38}$$

We now turn to the interpretation of σ_w^2. Most importantly, σ_w^2 does not depend on the variable set used in the VAR. It is the same whether Δp_t is projected onto only itself or onto a large collection of variables, including some irrelevant ones. In assessing the components of σ_w^2, it is useful partition Ω as $\Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{1Q} \\ \sigma_{Q1} & \Omega_Q \end{pmatrix}$ where $\sigma_1^2 = \text{Var}(\epsilon_{1t})$, the variance of the error associated with the Δp_t equation, and Ω_Q is the covariance matrix of the trade variables. We partition a accordingly as $a = (\,a_1 \quad a_Q\,)$.

In the case where $\sigma_{1Q} = 0$,

$$\sigma_w^2 = a_Q\,\Omega_Q\,a_Q' + a_1^2\,\sigma_1^2 \tag{13.c.39}$$

This identifies a dichotomy between trade-related and non-trade-related contributions to σ_w^2. At this level, resolution between these two components does not depend on resolving the contributions from the different components of the trade variable set Q_t. In the case where $\sigma_{1Q} \neq 0$, the contributions cannot be determined exactly, but they can be bounded using the Cholesky factorization approach.

When resolution of the σ_w^2 components is the main objective of the analysis, it is often useful to use the quote midpoint (bid-ask midpoint, BAM_t) as the price variable, i.e., replacing Δp_t with ΔBAM_t.

In the timing of the sequential trade models, recall that the market-maker revised the quote after a trade. Since the revision occurs after a trade, there is no ambiguity about contemporaneous causality. So we form $y_t = (\Delta BAM_t,\ Q_t')'$, where ΔBAM_t is the revision (if any) immediately after the trade, and estimate the VAR allowing for "contemporaneous" effects running from Q_t to ΔBAM_t.

How do we know that σ_w^2 implied by the ΔBAM_t is the same as that implied by the Δp_t? Intuitively, trade prices, bids and offers tend to move closely together over long periods. More formally, they are cointegrated. This property will be considered in a subsequent chapter.

13.d Summary of asymmetric information measures

The initial analysis of the sequential trade models suggested that the bid-ask spread might be a reasonable proxy for the extent of information asymmetry. From the multivariate dynamic analyses of prices and trades, other possible measures arise.

■ The trade impact coefficient, λ

The λ in models 1-4 is intuitively similar to λ in the Kyle model. It is a coefficient that measures how much a trade (or a trade innovation) moves the market price. λ is clearly identified in the structured models. The general VAR specifications that allow for multiple lags and multiple trade variables present a problem. It is not generally a good practice to pick one particular coefficient at one particular lag as "λ". Signed trade variables tend to be contemporaneously and serially correlated, leading to some multicollinearity and indeterminacy in particular VAR coefficients. A better practice is to compute, using the impulse response function, the cumulative price impact of an innovation corresponding to a representative trade.

■ Variance decomposition measures

Decomposition of the random-walk variance σ_w^2 can characterize trade-related contributions. Using the notation of the general VAR analysis, denote $\sigma_{w,x}^2 = a_Q\, \Omega_Q\, a_Q'$, an absolute measure of the trade contribution and $R_w^2 = \sigma_{w,x}^2 / \sigma_w^2$ as the relative measure.

13.e Case Study II

For your ticker symbol and date, using TaqAnalyze02.sas as a template:

1. Perform a VAR analysis of trades and prices.
   Assess the preliminary properties of trade direction indicator. 2. From the regression of price changes against current and lagged trades, determine c, λ, σ_w^2 and $R_{w,x}^2$ for the generalized Roll model.

2. Analyze the trade sign direction variable (mean, variance, autocorrelations). Fit a low-order ARMA model. (An AR(1) is a good start.) What is the innovation variance compared with Var(q_t)? (Compare a mean and demeaned model.) Comment on model fit.

3. Full VAR analysis: what proportion of the random-walk variance for your stock can be attributed to trades? Comment on difference relative to the generalized Roll model estimates in step 1.

Chapter 14. Prices and trades: structural models

The preceding section suggested approaches to broadly characterizing trade effects in microstructure models. There also exist many approaches based on structural models. This chapter discusses some representative examples.

14.a Glosten & Harris (1988)

The model is:

P_t^0	Price of transaction t
V_t	Number of shares in transaction t
T_t	Wall-clock time between transactions $t-1$ and t
Q_t	Buy/sell indicator (q_t in our notation)
m_t	Efficient price
e_t	Innovation in efficient price, $e_t = m_t - m_{t-1}$
Z_t	Adverse-selection
C_t	Transitory spread component.

(14.a.1)

$$m_t = m_{t-1} + e_t + Q_t Z_t \tag{14.a.2}$$

$$\begin{aligned} P_t &= m_t + Q_t C_t \\ P_t^0 &= \text{Round}\left(P_t, \frac{1}{8}\right) \\ Z_t &= z_0 + z_1 V_t \\ C_t &= c_0 + c_1 V_t \end{aligned} \tag{14.a.3}$$

There are a number of interesting features here.

- The change in the efficient price due to a trade is $Q_t Z_t = Q_t(z_0 + z_1 V_t)$: this reflects both directional and size effects.
- The transitory ("clerical and clearing") part of the cost in (1e) also contains a size effect.
- Price discreteness is explicitly modeled.

This model is also important because it estimated for U.S. equity data that contain trade prices and volumes, but not bid-ask quotes. This means that the trade direction indicator variables can't be constructed by comparing the trade price to the prevailing bid ask midpoint.

The Q_t are therefore unobserved ("latent") state variables. The estimation technique involves non-linear state-space filtering. When Glosten and Harris wrote the paper, this could only be carried out by numerical approximations to and integrations of the conditional state densities at each point in time. Non-linear state-space models are nowadays usually estimated by Bayesian Markov chain Monte Carlo (MCMC) methods that are easier to implement.

In applications involving U.S. equity data, the Glosten-Harris model has been superseded by approaches that use quote data, which are now widely available. There are many other markets, though, where quotes are also missing from the data record. In these cases, the Glosten-Harris model approach (except for the estimation technique) remains important.

14.b Madhavan, Richardson and Roomans (1997)

The model (in their notation):

- x_t is the trade-indicator variable, +1 if the trade is a "buy", -1 if a "sell", and zero if the trade occurred within the prevailing spread. $\Pr(x_t = 0) = \lambda$.
- The probability that a trade at the bid is followed by another trade at the bid (and similarly for the ask) is $\Pr(x_t = x_{t-1} \mid x_{t-1} \neq 0) = \gamma$.
- The first-order autocorrelation in x_t is $\mathrm{Corr}(x_t, x_{t-1}) = \rho = 2\gamma - (1 - \lambda)$.

  These assumptions imply that: $E[x_t \mid x_{t-1}] = \rho x_{t-1}$, i.e., that the innovation in the trade direction is $x_t - \rho x_{t-1}$.

The efficient price is:

$$\mu_t = \mu_{t-1} + \underbrace{\theta(x_t - E(x_t \mid x_{t-1}))}_{\text{Inferred private information}} + \underbrace{\epsilon_t}_{\text{public information}} \tag{14.b.4}$$

The trade price is:

$$p_t = \mu_t + \underbrace{\phi x_t}_{\text{Noninformational cost of trade}} + \underbrace{\xi_t}_{\text{Disturance due to price rounding}} \tag{14.b.5}$$

The model parameters are $\{\theta, \phi, \lambda, \rho, \sigma_\epsilon^2, \sigma_\xi^2\}$. However, moments can be constructed that use only the first four of these. Estimation proceeds via GMM.

14.c Huang and Stoll (1997)

The transaction price is

$$p_t = q_t + z_t \tag{14.c.6}$$

where q_t is the quote midpoint and z_t is

$$r_t^p = p_t - p_{t-1} = r_t^q + z_t - z_{t-1} \tag{14.c.7}$$

The quote midpoint return is

$$r_t^q = E[r_t^* \mid \Omega_{t-1}] + g(\Delta I_{t-1}) + \epsilon_t \tag{14.c.8}$$

where $E[r_t^* \mid \Omega_{t-1}]$ is the "consensus return" conditional on Ω_{t-1}, the public information set after the $t-1^{\text{st}}$ trade,

$g(\Delta I_{t-1})$ is an inventory-control term and ϵ_t arises from new public information (not contained in Ω_{t-1}).

$E[r_t^* \mid \Omega_{t-1}] = f\left(z_{t-1}, r_{t-1}^f\right)$ where r_t^f is the return on a stock index futures contract.

14.d The components of the spread

Historically, a prominent line of analysis in market microstructure has focused on modeling the bid-ask spread. In a dealer market, the (half) spread is an obvious and convenient measure of trading cost. Furthermore, the effects of clearing costs, inventory control and asymmetric information are in principle all reflected in the spread. It is therefore something a unifying feature in empirical analysis. Analyses along this line include Glosten (1987), Glosten and Harris (1988), Stoll (1989), George, Kaul, and Nimalendran (1991), Lin, Sanger, and Booth (1995), and Huang and Stoll (1997).

In most cases, these models have at their core a model of joint price-trade dynamics similar to the ones already considered. The specifications, however, often model cost parameters not in absolute terms, but rather relative to the spread.

The model of Huang and Stoll (1997) is illustrative.

The implicit efficient price, V_t, evolves as:

$$V_t = V_{t-1} + \underbrace{\underbrace{\left(\frac{\alpha S}{2}\right)}_{\text{Impact coefficient}} Q_{t-1}}_{\text{Asymmetric information}} + \underbrace{\epsilon_t}_{\text{Public information}} \tag{14.d.9}$$

Initially, the revision is driven by Q_{t-1}, implicitly assuming that this entire quantity is unanticipated. This is later generalized.

The quote equation contains an inventory control mechanism. The quote midpoint is M_t:

$$M_t = V_t + \underbrace{\left(\frac{\beta S}{2}\right) \sum_{i=1}^{t-1} Q_i}_{(-)\text{ Accumulated inventory}} \tag{14.d.10}$$

where $\beta > 0$. For example, after a run of positive Qs (customer buying), the MM will be short, and should adjust her quote midpoint upwards to encourage incoming sales.

$$\Delta M_t = \frac{(\alpha + \beta)\, S}{2}\, Q_{t-1} + \epsilon_t \tag{14.d.11}$$

The trade price and its first difference are:

$$P_t = M_t + \frac{S}{2}\, Q_t + \underbrace{\eta_t}_{\text{Price discreteness}} \tag{14.d.12}$$

$$\Delta P_t = \frac{S}{2}\, \Delta Q_t + \underbrace{\lambda}_{(\alpha+\beta)}\, \frac{S}{2}\, Q_{t-1} + \underbrace{e_t}_{\epsilon_t + \Delta\eta_t} \tag{14.d.13}$$

Note that while λ is identified, its individual components are not. Estimation proceeds via GMM.

The components of λ can be identified if we posit that the Q_t are autocorrelated. In this case, the trade innovation appears in the asymmetric information adjustment, while the full trade quantity appears in the inventory adjustment. The modified trade dynamics are those implied by $\Pr(Q_t \neq Q_{t-1}) = \pi$, where π is the reversal probability. Thus:

$$E[Q_t \mid Q_{t-1}] = (1 - 2\,\pi)\, Q_{t-1} \tag{14.d.14}$$

The unexpected component of the trade at time $t - 1$ is:

$$Q_{t-1} - E[Q_{t-1} \mid Q_{t-2}] = Q_{t-1} - (1 - 2\,\pi)\, Q_{t-2} \tag{14.d.15}$$

Proceeding, we obtain:

$$\begin{aligned}
\Delta V_t &= \left(\frac{\alpha S}{2}\right) Q_{t-1} - \left(\frac{\alpha S}{2}\right)(1 - 2\,\pi)\, Q_{t-2} + \epsilon_t \\
\Delta M_t &= \frac{(\alpha + \beta)\, S}{2}\, Q_{t-1} - \left(\frac{\alpha S}{2}\right)(1 - 2\,\pi)\, Q_{t-2} + \epsilon_t \\
\Delta P_t &= \frac{S}{2} Q_t + (\alpha + \beta - 1)\, \frac{S}{2}\, Q_{t-1} - \left(\frac{\alpha S}{2}\right)(1 - 2\,\pi)\, Q_{t-2} + e_t
\end{aligned} \tag{14.d.16}$$

Note that the trade process is exogenous to the price processes.

Estimation proceeds via GMM. All parameters are identified.

At present, models that concentrate on the spread and its components must contend with two developments. First, with decimalization, the spreads in U.S. equity markets have become narrow and sizes have dropped. Thus the quoted spread is less informative about the terms of trade that all but the smallest orders face. Second, with the increased prominence of electronic limit order books, the assumption that quotes and the spread are set by a dealer, or someone effectively acting as a dealer, has become less attractive.

Chapter 15. The probability of informed trading (*PIN*)

The papers to this point have assessed private information using trade/price impacts.

A series of papers (Easley, Kiefer, and O'Hara (1997), Easley, Kiefer, and O'Hara (1996), Easley, Kiefer, O'Hara, and Paperman (1996), and Easley, Hvidkjaer, and O'Hara (2002)), henceforth EHKOP, develops and implements methods of measuring information asymmetry that focuses on the trade (signed order flow) process. There are several variants, this discussion focuses on EHO (2002).

15.a Model structure

The model features information events that occur at (and only at) the beginning of the day with probability α. If an information event occurs, informed traders receive a Bernoulli signal $\delta \in \{\text{High, Low}\}$. Throughout the day, uninformed buyers arrive with Poisson intensity ϵ_b; uninformed sellers, with intensity ϵ_s. If an information event has occurred, informed traders arrive with intensity μ.

This is the event tree:

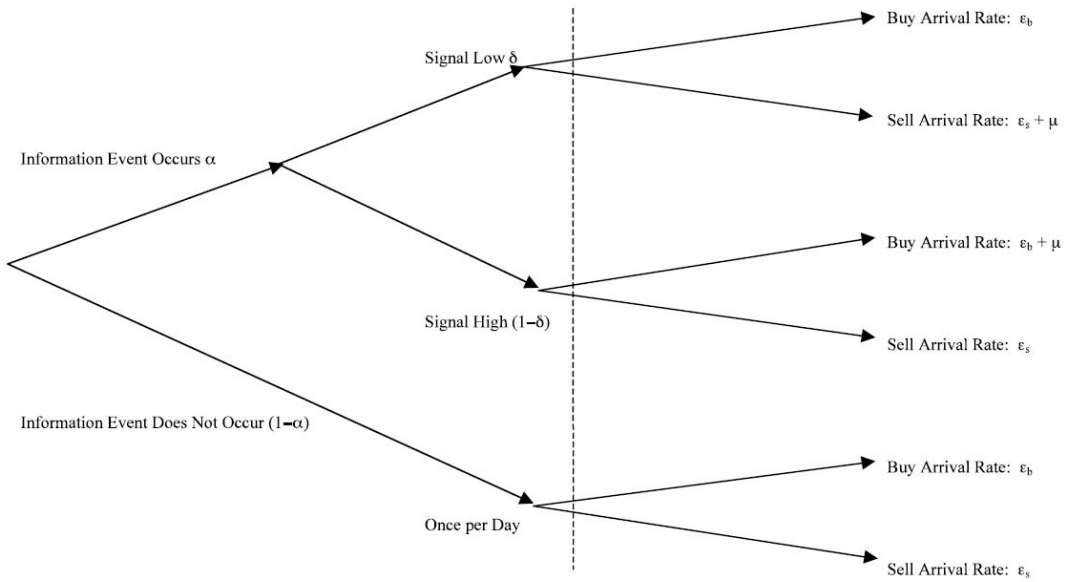


Figure 1. Tree diagram of the trading process. α is the probability of an information event, δ is the probability of a low signal, μ is the rate of informed trade arrival, ϵ_b is the arrival rate of uninformed buy orders, and ϵ_s is the arrival rate of uninformed sell orders. Nodes to the left of the dotted line occur once per day.

The summary proxy for asymmetric information is the probability of informed trading:

$$PIN = \frac{\alpha\mu}{E[B+S]} = \frac{\alpha\mu}{\alpha\mu+\epsilon_B+\epsilon_S} \tag{15.a.1}$$

This is the unconditional probability that a randomly selected trade originates from an informed trader.

Turning to inference, note first that the multivariate event arrival process is stationary within the day, but it is not ergodic. The effects of initial conditions (occurrence of the information event and signal realization) never die out. Estimation is based on the likelihood function for the number of buys (B) and sells (S) in a given day. Each day is essentially a separate observation.

The economic model is obviously a stylized one. No one would seriously suggest that information events occur only at the beginning of the day, and that signals are Bernoulli. The analysis is designed to capture certain characteristics of trade dynamics, not serve as a comprehensive model. Given these limitations, it is sensible to consider the characteristics of empirical (B, S) distributions that are likely to identify *PIN*.

The unconditional (B, S) distribution is a bivariate mixture of Poisson distributions. Defining $f(\lambda, n)$ to be the probability of n occurrences given a Poisson distribution with parameter λ, and assuming for simplicity that $\epsilon_S = \epsilon_B = \epsilon$, the unconditional density of buys and sells is:

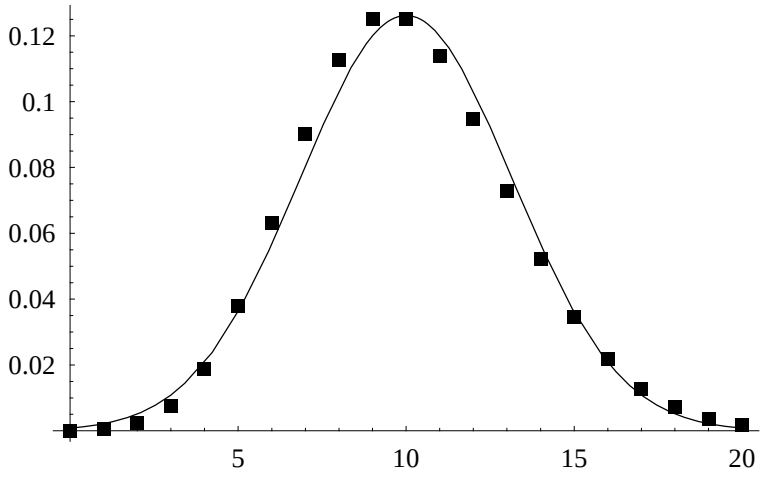
$$f(B,S) = (1-\alpha)\, f(\epsilon, B)\, f(\epsilon, S) + \alpha\,\delta\, f(\epsilon+\mu, B)\, f(\epsilon, S) + \alpha\,(1-\delta)\, f(\epsilon, B)\, f(\epsilon+\mu, S) \tag{15.a.2}$$

The Poisson distribution for n with parameter λ has:

Density	Mean	Std. Dev.
$\frac{e^{-\lambda}\lambda^n}{n!}$	λ	$\sqrt{\lambda}$

(15.a.3)

As long as λ is not too close to zero, the Poisson is approximately a discretized normal density. For example, if $\lambda = 10$ ("ten traders per day"), the Poisson and corresponding normal distributions are:



Accordingly, in what follows, we'll approximate the Poisson distribution with parameter λ by a normal distribution with mean λ and standard deviation $\sqrt{\lambda}$.

Before examining the EHKOP model in detail, it is useful to establish a few aspects of mixture distributions by examining the univariate case

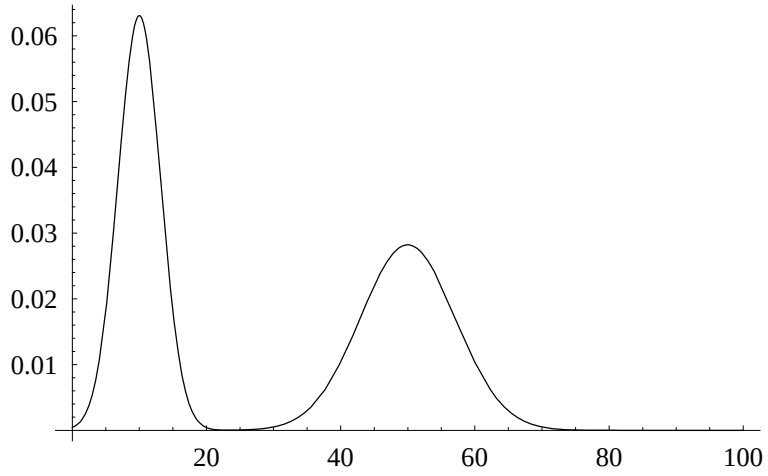
15.b A mixture of two Normal Poisson approximations

Each component density is a normal approximation to a Poisson distribution. Let $f(\lambda, x)$ denote a normal density with mean λ and standard deviation $\sqrt{\lambda}$. The mixture density is:

$$f(x) = \alpha\, f(\lambda_1, x) + (1 - \alpha)\, f(\lambda_2, x) \tag{15.b.4}$$

where α and $(1 - \alpha)$ are the mixture weights.

If λ_1 and λ_2 are very different, and $\alpha \approx 1/2$, the two component normals are distinct. With $\lambda_1 = 10,\ \lambda_2 = 50$ and $\alpha = 1/2$, the density of the mixture is:



Here, the two component distributions are clearly visible.

The components are not always so distinct. Suppose that we have a normal distribution that characterizes most sample observations, but there are a few "outliers" that we'd like to model with a more diffuse normal density. The latter is sometimes called the contaminating density. The prominence of the contaminating density in the final mixture depends on (a) how distinct it is from the base density, and (b) its mixing weight. There is a trade-off between these two features.

To take a homey analogy, suppose that we're mixing paints. We're starting with a bucket of pale yellow and want to end up with a pale orange. We can either mix in a very small amount of intense red, or a generous amount of pale pink. The results will (to my eye, at least) look very similar.

Back in the realm of probability densities, suppose that we start with $N\left(\lambda = 10,\ \sigma = \sqrt{10}\right)$ and consider mixtures constructed so that the mean remains at 10. The table below describes some mixtures and their properties.

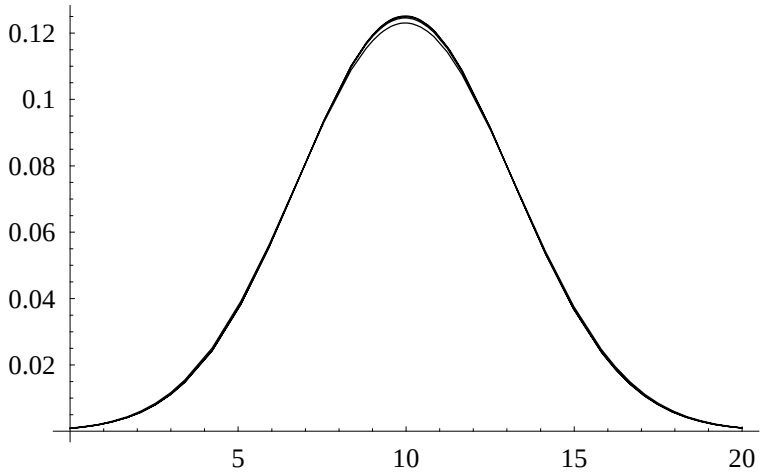
α	λ_1	λ_2	Mean	Variance
0.1	8	10.2222	10.	10.4444
0.2	9	10.25	10.	10.25
0.3	9.33333	10.2857	10.	10.1905
0.4	9.5	10.3333	10.	10.1667

(15.b.5)

α	Skewness	Kurtosis	$\alpha(10-\lambda_1)$	$(1-\alpha)(\lambda_2-10)$
0.1	0.0160931	2.97802	0.2	0.2
0.2	0.017141	2.99658	0.2	0.2
0.3	0.0153353	3.00088	0.2	0.2
0.4	0.0145673	3.00273	0.2	0.2

(15.b.6)

Recall that a normal density has zero skewness and a coefficient of kurtosis equal to three. Here is a plot of the four mixture densities:



In summary, despite the considerable variation in the mixture parameters $\{\alpha, \lambda_1, \lambda_2\}$, the mixture densities and the population moments are very similar. Statistical resolution of these mixtures on the basis of a data sample would be extremely difficult.

Recall, though, that the mean of the mixture is $\mu = \alpha\lambda_1 + (1-\alpha)\lambda_2$, implying $\alpha(\mu - \lambda_1) = (1-\alpha)(\lambda_2 - \mu)$. The parameter combinations were chosen not only to keep the mean constant, but also to hold constant these two components. They are reported as the last columns in the table.

These results suggest that for a given value of $\alpha(\mu - \lambda_1)$, we don't get much variation in sample moments by varying α or λ_1. Intuitively, the components of this product are:

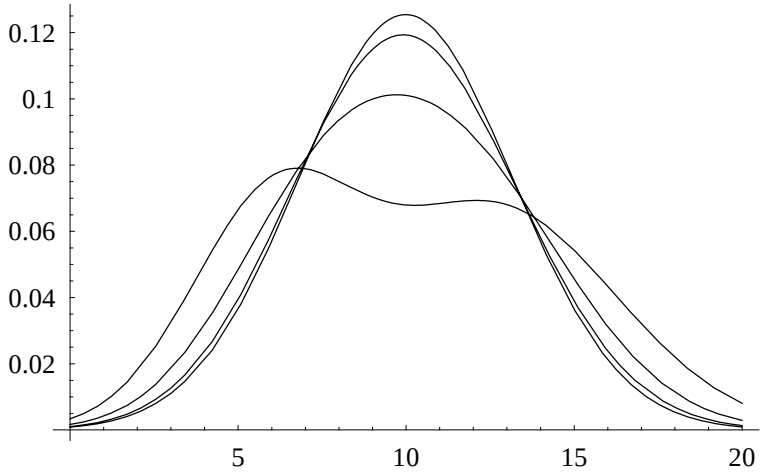
$$\underbrace{\alpha}_{\text{Mixing weight}} \times \underbrace{(\mu - \lambda_1)}_{\text{Degree of difference from base density}}$$

Is this really what's explaining the similarity of the densities, or is it merely a result of matching the mean? Here are parameter combinations that hold constant the mean, but without holding constant $\alpha(\mu - \lambda_1)$.

α	λ_1	λ_2	Mean	Variance	$\alpha(10-\lambda_1)$	$(1-\alpha)(\lambda_2-10)$
0.1	9	10.1111	10.	10.1111	0.1	0.1
0.2	8	10.5	10.	11.	0.4	0.4
0.3	7	11.2857	10.	13.8571	0.9	0.9
0.4	6	12.6667	10.	20.6667	1.6	1.6

(15.b.7)

The densities are much more distinct:



We've informally shown:

- When the parameters are varied in a way that keeps $\alpha(\mu - \lambda_1)$ fixed, the mixture distribution barely changes.
- When the parameters are varied in a way that does not keep $\alpha(\mu - \lambda_1)$ fixed, the mixture distribution varies substantially.

This suggests that the product $\alpha(\mu - \lambda_1)$, which involves the mixing parameter and a shape parameter, is likely to be better identified in a data sample. More precisely, in a GMM procedure based on matching the first four moments of the density, we'll typically find that the estimated precision of the parameter estimates is low and that the covariance matrix of parameter estimates exhibits strong positive correlation in estimation errors for α and λ_1 (as well as for α and λ_2). The estimates of $\alpha(\mu - \lambda_1)$ and $(1 - \alpha)(\lambda_2 - \mu)$, constructed as functions of the parameters will be relatively precise.

The importance of this point to the EHKOP model arises in the fact that, like $\alpha(\mu - \lambda_1)$ in above case, *PIN* is

essentially the product of a mixing weight and a parameter difference.This greatly enhances the precision with which *PIN* might be estimated.

We now return to the EHKOP model.

15.c Mixture aspects of EHKOP

For simplicity, we'll consider the case where $\epsilon_B = \epsilon_S = \epsilon$ and $\delta = \frac{1}{2}$. The approximation to the mixture of Poissons is then the mixture of normals:

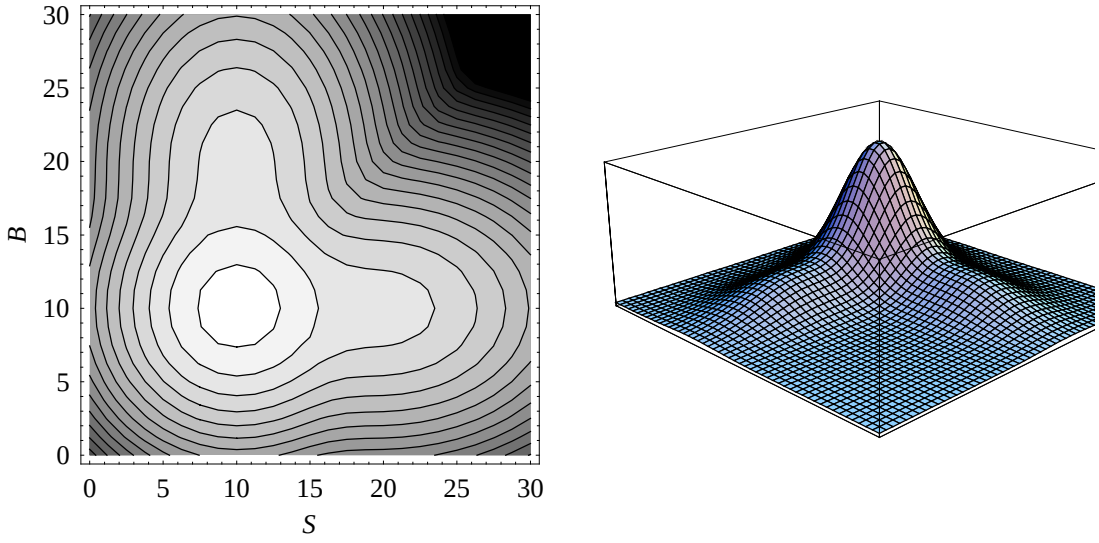
$$f(B,S) = (1-\alpha)\, f(\epsilon, B)\, f(\epsilon, S) + \frac{1}{2}\,\alpha\, f(\epsilon+\mu, B)\, f(\epsilon, S) + \frac{1}{2}\,\alpha\, f(\epsilon, B)\, f(\epsilon+\mu, S) \tag{15.c.8}$$

As a numerical illustration, we'll use the test values:

$$\{\alpha = 0.4,\ \epsilon = 10,\ \mu = 10\} \tag{15.c.9}$$

This implies that on a day with no information event, we expect to see ten buys and ten sells. With an information event, if the news is positive, we expect to see twenty buys and ten sells. The figures below depict contour and 3D plots of the unconditional distribution of buys and sells (using the normal approximation).

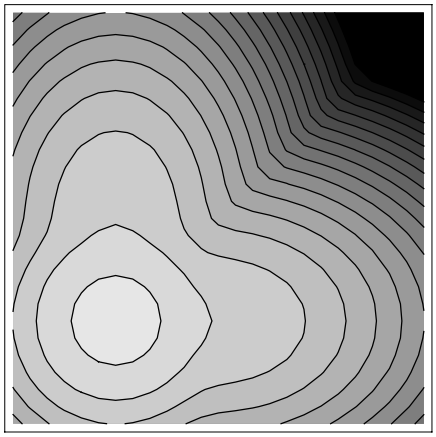
Probability density for the number of buys and sells



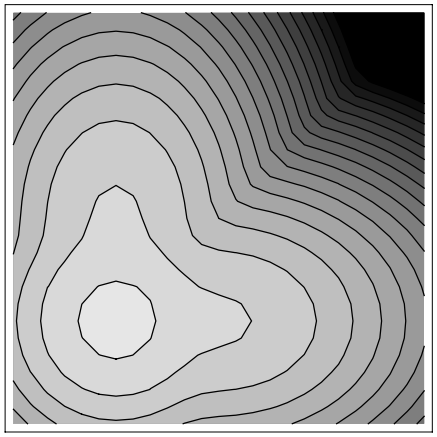
The "base" density here is a bivariate normal centered at ten buys and ten sells. The stretching of the density along the B and S axes reflects the higher arrival rates on information days. The canyon between the two lobes occurs because if an information event occurs, good and bad news are mutually exclusive.

The density above is representative. Changes in the parameter values can dramatically distort the picture, as the following examples show.

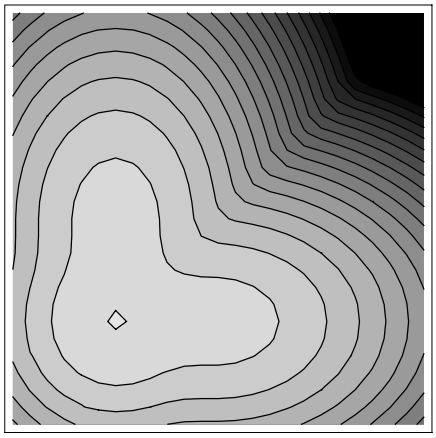
μ=10 α=0.1 ϵ=10 PIN=0.05



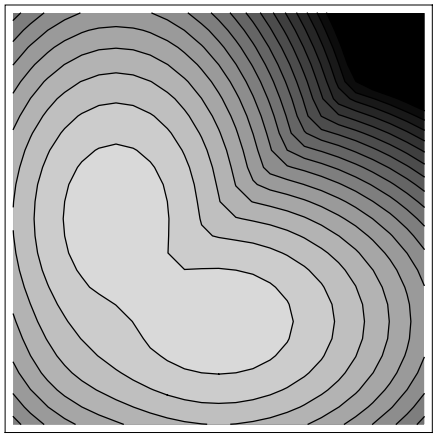
μ=10 α=0.2 ϵ=10 PIN=0.09



μ=10 α=0.5 ϵ=10 PIN=0.2



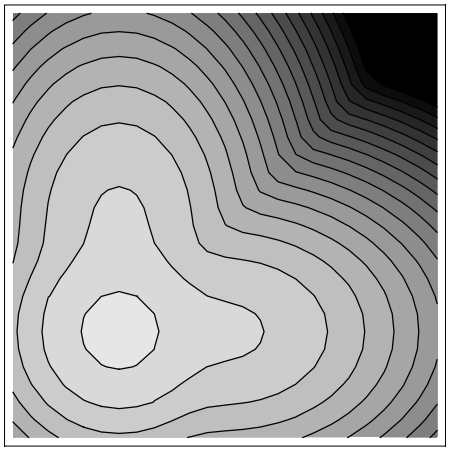
μ=10 α=0.99 ϵ=10 PIN=0.33



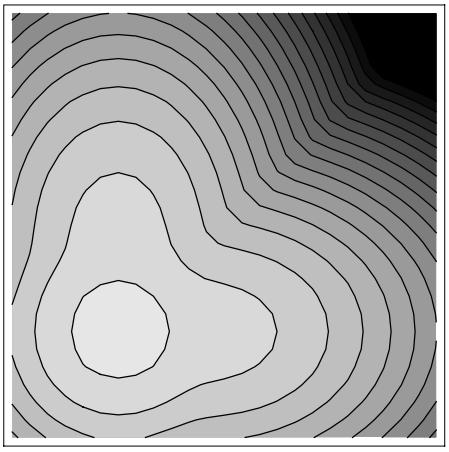
When the characteristics of a distribution are strongly dependent on parameter values, different sample distributions will imply different parameter values, i.e., the data are likely to informative in estimating the parameters with precision.

PIN, however, is a derived, summary quantity. Suppose that we investigate a set of distributions where *PIN* is held constant at 0.10.

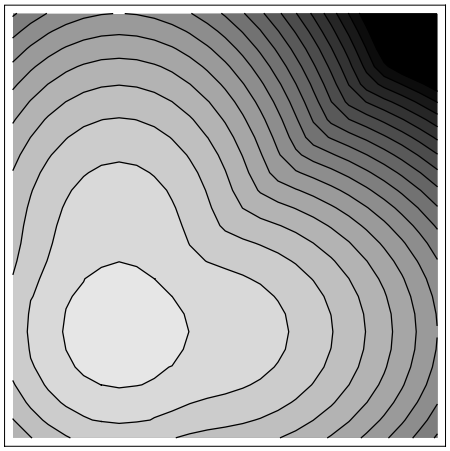
μ=10 α=0.22 ϵ=10 PIN=0.1



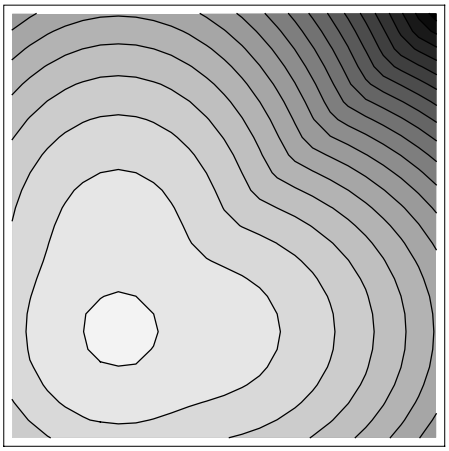
μ=9 α=0.25 ϵ=10 PIN=0.1



μ=8 α=0.28 ϵ=10 PIN=0.1



μ=7 α=0.32 ϵ=10 PIN=0.1



Although μ and α vary considerably, the distributions look quite similar. This suggests that *PIN* might be estimated precisely, even if this is not the case for the individual parameters. A large α and small μ have effects similar to small α and large μ.

15.d Summary

The EHKOP approach to characterizing information asymmetry is based solely on signed trades, and can be estimated from daily counts of the number of buy and sell orders (*B* and *S*). The summary statistic, probability of informed trading (*PIN*), is driven by the frequency and magnitudes of buy-sell imbalances $|S - B|$.

PIN is most strongly driven by the product $\alpha\,\mu$. The preceding analysis of the mixture distribution suggests that *PIN* is likely to be well-identified, even though α and μ might be less distinct. By the same reasoning, we'd also expect *PIN* to be well-identified if α and μ varied across the sample in such a way that relatively

high α was accompanied by low μ, and vice versa. In economic terms, this might arise if information events that were relatively frequent had fewer informed traders.

On non-information days, buys and sells will arrive randomly: there will be no autocorrelation in the signed order flow. On information days, though, the preponderance of buys or sells will imply greater likelihood of one-sided runs. Thus, a high *PIN* is equivalent to positive intraday autocorrelation in buys and sells.

Although positive autocorrelation of buys and sells is generally a feature of the sequential trade analyses , it is not a feature of the Kyle-type auction models. In the latter, the signed net order flow is serially uncorrelated. That is, if we have multiple market clearings during the day, the sequence of net orders at each clearing is uncorrelated. (The market maker does not observe buys and sells separately, however.)

In contrast, the joint dynamics of orders and price changes are common to both sequential trade and sequential auction approaches: a buy order (or net buy order) moves prices upwards (in expectation, permanently). This commonality suggests that specifications that focus on trade/price dynamics (such as the VAR approaches) might provide better characterizations of asymmetric information.

In any event, if a sequential trade model has implications for order price impacts, shouldn't we use these implications (and the price data) in our estimation? The answer is not obviously "yes". A more comprehensive statistical model should in principle lead to more precision in the estimates, but only if the model is correctly specified. It may be the case that inferences based solely on trades are more robust to misspecification that models based on joint dynamics.

As an additional consideration, one important mechanism may cause both VAR and *PIN* approaches to yield similar inferences. Many markets are characterized by quotes or limit orders that are not updated promptly in response to public announcements. In response to an announcement, market-order traders ("day traders") successively hit one side of the market until the stale orders have been exhausted. In a *PIN* analysis, this mechanism leads to high trade autocorrelation (and a large *PIN* estimate). A VAR analysis of the same data will typically attribute the quote changes to the incoming trades (rather than public information). A high price impact coefficient would also be viewed as evidence of information asymmetries. With respect to this mechanism, therefore, both VAR and *PIN* estimates are likely to lead to the same conclusion. Whether this conclusion is correct depends on whether one views the original announcement as public (because it was delivered by a broad medium like a newswire) or private (because only a subset of agents [the market order traders] had the opportunity to use the information in implementing their strategies).

Chapter 16. What do measures of information asymmetry tell us?

The exposition to this point has focused on how spreads, trade autocorrelations and price impact coefficients can all be used as microstructure-based proxies for information asymmetries.

Increasingly, these measures are being used in corporate finance and accounting studies where the need for such measures is compelling. A partial list of representative studies includes: Lee, Mucklow and Ready (1993); Dennis and Weston (2001); Ertimur (2003); Sunder (2003); Odders-White and Ready (2003).

At the same time, there are studies that suggest caution. Neal and Wheatley (1998) point out that spreads on closed end mutual funds are too large to be explained by information asymmetries (given the relatively transparent valuation of these portfolios). Saar and Yu (2003) examine spreads around revisions in the Russell indexes. These revisions are algorithmic and predictable. Saar and Yu suggest that this spread variation cannot, therefore, be linked to cash-flow uncertainty.

Furthermore, order flows in the Treasury bond and foreign exchange markets appear to have price impacts (Lyons (2001)). Shall these impacts be attributed to asymmetric information about interest rates? trade patterns? There have been a few cases in the U.S. of individuals trading on prior knowledge of government announcements, but these are relatively rare.

Most of the asymmetric information models focused, implicitly at least, on equity markets. In these markets, there are obviously large sources of uncertainty, and information endowments and production that are not uniform across agents. It is therefore natural to characterize private information as valuable to the extent that it predicts long-run, persistent changes in value.

Lyons points out that private information about *transient* price components may also have value. An agent who can buy low and sell high will make money even if these prices arose from "temporary" effects.

Chapter 17. Linked prices: cointegration and price discovery

The predictions of many interesting economic models concern multiple prices. "Multiple prices" in this context covers bid, ask, and trade prices, possibly for different securities, possibly for different markets. Often the economic hypotheses suggest arbitrage or other connections among the prices. This chapter discusses model specification in such situations.

The presentation is based on Engle and Granger (1987); Hasbrouck (1995); Lehmann (2002); de Jong (2002); Baillie, Booth, Tse, and Zabotina (2002); Harris, McInish, and Wood (2002a, 2002b); and, Hasbrouck (2002). Werner and Kleidon (1996), Hasbrouck (2003) and Chakravarty, Gulen, and Mayhew (2004) are representative applications.

17.a Two securities

Suppose that we have two securities that each behave in accordance with the simple Roll model, that is for $i = 1, 2$, we have:

$$\begin{aligned} m_{i,t} &= m_{i,t-1} + u_{i,t} \\ p_{i,t} &= m_{i,t} + cq_{i,t} \end{aligned} \tag{17.a.1}$$

What sort of joint dynamics are economically reasonable?

Both efficient prices follow random walks. The $u_{i,t}$ increments might be correlated, reflecting common dependence on macro or industry factors. But if the values of the two securities are subject to different firm-specific factors, then the correlation will be less than perfect. The two securities might appear to move together in the short-run. But in the long-run, the cumulative effect of the firm-specific factors will tend to cause the prices to diverge.

The behavior is different when the efficient prices are identical, $m_{1,t} = m_{2,t} = m_t$. This might occur when the two prices refer to the same security traded in different markets. With transaction costs and some degree of market separation, we would no longer expect arbitrage to ensure $p_{1,t} = p_{2,t}$. Hence the two prices might diverge in the short run. In the long-run, though, arbitrage and substitutability would almost certainly limit the divergence between the two prices. Thus, we'd expect the difference $p_{1,t} - p_{2,t}$ to be stationary.

When two variables are integrated of order one (i.e., contain random-walk components), they are said to be cointegrated if there exists a stationary linear combination of the variables. For example, the price of IBM on the NYSE and that on the Pacific exchange both contain random-walks, but the difference between the two prices does not diverge.

The econometrics of cointegrated systems are often quite complex. Much of this complexity arises because, in macro applications, we need to test whether the series contain random walks in the first place, then

whether they are in fact cointegrated, and finally we need to estimate the cointegrating vector (the weights in the stationary combination).

In microstructure applications, these issues are of distinctly secondary importance. The larger questions remain and are very pertinent. Do stock prices follow a random-walk, or are they trend-stationary? Are they cointegrated with consumption?, etc. It's not that we consider these issues settled or trivial. Instead, we admit at the outset that our data, high-frequency observations over short lengths of calendar time, are unlikely to have any power in resolving these larger questions. Microstructure models are best viewed as overlays on fundamental economic processes that capture short-term trading effects.

From this perspective, we assume at the outset that security prices are integrated. Obvious economic relationships or arbitrage principles dictate the cointegrating vectors. Accordingly, the main concern is representation and estimation.

17.b One security, two markets

For simplicity, we'll initially consider the case where a single security trades in different markets. We have a common efficient price, and the model becomes:

$$\begin{aligned} m_t &= m_{t-1} + u_t \\ \begin{pmatrix} p_{1,t} \\ p_{2,t} \end{pmatrix} &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} m_t + \begin{pmatrix} c_1\, q_{1,t} \\ c_2\, q_{2,t} \end{pmatrix} \end{aligned} \tag{17.b.2}$$

The cost parameters c_1 and c_2 are market-specific: the two markets might have different spreads. $\mathrm{Var}\begin{pmatrix} q_{1\,t} \\ q_{2\,t} \end{pmatrix} = \begin{pmatrix} 1 & \rho_q \\ \rho_q & 1 \end{pmatrix}$, reflecting the possibility that trade directions in the two markets are contemporaneously correlated.

It is easy to verify that the price changes Δp_t are jointly covariance stationary and that the autocovariances of order two or more are zero. Invoking the Wold result, we have a VMA of order 1:

$$\begin{pmatrix} \Delta p_{1,t} \\ \Delta p_{2,t} \end{pmatrix} = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} \begin{pmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \end{pmatrix} \tag{17.b.3}$$

Consider the forecast future prices:

$$E_t \begin{pmatrix} p_{1,t+1} \\ p_{2,t+1} \end{pmatrix} = \begin{pmatrix} p_{1,t} \\ p_{2,t} \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \tag{17.b.4}$$

Since dependencies in this model die out after the first lag, $E_t[p_{i,t+k}] = E_t[p_{i,t+1}]$ for $k \geq 1$.

By a generalization of Watson's argument, these forecasts are equal to the projection of m_t onto current and past prices. Since m_t is identical for both securities, these projections have be identical, i.e., we must have $E_t\, p_{1,t+1} = E_t\, p_{2,t+1}$. The revisions in these forecasts are:

$$E_t\begin{pmatrix} p_{1,t+1} \\ p_{2,t+1} \end{pmatrix} - E_{t-1}\begin{pmatrix} p_{1,t} \\ p_{2,t} \end{pmatrix} = \begin{pmatrix} \Delta p_{1,t} \\ \Delta p_{2,t} \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix}\left(\begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} - \begin{pmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \end{pmatrix}\right) = \left(I + \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix}\right)\begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \quad (17.b.5)$$

Now since the forecasts are identical, the revisions in the forecasts must also be identical. Thus:

$$(1 + \theta_{11} \quad \theta_{12}) = (\theta_{21} \quad 1 + \theta_{22}) \quad (17.b.6)$$

The variance of the (common) random-walk component is

$$\sigma_w^2 = \beta\, \Omega\, \beta' \text{ where } \beta = (\beta_1 \quad \beta_2) = (1 + \theta_{11} \quad \theta_{12}) = (\theta_{21} \quad 1 + \theta_{22}) \quad (17.b.7)$$

where $\Omega = \text{Var}(\epsilon_t)$.

The situation here is analogous to the one in which we attempted to decompose σ_w^2 into trade- and non-trade-related components. Here, though, the decomposition is between contributions attributable to each of the two markets.

For the case where Ω is diagonal, Hasbrouck (1996) defines the information share of the ith market to be

$$IS_i = \frac{\beta_i^2 \text{ Var}(\epsilon_{i,t})}{\sigma_w^2} \quad (17.b.8)$$

When Ω is nondiagonal, lower and upper bounds on the information share may be computed by investigating alternative Cholesky factorizations.

In the earlier single-security decomposition of σ_w^2, the trade-related component, $\sigma_{w,x}^2$ was viewed as reflecting the market's reaction to private information signalled by the trade, and the remainder, $\sigma_w^2 - \sigma_{w,x}^2$ was then attributable to public non-trade information.

Here, the variance attributions measure the relative amounts of information production in both markets. In the present case, among other things, the shares will depend on the relative magnitudes of c_1 and c_2. If $c_1 < c_2$, then the price in market 1 is effectively a higher precision signal, which is reflected in a higher information share.

One might hope, following the earlier developments that used VMAs, that one could specify a VAR for the Δps, which could then be estimated and inverted to obtain the above VMA. It turns out, though, that such a VAR representation does not exist: in the presence of cointegration, the VMA is non-invertible.

Fortunately, the VMA structure can be recovered from a vector specification that is a slight generalization of the VAR, specifically, an error-correction model. An error correction model includes, in addition to the usual lagged values, a term defined by the cointegrating vectors. In the present case, for example, the error correction model could be specified as:

$$\Delta p_t = \phi(L)\, \Delta p_t + \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} (p_{1,t-1} - p_{2,t-1}) + \epsilon_t \quad (17.b.9)$$

The middle r.h.s. term includes an "error" (i.e., deviation) and coefficients that reflect "correction" (i.e., adjustment response). A tendency for $p_{1,t}$ to move toward $p_{2,t}$ would suggest $\gamma_1 < 0$, while a tendency for

$p_{2,t}$ to move toward $p_{1,t}$ would suggest $\gamma_2 > 0$. There is an arbitrary normalization here. We could have defined the error as $p_{2,t-1} - p_{1,t-1}$, which would merely flip the signs on the coefficients.

We may estimate a truncated version of this system. Then compute the VMA representation by calculating the impulse response functions subsequent to orthogonal unit shocks (as was done for the price/trade VARs considered earlier).

17.c The general case of multiple prices

Let $p_t = (p_{1\,t} \quad p_{2\,t} \quad \cdots \quad p_{n,t})$ where all prices refer to the same security. At this level of generality, these prices might be trade prices, bids or asks in different markets. A general VECM specification is then:

$$\Delta p_t = \phi_1 \Delta p_{t-1} + \phi_2 \Delta p_{t-2} + + \phi_K \Delta p_{t-K} + \gamma(\alpha - z_{t-1}) + \epsilon_t \tag{17.c.10}$$

In the error correction term, $z_t = A p_t$ defines the cointegration vectors. Here, one possible choice for A is the coefficient matrix that defines the differences relative to the first price:

$$A p_t = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix} \begin{pmatrix} p_{1\,t} \\ p_{2\,t} \\ \vdots \\ p_{n\,t} \end{pmatrix} = \begin{pmatrix} p_{1\,t} - p_{2\,t} \\ p_{1\,t} - p_{3\,t} \\ \vdots \\ p_{1\,t} - p_{n\,t} \end{pmatrix} = z_t \tag{17.c.11}$$

Note that A is $(n-1) \times n$, and for conformability, α must be $(n-1) \times 1$ and γ must be $n \times (n-1)$.

An element of α, α_k has the interpretation of being the mean ("long run average") value of $p_{1\,t} - p_{k\,t}$. From the error correction perspective, α_k is the value of $p_{1\,t} - p_{k\,t}$ consistent with "equilibrium" or "stability", in the sense that if $z_{k\,t} = \alpha_k$, this component has no effect on the current dynamics. If the p_t are all trade prices, it would be reasonable to take $\alpha = 0$. But if the prices are bids and asks, possibly in different markets, then non-zero a_k will arise from the fact that bids are generally below ask prices, and the bid in one market might generally be higher (more aggressive) than the bid in another market.

A, α, and γ are not unique. If we take an arbitrary nonsingular square matrix of order $n-1$, denoted R, then the error correction term $\gamma(\alpha - z_{t-1}) = \gamma(\alpha - A p_{t-1}) = \gamma R^{-1}(R\,\alpha - RA p_{t-1})$. More formally, our choice of A is no more (or less) than a linear basis for the space of possible alternatives.

It is often cleaner to tell an economic story based on deviations relative to one particular price, such as the trade price in the presumptively "dominant" market. One might then describe dynamics in terms of adjustment toward or away from this price. From this, it is a small step to attributing this adjustment to the market or agents in the market. Beyond identifying a generally infinite set of possible adjustment mechanisms, however, the econometrics provide no support for such attributions.

Although in microstructure applications we can usually specify A a priori, γ and α must generally be estimated jointly. OLS is not, therefore, feasible, and we must employ nonlinear least square procedures instead.

The equivalent VMA representation may be derived by constructing the impulse response functions subsequent to one-unit shocks in each of the individual innovations. Letting this representation be:

$$\Delta p_t = \Theta(L)\,\epsilon_t \text{ where } \mathrm{Var}(\epsilon_t) = \Omega \tag{17.c.12}$$

Given a basis for the cointegrating vectors, the VMA is invariant to all linear transformations of this basis. This suggests that the VMA is a more reliable construct for inference than the γ or α. Among other things, the impulse response functions are invariant to basis rotations.

With a common efficient price, it can be shown that all of the rows of $\Theta(1)$ are equal, where $\Theta(1)$ is the sum of the VMA coefficient matrices. Defining β as any of these rows, the variance of the efficient price increments is $\sigma_w^2 = \beta\,\Omega\,\beta'$. This may be decomposed into absolute and relative contributions using methods already discussed.

■ Price discovery

In general, decompositions of σ_w^2 imply attributions of information origination.

When the elements of p_t are prices from different markets, the VECM describes the adjustment dynamics among the different markets. This is often of great interest from an industrial organization perspective.

When the elements of p_t are prices from different sets of traders, the decompositions may indicate who is (viewed by the market as) informed.

17.d Sources of cointegration

■ Linear arbitrage conditions

Often a set of prices is subject to an arbitrage relationship. Suppose, that $p_{1\,t},\ p_{2\,t},\ \ldots,\ p_{n\,t}$ are n components of an index given by $p_{I,t} = b\,p_t$, where b is the row vector of index weights. For an index futures contract with price $p_{f\,t}$, the no-arbitrage condition may be written as $p_{f\,t} = b\,p_t + c$ where c is the cost-of-carry (or the fair-value basis). A sensible VECM would be based on the augmented price vector $p_t^* = (\ p_{f\,t}\quad p_{1\,t}\quad \cdots\quad p_{n\,t}\ )'$

$$\Delta p_t^* = \Phi(L)\,\Delta p_t^* + \gamma(c - (\ 1\quad -b\ )\,p_{t-1}^*) + \epsilon_t \tag{17.d.13}$$

There are now multiple random-walk components:

$$\mathrm{Var}(w_t) = \begin{pmatrix} \mathrm{Var}(w_{f\,t}) & \mathrm{Cov}(w_{f\,t},\,w_{1\,t}) & & \\ \mathrm{Cov}(w_{1\,t},\,w_{f\,t}) & \mathrm{Var}(w_{1\,t}) & & \\ & & \ddots & \\ & & & \mathrm{Var}(w_{n\,t}) \end{pmatrix} = \Theta(1)\,\Omega\,\Theta(1)' \tag{17.d.14}$$

The rows of $\Theta(1)$ will not be identical. $\Theta(1)$ will be of rank n.

In these sorts of estimations, several issues typically arise. First, c is generally changing between days, reflecting forward-spot convergence, declining cost of carry, and a sudden upward jump when the contract is rolled over into the next maturity. Second, c is generally constant within the day, since none of its components (interest, dividends, etc.) accrue within the day. In practice, we can either determine a day's c by estimation of the model within a day, by modeling the cost determinants of c, or by fitting a time-trend (more properly, a step function that jumps overnight).

The error correction term in a VECM, $\gamma(\alpha - z_{t-1})$ reflects a speed of adjustment that depends on the magnitude of the error. In arbitrage situations, there may be no adjustment at all until the error reaches the transaction cost bounds of the marginal trader. Threshold error correction approaches are a way to model these dynamics.

A related situation involves a stock trading in two different currencies. The arbitrage relationship here involves the price of the stock in each currency, and the exchange rate. For most purposes, it is reasonable to assume the exchange rate exogenous to the stock dynamics. Going even further, since exchange rate variation is generally much smaller than stock variation, the foreign exchange rate might even be assumed to be fixed.

■ Nonlinear arbitrage conditions

The arbitrage relations linking underlying and derivative securities are often nonlinear in the prices. These may be accommodated by inverting the arbitrage relationship to restate all prices in terms of the price of one security (usually the underlying).

For example, suppose that the theoretical value of a call option is $C_t = f(S_t)$. An error correction model might be specified in terms of $S_t - f^{-1}(C_t)$, where $f^{-1}(C_t)$ is the stock price implied by the call value.

17.e Case Study III

For your symbol and date, using TaqAnalyze03.sas as a template:

1. Replicate VECM for NYSE bid and ask
2. Determine information shares for NYSE and Chicago bids.
3. In the set (NYSE bid, NYSE ask, Chicago bid, Chicago ask) estimate a bivariate VECM for any pair not consider above (e.g., NYSE bid, Chicago ask). Compute information shares.
4. (Optional) Estimate a VECM for all four bids and asks. Determine the *joint* information share of NYSE bids and asks. What are the min and max?

Part III: Limit orders

The models in Parts I and II are most clearly viewed in settings where dealers set the bid and ask quotes, and outside customers arrive and trade against these quotes. The models in this section examine limit orders and markets organized around them.

Chapter 18. Limit orders and dealer quotes

18.a Overview

A limit order is usually defined as a customer order that specifies quantity and price, and is subject to risk of non-execution. For example, if the market is 100 bid, offered at 101, a customer order to buy 100 shares at 100.2 won't normally be executed immediately. In US equity venues, under most circumstances, the customer bid would be publicly disseminated as the new prevailing market bid. There are two principal outcomes. A customer sell order might arrive that hits the original buy order, causing a transaction at 100.2. On the other hand, the market might "move away" from the order: bids and asks might rise, leaving the orginal order unexecuted.

The customer limit buy order is functionally the same as a dealer bid. In fact, limit buy orders are often simply called "bids". This similarity is extremely important because it implies that customer limit orders compete with dealer quotes. The tension between these two sorts of agents is an ongoing consideration in market structure evolution and regulation.

The similarity also facilitates a useful modeling fiction. The analyses developed in Parts I and II are for the most part models of risk-neutral dealer behavior. In applying these models (or their statistical counterparts) to markets in which limit orders play a large role, we sometimes assume that the limit order traders are identical to dealers, subject to the same costs and objectives.

In taking this position, though, it is usually necessary to broaden our concept of private information. Customers who place limit orders usually don't monitor their orders as closely as dealers. Accordingly, public limit orders are often "stale" in the sense that they don't reflect up-to-the-second public information. In response to a public new announcement (e.g., on a newswire), market orders will quickly "pick off" the stale orders. Should we view these market orders as motivated by private information?

I now turn to the differences between customer limit orders and dealer quotes. There are many, but the literature emphasizes two:

- The first difference between dealer and customer concerns the former's ability to condition on size of the incoming order. Dealers in US equity markets often post aggressive bids and asks for small quantities. By law the quotes must be firm for these quantities. But suppose that a customer order arrives for a larger quantity. After trading the small quantity at the posted quote, the dealer often has some discretion in how much to trade of the remainder, and at what price. This discretion is not absolute, owing to constraints of regulation and reputation. But it does exist. In any event, the dealer knows the full size of the order. The customers in the limit order book do not, and this (it will be shown) causes them to price their orders somewhat less aggressivley.

The second difference arises from the different objectives of customers and dealers. The dealer sets her bid so that, if hit, she can make a profit by quickly reversing the trade (selling) at a higher price. The customer's strategy is typically motivated by a need to acquire the security for reasons of hedging or long-term portfolio objectives. The dealer's alternative to placing the bid is to not participate in the market at all. The customer's alternative is accomplishing the trade with a market order.

Using a market order, the customer can buy immediately at 101. Using the limit order, the customer might buy at a lower price, but might also leave the market empty-handed. From the customer's viewpoint, then, execution uncertainty is an important aspect of the problem.

This chapter explores the first consideration; the next two chapters, the second. Finally I discuss equilibrium models.

18.b Limit order placement when faced with incoming orders of varying size

The framework here is a market with two sorts of traders.

- Market order traders. They are motivated by some combination of liquidity needs and superior information.
- Passive liquidity suppliers. These agents supply the limit orders that populate the book. They are risk-neutral agents who are subject to a zero-expected profit condition. They differ from the competitive dealers in the sequential trade models in the sort of price schedule they can quote. (Offering liquidity through the book, they can't condition on the size of the incoming order.)

The classic article here is Glosten (1994); the analysis below focuses on a special case due to Sandas (2001).

Sandas' framework is time-homogeneous. The security value (conditional on public information) is X_t, with dynamics:

$$X_t = X_{t-1} + d_t . \tag{18.b.1}$$

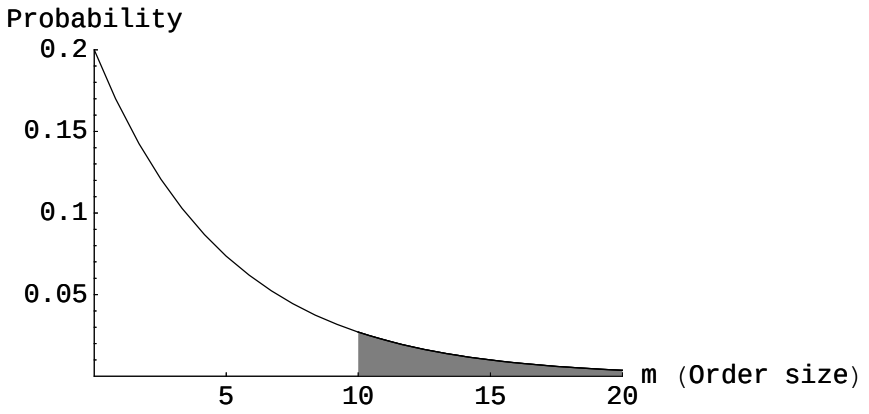
The increment d_t reflects the information content of orders that have arrived through time t and additional public non-trade information.

We'll analyze only the sell side of the book and arriving buy orders. Treatment of the bid side is symmetrical. The ask (sell) side of the book is described by a price vector $(p_1 \quad p_2 \quad \dots \quad p_k)$ ordered so that p_1 is the lowest (most aggressive). The associated vector of quantities is $(Q_1 \quad Q_2 \quad \dots \quad Q_k)$.

The incoming order is m ("shares") signed positively for a buy order (and negative for a sell order). Conditional on the incoming order being a buy, the distribution of m is:

$$f_{\text{Buy}}(m) = \frac{e^{-\frac{m}{\lambda}}}{\lambda} \tag{18.b.2}$$

The essential feature in this model can be illustrated as follows. With $\lambda = 5$, $f_{\text{Buy}}(m)$ is:



Consider the seller whose limit order is at the margin when the total quantity is 10. His order will execute when the incoming order size is 10 *or greater* (the shaded area). The price of his order must be set to recover the information costs associated with these larger orders.

The revision in beliefs subsequent to the order is given by:

$$E[X_{t+1} \mid X_t, m] = X_t + \alpha m$$

where $\alpha > 0$.

The order processing cost is γ. If a limit order priced at p_1 is executed, the profit (per unit traded) is.

$$p_1 - \gamma - E[X_{t+1} \mid X_t, m] = p_1 - X_t - \gamma - \alpha m \tag{18.b.3}$$

This will generally be positive for small m, but negative for large m. If we could condition on the size of the order, we'd impose a zero-expected profit condition for all m.

$$m = \frac{p - \gamma - X_t}{\alpha} \tag{18.b.4}$$

Suppose that the sell limit orders at the price p_1 are ordered in time priority, I wish to sell an infinitesimal amount at p_1 and that the cumulative quantity (my order plus everyone who's ahead of me) is q.

My order will execute if the incoming quantity is at least as high as q. Define $I(m \geq q)$ as the indicator function for this event (execution).

My expected profit conditional on execution is

$$\mathrm{E}\pi_1 = \mathrm{E}[p_1 - X_t - \gamma - \alpha m \mid m \geq q] = \int_q^\infty (p_1 - X - \gamma - \alpha m) f_{\mathrm{Buy}}(m) dm = -e^{-\frac{q}{\lambda}} (X + \gamma + \alpha (q + \lambda) - p_1) \tag{18.b.5}$$

I will be indifferent to adding my order to the queue at this price when $q = Q_1$ where

$$Q_1 = \frac{-X - \gamma - \alpha \lambda + p_1}{\alpha} \tag{18.b.6}$$

This might be negative for X just below p_1. In this case, $Q_1 = 0$.

Now suppose that I want to sell at p_2. $E[(p_2 - X - \gamma - \alpha m)\, I(m \geq Q_1 + q)]$

$$\mathrm{E}\pi_2 = \int_{Q_1+q}^\infty (p_2 - X - \gamma - \alpha m) f_{\mathrm{Buy}}(m) dm = -e^{-\frac{q+Q_1}{\lambda}} (X + \gamma + \alpha (q + \lambda) - p_2 + \alpha Q_1) \tag{18.b.7}$$

Which implies:

$$Q_2 = \frac{-X - \gamma - \alpha \lambda + p_2 - \alpha Q_1}{\alpha} \tag{18.b.8}$$

... and at p_3

$$\mathrm{E}\pi_3 = \int_{Q_1+Q_2+q}^\infty (p_2 - X - \gamma - \alpha m) f_{\mathrm{Buy}}(m) dm = e^{-\frac{q+Q_1+Q_2}{\lambda}} (-X - q\alpha - \gamma - \alpha \lambda + p_3 - \alpha Q_1 - \alpha Q_2) \tag{18.b.9}$$

and:

$$Q_3 = \frac{-X - \gamma - \alpha \lambda + p_3 - \alpha Q_1 - \alpha Q_2}{\alpha} \tag{18.b.10}$$

And so forth. In general:

$$Q_{k\_} = \mathrm{If}\left[k < \lfloor X \rfloor + 1,\ 0,\ -\lambda + \frac{-X - \gamma + p_k}{\alpha} - \sum_{j=\lfloor X \rfloor + 1}^{k-1} Q_j\right] \tag{18.b.11}$$

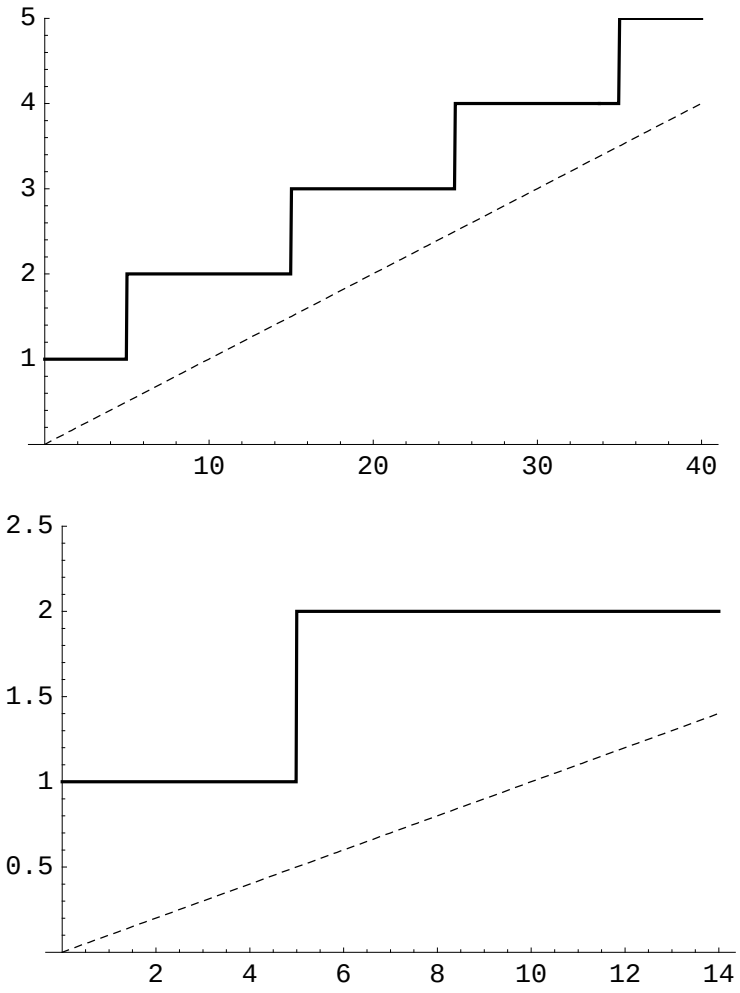
where $\lfloor X \rfloor$ is the floor of X, i.e., the largest integer less than or equal to X.

Normalize the price grid so that the tick size is unity: $p_k = k$.

As an example, consider the numerical values:

$$\{X = 0,\ \alpha = 0.1,\ \gamma = 0,\ \lambda = 5\} \tag{18.b.12}$$

Here are the book schedule and value revision function:

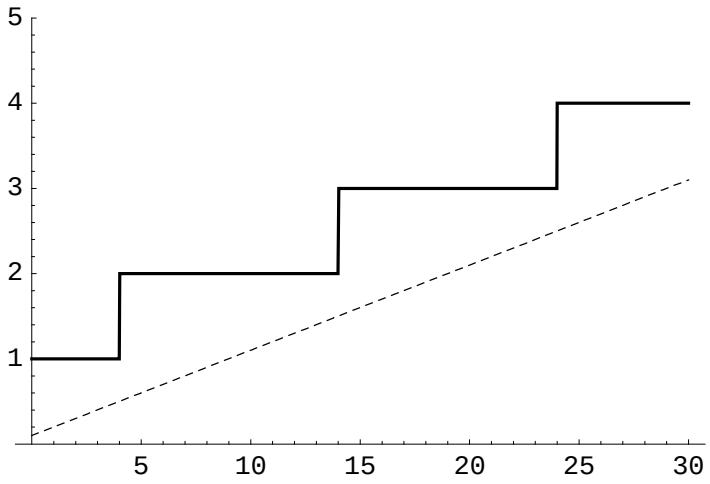


Notice that the limit order book price schedule lies entirely above the expectation revision function. This means that if my order is the last one to execute, I realize a profit.

Suppose that the initial valuation was slightly above zero:

$$\{\alpha = 0.1,\ \gamma = 0,\ \lambda = 5,\ X = 0.1\} \tag{18.b.13}$$

Then:

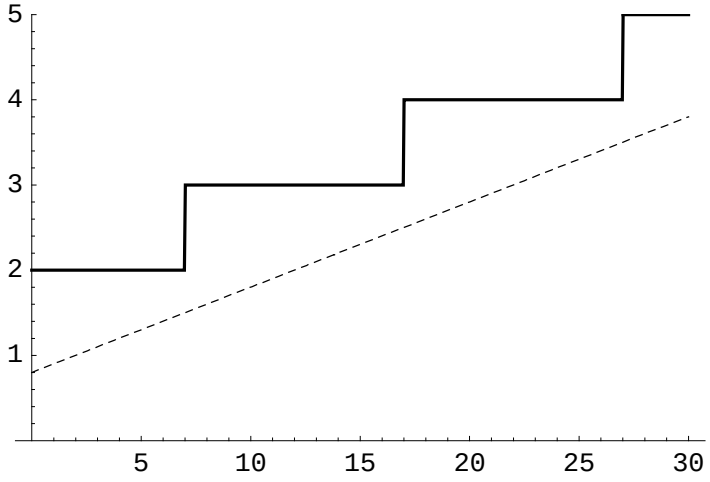


The apparent difference is that the quantities get reduced.

Does the book always start at the next higher tick above X ? Consider:

$$\{\alpha = 0.1,\ \gamma = 0,\ \lambda = 5,\ X = 0.8\} \tag{18.b.14}$$

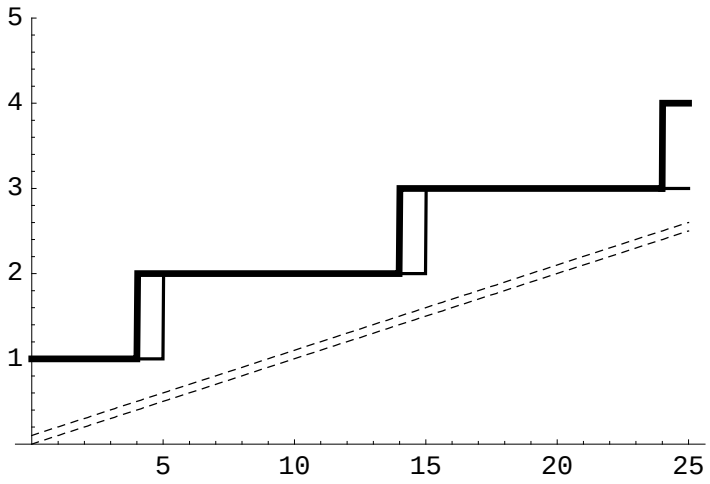
In this case:



Consider next the evolution of the book. Suppose that (starting from $\{\alpha = 0.1, \gamma = 0, \lambda = 5, \mathrm{X} = 0\}$) we get a small order of $m = 1$. The new value of $X = \alpha\, m = 0.1$, so the full set of parameters is now:

$$\{X = 0.1,\ \alpha = 0.1,\ \gamma = 0,\ \lambda = 5\} \tag{18.b.15}$$

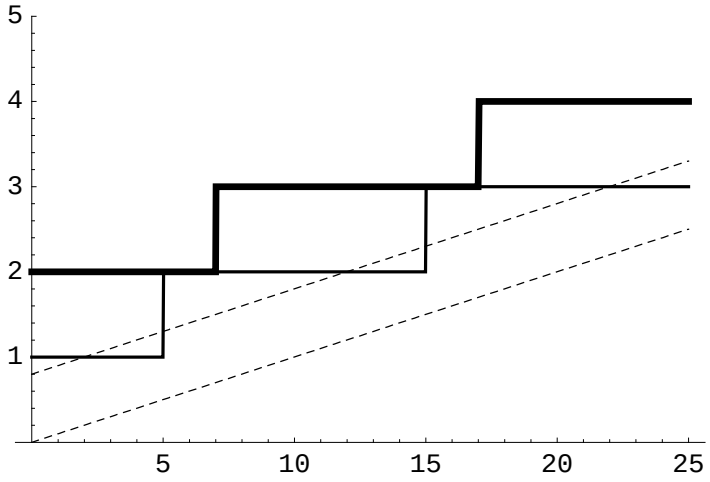
And:



Now suppose that the first order was $m = 8$, leading to revised parameters:

$$\{X = 0.8, \alpha = 0.1, \gamma = 0, \lambda = 5\} \tag{18.b.16}$$

and:



Originally, there were 10 shares available at a price of 2. The initial order of 8 shares left 7 shares at this price. In the new equilibrium, no additional shares were added.

Suppose we have an execution that leaves quantity q at the best price p. The book is said to "backfill" when, subsequent to the execution, additional limit orders arrive at p or better.

Conjecture 1: "backfilling" does not occur in this model.

Conjecture 2: "backfilling" might occur if we introduced event uncertainty.

18.c Empirical evidence

Sandas examines a sample of data from the Swedish Stock Exchange (and electronic limit order book market). His results may be summarized in the following graph.

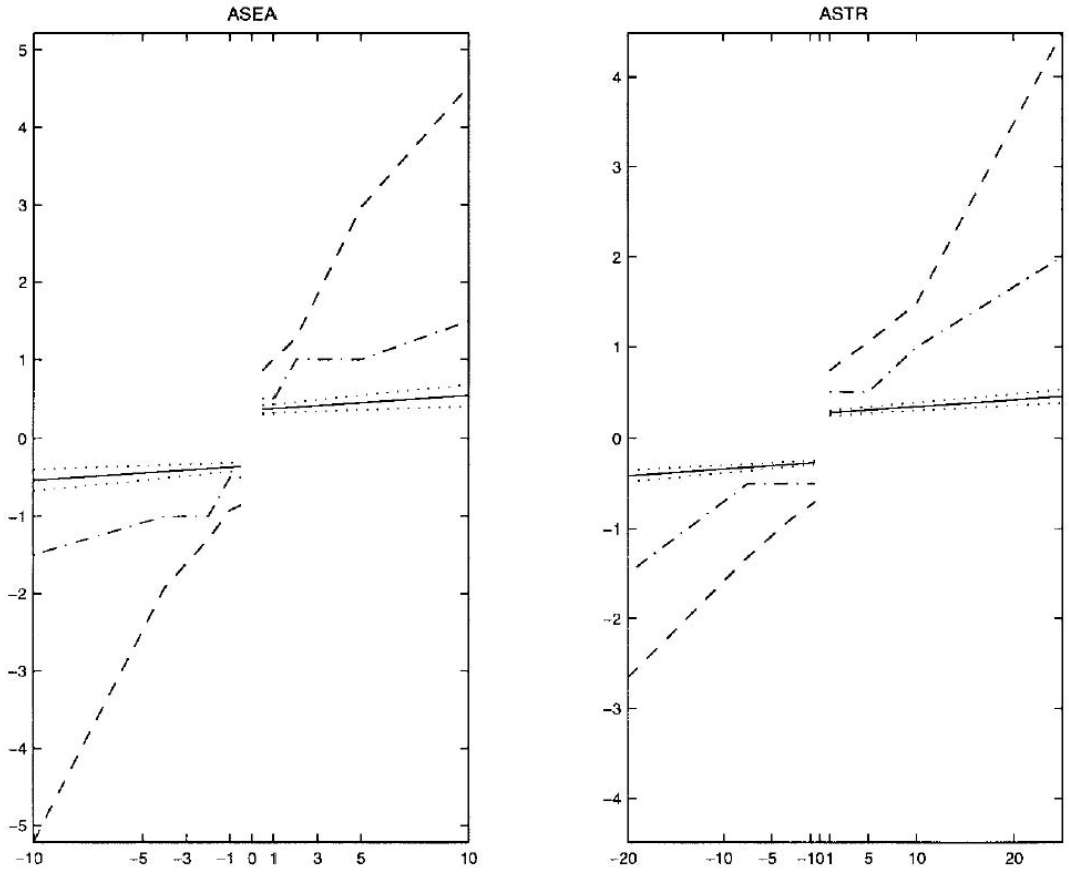


Figure 1
Comparison of Benchmark and Observed Price Schedules
The time-series mean (dashed line) and median (dash-dot line) of the markup or discount (measured in Swedish Crowns on the vertical axis) that a trader submitting hypothetical market orders of different sizes (measured in roundlots [100 shares] on the horizontal axis) would have paid or received relative to the mid-quote are plotted for two stocks, ASEA and ASTR. The markups and discounts are computed for the marginal unit of each market order size as follows. For each limit order book observed before a transaction, the markup (discount) that market orders of various sizes would pay or receive, respectively, are computed. The market order sizes are chosen to match the order flow by percentiles reported in Table 3. The solid line in each subplot corresponds to the implied price schedule based on the price impact regression results reported in Table 4, that is, $bI + cIq$ is plotted for $I = \pm 1$ and various values of q, with a 5% confidence interval (dotted line).

Sandas' Figure 1 illustrates the average shapes of the book for two stocks and an estimated price impact function. In principle if the the book represented the supply and demand curves of a single liquidity supplier who could condition on the size of the incoming order, these curves and the price impact function would coincide. The book is much steeper than the price impact functions.

Can this difference in slopes arise reflect the inability of liquidity suppliers to condition on the size?

Sandas uses the break-even conditions in the model as moment conditions in GMM estimation. Two sorts of moment conditions are used. First, the break-even conditions on the book at a point in time are sufficient the identify the parameters. Second, the model also implies conditions on the dynamics of book revisions that are sufficient to identify α.

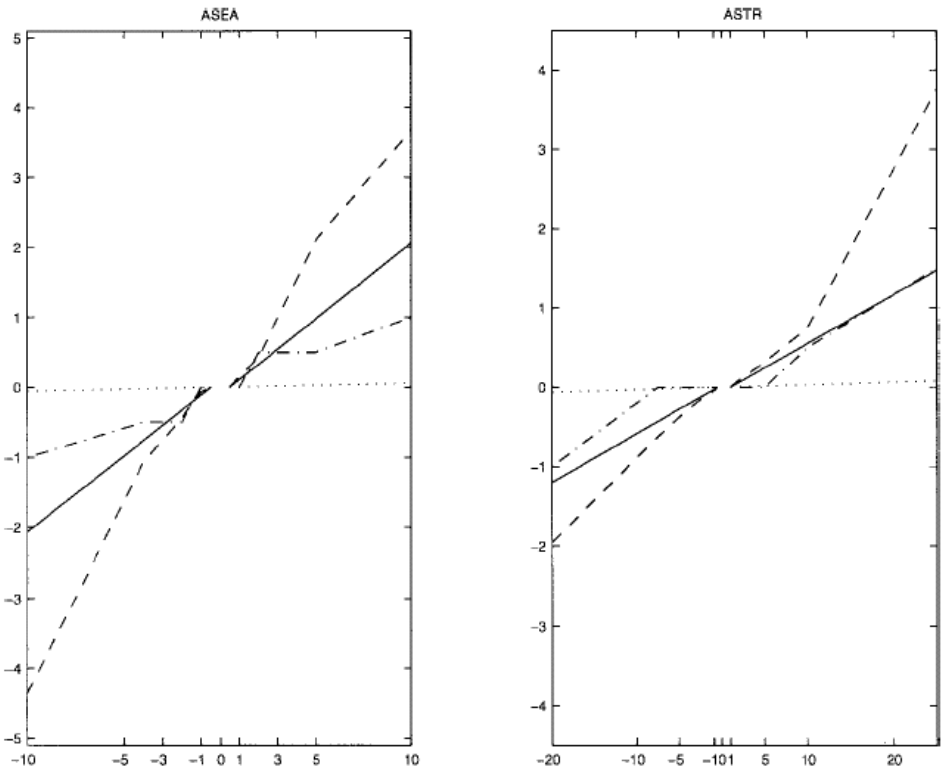


Figure 2
Comparison of Implied (Model) and Observed Price Schedules
The time-series mean (dashed line) and median (dash-dot line) of the markup or discount (measured in Swedish Crowns on the vertical axis) that a trader submitting hypothetical market orders of different sizes [measured in roundlots (100 shares) on the horizontal axis] would have paid or received relative to the mid-quote are plotted for two stocks, ASEA and ASTR. The markups and discounts are computed for the marginal unit of each market order size as follows. For each limit order book observed before a transaction, the markup (discount) that market orders of various sizes would pay or receive, respectively, are computed. The market order sizes are chosen to match the order flow by percentiles reported in Table 3. The solid line in each plot corresponds to the price schedules implied by the α estimates based on the break-even conditions (see Table 5). The dotted line represents the price schedules implied by the α estimates based on the updating conditions (see Table 5). The price schedules are normalized so that a one round-lot trade has a zero markup/discount.

Sandas' Figure 2 depicts book shapes and price impact functions implied by

1. α estimated using the break-even conditions (solid line)
2. α estimated using the dynamic conditions (dotted line)

These graphs suggest that the book price schedule is too steep relative to the dynamic price impact.

Sandas investigates several possible explanations for this finding. One possibility is that the exponential distribution assumed for the incoming orders is a poor approximation.

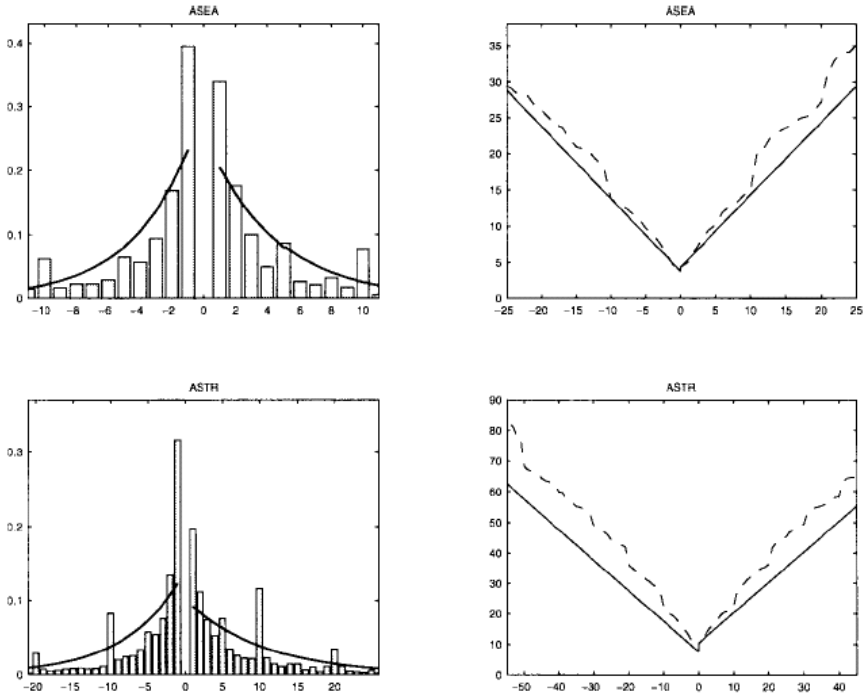


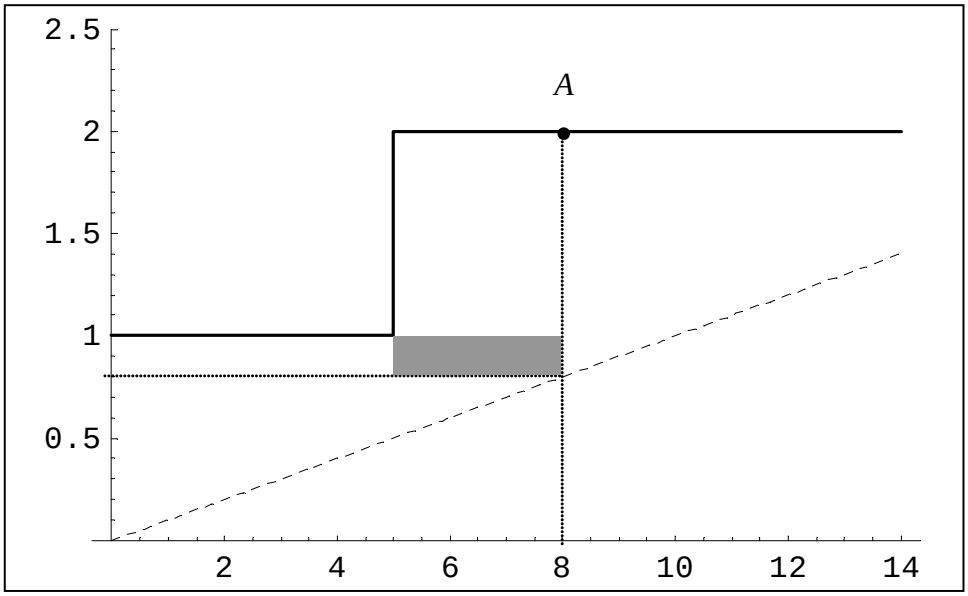
Figure 3
Distribution of Market Orders and "Upper (Lower)" Tail Expectations
The distribution of market order quantities implied by the parameter estimates (λ and ϕ) reported in Table 6 (solid line) and the empirical distribution of market order quantities (bars) are plotted on the left for ASEA and ASTR, respectively. The plots on the right show the "upper tail" expectations and the absolute value of the "lower tail" expectations computed based on (i) the estimated market order distribution (solid line) and (ii) the empirical distribution (dashed). The market sell order quantities appear with negative signs. The units for the market order quantities (x-axis) are 100 shares.

The left-hand graphs of Sandas' Figure 3 compare the actual and implied order distributions. Relative the exponential, there are too many small orders and too many large orders. Thus, when a midsize order is executed, the likelihood that it was blown through by a much larger order is higher than the estimated exponential distribution would suggest.

18.d Introduction of a dealer/specialist

Most markets (including US equity markets) are hybrids of electronic limit order books and dealers. Dealers in this context are defined by two features: (1) They can condition their trades on the total size of the incoming order; (2) They must yield to customer orders at the same price. Seppi (1997) suggests analysis on the follow lines.

To illustrate the situation, we'll take as a point of departure the ask side of the book from the Sandas model.



Suppose that (with the price impact parameter $\alpha = 0.1$), the incoming order is a purchase of 8 shares. If the book were the only liquidity supplier, the order would cross the sell schedule at point *A*: five shares would be sold at $p_1 = 1$ and three shares at $p_2 = 2$.

Now consider the dealer. Conditional on the order, the revised security value is 0.8. The dealer would make a profit on any shares he could sell at $p_1 = 1$. Customers have priority for five shares, but the dealer is free to offer more. In this case, he'll sell three shares. The shaded area is his profit. He can't sell at $p_2 = 2$ because other customers have priority at that price.

If the incoming order were for ten shares, he'd let the book take the full amount (five shares at p_1 and five shares at p_2).

Returning to the eight-share example, in actual markets, this is sometimes called "quantity improvement". Dealers in this situation typically claim, "The book was only showing five shares at p_1, so this was all the customer could expect. But I was able to give the customer a better deal, giving him all eight shares at that price. I saved the customer an amount equal to three shares $\times$ one price tick."

The dealer's claim is, as stated, correct. From a welfare viewpoint, however, there is an effect on the incentives for customers to post limit orders. The dealer's profit would otherwise accrue to the limit order sellers. Although the latter would make a profit on this particular order, a zero-expected profit condition holds over all of their executions. The profit on this order is offset by their losses on larger orders. Obviously, they will supply less liquidity (smaller quantities, higher prices).

Chapter 19. Bidding and offering with uncertain execution

To explore what happens with execution uncertainty, we will first explore how execution uncertainty affects the strategy of an agent who is already at her optimum portfolio position. This development is actually due to Stoll (1978). Stoll was interested in how risk aversion would affect dealers' quote setting behavior. The intuition is as follows. Consider a dealer who is at her portfolio optimum, and has posted bid and ask prices. If the bid is hit for one unit, she will be moved off of her optimum. The bid must be set to compensate her for this loss of utility. We won't develop Stoll's argument in its full generality. Instead, we'll examine a representative situation, and various extensions.

19.a Expected utility

Assume that the dealer has a negative exponential (constant absolute risk aversion, CARA) utility function $U(W) = -e^{-\alpha W}$, and that $W \sim N(\mu_W, \sigma_W^2)$. Then expected utility is

$$\mathrm{EU}(\mu_W,\sigma_W^2) = -e^{\frac{1}{2}\alpha^2\sigma_W^2-\alpha\mu_W} \tag{19.a.1}$$

This can be shown as follows. The characteristic function of a random variable W is defined as the expectation Ee^{itW} where $i = \sqrt{-1}$. If $W \sim N(\mu_W, \sigma_W^2)$, then the characteristic function is:

$$\mathrm{E}e^{itW} = e^{it\mu_W-\frac{1}{2}t^2\sigma_W^2} \tag{19.a.2}$$

Letting $t = i\alpha$, $-e^{itW} = -e^{-\alpha W} = U(W)$. Letting $t = i\alpha$ in the above gives the desired result.

19.b Setting the bid for a single risky security.

There is one risky asset that pays X. $X \sim N(\mu_X, \sigma_X^2)$. The dealer can borrow or lend at zero interest. The initial endowment of stock and cash is zero.

It will be useful in this analysis to employ the concept of a benchmark notional price, P. One interpretation of this is as the price that would obtain in a frictionless market. To establish a benchmark position, we assume that the dealer sets up her portfolio in this frictionless market. All purchases are made from cash borrowed at zero interest; all sales are short-sales. If n shares are purchased, then terminal wealth is given by $W = n(X - P)$, with expectation $\mathrm{EW} = n(\mu_X - P)$, and variance $\sigma_W^2 = n^2\sigma_X^2$. Stoll alternatively suggests that the price be viewed as the dealer's subjective valuation.

Expected utility is:

$$\mathrm{EU}_{\mathrm{Base}} = -e^{\frac{1}{2} n^2 \alpha^2 \sigma_X^2 - n\alpha(\mu_X - P)} \tag{19.b.3}$$

To find the optimal n, differentiate w.r.t. n, set to zero and solve:

$$n = \frac{\mu_X - P}{\alpha \sigma_X^2} \tag{19.b.4}$$

Expected utility at the optimum is:

$$\mathrm{EU}_{\mathrm{Base,Opt}} = -e^{-\frac{(P-\mu_X)^2}{2\sigma_X^2}} \tag{19.b.5}$$

This is the notional, benchmark utility. After it is established, the hypothetical frictionless market closes and the dealer opens operations in the "real" market.

That is, starting with n shares, she puts out a bid B. If she's hit, she buys at B, and her terminal wealth is $W = -B + X + n(X - P)$. So EW $= -B + \mu_X + n(\mu_X - P)$ and $\sigma_W^2 = (n+1)^2 \sigma_X^2$. Her expected utility is then:

$$\mathrm{EU}_{\mathrm{Buy}} = -e^{\frac{1}{2}(n+1)^2 \alpha^2 \sigma_X^2 - \alpha(-B + \mu_X + n(\mu_X - P))} \tag{19.b.6}$$

The expected utility of the new position (having just bought), assuming that n was originally optimal is:

$$\mathrm{EU}_{\mathrm{Buy,Opt}} = -e^{B\alpha + \frac{1}{2}\left(\sigma_X^2 \alpha^2 - 2P\alpha - \frac{(P-\mu_X)^2}{\sigma_X^2}\right)} \tag{19.b.7}$$

The key assertion is that the dealer sets the bid so that if hit, she achieves the same expected utility as at her base optimum. Setting $\mathrm{EU}_{\mathrm{Buy,Opt}} == \mathrm{EU}_{\mathrm{Base,Opt}}$ and solving for B gives:

$$B = P - \frac{\alpha \sigma_X^2}{2} \tag{19.b.8}$$

This is intuitively sensible. The bid is marked down from the notional frictionless price. The markdown increases with the risk of the asset and with the risk aversion of the dealer.

■ Extension: Bid as a function of quantity

Starting from n shares valued at notional price P, suppose that the dealer buys q more at price B (which will depend on q). The terminal wealth is $W = q(X - B) + n(X - P)$. So EW $= q(\mu_X - B) + n(\mu_X - P)$ and $\sigma_W^2 = (n+q)^2 \sigma_X^2$. Her expected utility is:

$$\mathrm{EU}_{\mathrm{Buy}} = -e^{\frac{1}{2}(n+q)^2 \alpha^2 \sigma_X^2 - \alpha(q(\mu_X - B) + n(\mu_X - P))} \tag{19.b.9}$$

The expected utility of the new position (having just bought), assuming that n was originally optimal is:

$$\mathrm{EU}_{\mathrm{Buy,Opt}} = -e^{\frac{1}{2} q^2 \sigma_X^2 \alpha^2 + (B-P) q \alpha - \frac{(P-\mu_X)^2}{2\sigma_X^2}} \tag{19.b.10}$$

The key assertion is that the dealer sets the bid so that if hit, she achieves the same expected utility as at her base optimum. Setting $\mathrm{EU}_{\mathrm{Buy,Opt}} == \mathrm{EU}_{\mathrm{Base,Opt}}$ and solving for B gives:

$$B = P - \frac{1}{2} q \alpha \sigma_X^2 \tag{19.b.11}$$

The bid is linear in quantity.

19.c Setting the bid with correlated risky assets

One of the strongest intuitions in modern finance is that the risk of security depends on how it interacts with the risk of other assets held. In a market making context, this might suggest that the bid price should depend on the covariance of the asset's payoff with the rest of the dealer's portfolio. We consider this as follows.

Suppose that we have two assets with payoffs $X \sim N(\mu, \Omega)$ bivariate normal. The expanded notation treats n, B, μ, and P as vectors:

$$\left\{n = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix}, B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, P = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}, \Omega = \begin{pmatrix} \omega_1^2 & \rho\,\omega_1\,\omega_2 \\ \rho\,\omega_1\,\omega_2 & \omega_2^2 \end{pmatrix}\right\} \tag{19.c.12}$$

Using vector notation, the expected utility of buying and holding n shares is:

$$\mathrm{EU}_{\mathrm{Base}} = -e^{\frac{1}{2}\alpha^2\, n^T.\Omega.n - \alpha\, n^T.(\mu - P)} \tag{19.c.13}$$

where T denotes transposition and "." denotes matrix multiplication. The optimal n are obtained by solving $\alpha\Omega n - (\mu - P) = 0 \Rightarrow n = \alpha^{-1}\,\Omega^{-1}(\mu - P)$.

Expanding this out gives the optimal n as:

$$n = \begin{pmatrix} \frac{-\rho\,P_2\,\omega_1 + \rho\,\mu_2\,\omega_1 + (P_1 - \mu_1)\,\omega_2}{\alpha\,(\rho^2-1)\,\omega_1^2\,\omega_2} \\ \frac{P_2\,\omega_1 - \mu_2\,\omega_1 + \rho\,(\mu_1 - P_1)\,\omega_2}{\alpha\,(\rho^2-1)\,\omega_1\,\omega_2^2} \end{pmatrix} \tag{19.c.14}$$

In the special case of $\mu_1 = \mu_2 = \mu$ and $P_1 = P_2 = P$,

$$n = \begin{pmatrix} -\frac{(P-\mu)\,(\rho\,\omega_1 - \omega_2)}{\alpha\,(\rho^2-1)\,\omega_1^2\,\omega_2} \\ \frac{(P-\mu)\,(\omega_1 - \rho\,\omega_2)}{\alpha\,(\rho^2-1)\,\omega_1\,\omega_2^2} \end{pmatrix} \tag{19.c.15}$$

Returning to the more general case, at the optimum, expected utility is:

$$\mathrm{EU}_{\mathrm{Base,Opt}} = -e^{\frac{P_2^2\,\omega_1^2+\mu_2^2\,\omega_1^2+2\,\rho\,(P_1-\mu_1)\,\mu_2\,\omega_2\,\omega_1-2\,P_2\,(\mu_2\,\omega_1+\rho\,(P_1-\mu_1)\,\omega_2)\,\omega_1+(P_1-\mu_1)^2\,\omega_2^2}{2\,(\rho^2-1)\,\omega_1^2\,\omega_2^2}} \tag{19.c.16}$$

■ The bid for asset 1:

Denote by S the quantity of stock that will be purchased if bid B is hit:

$$\mathrm{EU}_{\mathrm{Buy}} = -e^{\frac{1}{2}\,\alpha^2\,(n+S)^T.\Omega.(n+S)-\alpha\left(n^T.(\mu-P)+S^T.(\mu-B)\right)} \tag{19.c.17}$$

Initially:

$$S = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{19.c.18}$$

That is, if the bid for asset 1 is hit, we'll acquire one more unit of asset 1. Expected utility in this case is:

$$-e^{\frac{P_2^2\,\omega_1^2+\mu_2^2\,\omega_1^2+2\,\rho\,(P_1-\mu_1)\,\mu_2\,\omega_2\,\omega_1-2\,P_2\,(\mu_2\,\omega_1+\rho\,(P_1-\mu_1)\,\omega_2)\,\omega_1+\left(P_1^2-2\left(\alpha\left(\rho^2-1\right)\omega_1^2+\mu_1\right)P_1+\mu_1^2+\alpha\left(\rho^2-1\right)\omega_1^2\right.}{2\,(\rho^2-1)\,\omega_1^2\,\omega_2^2}} \tag{19.c.19}$$

where B_1 is the bid price of asset 1. Setting this equal to the optimal base utility and solving gives:

$$B_1 = P_1 - \frac{\alpha\,\omega_1^2}{2} \tag{19.c.20}$$

This is a surprising result: the bid does not depend on the payoff correlation, ρ.

What's going on is this: The correlation does indeed enter into the expected utility in both the base case and when the bid is hit. But it affects both in a similar fashion. From a comparative statics perspective, an increase in ρ causes the dealer to hold less of each security to begin with.

■ Bids for portfolios

In many markets, a dealer may be asked to provide a quote for a bundle of securities. An options market maker, for example, might put out a bid on a straddle (the combination of a put and a call). We've seen that correlation doesn't affect the bid on an individual security. Might it affect the bid on a package?

Suppose that the package is one unit of asset 1 and one unit of asset 2, purchased at bids B_1 and B_2, respectively. Expected utility is:

$$\mathrm{EU}_{\mathrm{Buy}} = -e^{\frac{P_2^2\,\omega_1^2+\mu_2^2\,\omega_1^2+2\,\rho\,(P_1-\mu_1)\,\mu_2\,\omega_2\,\omega_1-2\,P_2\left(\mu_2\,\omega_1+\omega_2\left(\rho\,P_1-\rho\,\mu_1+\alpha\left(\rho^2-1\right)\omega_1\,\omega_2\right)\right)\omega_1+\omega_2^2\left(P_1^2-2\left(\alpha\left(\rho^2-1\right)\omega_1^2+\right.}{2\,(\rho^2-1)\,\omega_1^2\,\omega_2^2}} \tag{19.c.21}$$

The bid for the package will be $B_{\text{Total}} = B_1 + B_2$:

$$B_{\text{Total}} = P_1 + P_2 - \frac{1}{2}\,\alpha\,(\omega_1^2 + 2\,\rho\,\omega_2\,\omega_1 + \omega_2^2) \tag{19.c.22}$$

Here, correlation affects things as we'd expect. As ρ increases, the package becomes riskier and the mark-down increases. The package essentially becomes a single asset with variance $\omega_1^2 + 2\,\rho\,\omega_2\,\omega_1 + \omega_2^2$.

Chapter 20. Limit order submission strategies

The last chapter considered execution uncertainty for an agent (a dealer) who was at her portfolio optimum and required compensation for acting as counterparty to customer orders that would drag her away from the optimum. Starting the agent at her optimum greatly simplified the analysis in that she would be indifferent between any execution outcome that had same expected utility.

In this chapter, we consider an agent (a "customer") who is not at her optimum. She is facing the choice between doing nothing, trading with a market order, and (maybe) trading with a limit order. This is a classic problem. The model here draws on Cohen, Maier, Schwartz and Whitcomb (1981), henceforth CMSW. As I did with the Stoll development in the last chapter, I'll explore a special case of the model.

As in the Stoll model, consider an agent who has a negative exponential utility function is $U(W) = -e^{-\alpha W}$. If terminal wealth is $W \sim N(\mu_W, \sigma_W^2)$,

$$\mathrm{EU}(\mu_W,\sigma_W^2) = -e^{\frac{1}{2}\alpha^2\sigma_W^2 - \alpha\mu_W} \tag{20.a.1}$$

There is one risky asset that pays $X \sim N(\mu_X, \sigma_X^2)$. There is unlimited borrowing and lending at zero interest. The notional price of the risky-asset is P (in the same sense as the Stoll model). If n shares are purchased, then terminal wealth is given by $W = n(X - P)$, with expectation $\mathrm{EW} = n(\mu_X - P)$, and variance $\sigma_W^2 = n^2\sigma_X^2$. Expected utility is:

$$\mathrm{EU} = -e^{\frac{1}{2}n^2\alpha^2\sigma_X^2 - n\alpha(\mu_X - P)} \tag{20.a.2}$$

Maximizing over n gives

$$n_{\mathrm{Optimum}} = \frac{\mu_X - P}{\alpha\sigma_X^2} \tag{20.a.3}$$

Without loss of generality, we normalize P to unity. Expected utility at the optimum is:

$$\mathrm{EU}_{\mathrm{Optimum}} = -e^{-\frac{(\mu_X - 1)^2}{2\sigma_X^2}} \tag{20.a.4}$$

Suppose that the trader enters the market one share short of her optimum. If she does nothing (the "null" strategy), her expected utility is:

$$\mathrm{EU}_{\mathrm{Null}} = -e^{\frac{(-\mu_X + \alpha\sigma_X^2 + 1)(\mu_X + \alpha\sigma_X^2 - 1)}{2\sigma_X^2}} \tag{20.a.5}$$

Suppose that the market ask price is A. If she buys a share at this price, her wealth becomes:

$$W = -A + \frac{(X-1)(\mu_X - 1)}{\alpha \sigma_X^2} + 1 \tag{20.a.6}$$

The expected terminal wealth is:

$$\mu_W = \frac{(\mu_X - 1)^2}{\alpha \sigma_X^2} - A + 1 \tag{20.a.7}$$

The variance of terminal wealth is:

$$\sigma_W^2 = \frac{(\mu_X - 1)^2}{\alpha^2 \sigma_X^2} \tag{20.a.8}$$

Buying a share at the ask price A follows as the outcome of a market order. The expected utility of this strategy is

$$\text{EU}_{\text{Market}} = -e^{(A-1)\alpha - \frac{(\mu_X - 1)^2}{2\sigma_X^2}} \tag{20.a.9}$$

By setting $\text{EU}_{\text{Market}} = \text{EU}_{\text{Null}}$ and solving for A, we find:

$$A_{\text{Critical}} = \frac{\alpha \sigma_X^2}{2} + 1 \tag{20.a.10}$$

If the market ask price $A < A_{\text{Critical}}$, the trader will use a market order in preference to doing nothing. If $A > A_{\text{Critical}}$, she'll do nothing. As risk (σ_X^2) and/or risk aversion (α), increase, A_{Critical} also increases. That is, the agent is more willing to pay up for the share.

To illustrate with a numerical example, take values:

$$\{\mu_X = 1.1,\ \sigma_X^2 = 1,\ \alpha = 1\} \tag{20.a.11}$$

These imply:

$$A_{\text{Critical}} = 1.5 \tag{20.a.12}$$

Recall that the notional asset price used to determine the optimum was unity. Thus, the agent is willing to pay up by half to acquire the share.

Now we turn to limit order strategies. The uncertainty of a limit order is that we don't know whether it will execute (be hit). So if we put in a limit buy order at price L.

$$\text{EU}_{\text{Limit}} = P_{\text{Hit}}\, \text{EU}_{\text{LimitHit}} + (1 - P_{\text{Hit}})\, \text{EU}_{\text{Null}} \tag{20.a.13}$$

where

$$\mathrm{EU}_{\mathrm{LimitHit}} = -e^{(L-1)\,\alpha - \frac{(\mu_X - 1)^2}{2\,\sigma_X^2}} \tag{20.a.14}$$

Now as long as the order is priced at $L < A_{\mathrm{Critical}}$, $\mathrm{EU}_{\mathrm{LimitHit}} > \mathrm{EU}_{\mathrm{Null}}$, so $\mathrm{EU}_{\mathrm{Limit}} \geq \mathrm{EU}_{\mathrm{Null}}$, with strict equality if $P_{\mathrm{Hit}} > 0$. Thus, we might as well put in some limit order, even if it is priced far away from the market and the probability of execution is near zero. But what is the *optimal* limit order?

We need to max $\mathrm{EU}_{\mathrm{Limit}}$ over L where both $\mathrm{EU}_{\mathrm{LimitHit}}$ and P_{Hit} depend on L. The dependence of $\mathrm{EU}_{\mathrm{LimitHit}}$ on L is given above. But how should we model P_{Hit}?

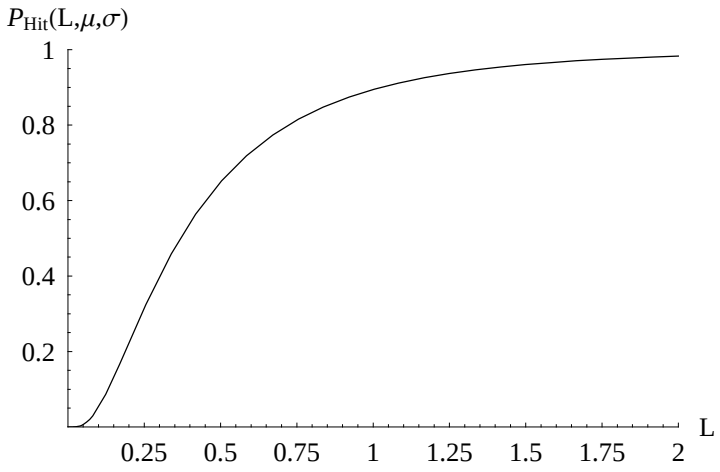
The modeling of limit order execution probabilities and durations is an active area for current research. One way of looking at things is to appeal to the lognormal diffusion process that underlies standard continuous time option pricing. A limit buy order priced at L will execute when S hits L from above. With a lognormal diffusion, the probability of hitting this barrier in a given time interval can be expressed as the lognormal distribution function. Lo, MacKinlay and Zhang (2002) demonstrate that this approach does not yield accurate predictions of time-to-execution durations. In the present context, though, we're simply trying to illustrate some qualitative features of the problem. Using the lognormal diffusion approach:

$$P_{\mathrm{Hit}}(L, \mu, \sigma) = \frac{1}{2}\left(\mathrm{erf}\left(\frac{\log(L) - \mu}{\sqrt{2}\,\sigma}\right) + 1\right) \tag{20.a.15}$$

where L is the limit price, and μ and σ are the parameters of the lognormal distribution. In what follows, it will be convenient to take numerical values:

$$\{\mu = -1,\ \sigma = 0.8\} \tag{20.a.16}$$

With these values, P_{Hit} looks like this:



With this hit probability function,

$$\mathrm{EU}_{\mathrm{Limit}} = \frac{1}{2}\, e^{-\frac{(\mu_X - 1)^2}{2\,\sigma_X^2}} \left(\left(e^{(L-1)\,\alpha} - e^{\frac{1}{2}\,\alpha^2\,\sigma_X^2} \right) \mathrm{erfc}\left(\frac{\log(L) - \mu}{\sqrt{2}\,\sigma} \right) - 2\, e^{(L-1)\,\alpha} \right) \tag{20.a.17}$$

It is not feasible to solve analytically for the L that maximizes this expression. For the model parameters given above, however, we may solve numerically:

$$L_{\mathrm{Optimal}} = 0.658584 \text{ and } \mathrm{EU}_{\mathrm{LimitOptimal}} = -0.924981 \tag{20.a.18}$$

Now what would the ask price have to be to make us indifferent between a market order and a limit order?

We solve numerically for the value of A that makes $\mathrm{EU}_{\mathrm{Market}} = \mathrm{EU}_{\mathrm{LimitOptimal}}$:

$$A = 0.927018 \tag{20.a.19}$$

At the trial values,

$$\mathrm{EU}_{\mathrm{Null}} = -1.6405 \tag{20.a.20}$$

There is a strict ordering $\mathrm{EU}_{\mathrm{Limit}} > \mathrm{EU}_{\mathrm{Null}}$. Consider a hypothetical market opening in which limit order *sellers* start high, setting a high initial ask price A, and then dropping it. As long as $A > 0.927018$, the limit will use a limit buy order priced at $L_{\mathrm{Optimal}} = 0.658584$. When A drops below this, she'll switch to a market order.

The switch point is well above the limit order price. A market order gives certainty of execution, and at some point this certainty induces a switch. CMSW refer to this as "gravitational pull", in the sense that as the ask drops into range, it "pulls" opposing limit bids toward it as market orders.

Note: Behavioral evidence suggests that individuals exhibit a "certainty preference" (cf. the Allais paradox). In practice, most individuals use limit order strategies relatively infrequently.

We can investigate (numerically) the sensitivity of limit order pricing to changes in model parameters. We've been using:

$$\{\mu_X = 1.1,\ \sigma_X^2 = 1,\ \alpha = 1\} \tag{20.a.21}$$

Suppose that we keep the asset characteristics the same, but consider a slightly higher degree of risk aversion:

$$\{\mu_X = 1.1,\ \sigma_X^2 = 1,\ \alpha = 1.1\} \tag{20.a.22}$$

With these parameters,

$$\mathrm{EU}_{\mathrm{Null}} = -1.82212 \tag{20.a.23}$$

$$L_{\mathrm{Optimal}} = 0.684141 \text{ and } \mathrm{EU}_{\mathrm{LimitOptimal}} = -0.948083 \tag{20.a.24}$$

The ask price at which the agent is indifferent between this limit order and a market order is:

$$A = 0.956079 \tag{20.a.25}$$

Thus, the more risk-averse trader submits a higher-priced (more aggressive) limit order, which will have a higher probability of execution. Furthermore, the switch point (to a market order) is also higher.

In general, the "gravitational pull" effect in limit orders refers to any mechanism that will cause a jump-type switch in order strategy (from limit buy to market buy) before the ask declines to the limit price. It is important because we can envision a market with large number of buyers and sellers, with a wide range of risk tolerances, motivated by varying degrees of suboptimality in their initial allocations.

The bid and ask in a market are determined by the *marginal* buyer and seller (among the those not previously matched). With a great diversity of trader characteristics and trading needs, we might expect the spread in a market with continuous prices to be infinitesimal. For example, we can envision a relatively risk-tolerant buyer with a minimal motive for trade to place a limit buy order at the market ask price "less epsilon".

What conditions might generate a finite spread, i.e., one uniformly bounded away from zero? A discrete price grid would obviously suffice. CMSW also point out that a discontinuity in the hit probability would also suffice. Suppose that for a buy order priced at L, we have $\lim_{L \to A^-} \Pr_{\text{Hit}} = \pi < 1$. In this case, limit order strategies for risk averse agents would not give rise to orders priced arbitrarily close to the ask.

■ Broader models of choice and strategy

The utility-based approach illustrates many of the features and trade-offs in the simple market vs. limit order choice. In practice, though, many realistic trading strategies (particularly ones used by institutions) are multiperiod (or continuous) and involve order revision. Angel (1994) and Harris (1998) model these strategies. Bertsimas and Lo (1998) consider order-splitting strategies.

The first step is defining the objective function. In the simple utility-based approach, expected utility is a unified objective, in the sense that it covers all sources of randomness, both the payoffs to the risky asset and the uncertainty of limit order execution. In general, though, trading strategies are generally formulated separately from the investment/portfolio strategy (cf. Perold (1988)). The latter problem is extremely complicated in its own right, and usually (but not always) involves decisions on longer time horizons. The representative case is an institutional equity fund manager with a long-term investment horizon. In this situation, the portfolio problem is solved first (possibly taking into account rough measures of trading costs). The trading problem is then one of achieving the desired portfolio.

At this stage, Harris identifies several "stylized" trading problems, specifically: an uninformed trader buying or selling a given quantity subject to a time constraint, an uninformed trader passively supplying liquidity, and an informed trader maximizing profits. The decision points are fixed in discrete time. Limit order execution probabilities are based on the beta distribution. The models are solved numerically.

The optimal strategies exhibit many characteristics of realistic behavior. For example, an uninformed trader facing a deadline will start by placing limit orders away from the market. As the deadline approaches, the trader will revise the orders, pricing them more aggressively. Finally, if nothing has executed by the deadline, the trade is accomplished with a market order.

Chapter 21. Dynamic equilibrium models

Some aspects of limit order book markets, notably social welfare considerations, can only be addressed in an equilibrium context. Even apart from welfare considerations, though, the need for equilibrium analysis arises directly. When we model an individual agent's choice of order type, we encounter obvious features of the problem that arise from the collective behavior of others facing similar problems. The quote that the agent faces coming into the market and the existing state of the book depend on past actions of other agents; the execution probabilities of a limit order depend on the agents arriving in the future, and so forth.

The dynamic models are stylized ones, but they nevertheless arrive at useful empirical predictions.

This chapter focuses on Parlour (1998) and Foucault (1999). Related work includes Parlour and Seppi (2003), Hollifield, Miller, Sandas, and Slive (2003), Goettler, Parlour, and Rajan (2003), Foucault, Kadan, and Kandel (2001).

■ Foucault (1999)

Structure

The model is set in discrete time, $t = 1, \ ..., \ T$ where T is the terminal payoff date. The underlying value of the security is $v_t = v_0 + \sum_{i=1}^{t} \epsilon_t$, where the ϵ_t are i.i.d. Bernoulli, taking on values of $\pm\sigma$ with equal probability.

T is not known by market participants. At the start of every period t, there is $1 - \rho$ probability that $t = T$: there is no more trading and the payoff is realized. With this modeling device, the problem (and solution) is identical in every period, greatly simplifying analysis. Were T known in advance, this would not be case.

At each time t (assuming that the game is not over), a trader arrives. The trader is characterized by the reservation price, R_t, he assigns to the security, a portion of which is idiosyncratic:

$$R_t = v_t + y_t \tag{21.a.1}$$

where $y_t \in \{+L, -L\}$ with probabilities k and $1 - k$, independent of the value process. y_t does not reflect private information. It arises from portfolio or liquidity considerations that are not explicitly modeled. y_t drives the direction of the agent's desired trade (buy or sell).

If a trade is executed at price P, a buyer will have utility $U(y_t) = V_T + y_t - P$. A seller will have utility $U(y_t) = P - V_T - y_t$.

The state of the book at the time of arrival is described by $s_t = \{A_t, \ B_t\}$. The no order (empty book) condition is indicated by setting $A_t = \infty$ and $B_t = -\infty$. The trader knows $s_t, \ v_t$ and y_t. The strategies open to him are as follows. If the book is not empty, he can hit either the bid or the ask with a market order.

Alternatively, he can place *both* a limit buy and a limit sell order. If the book is empty, this latter strategy is the only one available (apart from the suboptimal strategy of doing nothing).

A trader gets one shot at the market. He doesn't have the opportunity to return and revise his order. Furthermore, limit orders are valid only for one period. This implies that the book is either empty or full.

The probability of execution for a limit order depends on the limit price in the usual way. Here, though, the execution probability is not an ad hoc functional form, but instead arises endogenously. Specifically, the time-t trader knows the distribution of v_{t+1} and the distribution of the characteristics for the time $t+1$ trader. This enables him to derive the execution probability for any given limit price.

Despite the simplicity of the model, the strategic considerations regarding order choice are quite rich (and complicated!).

First consider execution risk of a limit order when there is no possibility of change in the underlying asset value ($\sigma = 0$). Part of the execution risk arises from the random characteristics of the next trader. If $y_t = +L$ ("a natural buyer") and $y_{t+1} = +L$ a trade is (in equilibrium) unlikely. So a limit order can fail to execute because the two parties wish to trade in the same direction. A limit order submitted at time t might also fail to execute, however, because $t+1 = T$ (the world ends).

Once we allow $\sigma \neq 0$, a buy limit order submitted at time t (for example) also faces the risk that $\epsilon_{t+1} = -\sigma$. This corresponds to the real-world situation of a limit order that can't be canceled promptly in response to a public news announcement. This is a form of the winner's curse. It increases the chance that my limit order will execute, but decreases my gain from the trade (and perhaps drives it negative). The limit order is said to be "picked off" subsequent to a "public" information event.

A move in the other direction, $\epsilon_{t+1} = +\sigma$, decreases my chance of execution (but increases my gains from an execution). This situation occurs in actual trading situations when the market "moves away" from a limit order, often leaving the trader (a) wishing he'd originally used a market order, and (b) "chasing the market" with more aggressively priced limit or market orders. (This strategy is not available in the Foucault model.)

Results

- As in the analyses of individual order choice, when the opposite side quote is distant, a trader is more likely to use a limit order.
- The fundamental risk of a security, σ, is a key variable. If σ increases ("higher fundamental risk") then a given limit order faces a higher pick-off risk. This causes limit order traders to fade their prices (make them less aggressive) and the spread widens. Market orders become more expensive, leading traders to favor limit orders. The order mix shifts in favor of limit orders, but fewer of them execute.

  This is a comparative statics result, and thus best viewed as a cross-sectional prediction (across firms) rather than dynamic one (what happens when the volatility changes over time).

■ Parlour (1998)

Model structure

Timing. Consumption can occur on day 1 or day 2. Trade can only take place at times $t = 0, \ ..., \ T$, all on day 1. Clearing occurs at the end of the day: all trades are settled in units of day-1 consumption. The security has a non-random payoff, V per share, realized at time 2.

Agents have preferences $U(C_1, C_2, \beta) = C_1 + \beta C_2$ where β is a continuous random variable distributed on the interval $(\underline{\beta}, \overline{\beta})$ where $0 < \underline{\beta} < 1 < \overline{\beta}$. That is, there is uncertainty and heterogeneity across agents in their relative valuations of C_1 and C_2. Agents also differ in their endowments. With probability π_S, the arriving trader has one unit, and is a (potential) seller. With probability π_B, the arriving trader is a potential buyer of one unit. With probability $1 - \pi_B - \pi_S$, the trader is neither a buyer nor a seller.

Variation in β is the sole source of randomness in the model.

The price grid is discrete. In fact, there are only two prices, a bid and an ask, B and A, and they are separated by one tick. There are dealers who are willing to buy an infinite amount at B and sell an infinite amount at A.

At each time t, a trader arrives. Using a market order, she may buy (at A) or sell (at B) a single share. Alternatively, she can enter a limit buy order (priced at B) or a limit sell order (priced at A). She may do nothing at all.

In the book, dealers must yield to customers. (All customer orders take priority over dealer orders.) The book is maintained in time priority. A customer's limit buy order will execute only if market sell orders arrive in the future that are sufficient to fill the customer's order and all limit buy orders that were placed earlier.

This is a model of queuing and quantities, therefore, rather than a model of prices.

Results

- Effect of same-side depth ("crowding out effect"). When the quantity is large on the ask side, an arriving seller is more likely to use a market order. This occurs because a new limit sell order would go to the end of a long queue (and have a low probability of execution).
- Effect of opposite-side depth. When the quantity is large on the bid side, an arriving seller is more likely to use a limit order. (Subsequent buyers are more likely to use market orders, so the execution probability of a limit sell order is higher.)
- The model also makes predictions about likelihoods of sequences of events.

Part IV: Microstructure and asset pricing

Chapter 22. Trading and asset pricing with fixed transaction costs

This chapter explores the links between microstructure and asset pricing.

22.a Theory

Modifications of standard equilibrium asset pricing models predict that nonstochastic transaction costs should generally have minor effects on expected returns. Agents typically adapt to these costs by trading infrequently and sparingly. In consequence, trading volumes and aggregate trading costs are small.

Standing against this prediction are two sorts of empirical evidence. One is the simple observation that trading volumes are much larger than these models would predict. Presumably aggregate trading costs are large as well. The second source of empirical evidence comes from empirical return specifications in which various measures of trading cost are introduced as explanatory variables. The evidence here is mixed, but is at least partially supportive of a positive cross-sectional relation between transaction costs and expected returns.

Recently, theoretical and empirical studies have started to examine stochastic transaction costs. This opens another avenue for transaction costs to affect expected returns. If transaction cost variation is not diversifiable, i.e., if the variation is at least in part systematic, then the common component becomes an aggregate risk factor. An individual asset's exposure (sensitivity) to this risk-factor should therefore be priced.

For starters, consider the Roll model, with log quote midpoint m_t, $a_t = \log(\text{ask price})_t = m_t + c$ and $b_t = \log(\text{bid price})_t = m_t - c$. An investor who buys at the ask, holds for one period and sells at the bid has a net return $m_{t+1} - m_t - 2c$, i.e.,

$$r_t^{\text{Net}} = r_t^{\text{Gross}} - 2c. \tag{22.a.1}$$

If c is now interpreted as impounding explicit trading costs (like commissions), it can easily be on the order of 1% or so for a small stock. Thus, $2c$ is of moderate importance relative to gross returns.

But if the agent holds for n years, the average annualized net return is

$$\overline{r_t^{\text{Net}}} = \overline{r_t^{\text{Gross}}} - \frac{2c}{n}. \tag{22.a.2}$$

Long holding periods can clearly reduce the impact of trading costs. If n is the same for all assets and investors, and if investors price assets to equate net expected returns (all else equal), then in a cross section of securities, gross returns are a linear function of spread.

■ Amihud and Mendelson (1986): The model

Amihud and Mendelson model an economy in which investors are heterogeneous in their expected holding periods.The key intuition of their model is that investors with longer horizons tend to hold assets with relatively high spreads.This induces a concave relationship between spread.

- There are $i = 1, \ ..., \ M$ investor types
- There are $j = 0, \ ..., \ N$ securities modeled as perpetuities with cash flows d_j (dollars per period).
- S_j is the relative spread: V_j is the ask price and $V_j(1 - S_j)$ is the bid. $S_0 = 0$: asset zero is something like cash or interest-bearing bank account. The assets are ordered by increasing spread: $S_0 = 0 \le S_1 \le \ldots \le S_{N-1} \le S_N < 1$. The vector of ask prices will be denoted $V = [V_j]$. The vector of bid prices is $B = [V_j(1 - S_j)]$.
- A type-i investor enters the market with wealth W_i (cash) and purchases a portfolio (at ask prices).

  Investor types are distinguished by their expected holding periods. The holding period of the portfolio is T_i. T_i is exponentially distributed with parameter μ_i: $E[T_i] = 1/\mu_i$. Investor types are ordered by increasing expected holding period: $\mu_1^{-1} \le \mu_2^{-1} \le \ldots \le \mu_{M-1}^{-1} \le \mu_M^{-1}$.
- Type i investors arrive randomly in continuous time with Poisson arrival intensity λ_i.

  The combination of a Poisson "birth" process plus an exponential "death" process implies that the population of type-i investors who are "alive" (i.e., holding assets) at any instant is on average $m_i = \lambda_i / \mu_i$.

Denote by x_i the vector of share holdings for a type-i agent. An agent of type i with risk-neutral time-additive utility max's:

$$E\Big[\underbrace{\int_0^{T_i} e^{-\rho_i y} x_i' \, d \, dy}_{\text{Present value of dividends}} + \underbrace{e^{-\rho_i T_i} x_i' B}_{\text{Present value of liquidation proceeds}}\Big] = (\mu_i + \rho_i)^{-1} x_i'(d + \mu_i B) \tag{22.a.3}$$

subject to a the initial wealth constraint $x_i' V \le W_i$ and $x \ge 0$ (no short sales). The simplicity of this expression arises in part from the exponentially distributed holding period. This ensures that a type-i's expected *remaining* holding period (measured from the present to liquidation) is $1/\mu_i$, irrespective of how long the individual has already held the portfolio. There are no life-cycle effects.

The quantity $m_i \, x_i$ is the total amount held (on average) by all type-i investors. If the supply of each asset is normalized to unity, then market clearing requires $\sum_{i=0}^{M} m_i \, x_i = \iota$ where ι is an $(N + 1) \times 1$ unit vector.

This market clearing condition equates supply and demand in the time-averaged sense. At any given time, the actual imbalance is absorbed by dealers. The dealer's compensation for this is presumably impounded in the bid-ask spread.

The model is now characterized by M linear optimizations subject to linear cons.traints and non-negativity

requirements.

$\underset{M\times(N+1)}{X^*}$ are the equilibrium allocations; V^* are the equilibrium ask prices. The spread-adjusted return is:

$$r_{ij} = \frac{d_j}{V_j} - \mu_i S_j. \tag{22.a.4}$$

Note: $\mu_i S_j$ is spread/expected holding period.

Which assets will a type-i investor hold? Her highest spread adjusted return is defined as $r_i^* = \max_j r_{ij}$.

The required gross return on asset j for a type-i investor is $r_i^* + \mu_i S_j$. In equilibrium:

$$\frac{d_j}{V_j^*} = \min_i \{r_i^* + \mu_i S_j\} \tag{22.a.5}$$

$$V_j^* = \max_i \left\{ \frac{d_j}{r_i^* + \mu_i S_j} \right\} \tag{22.a.6}$$

Assets with higher spreads are allocated to portfolios of investors with longer expected holding periods. The gross return is a concave function of spread.

Table 1

An example of the equilibrium relation between asset bid–ask spreads, returns and values (see section 2). There are 10 assets (j), each generating \$1 per period, with relative bid–ask spreads S_j (= dollar spread divided by asset value) ranging from 0 to 0.045 (column 2), and 4 investor types (i) with expected holding periods, μ_i^{-1}, of 1/12, 1/2, 1 and 5 periods.[a] The return on the zero-spread asset is ρ; all returns are measured in excess of ρ. A type-i investor chooses the assets j which maximize his spread-adjusted return, r_{ij}, given by the difference between the gross market return on asset j and its expected liquidation cost per unit time. The equilibrium solution gives the excess spread-adjusted returns, $r_{ij} - \rho$, in columns 3–6, where the boxes highlight the assets with the highest excess spread-adjusted return for each investor-type. The equilibrium portfolio for each investor-type is composed of the boxed assets. Column 7 shows the assets' equilibrium excess gross returns observed in the market, which include the expected liquidation cost to their holders. Column 8 shows the resulting asset values, obtained by discounting the perpetuity by the respective equilibrium market return, as a fraction of the value of the zero-spread asset.

		Investor type, i				Market return in excess of ρ, the return on the zero-spread asset	Value of asset j relative to that of the zero-spread asset, V_j/V_0
		1	2	3	4		
		Length of holding period, μ_i^{-1}					
Asset, j	Relative bid–ask spread, S_j	1/12	1/2	1	5		
		Excess spread-adjusted return, $r_{ij} - \rho$					
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	0	0	0	0	0	0	1
1	0.005	0	0.05	0.055	0.059	0.06	0.943
2	0.01	0	0.10	0.11	0.118	0.12	0.893
3	0.015	−0.05	0.10	0.115	0.127	0.13	0.885
4	0.02	−0.10	0.10	0.12	0.136	0.14	0.877
5	0.025	−0.155	0.095	0.12	0.140	0.145	0.873
6	0.03	−0.21	0.09	0.12	0.144	0.15	0.870
7	0.035	−0.265	0.085	0.12	0.148	0.155	0.866
8	0.04	−0.324	0.076	0.116	0.148	0.156	0.865
9	0.045	−0.383	0.067	0.112	0.148	0.157	0.864

[a] Investors have the same wealth, and the expected number of investors of each type is 1.

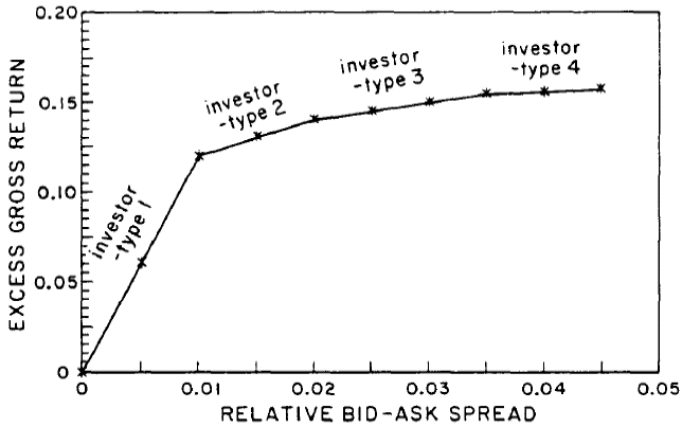


Fig. 1. An illustration of the relation between observed market return in excess of the return on the zero-spread asset (the excess gross return) and the relative bid–ask spread (see the numerical example of section 2 and table 1, column 7). There are 10 assets, each generating $1 per period, with relative bid–ask spreads (= dollar spread divided by asset value) ranging from 0 to 0.045, and 4 investor types with expected holding periods ranging from 1/12 to 5 periods. Investors have equal wealth, and the expected number of investors of each type is 1.
The relation between asset returns and bid–ask spreads is piecewise-linear, increasing and concave, with each linear section corresponding to the portfolio of a different investor type.

■ Constantinides (1986)

"The ... primary result ... is that transaction costs have only a second-order effect on equilibrium asset returns: investors accommodate large transaction costs by drastically reducing the frequency and volume of trade."

The model is cast as a modification of a continuous-time consumption-investment problem due to Merton (1973). There are two securities, $i = 0, 1$.

$$\begin{aligned} dP_0 / P_0 &= r\,dt \\ dP_1 / P_1 &= \mu\,dt + \sigma\,dw \end{aligned} \tag{22.a.7}$$

where prices are in units of consumption ("dollars").

The agent's wealth is W_t and the rate of consumption is c_t. Fraction α_t is invested in the risky-security, so wealth dynamics are:

$$dW_t = [((\mu - r)\,\alpha + r)\,W_t - c_t]\,dt + \sigma\alpha W_t\,dw \tag{22.a.8}$$

The initial endowment is W_0 and initial expected utility is $E_0 \int_0^\infty e^{-\rho t}\, c_t^\gamma / \gamma\, dt$. With no transaction costs, the optimal consumption-wealth ratio, c_t^* / W_t, and the optimal portfolio weight α^*, are both constant.

Transaction costs are introduced as follows. Suppose that the holdings of securities 0 and 1 are x_t and y_t. If v_t dollars of the risky security are bought, the holdings of the risk-free security become $x_t - v_t - |\,v_t\,|\,k$ where k is the proportional transaction cost.

Institutional commissions are currently about $0.05 per share. The bid-ask spread for a typical NYSE stock might be of similar magnitude, making the one-way half-spread $0.025. At a hypothetical share price of

\$50, $k = 0.075/50 = 0.0015$. For a thinly traded security, however, the spread might be as large as \$1.00, implying $k = 0.55/50 = 0.011$.

The optimal investment policy in this case is to keep the relative holdings y_t/x_t in the interval $[\underline{\lambda}, \bar{\lambda}]$. The interior of this interval is a "no-trade" region. Upon reaching $\underline{\lambda}$ or $\bar{\lambda}$, the individual only trades enough to remain in the interval.

Constantinides solves the problem numerically for a range of k and realistic or plausible values for the other parameters:

TABLE 1

OPTIMAL POLICY PARAMETERS AND LIQUIDITY PREMIA FOR DIFFERENT VALUES OF THE TRANSACTION COST RATE

	k									
	0	.005	.01	.02	.03	.04	.05	.10	.15	.20
$\underline{\lambda}$	1.667	1.450	1.377	1.277	1.202	1.140	1.087	.891	.754	.650
$\bar{\lambda}$	1.667	1.767	1.784	1.803	1.818	1.832	1.844	1.905	1.965	2.026
β	.2875	.2869	.2869	.2864	.2859	.2852	.2845	.2803	.2754	.2700
$\delta(k)$/year	0	.0008	.0014	.0025	.0037	.0049	.0061	.0130	.0216	.0347
$\delta(k)/k$	...	.16	.14	.13	.12	.12	.12	.13	.14	.17
$\hat{\delta}(k)$/year	0	.0013	.0024	.0046	.0068	.0091	.0114	.0250	>.0500	>.0500
$\hat{\delta}(k)/k$	...	.26	.24	.23	.23	.23	.23	.25	N.A.	N.A.

NOTE.—The table displays the lower ($\underline{\lambda}$) and upper ($\bar{\lambda}$) bounds of the risky to the riskless asset ratio in the region of no transactions, the optimal consumption rate (β), and the liquidity premia (δ, $\hat{\delta}$) on the risky asset, for different values of the transaction cost rate (k). The assumed parameter values are $\gamma = -1$, $\rho = .10$/year, $r = .10$/year, $\mu = .15$/year, and $\sigma^2 = .04$/year.

Note that as k increases, so does the width of the no-trade region.

Constantinides then introduces a hypothetical risky-security that is perfectly correlated with and has the same return variance as security 1, but can be traded with no cost. The liquidity premium, $\delta(k)$, is defined as the equalizer, an expected return component that, when added to μ, makes the investor indifferent between the actual costly-to-trade security and the costless-to-trade security, assuming that the agent starts out at his optimum portfolio. Table 1 shows that in terms of annual return, $\delta(k)/\text{year}$, the liquidity premium for small k is modest. At large k (0.10 and above) the premium starts to become visible relative to the gross return. They are, however, smaller than one might expect.

What are the implications for trading volume? Suppose that I have \$1 invested in the risk-free security. Interpreting the liquidity premium as the annual trading cost $\delta(k) = \text{Turnover} \times k$, a proportional trading cost of 0.05 implies an annual turnover of $0.0061/0.05 = 0.12$. The NYSE reports that annual turnover on its stocks, however, has recently been running about 100%.

Thus, although the Constantinides analysis suggests that liquidity premia are about one order of magnitude smaller than proportional trading costs, the analysis also implies a turnover that is about one order of magnitude smaller than observed.

■ Heaton and Lucas (1996)

The Heaton-Lucas model incorporates labor income, incomplete markets, and differential transaction costs on stocks and bonds. The agent may also face different borrowing and lending rates.

For stocks, the transaction cost function is:

$$\kappa(s_{t+1},\, s_t) = k_t[(s_{t+1} - s_t)\, p_t^s]^2 \tag{22.a.9}$$

where s_t is the number of shares held at the end of time t, p_t^s is the price per share and k_t is the quadratic cost factor. (Certain notation present in the Heaton-Lucas paper is suppressed here.) The quadratic dependence captures deterioration in the terms of trade associated with larger quantities.

As written, κ is an absolute cost. The proportional transaction cost is $\kappa / p_t^s \mid \Delta s_{t+1} \mid = k_t \mid \Delta s_{t+1} \mid p_t^s$.

In asset pricing models, bonds are generally not to be thought of as a specific security (like 30-year T-bonds), but instead more broadly as borrowing and lending opportunities. Accordingly, HL model bond trading costs in a variety of ways. Their emphasis is on an asymmetric quadratic specification:

$$\omega(b_{t+1}) = \Omega_t \min(0,\, b_{t+1}\, p_t^b)^2 \tag{22.a.10}$$

where Ω_t is the cost factor. The bonds have a one-period maturity, so the amount purchased is the same as the amount held during the period. A positive purchase/holding $(b_{t+1} > 0)$ is equivalent to lending. No cost is assessed in this direction: $\omega(b_{t+1}) = 0$. A negative purchase/holding $(b_{t+1} < 0)$ is equivalent to borrowing. The cost in this direction is $\omega(b_{t+1}) = \Omega_t(b_{t+1}\, p_t^b)^2 > 0$. Since only the borrower pays, for comparability with the stock case (where both sides pay), the proportional cost is measured as $\Omega_t \mid b_{t+1} \mid p_t^b / 2$.

The model is solved numerically, with parameters calibrated to US data. (In particular, the expected stock return is fixed at 8% per year.)

As in the Constantinides model, the optimal trading strategy is highly sensitive to transaction costs. If there are transaction costs in only one market (stock or bonds), then trading substantially shifts to the other market.

To investigate the effects of costs in both markets, HL present figures that summarize the dependence of outcomes on transaction cost parameters, set symmetrically for stock and bond markets, so that $\Omega = k/2$.

HL's Figure 2 maps Ω into percentage trading costs:

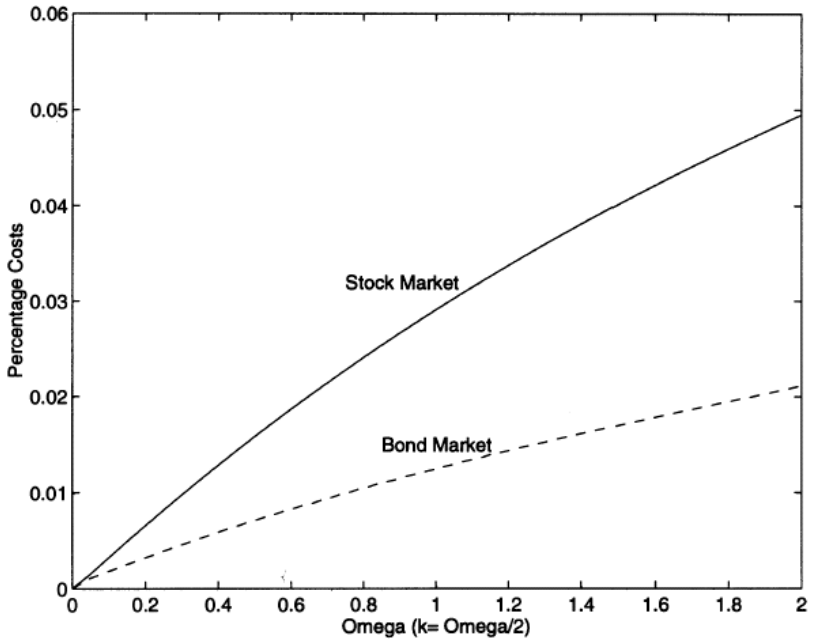


FIG. 2.—Base case, average costs

Thus, an average trading cost of 5% corresponds to $\Omega \approx 2$.

HL's Figure 1 depicts equilibrium expected returns:

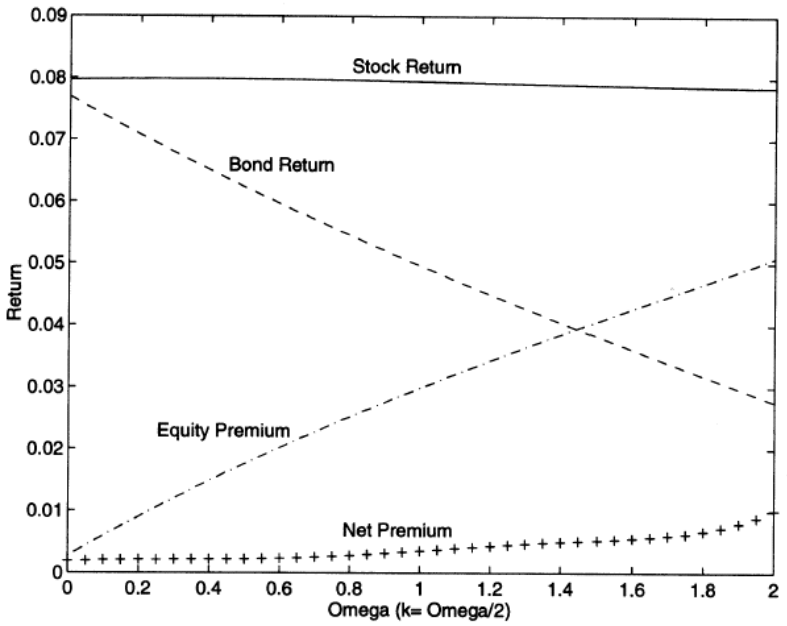


FIG. 1.—Base case, returns

With small transaction costs, the equity premium is near zero. This is the "equity premium puzzle". (The "net premium" measures the indirect effect of transactions costs associated with increased consumption volatility.)

Can costs explain the equity premium?

"If marginal stock market transactions costs of 6% are taken as a reasonable estimate, the model still predicts a substantial equity premium. ... [However] to obtain an equity premium as large as 5 percent requires a marginal stock market transactions cost of 10%, so that without strict borrowing constraints, very large costs are needed to produce a premium close to its observed average level." (p. 467)

HL's Figure 3 describes the trading volume:

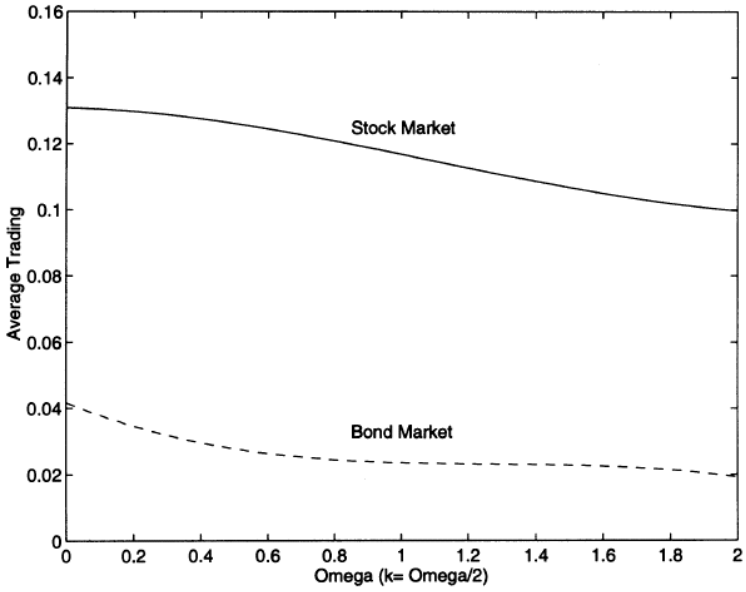


FIG. 3.—Base case, average trading

"Average trading" is the ratio, value of securities traded/consumption.

In 2002, total US personal consumption expenditure (from the national income and product accounts) was about $7.3 Trillion. At the end of 2002, household and nonprofit holdings of corporate equities had a market value of about $10 Trillion (Board of Governors Flow of Funds reports). Assuming that the 100% annual turnover figure for the NYSE is representative, the implied average trading is $10/7.3 = 1.37$. This roughly an order of magnitude higher than the model predicts.

The inability of normative models to explain trading volume is not limited to equity markets. Trading volume in foreign exchange markets also exceeds by an order of magnitude the level that explained by the requirements of trade in goods, services and financial assets.

22.b Empirical Analyses

■ Amihud and Mendelson (1986)

Elements of the analysis:

- CRSP monthly returns, 1960-1979.
- Cost measure is average of beginning of year and end of year relative spreads (Fitch data, last quote for last day of the year).
- Fama-MacBeth approach
- Estimate β over a five-year period (E_n=years 1-5)
- Portfolio formation over a five-year period (F_n = years $6-10$). Form 7 groups ranked by spread in year 10. Within each spread group, form 7 β groups (based on β estimates from years 1-5). This yields 49 portfolios. Estimate portfolio β in years 6-10. Compute average monthly excess returns for each portfolio in year 11.

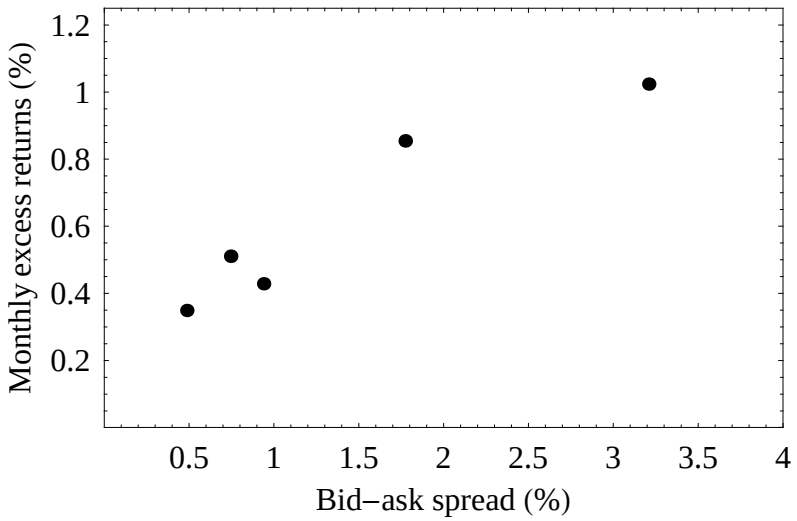
Here are the mean spreads and monthly excess returns for the groups formed by sorting on spread (from their Table 2):

	Spread (%)	Excess return (%)	Beta	Mkt.Cap.
1	0.486	0.349	0.799	2,333.
2	0.745	0.511	0.870	665.
3	0.939	0.429	0.884	418.
4	1.145	0.589	0.913	276.
5	1.396	0.669	0.932	184.
6	1.774	0.855	0.970	109.
7	3.208	1.024	1.115	55.

(22.b.11)

A move from group 7 to group 1 implies a 2.722 % drop in spread and a 0.675 % drop in excess monthly return. As an illustration, AM suggest a hypothetical stock in group 7 that has a total required monthly return of 2% and cash flows of \$1/month in perpetuity. The value of the stock is \$50. If managers could engineer a move to group 1, its new value would be $(0.02 - 0.00675)^{-1} = \$75.5$, a substantial increase.

Graphically:



Some concavity is evident from the graph. In expanded tests (that control for beta and size), AM find that it is statistically significant.

Related papers include:

- Eleswarapu and Reinganum (1993): Sample is NYSE, 1961-1990. Liquidity premium confounded with January seasonal. No evidence of liquidity premium in non-January months.
- Eleswarapu (1997). Sample is Nasdaq, 1973-1990. Spreads from CRSP. Here are Eleswarapu's mean spreads (in spread-ranked subgroups):

	Spread (%)
1	2.005
2	3.443
3	4.849
4	6.643
5	9.368
6	14.201
7	30.632

(22.b.12)

■ Brennan and Subrahmanyam (1996)

Two liquidity measures are used,

Glosten-Harris The λ coefficient from

$$\Delta p_t = \lambda q_t + \psi \Delta D_t + y_t \tag{22.b.13}$$

Hasbrouck-Foster-Viswanathan. The λ coefficient from the restricted VAR:

$$q_t = \alpha_q + \sum_{i=1}^{5} \beta_i \Delta p_{t-i} + \sum_{j=1}^{5} \gamma_j q_{t-j} + \tau_t$$
$$\Delta p_t = \alpha_p + \psi \Delta D_t + \lambda \tau_t + v_t \qquad (22.b.14)$$

These models are estimated for 1984 and 1988 for NYSE-listed firms.

How to scale the estimates?

Intuition from Kyle model where $\Delta p_t = \lambda(x_t + u_t)$. The expected total cost of trading x_t shares is λx_t^2. The marginal cost of the last share is $2\lambda x_t$ ($ per share). In terms of dollar volume of the trade, the marginal cost is $2\lambda x_t / p_t$.

Define $C_q \equiv 2\lambda q / p$ where q is the average trade size and p is the average price per share.

Alternatively: $C_n \equiv \lambda n / p$ where n is the number of shares outstanding.

Portfolio formation procedure sorts first on size, then on GH λ (5×5 portfolios).

Table 3
Intercepts from Fama–French OLS regressions for 30 portfolios of NYSE stocks sorted by size and the Glosten–Harris measure of illiquidity, λ, for the period 1984–1991

λ estimates the derivative transaction price ($/share) with respect to signed trade size (shares, positive for trades initiated by buyers). Portfolios are formed annually from all NYSE firms active at the beginning of the year. Within each calendar year, size is measured as market value of equity at the end of the preceding year. For the 1984–1987 period, λ and all other liquidity variables are estimated using 1984 data. For the 1988–1991 period, they are estimated from 1988 data. The portfolio labeled 0 in the λ group column denotes the portfolio for which data on λ are not available. The table presents intercepts from the following time-series regressions:

$$R_{it} = \alpha_i + \beta_i MKT_t + \delta_i SMB_t + \kappa_i HML_t + u_{it},$$

where R_{it} is the excess return on portfolio i in month t, and MKT_t, SMB_t, and HML_t denote the returns on the Fama and French (1993) factors related to the market, firm size, and the book-to-market ratio in month t. The bottom of the table presents the Gibbons, Ross, and Shanken (1989) test of the hypothesis that the intercepts jointly equal zero. Intercepts are reported in percentage terms (t-statistics are in parentheses).

	λ group					
Size group	0	1	2	3	4	5
1	−1.21	−2.14	−1.51	−1.34	−1.25	−0.45
	(−2.50)	(−3.62)	(−3.17)	(−3.34)	(−3.78)	(−1.76)
2	−0.39	−0.42	−0.19	0.13	0.37	0.80
	(−1.18)	(−1.76)	(−0.99)	(0.72)	(1.99)	(4.63)
3	0.03	0.02	−0.17	0.26	0.41	0.66
	(0.11)	(0.09)	(−1.04)	(1.39)	(2.57)	(4.55)
4	0.86	0.27	0.07	−0.07	0.41	0.63
	(2.24)	(1.54)	(0.51)	(−0.47)	(2.42)	(4.16)
5	0.71	0.29	0.33	0.38	0.38	0.60
	(1.47)	(2.05)	(3.03)	(3.96)	(3.11)	(4.74)

(1) F-value for the Gibbons, Ross, and Shanken test that the intercepts jointly equal zero is 4.77 (p-value = $9.08 * 10^{-8}$). (2) F-value for the Gibbons, Ross, and Shanken test that the intercepts equal zero (excluding portfolios with missing liquidity parameters) is 5.70 (p-value = $5.50 * 10^{-9}$).

22.c Alternative measures of "liquidity"

Empirical asset pricing studies generally require data samples longer than those needed for microstructure analyses. This is because expected asset returns are typically small relative to their variances, and a large sample is therefore needed to estimate the former with precision. (Recall the previous discussion on why the expected return is generally set to zero in microstructure analyses.)

Studies of based on US equity data, for example, usually use CRSP data, which begin in 1962 (daily) or 1926 (monthly). In contrast, the TAQ data begin in 1993. ISSM data go back about a decade earlier. The combined time span, therefore, is at best about half of CRSP's. Furthermore, these microstructure data are by no means homogeneous over this period. Institutions and reporting systems have greatly changed.

These considerations strongly motivate the need for liquidity and trading cost measures that involve only daily return and volume information. Here are some approaches.

■ Liquidity ratio

The Amivest liquidity ratio for a stock is

$$L = \overline{\left(\frac{\mathrm{Vol}_d}{|r_d|} \right)} \tag{22.c.15}$$

where r_d is the return on day d; Vol_d is the volume (dollar or share) on day d. The average is taken over all days in the sample where $r_d \neq 0$.

The originator of the ratio, Amivest, was a money management and broker/dealer concern. It was taken over by the North Fork Bank (New York) in 1998.

This measure has been used in cross-sectional studies of comparative liquidity across markets (see Cooper, Groth and Avera (1985)). Ideally, a liquidity measure should pick up only price changes that are associated with orders. Grossman and Miller (1987) point out that the liquidity ratio does not discriminate. If volatility driven by public information is accompanied by little or no volume, L will be low.

■ Illiquidity ratio

Proposed by Amihud (2002)

$$L = \overline{\left(\frac{|r_d|}{\mathrm{Vol}_d} \right)} \tag{22.c.16}$$

The average is taken over all days in the sample where $\mathrm{Vol}_d \neq 0$. Amihud finds that this measure is significantly and positively related to returns:

Table 2
Cross-section regressions of stock return on illiquidity and other stock characteristics[a]

The table presents the means of the coefficients from the monthly cross-sectional regression of stock return on the respective variables. In each month of year y, $y = 1964, 1965, \ldots, 1997$, stock returns are regressed cross-sectionally on stock characteristics that are calculated from data in year $y-1$. *BETA* is the slope coefficient from an annual time-series regression of daily return on one of 10 size portfolios on the market return (equally weighted), using the Scholes and Williams (1977) method. The stock's *BETA* is the beta of the size portfolio to which it belongs. The illiquidity measure *ILLIQ* is the average over the year of the daily ratio of the stock's absolute return to its dollar trading volume. *ILLIQ* is averaged every year across stocks, and *ILLIQMA* is the respective mean-adjusted variables, calculated as the ratio of the variable to its annual mean across stocks (thus the means of all years are 1). ln *SIZE* is the logarithm of the market capitalization of the stock at the end of the year, *SDRET* is the standard deviation of the stock daily return during the year, and *DIVYLD* is the dividend yield, the sum of the annual cash dividend divided by the end-of-year price. *R100* is the stock return over the last 100 days and *R100YR* is the return during the period between the beginning of the year and 100 days before its end.

The data include 408 months over 34 years, 1964–1997, (the stock characteristics are calculated for the years 1963–1996). Stocks admitted have more than 200 days of data for the calculation of the characteristics in year $y-1$ and their end-of-year price exceeds \$5. Excluded are stocks whose *ILLIQ* is at the extreme 1% upper and lower tails of the respective distribution for the year.

Variable	All months	Excl. January	1964–1980	1981–1997	All months	Excl. January	1964–1980	1981–1997
Constant	−0.364	−0.235	−0.904	0.177	1.922	1.568	2.074	1.770
	(0.76)	(0.50)	(1.39)	(0.25)	(4.06)	(3.32)	(2.63)	(3.35)
BETA	1.183	0.816	1.450	0.917	0.217	0.260	0.297	0.137
	(2.45)	(1.75)	(1.83)	(1.66)	(0.64)	(0.79)	(0.59)	(0.30)
ILLIQMA	0.162	0.126	0.216	0.108	0.112	0.103	0.135	0.088
	(6.55)	(5.30)	(4.87)	(5.05)	(5.39)	(4.91)	(3.69)	(4.56)
R100	1.023	1.514	0.974	1.082	0.888	1.335	0.813	0.962
	(3.83)	(6.17)	(2.47)	(2.96)	(3.70)	(6.19)	(2.33)	(2.92)
R100YR	0.382	0.475	0.485	0.279	0.359	0.439	0.324	0.395
	(2.98)	(3.70)	(2.55)	(1.59)	(3.40)	(4.27)	(2.04)	(2.82)
Ln *SIZE*					−0.134	−0.073	−0.217	−0.051
					(3.50)	(2.00)	(3.51)	(1.14)
SDRET					−0.179	−0.274	−0.136	−0.223
					(1.90)	(2.89)	(0.96)	(1.77)
DIVYLD					−0.048	−0.063	−0.075	−0.021
					(3.36)	(4.28)	(2.81)	(2.11)

[a] *t*-statistics in parentheses.

■ Reversal measures

Pastor and Stambaugh (2003) propose as an inverse measure of liquidity γ in the regression.

$$r^e_{d+1} = \theta + \phi\, r_d + \gamma\, \text{sign}(r^e_d)\, vol_d + \epsilon_{d+1} \tag{22.c.17}$$

where d runs over all days in the sample and r^e_d is the excess return (relative to the market). It's easiest to understand the intuition here by considering a variant based on returns and *signed* order flow, x_d, rather than volume:

$$r_{d+1} = \theta + \phi\, r_d + \gamma\, x_d + \epsilon_{d+1} \tag{22.c.18}$$

In this case, $\gamma > 0$ would suggest that the market did not fully respond to the preceding day's order flow. On the other hand, $\gamma < 0$ would suggest that the market over-reacted, perhaps due to limited capacity of market makers (broadly defined) to absorb the order flow.

The application in Pastor and Stambaugh calls for panel estimates of γ: a separate estimation for each stock in each month.

Pastor and Stambaugh validate this interpretation of γ by simulating the following market:

$$r_d = \underbrace{f_d}_{\text{Market factor}} + \underbrace{u_d}_{\text{Idiosyncratic factor}} + \underbrace{\phi\,(q_{d-1} - q_d)}_{\text{Order flow term}} + \underbrace{\eta_d - \eta_{d-1}}_{\text{Bid–ask bounce}} \tag{22.c.19}$$

where q_d is signed order flow on day d. This is in turn generated by a factor structure: $q_d = q^*_d + q^i_d$ where q^*_d is the market component of signed order flow and q^i_d is the idiosyncratic component.

Superficially, this resembles a microstructure specification. By way of comparison, consider the generalized Roll model:

$$\Delta p_t = (\lambda + c)\, q_t - c q_{t-1} + u_t \tag{22.c.20}$$

where t indexes transactions, c is the fixed cost of liquidity provision (clearing fees, etc.) and λ is the impact parameter. If we time-aggregate this over all trades on a given day, we'd have:

$$\sum_{t=1}^{N} \Delta p_t = \lambda \sum_{t=1}^{N} q_t + c(q_N - q_0) + \sum_{t=1}^{N} u_t \tag{22.c.21}$$

where q_0 is the trade direction indicator at the close of the previous day. To a point, we can establish a correspondence with the PS specification: Σu_t in the (time-aggregated, generalized) Roll model corresponds to $f_d + u_d$ in the PS specification; $c(q_N - q_0)$ in the Roll model corresponds to $\eta_d - \eta_{d-1}$ in the PS specification.

The day's aggregate order flows, though, appear in fundamentally different ways. In the Roll model, $\lambda \Sigma q_t$ represents the cumulative information content of the orders. This arises from quote-setters' beliefs about informed trading. The reaction to this occurs entirely within the day: there is no lagged term. In the PS model, $\phi(q_{d-1} - q_d)$ is a transient return component. There is an over-reaction to today's order flow. It is completely reversed, however, on the following day. There is no permanent impact of orders.

Here is a plot of estimated average γ_i:

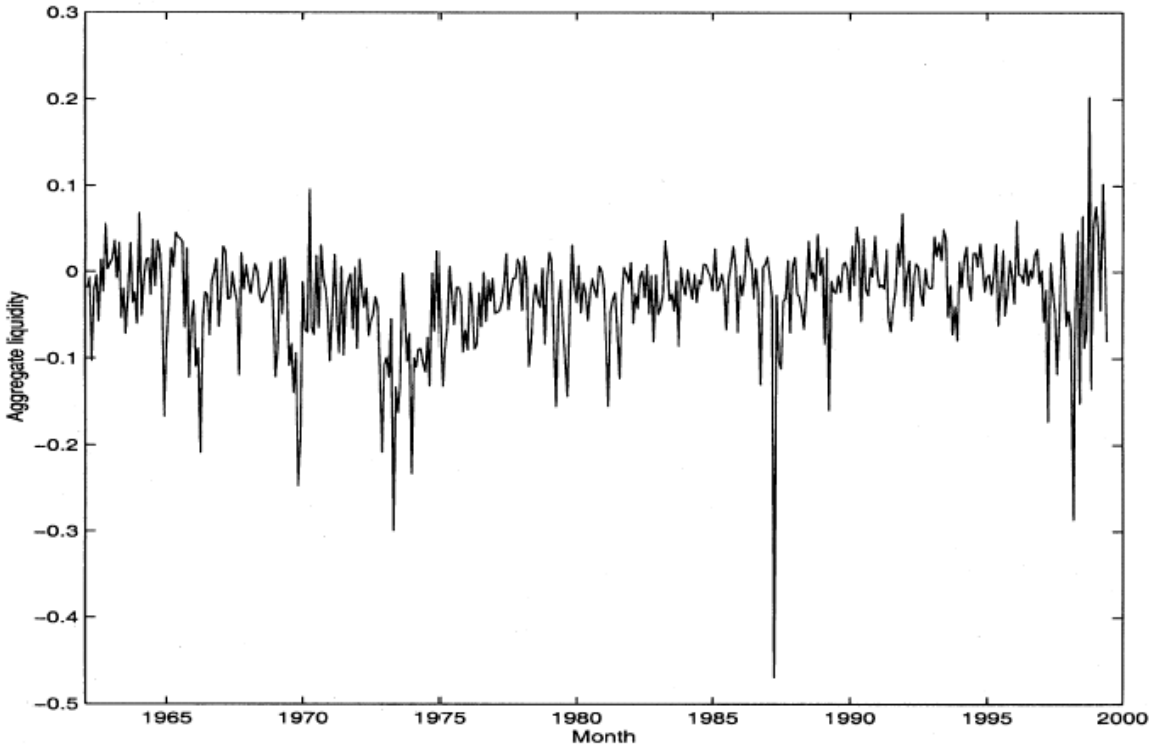


FIG. 1.—Aggregate liquidity. Each month's observation is constructed by averaging individual-stock measures for the month and then multiplying by m_t/m_1, where m_t is the total dollar value at the end of month $t-1$ of the stocks included in the average in month t, and month 1 corresponds to August 1962. An individual stock's measure for a given month is a regression slope coefficient estimated using daily returns and volume data within that month. Tick marks correspond to July of the given year.

Note the large variation relative to the mean.

22.d Stochastic liquidity

Most of the development to this point has assumed fixed transaction costs. Much recent work has studied time-varying liquidity, and common factors therein. See Chordia, Roll, and Subrahmanyam (2000); Hasbrouck and Seppi (2001); Huberman and Halka (2001); Pastor and Stambaugh (2003); Acharya and Pedersen (2002).

Appendix: US equity markets: overview and recent history

US Equity Markets: Overview and Recent History

Joel Hasbrouck

Kenneth G. Langone Professor of Business Administration
and Professor of Finance
Department of Finance
Stern School of Business
New York University
44 West 4th St. Suite 9-190
New York, NJ 10012-1126
212.998.0310

jhasbrou@stern.nyu.edu

Last updated on: 1/8/2004 1:52 PM

The latest version of this document and related materials are on my website at
http://www.stern.nyu.edu/~jhasbrou

Abstract

This paper summarizes the structure of the major US equity markets and the recent (roughly 1995 onwards) regulatory, economic and institutional developments.

1 Introduction

This is an overview of the current status and recent history of trading procedures in US equity markets. It is aimed at researchers and practitioners who work with current or historical market data, and seek to understand the institutions that give rise to these data.

2 Overview

Figure 1 depicts the main components of the US equity markets and their relationships. The connectors indicate only the primary relationships. (In a broader sense, every box is almost certainly connected to every other box.)

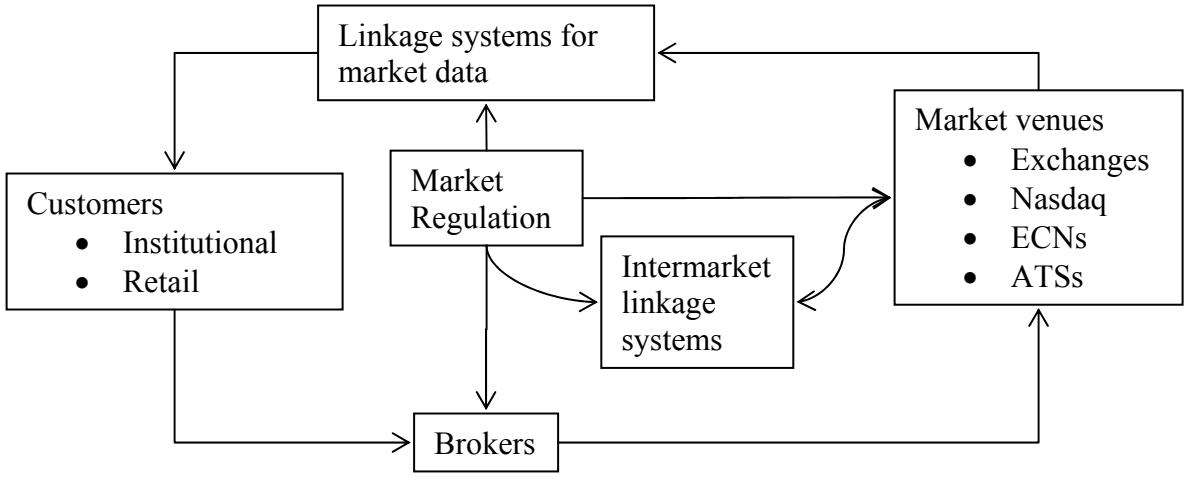


Figure 1. Market components and relationships

- The customers are the individuals and institutions (pension funds, mutual funds and other managed investment vehicles) who need to trade equities. The distinction between retail and institutional customers is mainly one of size: institutional customers simply need to trade larger amounts. The distinction should not be construed as "naïve vs. sophisticated". It should also be noted that the customers in this arena are not simply passive purchasers of the produced services. Increasingly, they are active participants, often effectively competing in the market-making process.

- Brokers have traditionally acted as agents for customer orders. They may provide other services as well (such as advice and research), but from the present perspective, their role is a narrower one.

- Dealers act as counterparties to customer trades.

  The dividing line between brokers and dealers is blurry. They are often the same people, and in US regulatory law are often referred to as "broker-dealers".

In many settings, when a single entity acts as agent for customer orders (broker) and trades against customer orders (dealer), this is called dual trading. Due to the obvious conflict of interest, dual trading is either strictly prohibited or regulated.

- The market venues are simply places (real or virtual) where trades occur. The principal venues are:
  - The exchanges (also known as the "listed" market). Most importantly, the New York Stock Exchange (NYSE), but also including the American Stock Exchange, Cincinnati and the regional stock exchanges (Philadelphia, Boston, Chicago, Pacific).
  - Nasdaq (a.k.a. the "unlisted" market).
  - The ECN's (Electronic Communication Networks). Formally (in law) the term merely describes an entity that displays quotes.
  - The ATS's (Alternative Trading Systems). Places where trades occur.

    "ECN" nowadays generally refers to a market organized as an electronic limit order book. The principal examples are Island and Instinet. Since these systems can also execute trades, they are also ATS's. There are also ATS's that are not ECN's (Liquidnet, Posit, etc.).
- Regulators. U.S. market regulation is somewhat diffuse.
  - In 1933, Congress created the Securities and Exchange Commission (SEC) and delegated to it the regulatory authority. The SEC has in turn delegated authority to industry entities, "Self-Regulatory Organizations" ("SRO's). The NYSE and NASD (National Association of Securities Dealers) are the principal SRO's.

    Legislative actions on security markets generally take the form of broad pronouncements that authorize the SEC's regulatory mandate, but leave the details of the rules up to the SEC.
  - The individual states regulate securities markets. Their involvement generally predates the creation of the SEC. Over time, it has waxed and waned. They have never been more prominent and active than they are at present.
  - In some cases, US law gives private parties a right of action. Improper disclosure by corporations is commonly enforced by private lawsuits ("10b-5 cases"). Private lawsuits in issues regarding trading practices are rarer, but there have been some important ones.
  - The SEC is not just a rule maker and enforcer. It has generally tried to foster open discussion and debate about the structure of markets. Their website (www.sec.gov) includes many useful and authoritative special studies. SEC rules are available online at http://www.law.uc.edu/CCL/sldtoc.html and http://www.sec.gov/rules/final.shtml.

- The SEC and state authorities primarily regulate corporate stock and bond markets. Other securities markets fall to other entities: futures markets (the U.S. Commodities Futures Trading Commission); T-bonds (the Treasury Department).

- Note: Figure 1 should not be construed as implying that customers' trading activities aren't subject to any regulatory oversight whatsoever. Their behavior is obviously governed by broad prohibitions on fraud and manipulation. Their trading activities do not normally, however, entail specific authorization from or reporting to market regulatory authorities.

Figure 1 indicates two sorts of linkage systems. Although they are drawn as connectors bridging customers and the markets, they also link the disparate market venues. As equity markets have grown more fragmented (i.e., trading activity is more dispersed), the linkage systems have become more important.

- Information links

  Information originating from the market venues (trade reports, quotes, etc.) are disseminated under complicated arrangements of consolidation and distribution. Organizations engaged in this activity must register as Securities Information Processors (SIPs). The most important SIPs are the CTA (Consolidated Tape Association) and Nasdaq. The CTA services primarily exchanges (but also includes some ECN's); Nasdaq reports the activity of its own dealers.

- Intermarket access links

  Access refers to the ability to send an order to, and obtain an execution from, a market venue.

The paper now turns to a more detailed discussion of the components, beginning with the most important and complex: the market venues.

3 The basic types of trading protocols

This section describes the general features of the three types of markets that dominate US equity trading. The specifics are discussed in the context of particular market venues.

3.1 Open outcry ("floor") markets

An open outcry market is a physically-centralized venue in which participants strike bilateral deals. This is still the dominant mechanism in most of the U.S. futures exchanges.

In an open outcry market, traders shout out the bids and asks ("24 bid for 1,000 shares" or "500 shares offered at 25"). Executions occur when another traders signals that he is hitting the bid (selling) or lifting the offer (buying).

Traders may act as brokers, i..e., as agent for the orders of off-floor customers. They may also trade for their own account (proprietary trading). When they trade against customers, they are effectively functioning as dealers.

The practice of simultaneously representing customer orders and trading against customer orders gives rise to an obvious conflict of interest. The phenomenon is called "dual trading" and most floor markets either regulate it or prohibit it altogether.

3.2 Dealer markets

Dealers are intermediaries who have established a reputation for standing ready to buy or sell. This reputation encourages customers to take their orders to a dealer. The costs of trading with a dealer are presumably lower than if the customer were to try to locate a potential counterparty on his own behalf.

Dealer markets are not generally physically centralized. Dealers often have electronic links to their customers, and also links to other dealers (which are not visible to customers).

Dealer markets are generally set up to discourage direct, disintermediated customer-to-customer trade. Customers cannot generally provide liquidity directly to the market.

3.3 Electronic limit order book markets

A limit order specifies direction (buy or sell), quantity and the worst acceptable price. If the limit price of a newly arriving buy order exceeds the limit price of a sell order already in the system, the buy order is said to be marketable and a trade occurs (at the limit price of the sell order).

If the incoming buy order is not marketable, it is added to the book of unexecuted buy orders. The book is maintained in price time priority. An order to buy at 100 won't be executed before an order to buy at 101. An order to buy submitted at 10:01 won't be executed before a buy order at the same price submitted at 10:00.

Modern limit order books are computerized, with electronic order entry and interfaces to reporting and clearing systems, and with public display of the prices and quantities on the book. Generally, these systems are anonymous.

In the electronic limit order book, a customer can buy or sell immediately only if there exists an unexecuted limit sell or buy order submitted by another customer. Thus, liquidity is said to be supplied by other customers.

There is nothing in principle to prevent a customer from acting as a dealer, i.e., continually posting buy and sell limit orders to maintain a market presence. But there is also no particular advantage to doing so. The anonymity of the market means that there is no possibility of sustaining a reputation. There are furthermore no barriers to entry. One's bid or ask may be undercut at any time by a new arrival.

4 The New York Stock Exchange

Historically, the NYSE has been the dominant US equity trading venue. An economist might describe it as a multiproduct firm, producing listing, regulatory and trading services. The present analysis focuses mainly on the trading services, i.e., how the NYSE operates as a market.

Basic background on the NYSE is given in Hasbrouck, Sofianos, and Sosebee (1993) and Teweles and Bradley (1998). At this point, however, both sources are somewhat dated. The NYSE Constitution and Rules is authoritative and complete in the details, but it is difficult to distill from this document an overall picture of how the market really functions.

NYSE trading protocols are complex because the Exchange is a hybrid market that embodies elements of an open outcry system, a dealer market and an electronic limit order book. These mechanisms are not simply run in parallel isolation, but are integrated in a fashion that attempts to merge the best features of each. It is perhaps easiest to approach these mechanisms and their interaction by reviewing them in the order in which they historically arose.

4.1 The NYSE as an open outcry ("floor" market)

The NYSE was founded in 1792 and first functioned as an open outcry market. In addition to the basic features of these markets described in the last section, the NYSE's procedures also embodied the following principles.

Price priority.

> For example, someone who is bidding 101 should have priority over someone who's bidding 100.
>
> In this example, it might be thought that self-interest of sellers would ensure price priority. Why would anyone sell at 100 when they could sell at 101? Why is a rule needed?
>
> A trade at 100 when a buyer is bidding 101 is called a "trade-through". (Or, as a verb, "The seller in the transaction traded through the 101 bid.")
>
> Hypothetically, a broker with a customer may care more about getting the order filled quickly than getting the best price, particularly if the customer can't easily monitor the market.
>
> The rule of price priority gives the other side (in this case, the bidder) the right to protest (and break) the trade.

Time priority

> First-come, first-served is a time-honored principle that rewards prompt action. In the present case, the first member to bid or offer at a price gets the first trade at that price.

> Beyond that, there is no time priority. In a crowd, it's possible to keep track of who was first. It's more difficult to keep track of who was second, third, etc.
>
> After the first trade at a price, all members bidding or offering at that price are said to be at parity. This means that they have equal claim to all counterparty interest at that price.

Size precedence

> This is a secondary priority rule. Normally, if *A* and *B* are both bidding $100 and are at parity, they will share arriving sellers equally. If an order to sell 300 shares at the price arrives, *A* and *B* will each buy 150 shares, or they might flip a coin for the whole amount. But if *A* is bidding for 300 shares and *B* is bidding for 100 shares, *A* would get the full amount based on size precedence. Size precedence is rarely invoked nowadays.

The practice of public last-sale reporting and dissemination of bids and offers dates from the floor phase of the NYSE's history (and predates by many years the establishment of any external regulatory authority).

4.2 The dealer (specialist).

The dealer part of the picture emerged in the 1870's. According to legend a member broke his leg and while constrained by immobility decided to specialize in certain selected stocks. The practice was adopted by more ambulatory brokers and the specialist system was born.

There is currently one specialist per stock. This has given rise to the expression in the academic literature (and elsewhere) of "monopolistic specialist". The specialist does enjoy some market power, but the qualifier greatly exaggerates the extent of it. The specialist participation rate (specialist purchase + specialist sales)/(2 x total volume) is about 15% (NYSE Fact Book).

Initially there might have been multiple specialists for a given stock. As recently as 1963, there were 35 listed stocks that had more than one specialist (Seligman (1995), citing the Special Study of the U.S. Securities and Exchange Commission (1963), p. 338). By 1967 there were none. Although the competition might have been thought beneficial to customers, the reality was somewhat different. Seligman quotes the Special Study:

> At present [1963] competition is unsatisfactory for several reasons. Commission firms are often confused as to who is quoting the best market in active stocks. The commission firms do not shop for the best service, but often give each competitor half their brokerage business. In addition, neither competitor accepts full market-making responsibilities, thus adding to the Exchange's regulatory problems.

The specialist's main responsibility is to maintain a "fair and orderly market". There a large number of rules that specify what the specialist must do (affirmative obligations) and what he can't do (negative obligations). But none of these rules supercedes the duty to maintain a fair and orderly market.

Affirmative obligations

The specialist must bid and offer (make a market) when nobody else is willing to do so. The specialist has the sole authority and responsibility for the quotes.

Agent for the limit order book. For many purposes, the book can be thought of as a single member. The specialist keeps the book and represents book interest in the crowd.

Agent for electronically-delivered market orders.

Maintain price continuity.

Negative obligations

A specialist (indeed no member) is allowed to trade ahead of a public customer at a price.

A specialist is discouraged from "trading in a destabilizing fashion" (buying on an uptick or selling on a downtick).

The specialist's role as agent for public orders has become more prominent with the prevalence of electronic delivery. The exchanges order delivery and routing systems (notably "SuperDOT") send virtually all orders that don't require a broker's human attention to the specialist's workstation ("DisplayBook").

4.3 The limit order book

The book is maintained by the specialist. When there were multiple specialists, each specialist could have his own limit order book. Now there is a single electronic book.

In acting as agent for limit order book, the specialist in a sense becomes the book, representing it as if it were a single floor trader. An important implication of this is that although price/time priority is strictly observed within the book, the book as a single entity might be at parity with floor traders that arrived considerably after the limit orders in the book were posted.

4.4 The bid and ask quotes

The specialist sets the bid and ask quotes, but in doing so he might be representing his own interest, orders on the book, bids or offers that a floor broker might want displayed, or a combination of all of these.

If there are orders on the book at the bid or ask, they must be represented (under the quote display rule), but the display is not automatic.

Historically, the specialist could exercise a fair amount of discretion in the display of customer limit orders. Presently, limit orders that better the existing quote must be executed or displayed within 30 seconds. (See the discussion of the SEC's Quote Display Rule in section 5.3).

4.5 Executions

4.5.1 Small orders

Order execution (particularly for small orders) is now often automatic. Any order (up to 1,099 shares) designated to go to the NYSE Direct+ system does not go to the specialist, but is executed automatically at the posted quote (against the specialist). See http://www.nyse.com/pdfs/NYSEDirect.pdf. A similar system is used for odd-lots (market orders smaller than 100 shares).

Market orders delivered to the specialist's post, however, do not execute automatically. Acting as agent for the order, the specialist effectively auctions it off. As an example, consider an incoming buy order. Here are some common (but by no means exhaustive) scenarios.

The simplest outcome is a trade at the posted ask price. The specialist might be selling for his own account at this price, but he can't do this if there are any public limit sell orders priced at the ask (or lower).

The only way for the specialist to sell from his own account when there are public sellers is to offer the buyer a better (lower) price. If the lowest public limit sell order price is 100, for example, the specialist can only sell from his own account at a price of 99.99 or lower.

This outcome results in the buyer receiving a price better than the posted ask, a phenomenon termed "price improvement". While this works to the benefit of the buyer, however, the limit order seller at 100, believing himself to be the most aggressive seller in the market, may feel disadvantaged.

The practice used by the specialist can be employed by any broker who is physically present at the specialist's post. The floor thus enjoys a last-mover advantage.

4.5.2 Large orders

Large executions are sometimes called block trades. The customary size threshold for a block trade is 10,000 shares, but nowadays many orders of that size would simply be allowed to follow the electronic route of small orders, as described above.

The terms of large orders are usually negotiated by customers and their brokers. Often the process involves the broker guaranteeing the customer a price and then working the order (feeding it to the market) slowly over time to minimize price impact. The general process is the same whether the stock is listed on an exchange or Nasdaq. (See below.)

When a broker has located both a buyer and seller for a block, he may, under certain circumstances, "cross" the block, i.e., execute the trade without other buyers and sellers stepping in to take part or all of one or both sides (the "clean-cross rule").

Trade sizes on the NYSE grew for many years, but then dropped. In 1963, the average trade size was 204 shares. The average grew rather steadily, peaking at 2,303 shares in 1988. In 2002, it was 606 shares, a level not seem since the late 1970's.

4.6 Opening and closing procedures

The opening procedure is effectively a single-price call auction. Brokers submit customer orders. The specialist tries to find a single price at which supply equals demand.

Some markets (Euronext) have a closing auction as well. In principle, the specialist can invoke the opening auction system one the close, but this is rarely (if ever) done.

5 Nasdaq

Historically, Nasdaq was primarily a dealer market with geographically dispersed dealers linked by electronic systems that displayed bid and ask quotes and (later) last sale prices. Smith, Selway, and McCormick (1998) discuss the history of Nasdaq up to 1998. The present discussion overlaps with this and extends it.

Nasdaq circa 1990 was distinctly a dealer market, essentially as described in Section 3.2. More so than the NYSE, it was transformed in the 1990's by economic, political and regulatory pressures. The changes greatly enhanced customer protection and reduced the power and profitability of Nasdaq members (brokers and dealers). The changes also weakened the authority and reach of Nasdaq as a central market operator.

5.1 The Manning Rules

Historically, Nasdaq (like most dealer markets) gave little protection to customer limit orders.

For example, suppose that the best market quote was 100 bid, offered at 102. A Nasdaq dealer who received a customer limit order to buy at 101 didn't have to display the order as a new, more aggressive quote. The dealer could furthermore buy for his own account at prices below 101 (thus trading through the customer order). The customer was only entitled to an execution when the market offer price dropped to 101 (essentially making the customer order marketable).

The "Manning" rules were adopted by NASD to prohibit brokers from trading ahead or through their customer limit orders. The name is explained in an SEC opinion on a subsequent administrative proceeding:

> "The term is a reference to E.F. Hutton & Co., 49 S.E.C. 829 (1988), in which the Commission held that a firm violated its fiduciary duties to a customer, William Manning, who had placed a limit order to sell a security, when the firm sold shares of that security at prices above the limit price."
>
> SEC Administrative proceedings file3-9941, available at:
> http://www.sec.gov/litigation/opinions/34-44357.htm#P58_4039

The Manning rules were adopted in two phases:

"Manning I" said that a dealer couldn't trade ahead of limit orders entrusted to them by their *own customers*.

"Manning II" said that a dealer couldn't trade ahead of customer limit orders that had been sent to them by other brokers.

It was and is common on Nasdaq (and other markets as well) for orders to ultimately be represented by brokers other than the one who originally received the order. Manning II simply says that the protection goes with the order, i.e., that it doesn't vanish when the order changes hands prior to representation.

5.2 The collusion charges

Christie and Schultz (1994) found that despite the ⅛ tick size used in US security markets, Nasdaq dealers tended to quote on a ¼-point grid. They suggested that this might be a coordination device to maintain spreads at ¼ . This would be profitable for dealers because most retail trades occurred at the bid or and the ask. Furthermore with weak limit order protection, there was little opportunity for customers to use limit orders to compete with dealer quotes.

Christie, Harris, and Schultz (1994) describe the events immediately surrounding the study. They had sought comments from industry participants, and, after the findings were accepted for publication, Vanderbilt University issued a press release (May 24, 1994).

Also on May 24, a meeting of major Nasdaq dealers was convened at the offices of Bear Sterns in New York. At this meeting a NASD official encouraged dealers to reduce their spreads. The stated reason for this exhortation was a an earlier rule change (January, 1994) to a Nasdaq automatic execution system (SOES), not the CS study. Whatever the motivation, on May 27, spreads began to drop dramatically. (See figure.)

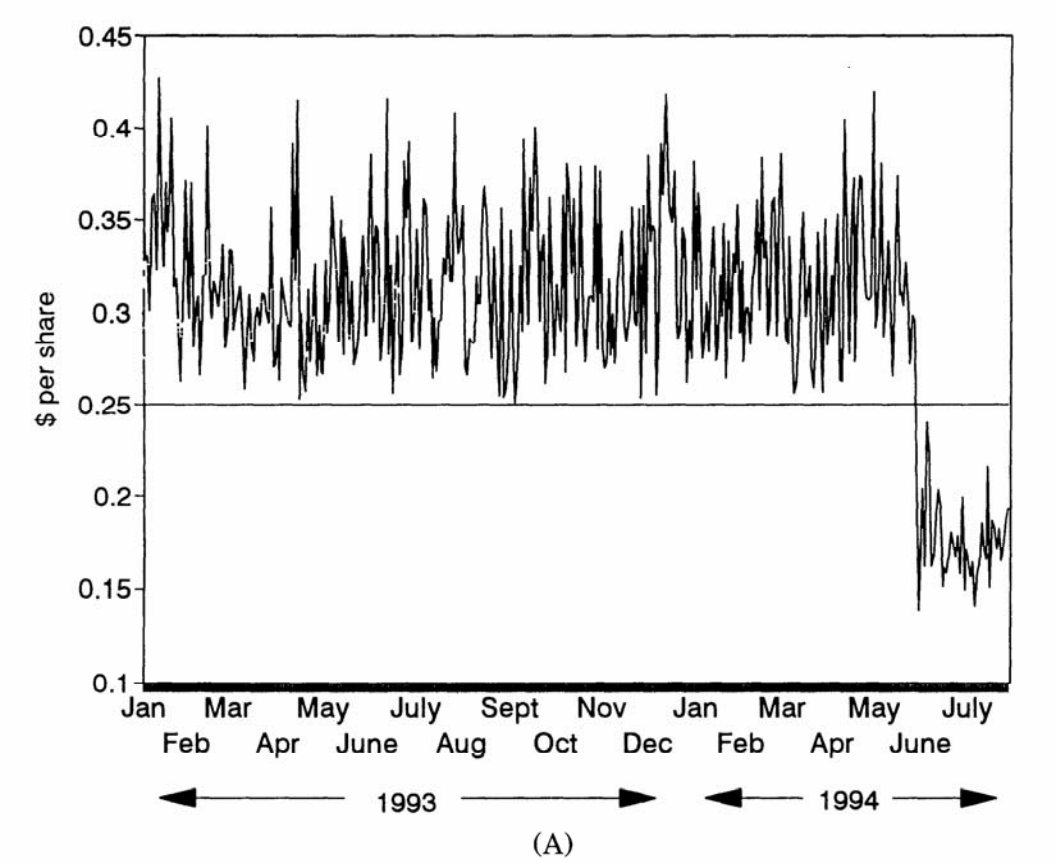


Figure 1. The time series of daily average inside spreads. For each stock, we obtain all the inside spreads during regular trading hours between January 1, 1993 and July 29, 1994, excluding July 15, 1994 when NASDAQ experienced difficulties installing new operating software that resulted in numerous locked and crossed quotes. The daily average inside spread is calculated by multiplying each inside spread by the proportion of the trading day that the spread is in effect. Panel A: Amgen, Inc.; Panel B: Apple Computer, Inc.; Panel C: Cisco Systems, Inc.; Panel D: Intel Corporation; and Panel E: Microsoft Corporation.

Nadaq authorized an external review of the matter (the Rudman Commission); the SEC and the Department of Justice opened investigations; and civil lawsuits were filed against the Nasdaq dealers (on behalf of customers). When the dust settled:

- The Rudman Commission examined NASD's governance and recommended that market operation and market regulation be separated. The latter was spun off as NASD-R (now identified on its website as "the world's leading private-sector provider of financial regulatory services").

  The separation was (and is) structurally complex, and is now enmeshed with Nasdaq's demutualization.

  One aspect of the arrangement, at least in retrospect, stands out as particularly significant. Nasdaq (the market operator) signed a long-term contract with NASD-R to provide regulatory services. As Nasdaq market share and revenues have declined, it is not clear

that this is a viable long-term arrangement. Market regulation is bedeviled with extensive free-rider problems, and it remains unclear how the costs should be borne.

- The SEC and DOJ investigations were concluded and settled. The SEC's "21(a)" report on the investigation is at http://www.sec.gov/litigation/investreport/nasdaq21a.htm.

- The civil law suits were consolidated and eventually settled in May, 1999 for about $1B.

5.3 The SEC's rule on Order Execution Obligations

While the SEC and DOJ investigations can be viewed as a straightforward attempt to hold Nasdaq market-makers responsible for their prior behavior, the practices uncovered by the investigations also served to support constructive reform going forward. The most striking examples of this are rules 11Ac1-4 and 11Ac1-1 on Order Execution Obligations (full text at http://www.sec.gov/rules/final/37619a.txt).

The extent of this rule is not limited to Nasdaq; it applies to all markets. However, it is best understood in the context of Nasdaq regulation. Although it had some ramifications for the NYSE, it had more profound effects on Nasdaq. It has two parts: the Display Rule and the Quote Rule.

From the SEC's summary:

> Specifically, the Commission is adopting new Rule 11Ac1-4 ("Display Rule") under the Securities Exchange Act of 1934 ("Exchange Act") to require the display of customer limit orders priced better than a specialist's or over-the-counter ("OTC") market maker's quote or that add to the size associated with such quote. The Commission also is adopting amendments to Rule 11Ac1-1 ("Quote Rule") under the Exchange Act to require a market maker to publish quotations for any listed security when it is responsible for more than 1% of the aggregate trading volume for that security and to make publicly available any superior prices that a market maker privately quotes through certain electronic communications networks ("ECNs") ("ECN amendment").

The Display Rule strengthened Nasdaq customer limit orders beyond the protections afforded by Manning. When display is required, a customer limit order can become the best bid or offer in the market.

The Quote Rule was designed to curb a practice whereby Nasdaq dealers would set wide quotes that were visible to the public, but narrow quotes in the interdealer and institutional markets that were not visible to the public.

This practice remains common in many other dealer markets (including FX and bonds).

The Quote Rule also marked the debut of the term "Electronic Communications Network" (ECN).

5.4 SuperMontage

Nasdaq is largely defined by its information systems. The first system (1960s?) imply allowed for display of dealer quotes (the quote "montage"). Systems subsequently added allowed for trade reporting and confirmation (ACT), interdealer communication of trading commitments (SelectNet), small retail order execution (SOES), etc. Smith, Selway, and McCormick (1998) describe the history and state of these systems up to 1998. Most systems underwent substantial modifications after that.

At present, most of the functionality in these disparate systems is now consolidated in one system, SuperMontage. Conceptually, SuperMontage comes closest to resembling an electronic limit order book for dealers. That is, the display and trading protocols are similar to what would be found in an electronic limit order book, except that customers are not permitted direct access to the system.

The system was designed to facilitate established Nasdaq practices like preferencing. At the SEC's request, the system was forced to include ECN's. As a result, the actual trading protocols are quite complex. (See material at http://www.nasdaqtrader.com/trader/hottopics/supermontage_hottopics.stm.)

SuperMontage's market share is (August, 2003) about 17% by volume.

6 ECN's

In US regulatory law, an ECN is simply a medium for the display of quotes. There is no requirement that it also provide a trading mechanism. Nevertheless, as a practical matter, virtually all ECN's are electronic limit order book markets, on which trades do in fact occur.

The claim as to whose system was the first electronic stock exchange is and always will be in dispute. Certainly, however, one strong contender is Instinet. It began operation in 1979 as an electronic limit order book for institutions (mutual funds, pension funds, etc.). It did not really take off, though, until it began to allow the entry of Nasdaq market makers. Until recently, it was the clearly the largest ECN (by trading volume).

The Nasdaq market makers used the system essentially as their interdealer market, and this clientele became a substantial, perhaps the dominant, group of Instinet participants.

Significantly, to avoid the regulatory overhead, Instinet went to some effort to avoid characterization as an "exchange," perhaps hoping to avoid the regulatory overhead such a designation would entail.

Perhaps more significantly, Instinet did not open itself to retail traders.

This proved to be a foregone opportunity, as newer entrants in the ECN business successfully sought and profited from the retail business.

The most successful of these was the Island ECN. Its volume grew until it met (and in some issues, surpassed) that of Instinet. Reuters (which owns Istinet) purchased Island in 2002 (2001?).

7 Alternative Trading Systems (ATSs)

With the advent of network and internet technology in the 1990's, there arose considerable interest in devising new electronic trading mechanisms. The structure of US regulation was not well-suited to this development. Regulation centers around institutions rather than functions, and the principal trading institution for regulatory purposes was the "national securities exchange". This concept is defined in narrow terms that the newer entrants did not really fit.

Initially, the SEC dealt with the new entrants using "No Action" letters that granted provisional permission for operation of trading systems. But it soon became clear that a more consistent and cohesive regulatory structure was called for.

The SEC's rule on the Regulation of Exchanges and Alternative Trading Systems ("Reg ATS," at http://www.sec.gov/rules/final/34-40760.txt) established a new framework. The key provision:

> To allow new markets to start, without disproportionate burdens, a system with less than five percent of the trading volume in all securities it trades is required only to: (1) file with the Commission a notice of operation and quarterly reports; (2) maintain records, including an audit trail of transactions; and (3) refrain from using the words "exchange," "stock market," or similar terms in its name.

Above the five percent threshold, the responsibilities of the market increase:

> If, however, an alternative trading system with five percent or more of the trading volume in any national market system security chooses to register as a broker-dealer -- instead of as an exchange -- the Commission believes it is in the public interest to integrate its activities into the national market system. In addition to the requirements for smaller alternative trading systems, Regulation ATS requires alternative trading systems that trade five percent or more of the volume in national market system securities to be linked with a registered market in order to disseminate the best priced orders in those national market system securities displayed in their systems (including institutional orders) into the public quote stream. Such alternative trading systems must also comply with the same market rules governing execution priorities and obligations that apply to members of the registered exchange or national securities association to which the alternative trading system is linked.

The ECNs constituted as electronic limit order books are also ATSs. ATSs that do not display quotes (and hence are not ECNs) include Liquidnet and ITG's Posit crossing.

Domowitz and Lee (2001) give further discussion.

8 Decimalization

Up to the 1990s, US stocks by longstanding practice had been priced in units of ⅛ of a dollar.

This was, by world-wide standards, archaic, as most exchanges quoted in decimals. The US practice was viewed mostly as a slightly annoying, but mostly neutral aberration.

Then the Nasdaq collusion investigations raised in public consciousness the possibility that a market's tick size might have a large effect on trading costs.

To take the simplistic view, if the bid and ask are set by "insiders" and outside customers and only trade at the these prices, insiders would seek to keep spreads wide. The tick size sets a floor on how narrow the spreads can become. It was conjectured that if the tick size were mandated to be smaller, spreads would fall to the cost of providing dealer services.

Positioned as a populist issue, the outcome was never in doubt. Congress passed the Common Cents Pricing Act of 1997. The NYSE switched to sixteenths and then, as required by the law, to pennies.

Two figures from Stoll and Schenzler (2002) describe the effects:

Figure 1: Monthly Quoted and Effective Half Spread MSFT 1993-2001.x
Based on Daily Averages

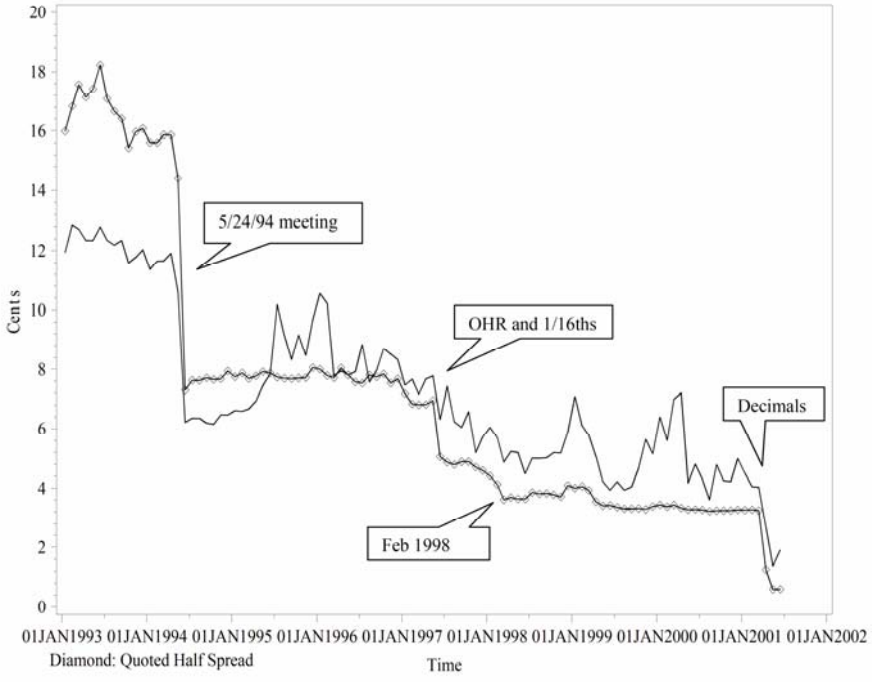
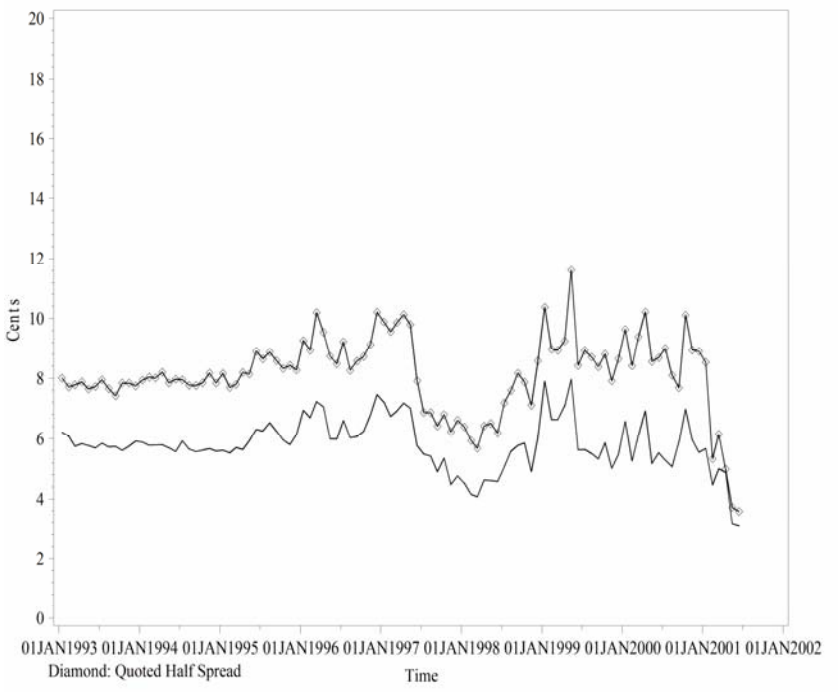


Figure 2: Monthly Quoted and Effective Half Spread IBM 1993-2001.x
Based on Daily Averages



9 The (ongoing) debate over consolidation and fragmentation

In all security markets there is a trade-off between consolidation and fragmentation. Consolidation or centralization brings all trading interest together in one place, thereby lessening the need for intermediaries, but as a regulatory principle it favors the establishment and perpetuation of a single market venue with consequent concern for monopoly power. Allowing new market entrants (like the ATSs) maximizes competition among trading venues, but at any given time the trading interest in a security is likely to be dispersed (fragmented) among the venues, leading to increased intermediation and price discrepancies among markets.

The growing role of alternative trading systems, increasing competition among market venues and the experience of the Nasdaq reforms brought these concerns to the fore.

The public debate was occasioned by the call for the repeal of the NYSE's Rule 390. This rule embodied the principle that NYSE members were prohibited from conducting trades off the floor of the Exchange. At one time, the rule had great force and reach, but by the 1990's, it had been weakened considerably. It nevertheless stood as strong symbol of the Exchange's anti-competitive power.

The relationship of the NYSE (as the self-regulatory organization) and the SEC (as the final authority on approval of rules) required the NYSE to propose the rule change to the SEC. The SEC then solicited comment on the proposal and finally took action (modification and approval, in this case).

In soliciting comment, the SEC took the occasion to address broad issues. In the "Rule 390 Concept Release" (February 23, 2000, at http://www.sec.gov/rules/sro/ny9948n.htm), the SEC laid out terms of the debate and raised the relevant policy questions.

At about the same time, the US Senate Banking Committee conducted hearings in New York on the Competitive Market Supervision Act (on February 28, 2000) and on the Financial Marketplace of the Future (February 29, 2000). Referred to at the time as the "World Trade Center hearings", these meetings are noteworthy because for a short time, it appeared that momentum was developing within the financial services industry in favor of a consolidated [electronic] limit order book ("CLOB"). For reasons apparently having to do with uncertainties over what such a system might do the institutions' competitive positions, however, the momentum suddenly abated. In retrospect, this marked the high-water mark for sentiments of centralization.

10 Market data

From a functional perspective, US equity markets provide trading services, listing sponsorship of corporations, information and regulatory services.

With the advent of the ECNs and ATSs, the provision of trading services, narrowly defined as covering only the entry and matching of orders, became extremely competitive and, therefore, for many players, only marginally profitable or worse. Attention then naturally shifted to the other roles as sources of revenue.

Listing arrangements are relatively stable. The "seal of approval" that listing is hypothesized to confer is not something that a new entrant can easily replicate.

The pricing, provision and payment for regulatory services is an important matter of ongoing debate, but regulatory services are not presently regarded as a "growth area".

This leaves market data. With market fragmentation and decimalization, the volume of market data has grown. Furthermore, with the advent of computerized trading systems, so too has the need for such data. At the same time, the costs of providing it have dropped.

It is difficult to define the property rights associated with market data. (Wherein does the value of a last sale report arise? Does it belong to the buyer and seller? The exchange where the trade took place? Or does it acquire value only when combined with all the other trade data for the security?) Furthermore, collection and dissemination of the data involve large economies of scale and scope, and large fixed costs.

The SEC appointed a panel to examine the issue (the "Seligman Committee", after the chairperson). The final report and associated documents are available at: http://www.sec.gov/divisions/marketreg/marketinfo.shtml.) It is an excellent summary of the issues and market participant views.

The committee arrived at no strong consensus for future policy, but it did at least recommend that regulators maintain the present policies. These include inexpensive dissemination of last sale prices and volumes and inside bids and asks (and sizes at these quotes). Beyond this, however,

venues would be allowed to develop and price information as they saw fit. Thus, the NYSE's OpenBook system (which reports the state of limit order book with a ten second delay) can be priced at what the market will bear.

11 Intermarket linkage systems

Market data systems allow participants to see bids, asks and last sale prices across the various market venues. They do not generally, however, permit the participants to send an order or otherwise interact with the venues. This characterizes what are sometimes called access systems.

The functionality is sometimes implemented at the broker level. A broker will often have electronic links to multiple venues, which the customer can access in a uniform fashion. At this level, the decision of where the order goes is made by the broker's routing algorithms or the customer.

Reflecting the complexity of the routing decision, these systems are generically known as SORT (smart order routing technology) or SOM (smart order management) systems.

Other systems are channels for passing orders ("commitments") between market venues.

The most venerable of these is the Intermarket Trading System (ITS). ITS links the NYSE and the regional stocks exchanges. It allows brokers at one exchange to send orders directly to another.

ITS was set up in response to a Congressional mandate (the 1975 Securities Act) to build a "national market system". The latter phrase has probably given rise to deeper exegesis than any other expression in market regulation. It has been interpreted vaguely as general support for widespread participation in security markets and at the extreme as a clear charge for a consolidated limit order book.

Although it embodies many other facilities, Nasdaq's SuperSOES system certainly qualifies. It can automatically generate executions between different market-makers, between a broker and a market-maker or even between a broker and an ECN.

In addition, some of the ECNs have set up bilateral links. Archipelago displays a consolidated limit order book, comprising both its own book and Island's.

Intermarket access systems have become flash points for inter-venue disputes. A major difference arises from the relative speeds of electronic and floor-based venues. To avoid trade-throughs, the former may be forced to send orders to the latter, at a penalty in response time that some traders feel is substantial.

These disputes would be moot but for forced participation. The latitude afforded the newer electronic venues by the SEC's Reg ATS vanishes when the a venue's market share exceeds 5%. At that point, linkage with a registered market becomes mandatory.

12 The SEC's initiative on disclosure of order routing and execution practices

When brokers handle customer orders, significant agency problems arise. These occur for the usual reasons.

- The broker has a private incentive to pursue courses of action that might deviate from the agent's interest.
- It is extremely difficult for the agent to monitor the principal's performance.

Regarding the first point, the private incentive is a rather direct and simple one. Retail brokers (who accept customer orders) often enter into arrangements with market makers whereby the market maker will pay the retail broker in cash or services for all customer orders sent to the market maker. This is called payment for order flow. Payment for order flow is legal, but it does not relieve the broker of the legal duty to provide "best execution" for the customer.

The monitoring difficulties arise because customers don't generally possess detailed level of market data that might enable them to estimate the quality of the brokerage service provided on the order. Moreover, the broker's performance on one order tells us little because individual order outcomes are noisy. Performance can really only be assessed by examining a sample of orders.

Payment for order flow has long bothered many observers because of its resemblance to commercial bribery. (Consider the case of a real estate agent who sells a house and accepts side payments from potential buyers.)

The SEC has not banned payment for order flow, but it has taken steps to make payment and performance more transparent.

These steps are codified in the rule on the Disclosure of Order Routing and Execution Practices (available at http://www.sec.gov/rules/final/34-44060.htm). There are two parts to the rule: 11Ac1-5 and 11Ac1-6.

Rule 11Ac1-5 ("Dash Five") requires all market centers to report summary measures of their execution quality. "Market centers" comprises exchanges, ECN's and Nasdaq dealers. For market orders, the measures are sensible ones, comparing the trade price relative to quote midpoints prevailing before and after the trade. The dash five statistics are widely reported on market center websites (as required by the law).

Calculation of the dash five statistics must be done according to very precise rules laid down by the SEC. Nevertheless, these statistics are not audited, and the data required to verify them are not publicly available. They must be taken, therefore, with a grain of salt.

Rule 11Ac1-6 ("Dash Six") requires all customer brokers to disclose their relationships with market makers to whom they send orders. The disclosure must be detailed (how many cents per share rebated, for example). The information must also be displayed on their web sites.

13 Time line

The preceding discussion has been organized by topic. The following chart summarizes the chronology. Links in the chart are to sections in the present paper.

(Other useful timelines include the NYSE's at http://www.nyse.com/about/1020656067766.html, and the Nasdaq's at http://www.nasdaq.com/about/about_nasdaq_long.stm.)

Date	Regulatory	NYSE	Nasdaq
Jan 24, 2002		OpenBook begins operation	
Sept. 14, 2001	Final report of the SEC's Advisory Committee for Market Information		
Jan 29, 2001		Decimal pricing fully implemented	
Jan 30, 2001	SEC's Rule on the Disclosure of Order Execution and Routing Practices.		
Aug. 28, 2000		Decimal pricing begins	
April 21, 1999	SEC's "Reg ATS" becomes effective		
June 24, 1997		Begins trading in sixteenths.	
Sept, 1996	SEC's rule on Order Execution Obligations ("display rule" and "quote rule")		
May 22, 1995			Manning II
June 24, 1994			Manning I
May 27, 1994			Nasdaq market makers begin reducing their spreads.
May 24, 1994			"Bear Sterns" meeting. (NASD officials and members)
May 24, 1994			Charges of collusion leveled by Christie and Schultz in a

Date	Regulatory	NYSE	Nasdaq
			Vanderbilt University press release.

14 References

Christie, William G., Jeffrey H. Harris, and Paul H. Schultz, 1994, Why did NASDAQ market makers stop avoiding odd-eighth quotes?, Journal of Finance 49, 1841-60.

Christie, William G., and Paul H. Schultz, 1994, Why do NASDAQ market makers avoid odd-eighth quotes?, Journal of Finance 49, 1813-1840.

Domowitz, I., Lee, R., 2001. On the road to reg ATS: A critical history of the regulation of automated trading systems. Unpublished working paper. Smeal College, Penn State University.

Hasbrouck, J., Sofianos, G., Sosebee, D., 1993. New York Stock Exchange trading systems and procedures. Unpublished working paper. New York Stock Exchange.

Seligman, Joel, 1995. *The transformation of Wall Street (revised)* (Boston: Northeastern University Press).

Smith, J. W., Selway, J. P. I., McCormick, D. T., 1998. The Nasdaq stock market: historical background and current operation. Unpublished working paper. Nasdaq working paper 98-01.

Stoll, H. R., Schenzler, C., 2002. Measuring market quality: the relation between quoted and effective spreads. Unpublished working paper. Owen School of Management, Vanderbilt University.

Teweles, Richard J., Bradley, Edward S., 1998. *The Stock Market, 7th Edition* (New York: John Wiley).

U.S. Securities and Exchange Commission. 1963. *Report of Special Study of Securities Markets of the Securities and Exchange Commission*.

Bibliography

Many of the papers below are available online. Some of the online links may be reachable only if you have access to JSTOR and/or Elsevier's EconBase. *Caveat surfor*.

Bibliography

Acharya, V. V., Pedersen, L. H., 2002. Asset pricing with liquidity risk. Unpublished working paper. Stern School of Business.

Admati, Anat, and Paul Pfleiderer, 1988, A theory of intraday trading patterns: Volume and price variability, Review of Financial Studies 1, 3-40.

Amihud, Yakov, 2002, Illiquidity and stock returns: cross-section and time-series effects, Journal of Financial Markets 5, 31-56.

Amihud, Yakov, and Haim Mendelson, 1980, Dealership markets: Market-making with inventory, Journal of Financial Economics 8, 31-53.

Amihud, Yakov, and Haim Mendelson, 1986, Asset pricing and the bid-ask spread, Journal of Financial Economics 17, 223-249.

Amihud, Yakov, and Haim Mendelson, 1987, Trading mechanisms and stock returns: An empirical investigation, Journal of Finance 42, 533-553.

Amihud, Yakov, and Haim Mendelson, 1991, Volatility, efficiency, and trading: Evidence from the Japanese stock market, Journal of Finance 46, 1765-1789.

Amihud, Yakov, Haim Mendelson, and Maurizio Murgia, 1990, Stock market microstructure and return volatility: Evidence from Italy, *Journal of Banking and Finance* 14, 423-440.

Angel, J. J., 1994. Limit versus market orders. Unpublished working paper. School of Business Administration, Georgetown University.

Ansley, Craig F., W. Allen Spivey, and William J. Wrobleski, 1977, On the structure of moving average prices, Journal of Econometrics 6, 121-134.

Back, Kerry, 1992, Insider trading in continuous time, *Review of Financial Studies* 5.

Back, K., Baruch, S., 2003. Information in securities markets: Kyle meets Glosten and Milgrom. Unpublished working paper.

Baillie, Richard T., G. Geoffrey Booth, Yiuman Tse, and Tatyana Zabotina, 2002, Price discovery and common factor models, Journal of Financial Markets 5, 309-322.

Beja, Avraham, and Barry Goldman, 1980, On the dynamics of behavior of prices in disequilibrium, *Journal of Finance* 35, 235-248.

Bertsimas, Dimitris, and Andrew Lo, 1998, Optimal control of execution costs, Journal of Financial Markets 1, 1-50.

Beveridge, Stephen, and Charles R. Nelson, 1981, A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle', Journal of Monetary Economics

7, 151-174.

Biais, B., Glosten, L., Spatt, C., 2002. The microstructure of stock markets. Unpublished working paper. Université de Toulouse .

Chakravarty, Sugato, H. Gulen, and Stewart Mayhew, 2004, Informed trading in stock and option markets, Journal of Finance, Forthcoming.

Choi, J. Y., Dan Salandro, and Kuldeep Shastri, 1988, On the estimation of bid-ask spreads: theory and evidence, Journal of Financial and Quantitative Analysis 23, 219-230.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2000, Commonality in liquidity, Journal of Financial Economics 56, 3-28.

Cochrane, John H., 2001. *Asset Pricing* (Princeton: Princeton University Press).

Cohen, Kalman J., Steven F. Maier, Robert A. Schwartz, and David K. Whitcomb, 1981, Transaction costs, order placement strategy, and existence of the bid-ask spread, Journal of Political Economy 89, 287-305.

Constantinides, George M., 1986, Capital market equilibrium with transaction costs, Journal of Political Economy 94, 842-862 .

de Jong, Frank, 2002, Measures of contributions to price discovery: A comparison, Journal of Financial Markets 5, 323-328.

Dennis, P. J., Weston, J. P., 2001. Who's informed? An analysis of stock ownership and informed trading. Unpublished working paper. McIntire School of Commerce, University of Virginia.

Duffie, Darrell, 2001. *Dynamic asset pricing theory, 3rd edition* (Princeton, NJ: Princeton University Press).

Easley, David, Soeren Hvidkjaer, and Maureen O'Hara, 2002, Is information risk a determinant of asset returns?, Journal of Finance 57, 2185-2221.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, 1996, Cream-skimming or profit-sharing? The curious role of purchased order flow, Journal of Finance 51, 811-33.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, 1997, One day in the life of a very common stock, Review of Financial Studies 10, 805-835.

Easley, David, Nicholas M. Kiefer, Maureen O'Hara, and Joseph Paperman, 1996, Liquidity, information and infrequently traded stocks, Journal of Finance 51, 1405-1436.

Easley, David, and Maureen O'Hara, 1987, Price, trade size, and information in securities markets, Journal of Financial Economics 19, 69-90.

Easley, David, and Maureen O'Hara, 1992, Time and the process of security price adjustment, Journal of Finance 47, 576-605.

Eleswarapu, Venkat R., 1997, Cost of transacting and expected returns in the Nasdaq market, Journal of Finance 52, 2113-2127.

Eleswarapu, Venkat R., and Marc R. Reinganum, 1993, The seasonal behavior of the liquidity premium in asset pricing, Journal of Financial Economics 34, 373-386.

Engle, Robert F., and Clive W. J. Granger, 1987, Co-integration and error correction: representation, estimation and testing, *Econometrica* 55, 251-276.

Ertimur, Y., 2003. Financial information environment of loss firms. Unpublished working paper. Department of Accounting, Stern School, NYU.

Euronext, 2003. Harmonized Market Rules I. Unpublished working paper. Euronext.

Foster, F. Douglas, and S. Viswanathan, 1990, A theory of the interday variations in volume, variance, and trading costs in securities markets, Review of Financial Studies 3, 593-624.

Foster, F. Douglas, and S. Viswanathan, 1995, Can speculative trading explain the volume-volatility relation?, *Journal of Business and Economic Statistics* 13, 379-396 .

Foucault, Thierry, 1999, Order flow composition and trading costs in a dynamic limit order market, Journal of Financial Markets 2, 99-134.

Foucault, T., Kadan, O., Kandel, E., 2001. Limit order book as a market for liquidity. Unpublished working paper. HEC School of Management.

Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley, 2003, A theory of power law distributions in financial market fluctuations, Nature 423, 267-270.

Garman, Mark, 1976, Market microstructure, Journal of Financial Economics 3, 257-275.

George, Thomas J., Gautam Kaul, and M. Nimalendran, 1991, Estimation of the bid-ask spread and its components: a new approach, Review of Financial Studies 4, 623-656.

Glosten, Lawrence R., 1987, Components of the bid-ask spread and the statistical properties of transaction prices, Journal of Finance 42, 1293-1307.

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable?, Journal of Finance 49, 1127-61.

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

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

Goettler, R. L., Parlour, C. A., Rajan, U., 2003. Welfare in a dynamic limit order market. Unpublished working paper. GSIA, Carnegie Mellon University.

Gourieroux, Christian, Jasiak, Joann, 2001. *Financial Econometrics* (Princeton: Princeton University Press).

Greene, William H., 2002 . *Econometric Analysis* (New York: Macmillan) 5th ed.

Grossman, Sanford J., 1976, On the efficiency of competitive stock markets where

traders have diverse information, Journal of Finance 31, 573-585.

Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, American Economic Review 70, 393-408.

Hamilton, James D., 1994 . *Time Series Analysis* (Princeton: Princeton University Press).

Harris, Frederick H. deB., Thomas H. McInish, and Robert A. Wood, 2002, Common factor components vs. information shares: A reply, Journal of Financial Markets 5, 341-348.

Harris, Frederick H. deB., Thomas H. McInish, and Robert A. Wood, 2002, Security price adjustment across exchanges: an investigation of common factor components for Dow stocks, Journal of Financial Markets 5, 277-308.

Harris, Lawrence, 1990, Statistical Properties of the Roll Serial Covariance Bid/Ask Spread Estimator, Journal of Finance 45, 579-590.

Harris, Lawrence, 1998, Optimal dynamic order submission strategies in some stylized trading problems, Financial Markets, Institutions and Instruments 7, 1-76.

Harris, Lawrence E., 2003. *Trading and Exchanges* (New York: Oxford University Press).

Hasbrouck, Joel, 1988, Trades, quotes, inventories, and information, *Journal of Financial Economics* 22, 229-52.

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

Hasbrouck, Joel, 1991, The summary informativeness of stock trades: An econometric analysis, Review of Financial Studies 4, 571-95.

Hasbrouck, Joel, 1993, Assessing the quality of a security market: A new approach to transaction-cost measurement, Review of Financial Studies 6, 191-212.

Hasbrouck, Joel, 1995, One security, many markets: Determining the contributions to price discovery, Journal of Finance 50, 1175-99.

Hasbrouck, Joel, 1996, Modeling microstructure time series, in G. S. Maddala and C. R. Rao, eds., *Handbook of Statistics 14: Statistical Methods in Finance,* (Elsevier North Holland, Amsterdam), 647-692.

Hasbrouck, Joel, 2002, Stalking the efficient price in empirical microstructure specifications, Journal of Financial Markets 5, 329-339.

Hasbrouck, Joel, 2003, Intraday price formation in US equity index markets, Journal of Finance 58, 2375-2399.

Hasbrouck, J., 2003. US equity markets: overview and recent history. Unpublished working paper. Stern School, New York Univesity.

Hasbrouck, Joel, and Thomas S. Y. Ho, 1987, Order arrival, quote behavior, and the return-generating process, Journal of Finance 42, 1035-48.

Hasbrouck, Joel, and Duane Seppi, 2001, Common factors in prices, order flows and liquidity, Journal of Financial Economics 59, 383-411.

Hasbrouck, Joel, and George Sofianos, 1993, The trades of market makers: An empirical examination of NYSE specialists, Journal of Finance 48, 1565-1593.

Hasbrouck, J., Sofianos, G., Sosebee, D., 1993. New York Stock Exchange trading systems and procedures. Unpublished working paper. New York Stock Exchange.

Heaton, John, and Deborah J. Lucas, 1996, Evaluating the effects of incomplete markets on risk sharing and asset pricing, Journal of Political Economy 104, 443-487.

Ho, Thomas S. Y., and Richard G. Macris, 1984, Dealer bid-ask quotes and transaction prices: An empirical study of some Amex options, Journal of Finance 39, 23-45.

Holden, C. W., and A. Subrahmanyam. 1994. Risk Aversion, Imperfect Competition, and Long-Lived Information. *Economics Letters* 44, no. 1-2: 181-90.

Hollifield, B., Miller, R. A., Sandas, P., Slive, J., 2003. Liquidity supply and demand in limit order markets. Unpublished working paper. GSIA, Carnegie Mellon University.

Huang, Chi-Fu, Litzenberger, Robert H., 1998. *Foundations for Financial Economics* (Pearson Education POD).

Huang, Roger, and Hans Stoll, 1997, The components of the bid-ask spread: a general approach, Review of Financial Studies 10, 995-1034.

Huberman, Gur, and Dominika Halka, 2001, Systematic liquidity, *Journal of Financial Research* 24, 161-178.

Ingersoll, Jonathan E. Jr., 1987. *Theory of Financial Decision Making* (Rowman and Littlefield).

Kyle, Albert S., 1985, Continuous auctions and insider trading, Econometrica 53, 1315-1336.

Lee, Charles M., Brenda Mucklow, and Mark J. Ready, 1993, Spreads, depths and the impact of earnings information: an intraday analysis, Review of Financial Studies 6, 345-374.

Lehmann, Bruce N., 2002, Some desiderata for the measurement of price discovery across markets, Journal of Financial Markets 5, 259-276.

Lin, Ji-Chai, Gary C. Sanger, and Geoffrey G. Booth, 1995, Trade size and the components of the bid-ask spread, Review of Financial Studies 8, 1153-1183.

Lo, Andrew W., A. Craig MacKinlay, and June Zhang, 2002, Econometric models of limit order execution, Journal of Financial Economics 65, 31-71.

Madhavan, Ananth, 2000, Market microstructure: A survey, Journal of Financial Markets 3, 205-258.

Madhavan, Ananth, Matthew Richardson, and Mark Roomans, 1997, Why do security prices change?, Review of Financial Studies 10, 1035-1064.

Madhavan, Ananth, and Seymour Smidt, 1991, A Bayesian model of intraday specialist pricing, Journal of Financial Economics 30, 99-134.

Madhavan, Ananth, and Seymour Smidt, 1993, An analysis of changes in specialist inventories and quotations, Journal of Finance 48, 1595-1628.

Merton, Robert, 1980, On estimating the expected rate of return on the market, Journal of Financial Economics 8, 323-362.

Neal, Robert, and Simon M. Wheatley, 1998, Adverse selection and bid-ask spreads: Evidence from closed-end funds, Journal of Financial Markets 1, 121-149.

O'Hara, Maureen, 1995. *Market Microstructure Theory* (Cambridge, MA: Blackwell Publishers).

O'Hara, Maureen, 2003, Presidential address: Liqudity and price discovery, Journal of Finance 58, 1335-1354.

O'Hara, Maureen, and George S. Oldfield, 1986, The microeconomics of market making, Journal of Financial and Quantitative Analysis 21, 361-376.

Odders-White, E. R., Ready, M. J., 2003. Credit ratings and stock liquidity. Unpublished working paper. University of Wisconsin - Madison, School of Business.

Parlour, Christine, 1998 , Price dynamics in limit order markets, Review of Financial Studies 11, 789-816 .

Parlour, Christine A., and Duane J. Seppi, 2003, Liquidity-based competition for order flow, Review of financial studies 16, 301-303.

Pastor, Lubos, and Robert F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642-685.

Perold, Andre, 1988, The implementation shortfall: Paper vs. reality, *Journal of Portfolio Management* 14, 4-9.

Reiss, Peter C., and Ingrid M. Werner, 1998, Does risk-sharing motivate interdealer trading?, Journal of Finance 53, 1657-1703.

Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, Journal of Finance 39, 1127-1139.

Ronen, Tavy, 1998, Trading structure and overnight information: A natural experiment from the Tel-Aviv Stock Exchange , *Journal of Banking and Finance* 22, 489-512 .

Saar, G., Yu, L., 2002. Information asymmetry about the firm and the permanent price impact of trades: Is there a connection. Unpublished working paper. Finance Department, Stern School, NYU.

Sandas, Patrik, 2001, Adverse selection and competitive market making: evidence from a pure limit order book, Review of Financial Studies 14, 705-734.

Sargent, Thomas J., 1979 . *Macroeconomic Theory* (New York: Academic Press).

Seppi, Duane J., 1997, Liquidity provision with limit orders and a strategic specialist, Review of Financial Studies 10, 103-150.

Smith, J. W., Selway, J. P. I., McCormick, D. T., 1998. The Nasdaq stock market: historical background and current operation. Unpublished working paper. Nasdaq

working paper 98-01.

Stoll, Hans, 1976, Dealer inventory behavior: An empirical examination of Nasdaq stocks, Journal of Financial and Quantitative Analysis 11, 359-380.

Stoll, Hans, 1978, The pricing of security dealer services: an empirical study of Nasdaq stocks, Journal of Finance 33, 1153-1172.

Stoll, Hans R., 1978, The supply of dealer services in securities markets, Journal of Finance 33, 1133-1151.

Stoll, Hans R., 1989, Inferring the components of the bid-ask spread: theory and empirical tests, Journal of Finance 44, 115-134.

Subrahmanyam, Avanidhar, 1991, Risk aversion, market liquidity, and price efficiency, Review of Financial Studies 4, 416-441.

Subrahmanyam, Avanidhar, 1991, A theory of trading in stock index futures, Review of Financial Studies 4, 17-51.

Sunder, S. V., 2003. Investor access to conference call disclosures: impact of Regulation Fair Disclosure on information asymmetry. Unpublished working paper. Kellogg School, Northwestern University.

Tsay, Ruey S., 2002. *Analysis of Financial Time Series* (New York: John Wiley and Sons).

Watson, Mark W., 1986, Univariate detrending methods with stochastic trends, *Journal of Monetary Economics* 18, 49-75.

Werner, Ingrid M., and Allan W. Kleidon, 1996, U.K. and U.S. trading of British cross-listed stocks: An intraday analysis of market integration, Review of Financial Studies 9, 619-664.