Trading Volume*

Andrew W. Lo and Jiang Wang†

First Draft: August 10, 2000

Latest Revision: January 24, 2002

Abstract

We develop a dynamic equilibrium model of an asset market with multiple securities in which investors trade to share risks and smooth consumption over time, and investigate the empirical implications for the cross-sectional characteristics of trading volume and the dynamic volume-return relation. We extend the model to include fixed transactions costs, and when calibrated to aggregate data, the model implies realistic levels of trading volume. We also evaluate the efficacy of technical analysis in capturing the relation between prices and volume heuristically.

*We thank Joon Chae, Jannette Papastaikoudi, Antti Petajisto, Jean-Paul Sursock, and Mila Getmansky for excellent research assistance, and Franklin Allen, Lars Hansen, and participants of the 8th World Congress of the Econometric Society for helpful comments. Financial support from the MIT Laboratory for Financial Engineering and the National Science Foundation (Grant No. SBR–9709976) is gratefully acknowledged.

†MIT Sloan School of Management, 50 Memorial Drive, Cambridge, MA 02142–1347, and NBER.

Contents

1 Introduction **1**

2 A Dynamic Equilibrium Model **2**
2.1 The Economy 3
2.2 Discussion, Notation, and Simplifications 5
2.3 The Equilibrium 7
2.4 Implications for Trading and Returns 10

3 The Data **13**
3.1 Volume Measures 13
3.2 MiniCRSP Volume Data 17
3.3 Turnover Indexes 18

4 Cross-Sectional Characteristics of Volume **20**
4.1 Theoretical Implications for Volume 20
4.2 The Cross Section of Turnover 22

5 Dynamic Volume-Return Relation **32**
5.1 Theoretical Implications for Volume-Return Relation 32
5.2 The Impact of Asymmetric Information 34
5.3 Empirical Evidence 35

6 Trading Volume and Transactions Costs **37**
6.1 Equilibrium under Fixed Transactions Costs 38
6.2 Volume under Fixed Transactions Costs 43
6.3 A Calibration Exercise 44

7 Technical Analysis **48**
7.1 Automating Technical Analysis 49
7.2 Statistical Inference 51
7.3 Empirical Results 53

8 Conclusion **56**

A Appendix **57**

References **60**

1 Introduction

One of the most fundamental notions of economics is the determination of prices through the interaction of supply and demand. The remarkable amount of information contained in equilibrium prices has been the subject of countless studies, both theoretical and empirical, and with respect to financial securities, several distinct literatures devoted solely to prices have developed.[1] Indeed, one of the most well-developed and most highly cited strands of modern economics is the asset-*pricing* literature.

However, the intersection of supply and demand determines not only equilibrium prices but also equilibrium quantities, yet quantities have received far less attention, especially in the asset-pricing literature (is there a parallel asset-*quantities* literature?). One explanation for this asymmetry is the fact that for most individual investors, financial markets have traditionally been considered close to perfectly competitive, so that the size of a typical investment has little impact on prices. For such scale-free investment opportunities, quantities are largely irrelevant and returns become the basic objects of study, not prices. But for large investors such as institutional investors, the securities markets are not perfectly competitive, at least not in the short run. Moreover, when investors possess private information—about price movements, their own trading intentions, and other market factors—perfect competition is even less likely to hold.

For example, if a large pension fund were to liquidate a substantial position in one security, that security's price would drop precipitously if the liquidation were attempted through a single sell-order, yielding a significant loss in the value of the security to be sold. Instead, such a liquidation would typically be accomplished over several days, with a professional trader managing the liquidation process by breaking up the entire order into smaller pieces, each executed at opportune moments so as to minimize the trading costs and the overall impact of the sale on the market price of the security.[2] This suggests that there is information to be garnered from quantities as well as prices; a 50,000-share trade has different implications than a 5,000-share trade, and the *sequence* of trading volume contains information as well. The fact that the demand curves of even the most liquid financial securities are downward-sloping for institutional investors, and that information is often

[1] For example, the *Journal of Economic Literature* classification system includes categories such as Market Structure and Pricing (D4), Price Level, Inflation, and Deflation (E31), Determination of Interest Rates and Term Structure of Interest Rates (E43), Foreign Exchange (F31), Asset Pricing (G12), and Contingent and Futures Pricing (G13).

[2] See Chan and Lakonishok (1995) for further discussion of the price impact of institutional trades.

revealed through the price-discovery process, implies that quantities are as fundamental as prices and equally worthy of investigation. Even in absence of market imperfections, quantities reveal important information about the underlying risks of the securities and their prices. After all, such risks constitute an important motive for trading in financial securities.

In this paper, we hope to provide some balance to the asset-pricing literature by developing quantity implications of a dynamic general equilibrium model of asset markets under uncertainty, and investigating those implications empirically. Through theoretical and empirical analysis, we seek to understand the motives for trade, the process by which trades are consummated, the interaction between prices and volume, and the roles that risk preferences and market frictions play in determining trading activity as well as price dynamics. In Section 2, we propose a continuous-time model of asset prices and trading activity in which a linear factor structure emerges not only for asset returns, but also for trading activity. We examine these implications empirically in Sections 3 and 4 using recently available trading volume data for individual US equity securities from 1962 to 1996, and find that substantial cross-sectional variation in trading volume can be explained by economic factors such as trading costs. In Section 5, we examine additional implications of the model for a dynamic volume-return relation, both theoretically and empirically. We focus on trading costs explicitly in Section 6 by incorporating them into our theoretical model, and show that even small fixed costs can induce the type of trading activity we observe in existing financial markets. In Section 7, we turn our attention to technical analysis, a controversial method of forecasting future price movements based on geometric patterns in past prices and volume. Despite its somewhat dubious academic standing, technical analysis has always emphasized the importance of volume in determining the impact of price movements over time, and the fact that many professional traders are still devoted to its practice today suggests that a deeper investigation is warranted. We conclude in Section 8 with several suggestions for future research directions for incorporating trading volume into a more complete understanding of the economics of financial markets.

2 A Dynamic Equilibrium Model

In this section, we develop a simple equilibrium model of asset trading and pricing in an dynamic setting. We restrict our attention to the case where investors have homogeneous

information about the economy; the extension to the case of heterogeneous information is discussed briefly at the end of this section. Since our motivation for developing this model is to derive qualitative implications for the behavior of volume and returns, we keep the model as parsimonious as possible.

2.1 The Economy

We consider an economy defined on a continuous time-horizon $[0, \infty)$. There is a single, perishable good, which is also used as the numeraire. The underlying uncertainty of the economy is characterized by an n-dimensional Brownian motion $B = \{B_t : t \geq 0\}$, defined on its filtered probability space $(\Omega, \mathcal{F}, F, P)$. The filtration $F = \{\mathcal{F}_t : t \geq 0\}$ represents the information revealed by B over time.

There are J risky assets in the economy which we call stocks. Each stock pays a stream of dividends over time. Let D_{jt} denote the cumulative dividend of stock j paid up to time t and P_{jt} its ex-dividend price. We assume that

$$D_{jt} = \mu_{Dj} t + \sigma_{Dj} B_t \qquad (j = 1, \ldots, J) \tag{1}$$

where $\mu_{Dj} > 0$. Without loss of generality, we assume that the total number of shares outstanding is unity for all stocks.

In addition to the stock, there is also a risk-free bond that yields a constant, positive interest rate r.

There are I investors in the economy, each initially endowed with equal shares of the stocks and zero bonds, and with a stream of non-financial income (e.g., income from labor, non-traded assets, etc.). Let N_t^i denote investor i's cumulative non-financial income up to t. We assume that

$$N_t^i = \int_0^t \left(X_s^i + Y_s^i + Z_s \right) \sigma_N dB_s \qquad (i = 1, \ldots, I) \tag{2}$$

where

$$X_t^i = \int_0^t \sigma_X^i dB_s \tag{3a}$$

$$Y_t^i = \int_0^t -\alpha_Y Y_s^i ds + \sigma_Y^i dB_s \tag{3b}$$

$$Z_t = \int_0^t -\alpha_Z Z_s ds + \sigma_Z dB_s \tag{3c}$$

where $\alpha_Y > 0$ and $\alpha_Z > 0$. We also assume that for all $1 \leq k, l \leq I$, there is perfect symmetry between $\left(X_t^k, Y_t^k\right)$ and $\left(X_t^l, Y_t^l\right)$ in the joint distribution of the state variables $(\{D_t, X_t^i, Y_t^i, Z_t; i = 1, \ldots, I\})$, and

$$\sum_{i=1}^{I} X_t^i = \sum_{i=1}^{I} Y_t^i = 0 \quad \forall\, t \geq 0. \tag{4}$$

From (2), investor i's non-financial income from t to $t + dt$ is $(X_t^i + Y_t^i + Z_t)\, \sigma_N dB_t$. Thus, $\sigma_N dB_t$ defines the risk in investors' non-financial income and $(X_t^i + Y_t^i + Z_t)$ gives the investor i's risk exposure. Since Z_t is common to all investors, it measures the aggregate exposure to the risk in non-financial income. $(X_t^i + Y_t^i)$ on the other hand measures the idiosyncratic exposure of investor i. From (3), it is clear that X_t^i measures the permanent component of the idiosyncratic exposure and Y_t^i measures the transitory component.[3]

Each investor chooses his consumption and stock trading policies to maximize his expected utility over lifetime consumption. Let C denote the investor's set of consumption policies and S the set of stock trading policies, respectively. The consumption and stock trading polices are F-adapted processes that satisfy certain technical conditions (see, for example, Huang and Pages, 1990). Under a particular consumption-trading policy within the policy set, investor i's financial wealth (the market value of his stock and bond holdings), denoted by W_t^i, satisfies

$$W_t^i = \int_0^t \left(r W_s^i - c_s^i\right) ds + \sum_{j=1}^{J} S_{js}^i \left(dD_{js} + dP_{js} - r P_{js} ds\right) + dN_t^i. \tag{5}$$

We denote the set of consumption-trading policies that satisfy the budget constraint (5) by Φ.

Investors are assumed to choose the consumption-trading policy within the set Φ to

[3]Here, we have omitted any non-risky component in investors' non-financial income. Also, by (3), we have assumed that the steady-state distribution of the investors' exposure has mean zero. Relaxing these assumptions have no impact on our results. We adopt them merely for parsimony in exposition. For example, under the assumption of constant absolute risk aversion for the investors, which we make later, there is no income effect in their demand of the stocks. Thus, the level of their mean non-financial income plays no role in their stock trading. Also, a non-zero mean in their idiosyncratic exposure to the non-financial risk may change their average stock demand, but not their trading, which is generated by changes in their exposures. See Section 2.2 for more discussion on this point.

maximize their expected lifetime utility of the following form:

$$\mathrm{E}_0\left[-\int_0^\infty e^{-\rho s-\gamma c_s^i}ds\right] \qquad (\rho>0 \text{ and } \gamma>0) \tag{6}$$

subject to the terminal wealth condition:

$$\lim_{t\to\infty}\mathrm{E}\left[-e^{-\rho t-r\gamma W_t^i}\right]=0 \tag{7}$$

where γ is his risk aversion coefficient, ρ is his time-discount coefficient, and $i=1,\ldots,I$.[4] In addition, we require $4\gamma^2\sigma_N^2\sigma_Z^2<1$ to ensure that the model is well behaved.

2.2 Discussion, Notation, and Simplifications

The economy defined above has several special features. For tractability, we assume that stock dividends are normally distributed and investors have constant absolute risk aversion. These assumptions have restrictive implications. For example, investors' stock demand becomes independent of their wealth, which makes the model tractable but less realistic. As a result, investors have no need to change their stock holdings as their wealth changes, an otherwise natural motive to trade.

To provide motivation for trading, we endow investors with non-financial income that is positively correlated with stock dividends. The investors' desire to manage their total risk exposure gives rise to their need to trade the stock. For example, when an investor's exposure to non-financial risks increases, he would like to reduce his exposure to financial risks from his stock holdings because the two risks are positively correlated. In other words, he sells stock from his portfolio. Each investor's non-financial risk is determined by two factors: his exposure to the risk, which is determined by $(X_t^i+Y_t^i+Z_t)$, and the risk itself, which is given be $\sigma_N B_t$. Moreover, $(X_t^i+Y_t^i+Z_t)$ gives the idiosyncratic exposure and Z_t the aggregate exposure. As their exposures change, they adjust their stock holdings, In particular, through trading, they mutually insure the idiosyncratic part of their non-financial risk.

Another feature of our model is that the interest rate on the bond is assumed to be constant, hence the bond market is required to clear. This is a modeling choice we have

[4]The terminal wealth condition is imposed on the strategies to prevent investors from running Ponzi schemes.

made to simplify our analysis and to focus on the stock market. As will become clear later, changes in the interest rate is not important for the issues we are concerned about in this paper. Also, from an empirical perspective, given the frequency we are interested in (daily or weekly), changes in interest rates are usually small.[5]

For convenience, we introduce some notational conventions. Given m scalar elements, $e_1, \ldots, e_m$, let $(e_1, \ldots, e_m)$ denote the row vector and $(e_1; \ldots; e_m)$ the column vector formed from the m elements. Let M' denote the transpose of a matrix M. Thus, $D_t \equiv (D_{1t}; \cdots; D_{Jt})$ is the column vector of cumulative stock dividends, $\mu_D \equiv (\mu_{D1}; \ldots; \mu_{DJ})$ is its expected growth rate, and $D_t = \mu_D t + \sigma_D B_t$, where $\sigma_D \equiv (\sigma_{D1}; \ldots; \sigma_{DJ})$.

A portfolio of stocks is given by the number of shares of each stock it contains. Let S_j be the number of shares of stock j in a portfolio, where $j = 1, \ldots, J$. Then, $S \equiv (S_1; \ldots; S_j; \ldots; S_J)$ defines the portfolio. A portfolio of particular importance is the market portfolio, denoted by S^M, which is given by:

$$S^M = \iota \tag{8}$$

where ι is a column vector of 1's with rank J.

For any two processes F_t and G_t, let $\langle F_t, G_t \rangle$ denote their cross-variation process and $\sigma_{F\sigma} \equiv d\langle F_t, G_t \rangle / dt$ denote their instantaneous cross-variation.

For expositional clarity, we make the following simplifying assumptions on the correlation (cross-variation) among the state variables:

$$\sigma_X^{i\prime} \sigma_D = \sigma_Y^{i\prime} \sigma_D = 0 \quad \text{and} \quad \sigma_N = \iota' \sigma_D. \tag{9}$$

The first condition states that the idiosyncratic components of investors' exposure to non-financial risk are uncorrelated with stock dividends. The second condition states that non-financial risk itself is perfectly correlated with the aggregate risk in stock dividends. We also assume that $n = J + 2I + 1$ and σ_{DD} has full rank to avoid redundancies.

[5] Endogenize the interest by clearing the bond market introduces additional risks (the interest rate risk). This creates additional risk-sharing motives. Also, bonds with longer maturities (if they are traded) provide additional vehicles for risk sharing.

2.3 The Equilibrium

We now define and derive the equilibrium of the economy described above. Let $P_t \equiv (P_{1t}; \ldots; P_{Jt})$ be the vector of (ex-dividend) stock prices and $S_t^i \equiv (S_{1t}^i; \ldots; S_{Jt}^i)$ the vector of investor i's stock holdings.

Definition 1 *An equilibrium is given by a price process* $\{P_t : t \geq 0\}$ *and the investors' stock holdings* $\{S_t^i : t \geq 0\} \in \Phi$, $i = 1, \ldots, I$, *such that:*

1. S_t^i *solves investor* i*'s optimization problem:*

$$\text{Max} \quad \mathrm{E}_0\left[-\int_t^\infty e^{-\rho t - \gamma c_t^i} dt\right] \tag{10}$$

$$\text{s. t.} \quad W_t^i = W_0^i + \int_0^t \left(rW_s^i - c_s^i\right) ds + S_s^{i\prime}\left(dD_s + dP_s - rP_s ds\right) + dN_t^i \tag{11}$$

$$\lim_{t\to\infty} \mathrm{E}\left[e^{-\rho t - r\gamma W_t^i}\right] = 0$$

2. *The stock market clears:*

$$\sum_{i=1}^{i} S_t^i = \iota \quad \forall\ t \geq 0. \tag{12}$$

This definition of equilibrium is standard, except that the clearing of the bond market is not imposed, as discussed above.

For expositional convenience, we define Q_t to be the vector of cumulative excess dollar returns on the stocks:

$$Q_t = \int_0^t \left(dD_s + dP_s - rP_t ds\right). \tag{13}$$

For the remainder of this section, returns on the stocks always refer to their excess dollar returns (13). The solution to the equilibrium is summarized in the following theorem (see the Appendix for its derivation):

Theorem 1 *The economy defined above has a linear equilibrium in which*

$$P_t = \frac{1}{r}\mu_D - a - bZ_t \tag{14}$$

and

$$S_t^i = I^{-1}\iota + \left(X_t^i + Y_t^i\right) S^I + \left(h_{II X} X_t^i + h_{II Y} Y_t^i\right) S^{II} \tag{15}$$

where

$$a = \bar{\gamma}\sigma_{QQ}\iota + \lambda_a \sigma_{QZ}, \quad b = \frac{r\gamma}{r+\alpha_Z}\sigma_{QN} + \lambda_b \sigma_{QZ}$$

$\bar{\gamma} = \gamma/I$, λ_a, λ_b, $h_{II X}$, $h_{II Y}$ *are constants given in the Appendix,* ι *is the market portfolio and*

$$S^I \equiv (\sigma_{QQ})^{-1}\sigma_{QN} \quad , \quad S^{II} \equiv (\sigma_{QQ})^{-1}\sigma_{QZ}$$

are two hedging portfolios.

The equilibrium has the following properties. First, stock prices equal the expected value of future dividends discounted at the risk-free rate (μ_D/r) minus a risk discount ($a + bZ_t$) which is an affine function of Z_t, the aggregate exposure to the non-financial risk. Stock prices are independent of the other state variables. Second, four-fund separation holds for the investors' portfolio choices, i.e., all investors hold combinations of the same four portfolios: the market portfolio, two hedging portfolios, and the riskless bond.[6] The first hedging portfolio S^I is used to hedge non-financial risks and the second hedging portfolio S^{II} is used to hedge changes in market conditions, which are driven by changes in Z_t. Third, stock returns are not IID over time; in particular, we have

$$dQ_t = [ra + (r+\alpha_Z)bZ_t]\,dt + \sigma_Q dB_t \tag{16}$$

where $\sigma_Q = \sigma_D - b\sigma_Z$. Thus, expected stock returns are affine functions of Z_t, which follows a AR(1) process.

To develop further intuition for the model and the nature of the equilibrium, consider the special case when $Z_t = 0$:

[6] As a matter of convention, we use the phrase "$(K+1)$-fund separation" to describe the situation in which investors hold combinations of the same $K+1$ funds—the riskless bond and K stock funds. Therefore, four-fund separation implies three stock funds.

Corollary 1 *When $Z_t = 0 \; \forall \; t \geq 0$, the equilibrium stock price is*

$$P_t = \frac{1}{r}\mu_D - \bar{\gamma}\sigma_{DD}\iota$$

and the investors' stock holdings are

$$S_t^i = \left(I^{-1} - X_t^i - Y_t^i\right)\iota\,.$$

In this case, the stock prices are constant over time and all investors hold a fraction of the market portfolio. Moreover, stock returns are given by

$$dQ_t = \left(r\bar{\gamma}\sigma_{DD}\iota\right)dt + \sigma_D dB_t$$

and the return on the market portfolio is

$$dQ_{Mt} = \iota' dQ_t = r\bar{\gamma}\left(\iota'\sigma_{DD}\iota\right)dt + \iota'\sigma_D dB_t.$$

Let $\sigma_M^2 = \iota'\sigma_{DD}\iota$ denote the squared volatility of the return on the market portfolio. We can define

$$\beta_M \equiv \left(1/\sigma_M^2\right) d\langle Q_t, Q_{Mt}\rangle/dt = \left(1/\sigma_M^2\right)\sigma_{DD}\iota$$

to be the vector of the stocks' betas with respect to the market portfolio. Then the expected returns (per unit of time) of the stocks are given by

$$\bar{Q} \equiv \mathrm{E}\left[dQ_t\right]/dt = r\bar{\gamma}\sigma_{DD}\iota = \beta_M\bar{Q}_M$$

where $\bar{Q}_M = r\bar{\gamma}\sigma_M^2$ is the expected return on the market portfolio. This is the well-known pricing relation of the Sharpe-Lintner Capital Asset Pricing Model (CAPM).

Thus, when $Z_t = 0 \; \forall \; t \geq 0$, each investor only holds a fraction of the market portfolio. In other words, two-fund separation holds. When the investors adjust their stock investments, as their exposure to the non-financial risk changes, they only trade in the market portfolio. Furthermore, stock returns are IID over time and the CAPM holds.

In the more general case when the aggregate exposure to the non-financial risk Z_t is changing over time, the situation is more complicated. First, in addition to the market portfolio, the investors invest in two other portfolios to hedge their non-financial risk and changes in market conditions, respectively. Second, stock returns are no long IID over time; in general, they are predictable. Third, the CAPM no longer holds. In addition to contemporaneous market risk (the risk with respect to the market portfolio), there is also the risk of changing market conditions. For convenience, we refer to these two risks as the static risk and the dynamic risk, respectively. Different stocks have different exposures to the dynamic risk as well as the static risk. The expected returns on the stocks now depend on their exposures to these two different risks.

2.4 Implications for Trading and Returns

In the remainder of this section, we explore in more detail the implications of our model for the behavior of trading and returns.

Trading Activity

As (15) shows, the investors' portfolio choices satisfy four-fund separation where the three stock funds are: the market portfolio ι, the hedging portfolio S^I (which allows investors to hedge their current non-financial risk), and the hedging portfolio S^{II} (which allows investors to hedge against changes in market conditions, driven by changes in the aggregate exposure to non-financial risk Z_t). The fact that investors only hold and trade in a few funds has strong implications about the behavior of trading activities in the market, especially their cross-sectional characteristics.

As an illustration, consider the special case when $Z_t = 0$ for all $t \geq 0$. In this case, investors trade only in the market portfolio, implying that when an investor reduces his stock investments, he sells stocks in proportion to their weights in the market portfolio. Under our normalization, he sells an equal number of shares of each stock. Consequently, the turnover ratio must be numerically identical across all stocks, i.e., turnover exhibits an exact one-factor structure.

In the more general case of changing market conditions, investors trade in three stock portfolios, which, in general, can give rise to complex patterns in the turnover across stocks. However, when the trading in the market portfolio dominates the trading in the other two portfolios, turnover exhibits an approximate three-factor structure. There is a dominating

factor—representing the trading in the market portfolio—and two minor factors, representing the trading in the other two portfolios. Furthermore, when the trading in the market portfolio dominates and the approximate three-factor structure holds for the cross-section of turnover, the loadings on the two minor factors are proportional to the share weights of the two hedging portfolios. This provides a way to empirically identify the hedging portfolio from the data on the turnover of individual stocks. In Sections 3 and 4, we discuss the empirical implications for volume in more detail and present some supporting empirical evidence for the cross-sectional behavior of volume.

Stock Returns

The identification of the hedging portfolio allows us to further explore the predictions of the model for the behavior of returns. For example, since the stock returns are changing over time as Z_t changes, the second hedging portfolio $S^{II} = (\sigma_{QQ})^{-1} \sigma_{QZ}$ allows investors to hedge the risk of changing expected returns. As Merton (1971) has shown, among all the portfolios, the return on the hedging portfolio S^{II} has the highest correlation with changes in expected returns. In other words, it best predicts future stock returns. Moreover, the return on S^{II} serves as a proxy for the dynamic risk, while the return on the market portfolio serves as a proxy for the static risk. Consequently, the returns on these two portfolios give the two risk factors. The stocks' expected returns only depend on their loadings on these two portfolio returns. We obtain a two-factor pricing model, which extends the static CAPM when returns are IID into an intertemporal CAPM when returns are changing over time.

In contrast to the approach of Fama and French (1992), which seeks purely empirically based factor models, the approach here is to start with a structural model which specifies the risk factors, identify the risk factors empirically using the predictions of the model on trading activities, and further test its pricing relations. We explore this approach in more detail in Lo and Wang (2000b).

Volume-Return Relations

The previous discussion mainly focused on the behavior of trading activity and the behavior of returns separately. We now consider the joint behavior of return and trading activities. To fix ideas, we divide the investors into two groups: those for whom the correlation between their idiosyncratic exposure to the non- financial risk (X_t^i and Y_t^i) and the aggregate exposure (Z_t) is positive and those for whom the correlation is negative. For now, we call them group

A and group B.

When the investors' exposure to non-financial risk changes, their stock demand also shifts. In equilibrium, they trade with each other to revise their stock holdings and the stock prices adjust to reflect the change in demand. Thus, an immediate implication of our model is that the absolute changes in prices are positively correlated with contemporaneous trading activities, which has been documented empirically (for a survey of earlier empirical literature, see Karpoff, 1987).

Another implication of our model involves the dynamic relation between return and volume: current returns and volume can predict future returns. Campbell, Grossman and Wang (1993) and Wang (1994) have examined this relation (see also Antoniewicz, 1993; Llorente, Michaely, Saar, and Wang, 2000; LeBaron, 1992). The intuition for such a dynamic volume-return relation is as follows. When the investors' exposure to non-financial risk changes, they wish to trade. Stock prices must adjust to attract other investors to take the other side. However, these price changes are unrelated to the stocks' future dividends. Thus, they gives rise to changes in expected future returns. For example, when group A investors' exposure to non-financial risk increases, they want to sell the stocks in their portfolios. To attract group B investors to buy these stocks, stock prices have to decrease, yielding a negative return in the current period. Since there is no change in expectations about future dividends, the decrease in current stock prices implies an increase in expected future returns. In fact, it is this increase in expected returns that induces group B investors to increase their stock holdings. Thus, low current returns accompanied by high volume portends high future returns. In the opposite situation when group B investors' exposure to non-financial risk decreases, we have high current returns (group A investors want to buy and drives up the prices) accompanied by high volume, which predicts lower future returns (with unchanged expectation of future dividends but higher prices). Hence, our model leads to the following dynamic volume-return relation: returns accompanied by high volume are more likely to exhibit reversals in the future. In Section 5, we discuss the empirical analysis of this relation.

Merton's ICAPM

Our model bears an important relation to the intertemporal CAPM (ICAPM) framework of Merton (1973). Based on a class of assumed price processes, Merton has characterized the equilibrium conditions—including mutual-fund separation—for the investors' portfolios and

the risk factors in determining expected returns. But except for one special case when the static CAPM is obtained, Merton does not provide any specification of the primitives of the economy that can support the assumed price processes in equilibrium. Several authors, e.g., Ohlson and Rosenberg (1976), have argued that it is difficult to construct an economy capable of supporting Merton's price processes. Our model provides a concrete example in which equilibrium prices indeed fit Merton's specification.[7] Obviously, restrictive assumptions are required to achieve this specification.

3 The Data

One of the most exciting aspects of the recent literature on trading volume is the close correspondence between theory and empirical analysis, thanks to newly available daily volume data for individual US securities from the Center for Research in Securities Prices (CRSP). In this section, we describe some of the basic characteristics of this dataset as a prelude to the more formal econometric and calibration analyses of Sections 4–7. We begin in Section 3.1 by reviewing the various measures of volume, and in light of the implications of the model in Section 2, we argue that turnover is the most natural measure of trading activity. In Section 3.2, we describe the construction of our turnover database which is an extract of the CRSP volume data, and we report some basic summary statistics for turnover indexes in Section 3.3.

3.1 Volume Measures

To analyze the behavior of volume empirically, a prerequisite is an appropriate definition of volume. In the existing literature, many different volume measures have been used, including share volume, dollar volume, share turnover ratio, dollar turnover ratio, number of trades, etc. Certainly, the choice of volume measure should depend on what is supposed to be measured, i.e., what information we would like the measure to convey. A good measure of volume should contain useful information about underlying economic conditions, especially why investors trade and how they trade. Thus, the appropriate measure of volume is intimately tied to the particular model and motives for trading.

[7] Wu and Zhou (2001) considers a model that has many features similar to ours.

A Numerical Example

The theoretical model we developed in Section 2 suggests that the turnover ratio provides a good measure, in the sense that it clearly reflects the underlying economic regularities in trading activities. To illustrate this point, we consider a simple numerical example in the special case of our model when investors only trade in the market portfolio (i.e., when $Z_t = 0$ for all $t \geq 0$).

Suppose there are only two stocks, 1 and 2. For concreteness, assume that stock 1 has 10 shares outstanding and is priced at \$100 per share, yielding a market value of \$1000, and stock 2 has 30 shares outstanding and is priced at \$50 per share, yielding a market value of \$1500. Let N_{1t} and N_{2t} denote the numbers of shares outstanding for the two stocks, respectively (for expositional convenience, we do not normalize the number of shares outstanding to one in this example). We have $N_{1t} = 10$, $N_{2t} = 30$, $P_{1t} = 100$, $P_{2t} = 50$. In addition, suppose there are only two investors in this market—investor 1 and 2—and two-fund separation holds for their stock investments so that both investors hold different amounts of the market portfolio. Specifically, let investor 1 hold 1 share of stock 1 and 3 shares of stock 2, and let investor 2 hold 9 shares of stock 1 and 27 shares of stock 2. Thus, $S^1_{1t-1} = 1$, $S^1_{2t-1} = 3$, $S^2_{1t-1} = 9$, and $S^2_{2t-1} = 27$. In this way, all shares are held and both investors hold a (fraction of) the *market* portfolio (10 shares of stock 1 and 30 shares of stock 2).

Now suppose that investor 2 liquidates \$750 of his portfolio—3 shares of stock 1 and 9 shares of stock 2—and investor 1 is willing to purchase exactly this amount from investor 2 at the prevailing market prices.[8] After completing the transaction, investor 1 owns 4 shares of stock 1 and 12 shares of stock 2, and investor 2 owns 6 shares of stock 1 and 18 shares of stock 2. In other words, $S^1_{1t} = 4$, $S^1_{2t} = 12$, $S^2_{1t} = 6$, $S^2_{2t} = 18$. What kind of trading activity does this transaction imply?

For individual stocks, we can construct the following measures of trading activity:

- Number of trades per period
- Share volume, $V_{jt} \equiv \frac{1}{2}\sum_{i=1}^{2} |S^i_{jt} - S^i_{jt-1}|$
- Dollar volume, $P_{jt}V_{jt}$
- Relative dollar volume, $P_{jt}V_{jt} / \sum_j P_{jt}V_{jt}$

[8] In our model, in absence of aggregate exposure to the non-financial risk, the trading needs of the two investors are exactly the opposite of each other. Thus, trade occurs without any impact on prices.

- Share turnover, $\tau_{jt} \equiv V_{jt}/N_{jt}$
- Dollar turnover, $\nu_{jt} \equiv (P_{jt}V_{jt})/(P_{jt}N_{jt}) = \tau_{jt}$

where $j = 1, 2$.[9] To measure aggregate trading activity, we can define similar measures:

- Number of trades per period
- Total number of shares traded, $V_{1t}+V_{2t}$
- Dollar volume, $P_{1t}V_{1t}+P_{2t}V_{2t}$
- Share-weighted turnover, $\tau_t^{SW} \equiv \omega_1^{SW}\tau_{1t} + \omega_2^{SW}\tau_{2t}$, where $\omega_j^{SW} \equiv N_j/(N_1 + N_2)$ and $j = 1, 2$
- Equal-weighted turnover, $\tau_t^{EW} \equiv \frac{1}{2}(\tau_{1t} + \tau_{2t})$
- Value-weighted turnover, $\tau_t^{VW} \equiv \omega_{1t}^{VW}\tau_{1t} + \omega_{2t}^{VW}\tau_{2t}$, where $\omega_j^{VW} \equiv P_{jt}N_j/(P_{1t}N_1 + P_{2t}N_2)$ and $j = 1, 2$.

Table 1 reports the values that these various measures of trading activity take on for the hypothetical transaction between investors 1 and 2. Though these values vary considerably—2 trades, 12 shares traded, \$750 traded—one regularity does emerge: the turnover measures are all identical. This is no coincidence, but is an implication of two-fund separation from our equilibrium model. If all investors hold the same relative proportions of stocks at all times, then it can be shown that trading activity—as measured by turnover—must be identical across all stocks. Although the other measures of volume do capture important aspects of trading activity, if the focus is on the relation between volume and equilibrium models of asset markets, such as the ICAPM we developed in Section 2, turnover yields the sharpest empirical implications and is the most natural measure. For this reason, we use turnover in our empirical analysis.

Defining Individual and Portfolio Turnover

For each individual stock j, its share volume at time t is defined by

$$V_{jt} = \frac{1}{2}\sum_{i=1}^{I} |S_{jt}^i - S_{jt-1}^i|. \tag{17}$$

[9] Although the definition of dollar turnover may seem redundant since it is equivalent to share turnover, it will become more relevant in the portfolio case below.

Its turnover is defined by:

$$\tau_{jt} \equiv \frac{V_{jt}}{N_j} \tag{18}$$

where N_j is the total number of shares outstanding of stock j.[10]

For the specific purpose of investigating the volume implications of our model, we also introduce a measure of portfolio trading activity, defined as follows: For any portfolio p defined by the vector of shares held $S_t^p = (S_{1t}^p; \ldots; S_{Jt}^p)$ with non-negative holdings in all stocks, i.e., $S_{jt}^p \geq 0$ for all j, and strictly positive market value, i.e., $S_t^{p\prime} P_t > 0$, let $\omega_{jt}^p \equiv S_{jt}^p P_{jt}/(S_t^{p\prime} P_t)$ be the fraction invested in stock j, $j = 1, \ldots, J$. Then its turnover is defined to be

$$\tau_t^p \equiv \sum_{j=1}^{J} \omega_{jt}^p \tau_{jt} \ . \tag{19}$$

Under this definition, the turnover of value-weighted and equal-weighted indexes are well-defined

$$\tau_t^{VW} \equiv \sum_{j=1}^{J} \omega_{jt}^{VW} \tau_{jt} \qquad , \qquad \tau_t^{EW} \equiv \frac{1}{J} \sum_{j=1}^{J} \tau_{jt} \tag{20}$$

respectively, where $\omega_{jt}^{VW} \equiv N_j P_{jt} / \left(\sum_j N_j P_{jt}\right)$, for $j = 1, \ldots, J$.

Although (19) gives a reasonable definition of portfolio turnover within the context of our model, care must be exercised in interpreting it and generalizing it into a different context. While τ_t^{VW} and τ_t^{EW} are relevant to the volume implications of our model, they should be viewed in a more general context only as particular weighted averages of individual turnover, not necessarily as the turnover of any specific trading strategy. In particular, our definition for portfolio turnover cannot be applied too broadly. For example, when shortsales are allowed, some portfolio weights can be negative and (19) can be quite misleading since the turnover of short positions will offset the turnover of long positions. In general, the appropriate turnover measure for portfolio trading crucially depends on why these portfolios are traded and how they are traded. See Lo and Wang (2000a) for further discussion.

[10] Although we define the turnover ratio using the total number of shares traded, it is obvious that using the total dollar volume normalized by the total market value gives the same result.

Time Aggregation

Given our choice of turnover as a measure of volume for individual securities, the most natural method of handling time aggregation is to sum turnover across dates to obtain time-aggregated turnover. Formally, if the turnover for stock j at time t is given by τ_{jt}, the turnover between $t-1$ to $t+q$, for any $q \geq 0$ is given by:

$$\tau_{jt}(q) \equiv \tau_{jt} + \tau_{jt+1} + \cdots + \tau_{jt+q}. \tag{21}$$

3.2 MiniCRSP Volume Data

Having defined our measure of trading activity as turnover, we use the CRSP Daily Master File to construct *weekly* turnover series for individual NYSE and AMEX securities from July 1962 to December 1996 (1,800 weeks) using the time-aggregation method discussed in Section 3.1.[11] We choose a weekly horizon as the best compromise between maximizing sample size while minimizing the day-to-day volume and return fluctuations that have less direct economic relevance. Since our focus is the implications of portfolio theory for volume behavior, we confine our attention to ordinary common shares on the NYSE and AMEX (CRSP sharecodes 10 and 11 only), omitting ADRs, SBIs, REITs, closed-end funds, and other such exotica whose turnover may be difficult to interpret in the usual sense.[12] We also omit NASDAQ stocks altogether since the differences between NASDAQ and the NYSE/AMEX (market structure, market capitalization, etc.) have important implications for the measurement and behavior of volume (see, for example, Atkins and Dyl, 1997), and this should be investigated separately.

Throughout our empirical analysis, we report turnover and returns in units of percent per week—they are *not* annualized.

[11] To facilitate research on turnover and to allow others to easily replicate our analysis, we have produced daily and weekly "MiniCRSP" dataset extracts comprised of returns, turnover, and other data items for each individual stock in the CRSP Daily Master file, stored in a format that minimizes storage space and access times. We have also prepared a set of access routines to read our extracted datasets via either sequential and random access methods on almost any hardware platform, as well as a user's guide to MiniCRSP (see Lim et al., 1998). More detailed information about MiniCRSP can be found at the website `http://lfe.mit.edu/volume/`.

[12] The bulk of NYSE and AMEX securities are ordinary common shares, hence limiting our sample to securities with sharecodes 10 and 11 is not especially restrictive. For example, on January 2, 1980, the entire NYSE/AMEX universe contained 2,307 securities with sharecode 10, 30 securities with sharecode 11, and 55 securities with sharecodes other than 10 and 11. Ordinary common shares also account for the bulk of the market capitalization of the NYSE and AMEX (excluding ADRs of course).

Finally, in addition to the exchange and sharecode selection criteria imposed, we also discard 37 securities from our sample because of a particular type of data error in the CRSP volume entries.[13]

3.3 Turnover Indexes

Although it is difficult to develop simple intuition for the behavior of the entire time-series/cross-section volume dataset—a dataset containing between 1,700 and 2,200 individual securities per week over a sample period of 1,800 weeks—some gross characteristics of volume can be observed from value-weighted and equal-weighted turnover indexes.[14] These characteristics are presented in Figures 1 and 2, and in Table 2.

Figure 1a shows that value-weighted turnover has increased dramatically since the mid-1960's, growing from less than 0.20% to over 1% per week. The volatility of value-weighted turnover also increases over this period. However, equal-weighted turnover behaves somewhat differently: Figure 1b shows that it reaches a peak of nearly 2% in 1968, then declines until the 1980's when it returns to a similar level (and goes well beyond it during October 1987). These differences between the value- and equal-weighted indexes suggest that smaller-capitalization companies can have high turnover.

Table 3 reports various summary statistics for the two indexes over the 1962–1996 sample period as well as over five-year subperiods. Over the entire sample the average weekly turnover for the value-weighted and equal-weighted indexes is 0.78% and 0.91%, respectively. The standard deviation of weekly turnover for these two indexes is 0.48% and 0.37%, respectively, yielding a coefficient of variation of 0.62 for the value-weighted turnover index and 0.41 for the equal-weighted turnover index. In contrast, the coefficients of variation for the value-weighted and equal-weighted *returns* indexes are 8.52 and 6.91, respectively. Turnover

[13]Briefly, the NYSE and AMEX typically report volume in round lots of 100 shares—"45" represents 4500 shares—but on occasion volume is reported in shares and this is indicated by a "Z" flag attached to the particular observation. This Z status is relatively infrequent, usually valid for at least a quarter, and may change over the life of the security. In some instances, we have discovered daily share volume increasing by a factor of 100, only to decrease by a factor of 100 at a later date. While such dramatic shifts in volume is not altogether impossible, a more plausible explanation—one that we have verified by hand in a few cases—is that the Z flag was inadvertently omitted when in fact the Z status was in force. See Lim et al. (1998) for further details.

[14]These indexes are constructed from weekly individual security turnover, where the value-weighted index is re-weighted each week. Value-weighted and equal-weighted return indexes are also constructed in a similar fashion. Note that these return indexes do not correspond exactly to the time-aggregated CRSP value-weighted and equal-weighted return indexes because we have restricted our universe of securities to ordinary common shares. However, some simple statistical comparisons show that our return indexes and the CRSP return indexes have very similar time series properties.

is not nearly so variable as returns, relative to their means.

Table 3 also illustrates the nature of the secular trend in turnover through the five-year subperiod statistics. Average weekly value-weighted and equal-weighted turnover is 0.25% and 0.57%, respectively, in the first subperiod (1962–1966); they grow to 1.25% and 1.31%, respectively, by the last subperiod (1992–1996). At the beginning of the sample, equal-weighted turnover is three to four times more volatile than value-weighted turnover (0.21% versus 0.07% in 1962–1966, 0.32% versus 0.08% in 1967–1971), but by the end of the sample their volatilities are comparable (0.22% versus 0.23% in 1992–1996).

The subperiod containing the October 1987 crash exhibits a few anomalous properties: excess skewness and kurtosis for both returns and turnover, average value-weighted turnover slightly higher than average equal-weighted turnover, and slightly higher volatility for value-weighted turnover. These anomalies are consistent with the extreme outliers associated with the 1987 crash (see Figure 1).

Table 3 also reports the percentiles of the empirical distributions of turnover and returns which document the skewness in turnover that Figure 1 hints at, as well as the first 10 autocorrelations of turnover and returns and the corresponding Box-Pierce Q-statistics. Unlike returns, turnover is highly persistent, with autocorrelations that start at 91.25% and 86.73% for the value-weighted and equal-weighted turnover indexes, respectively, decaying very slowly to 84.63% and 68.59%, respectively, at lag 10. This slow decay suggests some kind of nonstationarity in turnover—perhaps a stochastic trend or *unit root* (see, for exampe, Hamilton, 1994). For these reasons, many empirical studies of volume use some form of detrending to induce stationarity. This usually involves either taking first differences or estimating the trend and subtracting it from the raw data. However, Lo and Wang (2000a) show that detrending methods can alter the time series properties of the data in significant ways, and that without further economic structure for the source of the trend, there is no canonical method for eliminating the trend (see Lo and Wang, 2000a, Table 4, for a more detailed analysis of detrending). Therefore, we shall continue to use raw turnover rather than its first difference or any other detrended turnover series in much of our empirical analysis (the sole exception is the eigenvalue decomposition of the first differences of turnover in Table 5). To address the problem of the apparent time trend and other nonstationarities in raw turnover, the empirical analysis of Section 4.2 is conducted within five-year subperiods only.

4 Cross-Sectional Characteristics of Volume

The theoretical model of Section 2 leads to sharp predictions about the cross-sectional behavior of trading activity. In this section, we examine the empirical support for these predictions. The empirical results presented here are from Lo and Wang (2000a), in which the starting point was mutual-fund separation, and our dynamic equilibrium model of Section 2 provides the theoretical foundations for such a starting point.

4.1 Theoretical Implications for Volume

The dynamic equilibrium model of Section 2 leads to (K+1)-fund separation for the investors' stock investments. In the special case with IID stock returns, we have two-fund separation. In the general case with time-varying stock returns, we have four-fund separation. In this case, expected stock returns are driven by a one-dimensional state variable Z_t. However, our model can easily be generalized to allow Z_t to be multi-dimensional, in which case, more than three funds would emerge as the separating stock funds. Thus, in the following discussion, we leave K unspecified, except that it is a small number when compared to the total number of stocks, J.

Let $S_t^k = (S_1^k; \ldots; S_J^k)$, $k = 1, \ldots, K$, denote the K separating stock funds, where the separating funds are expressed in terms of the number of shares of their component stocks. Each investors stock holdings can be expressed in term of their holdings of the K separating funds:

$$S_t^i = \sum_{k=1}^{K} h_{kt}^i S^k, \qquad i = 1, \ldots, I. \tag{22}$$

It should be emphasized that from our theoretical model, the separating stock funds are constant over time, which leads to much simpler behavior in volume. Since in equilibrium, $\sum_{i=1}^{I} S_{i,t} = S^M$ for all t, we have

$$\sum_{k=1}^{K} \left(\sum_{i=1}^{I} h_{kt}^i \right) S^k = S^M \ .$$

Thus, without loss of generality, we can assume that the market portfolio S^M is one of the separating stock funds, which we label as the first fund. The remaining stock funds *hedging*

portfolios (see Merton (1973)).[15]

From (22), investor i's holding in stock j is $S^i_{jt} = \sum_{k=1}^K h^i_{kt} S^k_j$. Therefore, the turnover of stock j at time t is

$$\tau_{jt} = \frac{1}{2}\sum_{i=1}^{I} |S^i_{jt} - S^i_{jt-1}| = \frac{1}{2}\sum_{i=1}^{I} \left| \sum_{k=1}^{K} \left(h^i_{kt} S^k_j - h^i_{kt-1} S^k_j \right) \right|, \quad j = 1, \ldots, J. \tag{23}$$

To simplify notation, we define $\tilde{h}^i_{kt} \equiv h^i_{kt} - h^i_{kt-1}$ as the change in investor i's holding of fund k from $t-1$ to t.

We now impose the assumption that the amount of trading in the hedging portfolios is small (relative to the trading in the market portfolio) for all investors:

Assumption 1 *For $k = 1, \ldots, K$, and $i = 1, \ldots, I$, $|\tilde{h}^i_{1t}| < H < \infty$ and $|\tilde{h}^i_{kt}| \leq \lambda H < \infty$ for $1 < k \leq K$, where $0 < \lambda \ll 1$, and $\tilde{h}^i_{1t}$, $\tilde{h}^i_{2t}$, ..., $\tilde{h}^i_{Jt}$ have a continuous joint probability density.*

We then have the following proposition (Lo and Wang, 2000a):

Proposition 1 *Under Assumption 1, the turnover of stock j at time t can be approximated by*

$$\tau_{jt} \approx \frac{1}{2}\sum_{i=1}^{I} |\tilde{h}^i_{1t}| + \frac{1}{2}\sum_{k=2}^{K} \left[\sum_{i=1}^{I} \operatorname{sgn}\left(\tilde{h}^i_{1t}\right) \tilde{h}^i_{kt} \right] S^k_j, \quad j = 1, \ldots, J \tag{24}$$

and the n-th absolute moment of the approximation error is $o(\lambda^n)$.

Now define the following "factors":

$$\begin{aligned} F_{1t} &\equiv \frac{1}{2}\sum_{i=1}^{I} \left| \tilde{h}^i_{1t} \right| \\ F_{kt} &\equiv \frac{1}{2}\sum_{i=1}^{I} \operatorname{sgn}\left(\tilde{h}^i_{1t}\right) \tilde{h}^i_{kt}, \qquad k = 2, \ldots, K. \end{aligned}$$

[15] In addition, we can assume that all the separating stock funds are mutually orthogonal, i.e., $S^{k\prime} S^{k'} = 0$, $k = 1, \ldots, K$, $k' = 1, \ldots, K$, $k \neq k'$. In particular, $S^{M\prime} S_k = \sum_{j=1}^{J} S^k_j = 0$, $k = 2, \ldots, K$, hence the total number of shares in each of the hedging portfolios sum to zero under our normalization. For this particular choice of the separating funds, h^i_{kt} has the simple interpretation that it is the projection coefficient of S^i_t on S^k. Moreover, $\sum_{i=1}^{I} h^i_{1t} = 1$ and $\sum_{i=1}^{I} h^i_{kt} = 0$, $k = 2, \ldots, K$.

Then the turnover of each stock j can be represented by an approximate K-factor model

$$\tau_{jt} = F_{1t} + \sum_{k=2}^{K} S_j^k F_{kt} + o(\lambda) \ , \qquad j = 1, \ldots, J. \tag{25}$$

(25) summarizes the implication of our model on the cross- sectional behavior of volume, which we now examine empirically.

4.2 The Cross Section of Turnover

To develop a sense for cross-sectional differences in turnover over the sample period, we turn our attention from turnover indexes to the turnover of individual securities. Since turnover is, by definition, an asymmetric measure of trading activity—it cannot be negative—its empirical distribution is nautrally skewed. Taking natural logarithms may provide a clearer visual representation of its behavior, hence we plot in Figure 2a the weekly deciles for the cross section of the logarithm of weekly turnover for each of the 1,800 weeks in the sample period. Figure 2b simplifies this by plotting the deciles of the cross section of *average* log-turnover, averaged within each year.

Figure 2b shows that the median log-turnover (the horizontal bars with vertical sides in Figure 2b) has a positive drift over time, but the cross-sectional dispersion is relatively stable. This suggests that the cross-sectional distribution of log-turnover is similar over time up to a location parameter, an implies a potentially useful "reduced-form" description of the cross-sectional distribution of turnover: an identically distributed random variable multiplied by a time-varying scale factor.

An important aspect of individual turnover data that is not immediately obvious from Figure 2 is the frequency of turnover outliers among securities and over time. To develop a sense for the magnitude of such outliers, we plot in Figure 3 the turnover and returns of a stock in the 1992–1996 subperiod that exhibited large turnover outliers: UnionFed Financial Corporation. Over the course of just a few weeks in the second half of 1993, Unionfed's weekly turnover jumped to a level of 250%. This was likely the result of significant news regarding the company's prospects—on June 15, 1993, the Los Angeles Times reported the fact that UnionFed, a California savings and loan, "has become 'critically undercapitalized' and subject to being seized within 90 days by federal regulators". In such cases, turnover outliers are not surprising, yet it is interesting that the returns of UnionFed during the same

period do not exhibit the same extreme behavior.

Figure 4 displays the time series of turnover and return for four randomly selected stocks during the same 1992–1996 subperiod, and although the outliers are considerably smaller here, nevertheless, there are still some rather extreme turnover values in their time series. For example, for most of 1996, Culligan Water Technologies Inc. exhibits weekly turnover in the 2–3% range, but towards the end of the year, there is one week in which the turnover jumped to 13%. Other similar patterns in Figure 4 seem to suggest that for many stocks there are short bursts of intense trading activity lasting only a few weeks, but with turnover far in excess of the "typical" values during the rest of the time. This characteristic of individual turnover is prevalent in the entire database, and must be taken into account in any empirical analysis of trading activity.

Cross-Sectional Regressions

The volume implications of our theoretical model provide a natural direction for empirical analysis: look for linear factor structure in the turnover cross- section. If two-fund separation holds, turnover should be identical across all stocks, i.e., a one-factor linear model where all stocks have identical factor loadings. If $(K+1)$-fund separation holds, turnover should satisfy a K- factor linear model. We examine these hypotheses in this section.

It is clear from Figure 2 that turnover varies considerably in the cross section, hence two-fund separation may be rejected out of hand. However, the turnover implications of two-fund separation might be *approximately* correct in the sense that the cross-sectional variation in turnover may be "idiosyncratic" white noise, e.g., cross-sectionally uncorrelated and without common factors. We shall test this and the more general $(K+1)$-fund separation hypothesis below but before doing so, we first consider a less formal, more exploratory analysis of the cross-sectional variation in turnover. In particular, we wish to examine the explanatory power of several economically motivated variables such as expected return, volatility, and trading costs in explaining the cross section of turnover.

To do this, we estimate cross-sectional regressions over five-year subperiods where the dependent variable is the median turnover $\tilde{\tau}_j$ of stock j. We use median turnover instead of mean turnover to minimize the influence of outliers, which can be substantial in this dataset (see Figures 3 and 4 and the corresponding discussion above).[16] The explanatory variables

[16]Also, within each five-year period we exclude all stocks that are missing turnover data for more than two-thirds of the subsample.

are the following stock-specific characteristics:[17]

$\hat{\alpha}_{r,j}$:	Intercept coefficient from the time-series regression of stock j's return on the value-weighted market return.
$\hat{\beta}_{r,j}$:	Slope coefficient from the time-series regression of stock j's return on the value-weighted market return.
$\hat{\sigma}_{\epsilon,r,j}$:	Residual standard deviation of the time-series regression of stock j's return on the value-weighted market return.
v_j:	Average of natural logarithm of stock j's market capitalization.
p_j:	Average of natural logarithm of stock j's price.
d_j:	Average of dividend yield of stock j, where dividend yield in week t is defined by $$d_{jt} = \max\left[\, 0\ ,\ \log\Big((1+R_{jt})V_{jt-1}/V_{jt}\Big)\,\right]$$ and V_{jt} is j's market capitalization in week t.
SP500_j	Indicator variable for membership in the S&P 500 Index.
$\hat{\gamma}_{r,j}(1)$	First-order autocovariance of returns.

The inclusion of these regressors in our cross-sectional analysis is loosely motivated by various intuitive "theories" that have appeared in the volume literature.

The motivation for the first three regressors comes partly from linear asset-pricing models such as the CAPM and APT; they capture excess expected return ($\hat{\alpha}_{r,j}$), systematic risk ($\hat{\beta}_{r,j}$), and residual risk ($\hat{\sigma}_{\epsilon,r,j}$), respectively. To the extent that expected excess return ($\hat{\alpha}_{r,j}$) may contain a premium associated with liquidity (see, for example, Amihud and Mendelson, 1986a,b, and Hu, 1997) and heterogeneous information (see, for example, He and Wang, 1995 and Wang, 1994), it should also give rise to cross-sectional differences in turnover.

[17] We use median turnover instead of mean turnover to minimize the influence of outliers, which can be substantial in this dataset (see Figures 3 and 4 and the corresponding discussion above). Also, within each five-year period we exclude all stocks that are missing turnover data for more than two-thirds of the subsample.

Although a higher premium from lower liquidity should be inversely related to turnover, a higher premium from heterogeneous information can lead to either higher or lower turnover, depending on the nature of information heterogeneity. The two risk measures of an asset, $\hat{\beta}_{r,j}$ and $\hat{\sigma}_{\epsilon,r,j}$, also measure the volatility in its returns that is associated with systematic risk and residual risk, respectively. Given that realized returns often generate portfolio-rebalancing needs, the volatility of returns should be positively related to turnover.

The motivation for log-market-capitalization (v_j) and log-price (p_t) is two-fold. On the theoretical side, the role of market capitalization in explaining volume is related to Merton's (1987) model of capital market equilibrium in which investors hold only the assets they are familiar with. This implies that larger-capitalization companies tend to have more diverse ownership, which can lead to more active trading. The motivation for log-price is related to trading costs. Given that part of trading costs comes from the bid-ask spread, which takes on discrete values in dollar terms, the actual costs in percentage terms are inversely related to price levels. This suggests that volume should be positively related to prices.

On the empirical side, there is an extensive literature documenting the significance of log-market-capitalization and log-price in explaining the cross-sectional variation of expected returns, e.g., Banz (1981), Black (1976), Brown, Van Harlow, and Tinic (1993), Marsh and Merton (1987), and Reinganum (1992). If size and price are genuine factors driving expected returns, they should drive turnover as well (see Lo and Wang (2000b) for a more formal derivation and empirical analysis of this intuition).

Dividend yield (d_j) is motivated by its (empirical) ties to expected returns, but also by *dividend-capture* trades—the practice of purchasing stock just before its ex-dividend date and then selling it shortly thereafter.[18] Often induced by differential taxation of dividends versus capital gains, dividend-capture trading has been linked to short-term increases in trading activity, e.g., Karpoff and Walking (1988, 1990), Lakonishok and Smidt (1986), Lakonishok and Vermaelen (1986), Lynch-Koski (1996), Michaely (1991), Michaely and Murgia (1995), Michaely and Vila (1995, 1996), Michaely, Vila and Wang (1996), and Stickel (1991). Stocks with higher dividend yields should induce more dividend-capture trading activity, and this may be reflected in higher median turnover.

The effects of membership in the S&P 500 have been documented in many studies, e.g.,

[18]Our definition of d_j is meant to capture net corporate distributions or outflows (recall that returns R_{jt} are inclusive of all dividends and other distributions). The purpose of the non-negativity restriction is to ensure that inflows, e.g., new equity issues, are not treated as negative dividends.

Dhillon and Johnson (1991), Goetzmann and Garry (1986), Harris and Gurel (1986), Jacques (1988), Jain (1987), Lamoureux and Wansley (1987), Pruitt and Wei (1989), Shleifer (1986), Tkac (1996), and Woolridge and Ghosh (1986). In particular, Harris and Gurel (1986) document increases in volume just after inclusion in the S&P 500, and Tkac (1996) uses an S&P 500 indicator variable to explain the cross-sectional dispersion of relative turnover (relative dollar-volume divided by relative market-capitalization). The obvious motivation for this variable is the growth of indexation by institutional investors, and by the related practice of *index arbitrage*, in which disparities between the index futures price and the spot prices of the component securities are exploited by taking the appropriate positions in the futures and spot markets. For these reasons, stocks in the S&P 500 index should have higher turnover than others. Indexation began its rise in popularity with the advent of the mutual-fund industry in the early 1980's, and index arbitrage first became feasible in 1982 with the introduction of the Chicago Mercantile Exchange's S&P 500 futures contracts. Therefore, the effects of S&P 500 membership on turnover should be more dramatic in the later subperiods. Another motivation for S&P 500 membership is its effect on the publicity of member companies, which leads to more diverse ownership and more trading activity in the context of Merton (1987).

The last variable, the first-order return autocovariance ($\hat{\gamma}_{r,j}(1)$), serves as a proxy for trading costs, as in Roll's (1984) model of the "effective" bid/ask spread. In that model, Roll shows that in the absence of information-based trades, prices bouncing between bid and ask prices implies the following approximate relation between the spread and the first-order return autocovariance:

$$\frac{s_{r,j}^2}{4} \quad \approx \quad -\operatorname{Cov}[R_{jt}, R_{jt-1}] \quad \equiv \quad -\gamma_{r,j}(1) \tag{26}$$

where $s_{r,j} \equiv s_j/\sqrt{P_{aj}P_{bj}}$ is the percentage effective bid/ask spread of stock j as a percentage of the geometric average of the bid and ask prices P_{bj} and P_{aj}, respectively, and s_j is the dollar bid/ask spread.

Rather than solve for $s_{r,j}$, we choose instead to include $\hat{\gamma}_{r,j}(1)$ as a regressor to sidestep the problem of a positive sample first-order autocovariance, which yields a complex number for the effective bid/ask spread. Of course, using $\hat{\gamma}_{r,j}(1)$ does not eliminate this problem, which is a symptom of a specification error, but rather is a convenient heuristic that allows us to estimate the regression equation (complex observations for even one regressor can yield

complex parameter estimates for all the other regressors as well!). This heuristic is not unlike Roll's method for dealing with positive autocovariances, however, it is more direct.[19]

Under the trading-cost interpretation for $\hat{\gamma}_{r,j}(1)$, we should expect a positive coefficient in our cross-sectional turnover regression—a large negative value for $\hat{\gamma}_{r,j}(1)$ implies a large bid/ask spread, which should be associated with lower turnover. Alternatively, Roll (1984) interprets a positive value for $\hat{\gamma}_{r,j}(1)$ as a negative bid/ask spread, hence turnover should be higher for such stocks.

These eight regressors yield the following regression equation to be estimated:

$$\begin{aligned}\tilde{\tau}_j \;=\; \gamma_0 \;+\; \gamma_1\hat{\alpha}_{r,j} \;+\; \gamma_2\hat{\beta}_{r,j} \;+\; \gamma_3\hat{\sigma}_{\epsilon,r,j} \;+\; \gamma_4 v_j \;+\; \gamma_5 p_j \;+\; \gamma_6 d_j \;+\\ \gamma_7 \mathrm{SP500}_j \;+\; \gamma_8\hat{\gamma}_{r,j}(1) \;+\; \epsilon_j\ . \end{aligned} \tag{27}$$

Table 3 contains the estimates of the cross-sectional regression model (27). We estimated three regression models for each subperiod: one with all eight variables and a constant term included, one excluding log market-capitalization, and one excluding log price. Since the log price and log market-capitalization regressors are so highly correlated (see Lim et al., 1998), regressions with only one or the other included were estimated to gauge the effects of multicollinearity. The exclusion of either variable does not affect the qualitative features of the regression—no significant coefficients changed sign other than the constant term—though the quantitative features were affected to a small degree. For example, in the first subperiod v_j has a negative coefficient (-0.064) and p_j has a positive coefficient (0.150), both significant at the 5% level. When v_j is omitted the coefficient of p_j is still positive but smaller (0.070), and when p_j is omitted the coefficient of v_j is still negative and also smaller in absolute magnitude (-0.028), and in both these cases the coefficients retain their significance.

The fact that size has a negative impact on turnover while price has a positive impact is an artifact of the earlier subperiods. This can be seen heuristically in the time-series plots of Figure 1—compare the value-weighted and equal-weighted turnover indexes during the first two or three subperiods. Smaller-capitalization stocks seem to have higher turnover than larger-capitalization stocks.

This begins to change in the 1977–1981 subperiod: the size coefficient is negative but

[19] In a parenthetical statement in footnote *a* of Table I, Roll (1984) writes "The sign of the covariance was preserved after taking the square root".

not significant, and when price is excluded, the size coefficient changes sign and becomes significant. In the subperiods after 1977–1981, both size and price enter positively. One explanation of this change is the growth of the mutual fund industry and other large institutional investors in the early 1980's. As portfolio managers manage larger asset bases, it becomes more difficult to invest in smaller- capitalization companies because of liquidity and corporate-control issues. Therefore, the natural economies of scale in investment management coupled with the increasing concentration of investment capital make small stocks less actively traded than large stocks. Of course, this effect should have implications for the equilibrium return of small stocks versus large stocks, and we investigate such implications in ongoing research.

The first-order return autocovariance has a positive coefficient in all subperiods except the second regression of the last subperiod (in which the coefficient is negative but insignificant), and these coefficients are significant at the 5% level in all subperiods except 1972–1976 and 1992–1996. This is consistent with the trading-cost interpretation of $\hat{\gamma}_{r,j}(1)$: a large negative return autocovariance implies a large effective bid/ask spread which, in turn, should imply lower turnover.

Membership in the S&P 500 also has a positive impact on turnover in all subperiods as expected, and the magnitude of the coefficient increases dramatically in the 1982–1986 subperiod—from 0.013 in the previous period to 0.091—also as expected given the growing importance of indexation and index arbitrage during this period, and the introduction of S&P 500 futures contracts in April 1982. Surprisingly, in the 1992–1996 subperiod, the S&P 500 coefficient declines to 0.029, perhaps because of the interactions between this indicator variable and size and price (all three variables are highly positively correlated with each other; see Lim et al. (1998) for further details). When size is omitted, S&P 500 membership becomes more important, yet when price is omitted, size becomes more important and S&P 500 membership becomes irrelevant. These findings are roughly consistent with those in Tkac (1996).[20]

Both systematic and idiosyncratic risk—$\hat{\beta}_{r,j}$ and $\hat{\sigma}_{\epsilon,r,j}$—have positive and significant impact on turnover in all subperiods. However, the impact of excess expected returns $\hat{\alpha}_{r,j}$ on

[20]In particular, she finds that S&P 500 membership becomes much less significant after controlling for the effects of size and institutional ownership. Of course, her analysis is not directly comparable to ours because she uses a different dependent variable (monthly relative dollar-volume divided by relative market-capitalization) in her cross-sectional regressions, and considers only a small sample of the very largest NYSE/AMEX stocks (809) over the four year period 1988–1991.

turnover is erratic: negative and significant in the 1977–1981 and 1992–1996 subperiods, and positive and significant in the others.

The dividend-yield regressor is insignificant in all subperiods but two: 1982–1986 and 1992–1996. In these two subperiods, the coefficient is negative, which contradicts the notion that dividend-capture trading affects turnover.

In summary, the cross-sectional variation of turnover does seem related to several stock-specific characteristics such as risk, size, price, trading costs, and S&P 500 membership. The explanatory power of these cross-sectional regressions—as measured by R^2—range from 29.6% (1992–1996) to 44.7% (1967–1971), rivaling the R^2's of typical cross-sectional return regressions. With sample sizes ranging from 2,073 (1962–1966) to 2,644 (1982–1986) stocks, these R^2's provide some measure of confidence that cross-sectional variations in median turnover are not purely random but do bear some relation to economic factors.

Tests of $(K+1)$-Fund Separation

Since two-fund and $(K+1)$-fund separation imply an approximately linear factor structure for turnover, we can investigate these two possibilities by using principal components analysis to decompose the covariance matrix of turnover (see Muirhead (1982) for an exposition of principal components analysis). If turnover is driven by a linear K-factor model, the first K principal components should explain most of the time-series variation in turnover. More formally, if

$$\tau_{jt} = \alpha_j + \delta_1 F_{1t} + \cdots + \delta_K F_{Kt} + \epsilon_{jt} \tag{28}$$

where $\mathrm{E}[\epsilon_{jt}\epsilon_{j't}] = 0$ for any $j \neq j'$, then the covariance matrix Σ of the vector $\tau_t \equiv (\tau_{1t}; \cdots; \tau_{Jt})$ can be expressed as

$$\mathrm{Var}[\tau_t] \equiv \Sigma = H\Theta H' \tag{29}$$

$$\Theta = \begin{pmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \theta_N \end{pmatrix} \tag{30}$$

where Θ contains the eigenvalues of Σ along its diagonal and H is the matrix of corresponding eigenvectors. Since Σ is a covariance matrix, it is positive semidefinite hence all the eigenvalues are nonnegative. When normalized to sum to one, each eigenvalue can be interpreted as the fraction of the total variance of turnover attributable to the corresponding principal component. If (28) holds, it can be shown that as the size N of the cross section increases without bound, exactly K normalized eigenvalues of Σ approach positive finite limits, and the remaining $N-K$ eigenvalues approach 0 (see, for example, Chamberlain, 1983 and Chamberlain and Rothschild, 1983). Therefore, the plausibility of (28), and the value of K, can be gauged by examining the magnitudes of the eigenvalues of Σ.

The only obstacle is the fact that the covariance matrix Σ must be estimated, hence we encounter the well-known problem that the standard estimator

$$\widehat{\Sigma} \equiv \frac{1}{T}\sum_{t=1}^{T}(\tau_t - \bar{\tau})(\tau_t - \bar{\tau})'$$

is singular if the number of securities J in the cross section is larger than the number of time series observations T.[21] Since J is typically much larger than T—for a five-year subperiod T is generally 261 weeks, and J is typically well over 2,000—we must limit our attention to a smaller subset of stocks. We do this by following the common practice of forming a small number of portfolios (see Campbell, Lo, and MacKinlay, 1997, Chapter 5), sorted by turnover beta to maximize the dispersion of turnover beta among the portfolios.[22] In particular, within each five-year subperiod we form ten turnover-beta-sorted portfolios using betas estimated from the previous five-year subperiod, estimate the covariance matrix $\widehat{\Sigma}$ using 261 time-series observations, and perform a principal-components decomposition on $\widehat{\Sigma}$. For purposes of comparison and interpretation, we perform a parallel analysis for returns,

[21] Singularity by itself does not pose any problems for the computation of eigenvalues—this follows from the singular-value decomposition theorem—but it does have implications for the statistical properties of estimated eigenvalues. In some preliminary Monte Carlo experiments, we have found that the eigenvalues of a singular estimator of a positive-definite covariance matrix can be severely biased. We thank Bob Korajczyk and Bruce Lehmann for bringing some of these issues to our attention and plan to investigate them more thoroughly in ongoing research.

[22] Our desire to maximize the dispersion of turnover beta is motivated by the same logic used in Black, Jensen, and Scholes (1972): a more dispersed sample provides a more powerful test of a cross-sectional relationship driven by the sorting characteristic. This motivation should not be taken literally in our context because the theoretical implications of Section 2 need not imply a prominent role for turnover beta (indeed, in the case of two-fund separation, there is no cross-sectional variation in turnover betas!). However, given the factor structure implied by $(K+1)$-fund separation (see Section 4.1), sorting by turnover betas seems appropriate.

using ten return-beta-sorted portfolios. The results are reported in Table 4.

Table 4 contains the principal components decomposition for portfolios sorted on out-of-sample betas, where the betas are estimated in two ways: relative to value-weighted indexes (τ^{VW} and R^{VW}) and equal-weighted indexes (τ^{EW} and R^{EW}).[23] The first principal component typically explains between 70% to 85% of the variation in turnover, and the first two principal components explain almost all of the variation. For example, the upper-left subpanel of Table 4 shows that in the second five-year subperiod (1967–1971), 85.1% of the variation in the turnover of turnover-beta-sorted portfolios (using turnover betas relative to the value-weighted turnover index) is captured by the first principal component, and 93.6% is captured by the first two principal components. Although using betas computed with value-weighted instead of equal-weighted indexes generally yields smaller eigenvalues for the first principal component (and therefore larger values for the remaining principal components) for both turnover and returns, the differences are typically not large.

The importance of the second principal component grows steadily through time for the value-weighted case, reaching a peak of 15.6% in the last subperiod, and the first two principal components account for 87.3% of the variation in turnover in the last subperiod. This is roughly comparable with the return portfolios sorted on value-weighted return-betas—the first principal component is by far the most important, and the importance of the second principal component is most pronounced in the last subperiod. However, the lower left subpanel of Table 4 shows that for turnover portfolios sorted by betas computed against equal-weighted indexes, the second principal component explains approximately the same variation in turnover, varying between 6.0% and 10.4% across the six subperiods.

Of course, one possible explanation for the dominance of the first principal component is the existence of a time trend in turnover. Despite the fact that we have limited our analysis to five-year subperiods, within each subperiod there is a certain drift in turnover; might this account for the first principal component? To investigate this conjecture, we perform eigenvalue decompositions for the covariance matrices of the *first differences* of turnover for the 10 turnover portfolios.

These results are reported in Table 5 and are consistent with those in Table 4: the first principal component is still the most important, explaining between 60% to 88% of the

[23] In particular, the portfolios in a given period are formed by ranking on betas estimated in the immediately preceding subperiod, e.g., the 1992–1996 portfolios were created by sorting on betas estimated in the 1987–1991 subperiod, hence the first subperiod in Table 4 begins in 1967, not 1962.

variation in the first differences of turnover. The second principal component is typically responsible for another 5% to 20%. And in one case—in-sample sorting on betas relative to the equal-weighted index during 1987–1991—the third principal component accounts for an additional 10%. These figures suggest that the trend in turnover is unlikely to be the source of the dominant first principal component.

In summary, the results of Tables 4 and 5 indicate that a one-factor model for turnover is a reasonable approximation, at least in the case of turnover-beta-sorted portfolios, and that a two-factor model captures well over 90% of the time-series variation in turnover. This lends some support to the practice of estimating "abnormal" volume by using an event-study style "market model", e.g., Bamber (1986), Jain and Joh (1988), Lakonishok and Smidt (1986), Morse (1980), Richardson, Sefcik, Thompson (1986), Stickel and Verrecchia (1994), and Tkac (1996).

As compelling as these empirical results seem to be, several qualifications should be kept in mind. First, we have provided little statistical inference for our principal components decomposition. In particular, the asymptotic standard errors reported in Tables 4 and 5 were computed under the assumption of IID Gaussian data, hardly appropriate for weekly US stock returns and even less convincing for turnover (see Muirhead, 1982, Chapter 9 for further details). In particular, Monte Carlos simulations should be conducted to check the robustness of these standard errors under a variety of data-generating processes.

5 Dynamic Volume-Return Relation

In this section, we examine the empirical support for the implications of our model for the joint behavior of returns and volume. We first use a simple version of the model to derive its specific implications for a dynamic volume-return relation. We then present the empirical evidence based on the work of Campbell, Grossman and Wang (1993) (CGW) and Llorente, Michaely, Saar, and Wang (2001) (LMSW).

5.1 Theoretical Implications for Volume-Return Relation

Our model leads to a set of predictions about the joint behavior of volume and returns, especially how they are related intertemporally. A specific relation we want to examine is how current volume and returns can forecast future returns. In the context of our model, this relation can be formally expressed by the following conditional expectation: $\mathrm{E}\left[\tilde{Q}_{t+1}|\tilde{Q}_t, \tau_t\right]$,

where $\tilde{Q}_t \equiv Q_t - Q_{t-1}$ is the (excess dollar) return on the stocks from $t-1$ to t and τ_t is the vector of turnover at t.

To simplify our analysis, we consider the special case of a single stock ($J = 1$) and two investors ($I = 2$).[24] Furthermore, we let $X_t^1 = Z_t = -X_t^2$, $Y_t^1 = Y_t^2 = 0$. In this case, the return on the stock and the investors' stock holdings can be expressed as follows

$$\tilde{Q}_{t+1} = [ra + (r + \alpha_Z)bZ_t] + \epsilon_{Qt+1} \tag{31a}$$

$$S_t^i = \left(\frac{1}{2} - X_t^i\right) \quad (i = 1, 2) \tag{31b}$$

where ϵ_{Qt+1} is normally distributed and uncorrelated with Z_t.[25] The turnover of the stock is then

$$\tau_t \equiv \frac{1}{2}\left(|X_t^1 - X_{t-1}^1| + |X_t^2 - X_{t-1}^2|\right) = |Z_t - Z_{t-1}|. \tag{32}$$

It should be emphasized that in our continuous-time model, in absence of any frictions, there is no natural timescale and the share volume over any finite time interval is not well defined. However, a natural timescale does emerge and the share volume is well defined once transactions costs are introduced; we return to this issue in Section 6. For the moment, we assume that trading occurs at certain time intervals (at least in expectation), and volume and return are measured over this interval, which is taken to be the unit of time.

We can now compute the expected return on the stock conditional on the current return and volume. The result is summarized in the following proposition (its proof can be found in Wang, 1994):

Proposition 2 *From (31) and (32), we have*

$$\mathrm{E}\left[\tilde{Q}_{t+1}|\tilde{Q}_t, \tau_t\right] = \theta_0 + \theta_1\tilde{Q}_t + \theta_2\tau_t\tilde{Q}_t + \text{higher order terms in } \tilde{Q}_t \text{ and } \tau_t \tag{33}$$

and $\theta_2 \leq 0$.

[24] In our model, even with multiple stocks, when $\sigma_D'\sigma_Z = 0$ and σ_Z^2 is small, $S^I \approx S^M$ and we obtain approximate two-fund separation. Effectively, the model reduces to the one-stock case we consider here, where the market portfolio plays the role of the single stock.

[25] From Theorem 1, $(Q_t; Z_t)$ is a Gaussian process. Furthermore, $\mathrm{E}[\tilde{Q}_{t+1}|Z_t] = ra + (r + \alpha_s z)bZ_t$ gives the expectation of $\tilde{Q}_{t+1}$ conditional on Z_t and U_{t+1} is the residual, which is also a normally distributed and independent of Z_t.

In other words, returns accompanied by high volume are more likely to exhibit reversals. CGW and LMSW have explicitly tested this dynamic relation between volume and returns in the form of (33). We discuss their empirical findings in Section 5.3.

5.2 The Impact of Asymmetric Information

An important aspect of financial markets that our dynamic equilibrium model of Section 2 neglected is the possibility of asymmetric information, which is critical to an understanding of dynamic volume-return relations.

Without information asymmetry, returns accompanied with higher volume are more likely to reverse themselves, as Proposition 2 states. The intuition behind this result is simple: Suppose (a subset of) investors sell stocks to adjust their risk profile in response to exogenous shocks to their risk exposure or preferences (e.g., their exposure to the non-financial risk), stock prices must decrease to attract other investors to take the other side of the trade. Consequently, we have a negative current return accompanied by high volume. Since the expectation of future stock dividends has not changed, the decrease in current prices corresponds to an increase in expected future returns.

In the presence of information asymmetry, especially when some investors have superior information about the stocks relative to other investors, the dynamic volume-return relation can be different. Suppose, for example, better informed investors reduce their stock positions in response to negative private information about future dividends. Current prices have to decrease to attract other investors to buy. However, prices will not drop by the amount that fully reflects the negative information since the market, in general, is not informationally efficient (in our model, due to incompleteness). As a result, they drop further later as more of the negative information gets impounded into prices (through additional trading or public revelation). In this case, we have negative current returns accompanied by high volume, followed by lower returns later.

Of course, investors may still trade for non-informational reasons (e.g., hedging their non-financial risk), in which case the opposite is true, as discussed before. The net effect, i.e., the expected future return conditioned on current return and volume, depends on the relative importance of asymmetric information. Wang (1994) has shown that in the presence of severe information asymmetry, it is possible that θ_2 becomes negative. Under the simplifying assumption that private information is short-lived and investors are myopic, LMSW further show that θ_2 decreases monotonically with the degree of information asymmetry.

The differences in the degree of information asymmetry can give rise to differences in the dynamic volume-return relation across stocks.

5.3 Empirical Evidence

Many authors have explored the dynamic volume-return relation (33). For example, CGW and LeBaron (1992) provided supporting evidence based on market indices and aggregate volume measures. Antoniewicz (1993) and Stickel and Verrecchia (1994) examine the same relation using pooled returns and volume of individual stocks and find conflicting evidence. Wang (1994) develops a model to incorporate the effect of information asymmetry on volume-return relations, which provides a framework to explain the difference between market indices and individual stocks. LMSW sharpened the results in Wang (1994) and empirically examined the cross-sectional variations in the volume-return relation predicted by the model. Their results reconcile the previous empirical evidence on the volume-return relation.

We now discuss the empirical results presented in CGW and LMSW in testing (33) for market indices and individual stocks, respectively. In both studies, turnover for the market and individual stocks are used as a measure of trading volume. In particular, following the existing literature, detrended log-turnover is used:[26]

$$\tilde{\tau}_t \quad = \quad \log\left(\tau_t + \varepsilon\right) - \frac{1}{N}\sum_{s=-N}^{-1} \log\left(\tau_{t+s} + \varepsilon\right) \ . \tag{34}$$

Here, a moving average of past turnover is used for detrending. The average is taken over N days, where N is set at 250 in CGW and 200 in LMSW. For individual stocks, daily trading volume is often zero, in which case, a small constant ε, which is set at 0.00000255 in LMSW, is added to the turnover before taking logarithms.[27] For market indices, daily volume is always positive and ε is set to zero.

Both CGW and LMSW estimate the dynamic volume-return relation of the following form:

$$R_{t+1} = \theta_0 + \theta_1 R_t + \theta_2 \tilde{\tau}_t R_t + \epsilon_{t+1} \tag{35}$$

[26] As discussed in Section 3, detrending can affect the time-series properties of the data. The detrending scheme in CGW and LMSW merely to make the results comparable to theirs or to report their results.

[27] The value of the constant is chosen to maximize the normality of the distribution of daily trading volume. See Richardson, Sefcik and Thompson (1986), Cready and Ramanan (1991), and Ajinkya and Jain (1989) for an explanation.

for market indices and individual stocks, respectively, where R_t is the rate of return on an index or an individual stock. LMSW further examine the cross-sectional variation in the coefficient θ_2. Note that there is a slight gap between the variables proposed in the theoretical model and those used in the empirical part. In particular, the model considers excess dollar returns per share and turnover while the empirical analysis considers returns per dollar and detrended log-turnover. The difference between the theoretical and corresponding empirical variables is mainly a matter of normalization. At the daily frequency that we focus on, the relation among these variables should not be very sensitive to the normalizations used here.

CGW used the value-weighted return on stocks traded on NYSE and AMEX, as provided by CRSP, from July 1962 to December 1988. For the turnover measure of the market, they used the total number of shares traded for all the stocks (in the sample), normalized by the total number of shares outstanding. As shown in Lo and Wang (2000a), this turnover measure is equivalent to the weighted average of individual turnover, where the weight for each stock is proportional to its number of shares outstanding. To be more compatible with our current analysis, we repeat the CGW procedure using our data set which is more complete (including the additional period from January 1989 to December 1996), and using the value-weighted turnover index as a measure of market turnover. The results are reported in Table 6. We find that for market indices, θ_2 is negative for the whole sample period and for each of the subperiods, confirming the prediction of the model.

LMSW examine (35) at the level of individual stocks and its cross-sectional differences. From the CRSP dataset, they have chosen the sample to be stocks traded on the NYSE and AMEX between January 1, 1983 and December 31, 1992.[28] Stocks with more than 20 consecutive days of missing values or more than 20 consecutive days without a single trade are excluded from the sample. The characteristics of the sample are given in Table 7, which reports the summary statistics of the average market capitalization, average daily share volume, average daily turnover and average price for the whole sample and for the five quintiles.

Their results on (35) are reproduced in Table 8. It shows that for stocks in the largest three quintiles, the average of their estimates for θ_2 is positive, as our model predicts. This is consistent with the Results of CGW and LeBaron (1992) based on market indices or large

[28]The main reason for their choosing this ten-year period it because it overlaps with the TAQ data on bid-ask spreads of individual stocks, which they use as a potential measure of information asymmetry among investors. In the published version of LMSW, they use a more recent but shorter sample, January 1, 1993 to December 31, 1998.

stocks. However, for the two smallest quintiles, the average of the θ_2 estimate is negative. As LMSW suggest, if market capitalization can be used an negative proxy for the degree of information asymmetry about a stock, the decrease in the value of θ_2 as market capitalization decreases is consistent with the above theoretical discussion. The negative θ_2 estimates for many small stocks are consistent with the negative θ_2 Antoniewicz (1993) reported based on pooled individual returns and volume.

What we conclude from the discussion in this section is that our theoretical model leads to important implications about the joint behavior of volume and returns. In particular, its implication on the dynamic volume-return relation is generally supported by the empirical findings. However, the model ignores many other factors in the market, such as the existence of private information, frictions, and a rich set of trading motives. These factors can be important in developing a more complete understanding the behavior of volume and its relation with returns. For the specific dynamic volume-return relations examine by LMSW, for example, the existence of information asymmetry is crucial. In this sense, we should view our model as a merely a starting point from which we can build a more complete model. Our discussion in the next two sections, which deals with transactions costs and trading behavior based on technical analysis, is one approach to pursuing this broader agenda.

6 Trading Volume and Transactions Costs

The theoretical model of Section 2 assumes that investors can trade in the market continuously at no cost. Consequently, the volume of trade is unbounded over any finite time interval. In our empirical analysis, we have avoided this issue by implicitly assuming that investors trade at a finite frequency. Trading volume over the corresponding trading interval is determined by the desired changes in the investors' stock positions over the interval, which is finite. This shortcut is intuitive. In practice, transactions costs would prevent any investor from trading continuously. They only trade discretely over time. However, the actual trading interval does depend on the nature of the costs as well as investors' motives to trade.

In this section, we address this issue directly within the framework of our theoretical model by incorporating transactions costs explicitly into our analysis. Our objective is two-fold. The narrow objective is to provide a theoretical justification for the shortcut in our empirical analysis. The broader objective is to provide a benchmark for understanding the level of volume. It has often been argued that the level of volume we see in the market is too

high to be justifiable from a rational asset-pricing perspective, in which investors trade to share risk and to smooth consumption over time (see, for example, Ross, 1986). Yet, in the absence of transactions costs, most dynamic equilibrium models imply that trading volume is infinite when the information flow to the market is continuous (i.e., a diffusion). Thus, the level of volume crucially depends on the nature and the magnitude of transactions costs. By considering the impact of transactions costs on the investors' trading behavior as well as the stock prices in our theoretical model, we hope to better understand the level of trading volume.

Our discussion in this section is based on Lo, Mamaysky and Wang (2000a) (LMW1, thereafter). Only the relevant results are presented here and the readers are referred to LMW1 for details.

6.1 Equilibrium under Fixed Transactions Costs

We start by introducing fixed transactions costs into our model. It is obvious that fixed costs makes continuous trading prohibitively costly. Consequently, investors do not adjust their stock positions continuously as new shocks hit the economy. Instead, as market conditions and their own situations change, investors will wait until their positions are far from the optimal ones under the new conditions and then trade discretely toward the optimal positions. As a result, they only trade infrequently, but by discrete amounts when they do trade. How frequent and how much they trade depend on the magnitude of the cost and the nature of their trading needs.

For tractability, we again consider a simple case of our model with only one stock and two investors. We also assume that $Z_t = 0$ for all $t \geq 0$. In this case there is no aggregate exposure to non-financial risk and the stock returns are IID over time. It should be pointed out that in our model, it is the difference in the investors' exposures to the non-financial risk that gives rise to the trading between them. The aggregate exposure does not give rise to trading needs since it affects all investors in the same way. In addition, we let $Y_t^i = 0$ for all $t \geq 0$ $(i = 1, 2)$ and $\alpha_X = 0$.[29]

In addition, we assume that investors have to pay a fixed cost each time they trade the stock. In particular, for each stock transaction, the two sides have to pay a total fixed cost

[29]There is no loss of generality by setting $Y_t^i = 0$ since X_t^i captures the same effect. Setting $\alpha_X = 0$ implies that shocks to investors' exposure to non-financial risk are permanent, not transitory, which simplifies the analysis (see LMW1 for more details).

of κ, which is exogenously specified and independent of the amount transacted. This cost is allocated between the buyer and seller as follows. For a trade δ, the transactions cost is given by

$$\kappa(\delta) = \begin{cases} \kappa^+ & \text{for } \delta > 0 \\ 0 & \text{for } \delta = 0 \\ \kappa^- & \text{for } \delta < 0 \end{cases} \tag{36}$$

where δ is the signed volume (positive for purchases and negative for sales), κ^+ is the cost for purchases, κ^- is the cost for sales, and the sum $\kappa^+ + \kappa^- = \kappa$. The allocation of the fixed cost, given by κ^+ and κ^-, is determined endogenously in equilibrium.

Under fixed transactions costs, investors only trade infrequently. We limit their stock trading policy set to impulse policies defined as follows:

Definition 2 *Let* $\mathbf{N}_+ \equiv \{1, 2, \ldots\}$. *An impulse trading policy* $\{(\tau_k, \delta_k) : k \in \mathbf{N}_+\}$ *is a sequence of trading times* τ_k *and trade amounts* δ_k *with*

(1) $0 \leq \tau_k \leq \tau_{k+1}$ *a.s.* $\forall\ k \in \mathbf{N}_+$

(2) τ_k *is a stopping time of* F

(3) δ_k *is progressively measurable with respect to* F_{τ_k}.

Following an impulse trading policy, investor i's stock holding at time t is S_t^i, given by

$$S_t^i = S_{0^-}^i + \sum_{\left\{k:\, \tau_k^i \leq t\right\}} \delta_k^i \tag{37}$$

where $S_{0^-}^i$ is his initial endowment of stock shares, which is assumed to be $\bar{S}$.

We denote the set of impulse trading policies by S_Δ and the set of consumption-trading policies by Φ_Δ. For the remainder of this section, we restrict the investors policies to the set Φ_Δ.

We also need to revise the notion of equilibrium in order to accommodate the situation of infrequent trading:

Definition 3 *In the presence of fixed transactions costs, an equilibrium in the stock market is defined by:*

(a) a price process $P = \{P_t : t \geq 0\}$ progressively measurable with respect to F

(b) an allocation of the transaction cost (κ^+, κ^-) as defined in (36)

(c) agents' consumption-trading policies $(c^i, S^i) \in \Phi_\Delta$, $i = 1, 2$

such that:

(i) each agent's consumption-trading policy solves his optimization problem as defined in (10)

(ii) the stock market clears:

$$\forall\ k \in \mathbf{N}_+ : \qquad \tau_k^1 = \tau_k^2 \tag{38a}$$

$$\delta_k^1 = -\delta_t^2. \tag{38b}$$

The solution to the equilibrium involves two standard steps: first to solve for the optimal consumption-trading policy for the two investors, given a stock price process and an allocation of the fixed transaction cost, and next to find the stock price process and cost allocation such that the stock market clears. However, in the presence of transactions costs, the market-clearing condition consists of two parts: investors' trading times always match, which is (38), and their desired trade amount also match, which is (38). In other words, "double-coincidence of wants" must always be guaranteed in equilibrium, which is a very stringent condition when investors trade only occasionally.

Before solving it, we make several comments about the equilibrium (assuming it exists). We have skipped technical details in supporting some of the comments and refer the readers to LMW1 for formal derivations.

First, in the absence of transaction costs, our model reduces to (a special version of) the model considered in Section 2. Investors trade continuously in the stock market to hedge their non-financial risk. Since their non-financial risks offset each other, the investors can eliminate their non-financial risk through trading. Consequently, the equilibrium price remains constant over time, independent of the idiosyncratic non-financial risk as characterized by $X_t^1 = -X_t^2$. In particular, the equilibrium price has the following form:

$$P_t = \frac{\mu_D}{r} - a \quad \forall\ t \geq 0 \tag{39}$$

where $a \equiv \bar{a} = \gamma\sigma_D^2\bar{S}$ gives the risk discount on the price of the stock to compensate for its risk. The investors' optimal stock holding is linear in his exposure to the non-financial risk:

$$S_t^i = \bar{S} - X_t^i \tag{40}$$

where $\bar{\theta}$ is a number of stock shares per capita.

Second, in the presence of transaction costs, investors only trade infrequently. However, whenever they trade, we expect them to reach optimal risk-sharing. This implies, as in the case of no transactions costs, that the equilibrium price at all trades should be the same, independent of the idiosyncratic non-traded risk X_t^i $(i = 1, 2)$. Thus, we consider the candidate stock price processes of the form (39) even in the presence of transaction costs.[30]The discount a now reflects the price adjustment of the stock for both its risk and illiquidity.

Third, since the stock price stays constant (in the conjectured equilibrium, which is verified later), changes in investors' stock demand is purely driven by changes in their exposure to the non-financial risk. Given the correlation between the non-financial risk and the stock risk, which becomes perfect when $Z_t = 0$ as assumed here, each investor can completely hedge their non-financial risk by trading the stock. The net risk an investor actually bears is the difference between his stock position and his exposure to the non-financial risk. In other words, what he would like to control is really $z_t^i = S_t^i - X_t^i$, which determines his net risk exposure. Thus, z_t^i becomes the effective state variable (in addition to his wealth) that drives investor i's optimal policy and determines his value function.

Investor i can control z_t^i by adjusting his stock holding S_t^i, from time to time. In particular, his optimal trading policy falls into the class of bang-bang policies:[31] He maintains z_t^i within a certain band, characterized by three parameters, (z_l, z_m, z_u). When z_t^i hits the upper bound of the band z_u, investor i sells $\delta^- \equiv z_u - z_m$ shares of the stock to move z_t^i down to z_m. When z_t^i hits the lower bound of the band z_l, investor i buys $\delta^+ \equiv z_m - z_l$ shares of the stock to move z_t^i up to z_m. Thus, trading occurs at the first passage time when z_t^i hits z_l or z_m and the trade amount is δ^+ or δ^-, correspondingly. The optimal policy is then further determined by the optimal choice of (z_l, z_m, z_u), which depends on the stock price (i.e., a)

[30]Given the perfect symmetry between the two agents, the economy is invariant under the following transformation: $X_t^1 \to -X_t^1$. This implies that the price must be an even function of X_t. A constant is the simplest even function.

[31]see LMW1 for more discussion on the optimal policy and references therein.

and the allocation of transaction costs.

Fourth, given nature of the investors' trading policies, an equilibrium can be achieved by choosing a and $\kappa^{\pm}$, such that (1) $\delta^+ = \delta^-$ (i.e., $z_u - z_m = z_m - z_l$), and (2) $z_m = \bar{S}$. The first condition guarantees match of trading timing and amount for both investors (noting that $z_t^1 = z_t^2$. The second condition guarantees that the two investors always hold all the shares of the stock at the prevailing prices.

In light of the discussion above, the solution to the optimal trading policy becomes solving for z_m, δ^+ and δ^- given a and $\kappa^{\pm}$, and the solution to the equilibrium reduces to finding the right a and $\kappa^{\pm}$ such that $\delta^+ = \delta^-$ and $z_m = \bar{S}$.

For an arbitrary fixed cost, only numerical solutions to the equilibrium are available (see LMW1). However, when the fixed cost is small, approximate analytical solution to the equilibrium can be obtained. In particular, we seek the solution to each agent's optimal trading policy, the equilibrium stock price and cost allocation and stock price that can be approximated by powers of $\varepsilon \equiv \kappa^{\alpha}$ where α is a positive constant. Especially, $\kappa^{\pm}$ takes the form:

$$\kappa^{\pm} = \kappa \left(\tfrac{1}{2} \pm \sum_{n=1}^{\infty} k^{(n)} \varepsilon^n \right). \tag{41}$$

The following theorem summarizes the equilibrium (see LMW1 for the proof):

Theorem 2 *Let $\varepsilon \equiv \kappa^{\frac{1}{4}}$. For (a) κ small and $\kappa^{\pm}$ in the form of (41), and (b) the value function is analytic for small z and ε, the investors optimal trading policy is given by*

$$\delta^{\pm} = \phi \kappa^{\frac{1}{4}} + o(\kappa^{\frac{1}{2}}) \tag{42a}$$

$$z_m = \bar{S}. \tag{42b}$$

In equilibrium, the stock price and allocation of transaction cost are given by

$$a = \bar{a} \left(1 + \tfrac{1}{6} r \gamma^2 \sigma_D^2 \phi^2 \kappa^{\frac{1}{2}} \right) + o(\kappa^{\frac{1}{2}}) \tag{43a}$$

$$\kappa^{\pm} = \kappa \left[\tfrac{1}{2} \pm \tfrac{2}{15} r \gamma a \phi \kappa^{\frac{1}{4}} + o(\kappa^{\frac{1}{4}}) \right] \tag{43b}$$

where

$$\phi = \left(\frac{6\sigma_z^2}{r\gamma\sigma_D^2} \right)^{\frac{1}{4}}.$$

Here, $o(\kappa^\alpha)$ denotes terms of higher order of κ^α.

Two things are worth pointing out. Investors now indeed trade only infrequently. Yet, in equilibrium the coincidence of the timing and the size of their trade is guaranteed. Moreover, the transactions costs lower the stock price, giving rise to an illiquidity discount, in addition to the risk discount given by $\bar{a}$.

The results above are derived with one pair of investors, who have offsetting trading needs. Extending these results to allow more pairs is straightforward. However, when the heterogeneity in investors' trading needs takes more general forms, the solution to the equilibrium becomes much harder. Yet, as argued in LMW1, the qualitative features of investors' optimal trading policy are quite robust. Thus, we may expect the results on the trading frequency, trade size, volume to survive. But the robustness of the results on the equilibrium price with general forms of heterogeneity in trading needs is less clear.

6.2 Volume under Fixed Transactions Costs

Given the optimal trading policies in equilibrium, we can now analyze how the level of trading volume depends on the transactions costs. Intuitively, an increase in transaction costs must reduce the volume of trade. Our model suggests a specific form for this relation. In particular, the equilibrium trade size is a constant. From our solution to equilibrium, the volume of trade between time interval t and $t+1$ is given by:

$$V_{t+1} = \sum_{\{k:\, t<\tau_k\leq t+1\}} |\delta_k^i| \tag{44}$$

where $i = 1$ or 2. The average trading volume per unit of time is

$$\mathrm{E}\left[V_{t+1}\right] = \mathrm{E}\left[\sum_k 1_{\{\tau_k\in(t,\,t+1]\}}\right]\delta \equiv \omega\delta$$

where ω denotes the expected frequency of trade (i.e., the number of trades per unit of time). For convenience, we define another measure of average trading volume as the number

of shares traded over the average time between trades, or

$$V = \frac{\delta}{\Delta\tau} = \sigma_X^2/\delta \tag{45}$$

where $\Delta\tau \equiv \mathrm{E}\left[\tau_{k+1} - \tau_k\right] \approx \delta^2/\sigma_z^2$ is the average time between trades.[32] From (42), we have

$$V = \sigma_Z^2 \phi^{-1} \kappa^{-\frac{1}{4}} \left[1 + O\left(\kappa^{\frac{1}{4}}\right)\right]$$

where $O(\kappa^\alpha)$ denotes terms of the same order of κ^α. Clearly, as κ goes to zero, trading volume goes to infinity. However, we also have

$$\frac{\Delta V}{V} \approx -\tfrac{1}{4}\frac{\Delta\kappa}{k}.$$

In other words, (for positive transaction costs) one percentage increase in the transaction cost only decreases trading volume by a quarter of a percent. In this sense, within the range of positive transaction costs, an increase in the cost only reduce the volume mildly at the margin.

Figure 5 plots the average volume measure V versus different values of transaction cost κ as well as the appropriate power laws. Clearly, as κ approaches zero, volume diverges.

6.3 A Calibration Exercise

Our model shows that even small fixed transactions costs imply a significant reduction in trading volume and an illiquidity discount in asset prices. To further examine the impact of fixed costs in equilibrium, we calibrate our model using historical data and derive numerical implications for the illiquidity discount, trading frequency, and trading volume. From (43), for small fixed costs κ we can re-express the illiquidity discount π as:

$$\pi \equiv a - \bar{a} \approx \tfrac{1}{\sqrt{6}} r^{-\frac{1}{2}} \gamma^{\frac{3}{2}} \sigma_X \bar{a} \kappa^{\frac{1}{2}} \tag{46}$$

Without loss of generality, we set $\sigma_N = 1$, hence the remaining parameters to be calibrated are: the interest rate r, the risk discount $\bar{a}$, the volatility of the idiosyncratic non-traded risk σ_X, the agents' coefficient of absolute risk aversion γ, and the fixed transaction cost κ.

[32] Of course, V is different from $\mathrm{E}[V_{t+1}]$ by Jensen's inequality. The calculation of $\Delta\tau$ is straightforward (see LMW1).

The starting point for our calibration exercise is a study by Campbell and Kyle (1993). In particular, they propose and estimate a detrended stock-price process of the following form:

$$P_t = A_t - \frac{\lambda}{r} - Z_t \tag{47}$$

where A_t (the present value of future dividends discounted at the risk-free rate) is assumed to follow a Gaussian process, Z_t (fluctuations in stock demand) is assumed to follow an AR(1) Gaussian process, r is the risk-free rate, and λ/r is the risk discount.[33] In Section 2, we have seen that in the absence of transactions costs, our model yields the same price process as (47). Moreover, in our model λ/r is denoted by $\bar{a}$ and Z_t is the aggregate exposure of non-traded risk, which generates changes in stock demand. Therefore, we can obtain values for r, $\bar{a}$, γ, and σ_Z (the instantaneous volatility of Z_t) from their parameter estimates.

Campbell and Kyle based their estimates on annual time series of US real stock prices and dividends from 1871 to 1986. The real stock price of each year is defined by the Standard & Poors Composite Stock Price Index in January, normalized by the Producer Price Index in the same month. The real dividend each year is taken to be the annual dividend per-share normalized by the Producer Price Index (over this sample period, the average annual dividend growth rate is 0.013). The price and dividend series are then detrended by an exponential detrending factor $\exp(-0.013\,t)$ and the detrended series are used to estimate (47) via maximum likelihood estimation. In particular, they obtain the following estimates for the price process:

$$r = 0.037, \quad \lambda = 0.0210, \quad \bar{A} = 1.3514, \quad \alpha_Z^{\text{CK}} = 0.0890$$
$$\sigma_Z^{\text{CK}} = 0.1371, \quad \sigma_P^{\text{CK}} = 0.3311, \quad \rho_{PZ}^{\text{CK}} = -0.5176$$

where $\bar{A}$ denotes the unconditional mean of A_t^{CK}, α_Z^{CK} and σ_Z^{CK} denote the mean-reversion coefficient and the instantaneous volatility of Z_t^{CK}, σ_P^{CK} denotes the instantaneous volatility of P_t^{CK}, and ρ_{PY}^{CK} denotes the instantaneous correlation between P_t^{CK} and Z_t^{CK}.[34] From these estimates, we are able to compute values for the following parameters in our model (in

[33] See Campbell and Kyle (1993, equation (2.3), p. 3).
[34] See Campbell and Kyle's (1993, p. 20) estimates for "Model B".

addition to the value of r):

$$\bar{a}_D = 0.0500, \quad \bar{p}_0 = 0.5676, \quad \gamma\sigma_Z = 1.3470, \quad \sigma_D = 0.2853, \quad \bar{P} = 0.6486$$

(see LMW1 for the computation of these parameter values of the model from the estimates of Campbell and Kyle). These estimates do not allow us to fully specify the values of γ and σ_Z. However, they do allow us to fix the product of the two. Thus, a choice of γ uniquely specifies the value of σ_Z.

Our model also contains the parameter σ_X, the volatility of idiosyncratic non-traded risk. Because it is the *aggregate* non-traded risk that affects prices, Campbell and Kyle (1993) only provides an estimate for the volatility σ_Z of *aggregate* non-traded risk as a function of the coefficient of absolute risk aversion γ.

Obtaining an estimate for the magnitude of σ_X requires data at a more disaggregated level, which has been performed by Heaton and Lucas (1996) using PSID data. Their analysis shows that the residual variability in the growth rate of individual income—the variability of the component that is uncorrelated with aggregate income—is 8 to 13 times larger than the variability in the growth rate of aggregate income. Based on this result, we use values for σ_X to be 4 times the value of σ_Z to be conservative.[35]

The two remaining parameters to be calibrated are the coefficient of absolute risk aversion γ and the fixed cost κ. Since there is little agreement as to what the natural choices are for these two parameters, we calibrate our model for a range of values for both.

Table 9 reports the results of our calibrations. The table contains five sub-panels. The first sub-panel reports the fixed transaction cost as the percentage of average trade amount, the second sub-panel reports the expected time between trades τ (in years), the third sub-panel reports the illiquidity discount in the stock price (as a percentage of the price $\bar{P} \equiv \mu_D/r - \bar{a}$ in the absence of transaction costs, the fourth panel reports the return premiums on the stock (due to its risk as well as illiquidity), the fifth panel reports the average turnover per year, all as functions of the transactions cost κ, which ranges from 1 basis point to 5 percent of $\bar{P}$, and the absolute risk aversion coefficient γ, which ranges from 0.001 to 5.000.[36]

The entries in Table 9 show that our model is capable of yielding empirically plausible

[35]Other values for the ratio of σ_X and σ_Z are considered in LMW1.

[36]We display transactions costs as a percentage of $\bar{P}$ simply to provide a less scale-dependent measure of their magnitudes. Since κ is a fixed cost, its value is, by definition, scale-dependent and must therefore be considered in the complete context of the calibration exercise.

values for trading frequency, trading volume, and the illiquidity discount. In contrast to much of the existing literature, e.g., Huang (1998), Schroeder (1998), Vayanos (1998), we find that transactions costs can have very large impact on both the trading frequency as well as the illiquidity discount in the stock price. For example, Schroeder (1998) finds that when faced with a fixed transactions cost of 0.1%, individuals in his model trade only once every 10 years! In Table 9, we see that for a 0.1% fixed cost, individuals in our model trade anywhere between $1/0.002 = 600$ and $1/0.148 = 6.8$ times per year as the risk aversion parameter varies from 0.001 to 5.000, respectively. This contrast between our results and those of the existing literature stems from the fact that our investors have a strong need to trade frequently. The high-frequency changes in their risk exposure to non-financial risk imply that not trading can be very costly. Furthermore, not trading means that the risk exposure from holding market-clearing levels of the stock is much greater. Models of transactions cost often fail to account for a high-frequency component in trading needs.[37]

As the risk aversion parameter increases while holding κ fixed, trading becomes less frequent, the illiquidity discount increases, and the trade size also declines. For example, a risk aversion parameter of 5.00 and a fixed cost of 1% of $\bar{P}$ implies that the investor will trade approximately once every two years, each trade consisting of only 1.297 shares, with an illiquidity discount of 1.547% of $\bar{P}$.

What seems very striking is that for reasonable magnitudes of investors trading needs (measured by σ_X) and transaction cost, our model produces reasonable levels of volume. For example, a risk aversion parameter of 1 and a fixed cost of 0.50% of $\bar{P}$ imply a trading frequency of $1/0.148 = 6.8$ trades per year, an illiquidity discount of 1.97%, and a turnover of 352%, which is higher than the observed turnover (see Section 4). These results suggest that existing levels of trading frequency and volume in financial markets may not be as unusual or as irrational as many have thought. The calibration results in Table 9 shows that our dynamic equilibrium model is clearly capable of generating empirically plausible implications.

[37] While many partial equilibrium models do contain a high-frequency component in the uncertainty faced by their investors, such as Constantinides (1986) and Amihud and Mendelson (1986a), they still miss these equilibrium effects because they do not take into account the unwillingness of investors to hold large amounts of the risky asset in the presence of transactions costs.

7 Technical Analysis

Although the interest in trading volume is a relatively recent development in the academic finance literature, it has been a long-standing tradition among finance professionals engaging in "technical analysis" or "charting", the practice of forecasting future price movements in financial securities based on geometric patterns in the time series plots of past prices and volume. Historically ridiculed by academics as "voodoo finance", technical analysis has never enjoyed the widespread acceptance among academics and industry practitioners that fundamental analysis and quantitative finance have. However, several recent academic studies suggest that historical prices may contain incremental information that has not already been incorporated into current market prices, raising the possibility that technical analysis can add value to the investment process.[38] Moreover, a closer reading of the early technical analysis literature, e.g., Hamilton (1922), reveals a surprisingly contemporary view of the market forces that influence prices and price dynamics. In particular, the importance of supply and demand, buying and selling pressure, and the risk preferences of market participants were acknowledged by technical analysts long before financial economists developed similar interests (albeit with different tools and terminology). And the emphasis that technical analysts place on trading volume is the motivation for our interest in technical analysis.

In this section, we review the results of Lo, Mamaysky, and Wang (2000b) (hereafter, "LMW2") in which the information content of technical indicators is measured by first developing an automated procedure for detecting certain types of patterns, e.g., head-and-shoulders, and then applying this procedure to historical prices of US stocks to measure the impact of these patterns on post-pattern return distributions. By comparing the unconditional empirical distribution of daily stock returns to the conditional distribution—conditioned on the occurrence of specific technical indicators such as head-and-shoulders or double-bottoms—they find that over the 35-year sample period, several technical indicators do provide incremental information and may have some practical value.

[38] For example, in rejecting the Random Walk Hypothesis for weekly US stock indexes, Lo and MacKinlay (1988, 1999) have shown that past prices may be used to forecast future returns to some degree, a fact that all technical analysts take for granted. Studies by Tabell and Tabell (1964), Treynor and Ferguson (1985), Brown and Jennings (1989), Jegadeesh and Titman (1993), Blume, Easley, and O'Hara (1994), Chan, Jegadeesh, and Lakonishok (1996), Lo and MacKinlay (1997), Grundy and Martin (1998), and Rouwenhorst (1998) have also provided indirect support for technical analysis, and more direct support has been given by Pruitt and White (1988), Neftci (1991), Brock, Lakonishok, and LeBaron (1992), Neely, Weller, and Dittmar (1997), Neely and Weller (1998), Chang and Osler (1994), Osler and Chang (1995), and Allen and Karjalainen (1999).

In Section 7.1 we describe the pattern-detection algorithm of LMW2, Section 7.2 discusses the statistical methods for gauging the information content of the patterns detected, and Section 7.3 reports the empirical results of the pattern-detection algorithm applied to a large sample of individual US stocks from 1962 to 1996.

7.1 Automating Technical Analysis

To determine the efficacy of technical analysis, we must be able to apply it in a consistent fashion over an extended period of time and across a broad sample of securities, and then assess its performance statistically. Therefore, we must first develop a method for automating the identification of technical indicators, i.e., we require a pattern-recognition algorithm. Once such an algorithm is developed, it can be applied to a large number of securities over many time periods to quantify the information content of various technical indicators. Moreover, quantitative comparisons of the performance of several indicators can be conducted, and the statistical significance of such performance can be assessed through Monte Carlo simulation and bootstrap techniques. This is the approach taken by LMW2.[39]

The starting point of LMW2's analysis is the assumption that prices P_t can be represented by the following expression:

$$P_t = m(\cdot) + \epsilon_t \tag{48}$$

where $m(\cdot)$ is a nonlinear function of time (and perhaps other state variables) and ϵ_t is white noise. LMW2 argue that technical analysts estimate $m(\cdot)$ visually by attempting to discern geometric regularities in the raw price series $\{P_t\}$, and that this process is similar in spirit to *smoothing estimators* in which sophisticated forms of local averaging are used to estimate $m(\cdot)$ by averaging out the noise ϵ_t. Specifically, LMW2 propose the following algorithm for detecting the occurrence of various technical patterns:

1. Define each technical pattern in terms of its geometric properties, e.g., local extrema (maxima and minima) of $m(\cdot)$.

2. Construct a kernel estimator $\hat{m}(\cdot)$ of a given time series of prices so that its extrema

[39] A similar approach has been proposed by Chang and Osler (1994) and Osler and Chang (1995) for the case of foreign-currency trading rules based on a head-and-shoulders pattern. They develop an algorithm for automatically detecting geometric patterns in price or exchange data by looking at properly defined local extrema.

can be determined numerically.

3. Analyze $\hat{m}(\cdot)$ for occurrences of each technical pattern.

LMW2 focus on five pairs of technical patterns that are among the most popular patterns of traditional technical analysis (see, for example, Edwards and Magee, 1966, Chapters VII–X): head-and-shoulders (HS) and inverse head-and-shoulders (IHS), broadening tops (BT) and bottoms (BB), triangle tops (TT) and bottoms (TB), rectangle tops (RT) and bottoms (RB), and double tops (DT) and bottoms (DB). Specifically, denote by $E_1, E_2, \ldots, E_n$ the n extrema of $m(\cdot)$ and $t_1^*, t_2^*, \ldots, t_n^*$ the dates on which these extrema occur. Then LMW2 propose the following definitions for the head-and-shoulders and inverted head-and-shoulders patterns:

Definition 1 (Head-and-Shoulders) *Head-and-shoulders (HS) and inverted head-and-shoulders (IHS) patterns are characterized by a sequence of five consecutive local extrema $E_1, \ldots, E_5$ such that:*

$$HS \equiv \begin{cases} E_1 \text{ a maximum} \\ E_3 > E_1 \ , \ E_3 > E_5 \\ E_1 \text{ and } E_5 \text{ within } 1.5 \text{ percent of their average} \\ E_2 \text{ and } E_4 \text{ within } 1.5 \text{ percent of their average} \end{cases}$$

$$IHS \equiv \begin{cases} E_1 \text{ a minimum} \\ E_3 < E_1 \ , \ E_3 < E_5 \\ E_1 \text{ and } E_5 \text{ within } 1.5 \text{ percent of their average} \\ E_2 \text{ and } E_4 \text{ within } 1.5 \text{ percent of their average} \end{cases}$$

Note that only five consecutive extrema are required to identify a head-and-shoulders pattern, which follows from the formalization of the geometry of a head-and-shoulders pattern: three peaks, with the middle peak higher than the other two. Because consecutive extrema must alternate between maxima and minima for smooth functions,[40] the three-peaks pattern corresponds to a sequence of five local extrema: maximum, minimum, highest maximum, minimum, and maximum. The inverse head-and-shoulders is simply the mirror image of the head-and-shoulders, with the initial local extrema a minimum.

LMW2 develop similar definitions for broadening, rectangle, and triangle patterns, each with two possible versions depending on whether the initial extremum is a local maximum or minimum, yielding a total of ten patterns in all.

[40] After all, for two consecutive maxima to be local maxima, there must be a local minimum in between, and vice versa for two consecutive minima.

Given a sample of prices $\{P_1, \ldots, P_T\}$, kernel regressions for rolling subsamples or *windows*, and within each window, local extrema of the estimated function $\hat{m}(\tau)$ can be readily identified by finding times τ such that $\text{Sgn}(\hat{m}'(\tau)) = -\text{Sgn}(\hat{m}'(\tau+1))$, where $\hat{m}'$ denotes the derivative of $\hat{m}$ with respect to τ and $\text{Sgn}(\cdot)$ is the signum function. If the signs of $\hat{m}'(\tau)$ $\hat{m}'(\tau+1)$ are $+1$ and -1, respectively, then we have found a local maximum, and if they are -1 and $+1$, respectively, then we have found a local minimum. Once such a time τ has been identified, we proceed to identify a maximum or minimum in the original price series $\{P_t\}$ in the range $[t-1, t+1]$, and the extrema in the original price series are used to determine whether or not a pattern has occurred according to the definitions of the 10 technical patterns.[41] One useful consequence of this algorithm is that the series of extrema which it identifies contains alternating minima and maxima. That is, if the k^{th} extremum is a maximum, then it is always the case that the $(k+1)^{\text{th}}$ extremum is a minimum, and vice versa.

An important advantage of using this kernel regression approach to identify patterns is the fact that it ignores extrema that are "too local." For example, a simpler alternative is to identify local extrema from the raw price data directly, i.e., identify a price P_t as a local maximum if $P_{t-1} < P_t$ and $P_t > P_{t+1}$, and vice versa for a local minimum. The problem with this approach is that it identifies too many extrema, and also yields patterns that are not visually consistent with the kind of patterns that technical analysts find compelling.

Once all of the local extrema in a given window have been identified, the presence of the various technical patterns can be determined using definitions such as 1. This procedure is then repeated for the next window and continues until the end of the sample is reached.

7.2 Statistical Inference

Although there have been many tests of technical analysis over the years, most of these tests have focused on the profitability of technical trading rules.[42] While some of these studies do

[41] If $\hat{m}'(\tau) = 0$ for a given τ, which occurs if closing prices stay the same for several consecutive days, we need to check whether the price we have found is a local minimum or maximum. We look for the date s such that $s = \inf\ \{\ s > \tau\ :\ \hat{m}'(s) \neq 0\ \}$. We then apply the same method as discussed above, except here we compare $\text{Sgn}\big(\hat{m}'(\tau-1)\big)$ and $\text{Sgn}\big(\hat{m}'(s)\big)$. See LMW2 for further details.

[42] For example, Chang and Osler (1994) and Osler and Chang (1995) propose an algorithm for automatically detecting head-and-shoulders patterns in foreign exchange data by looking at properly defined local extrema. To assess the efficacy of a head-and-shoulders trading rule, they take a stand on a class of trading strategies and compute the profitability of these across a sample of exchange rates against the U.S. dollar. The null return distribution is computed by a bootstrap that samples returns randomly from the original data so as to induce temporal independence in the bootstrapped time series. By comparing the actual returns from

find that technical indicators can generate statistically significant trading profits, they beg the question of whether or not such profits are merely the equilibrium rents that accrue to investors willing to bear the risks associated with such strategies. Without specifying a fully articulated dynamic general equilibrium asset-pricing model, it is impossible to determine the economic source of trading profits.

Instead, LMW2 proposes a more fundamental test in their study, one that attempts to gauge the information content in the technical patterns of Section 7.1 by comparing the unconditional empirical distribution of returns with the corresponding conditional empirical distribution, conditioned on the occurrence of a technical pattern. If technical patterns are informative, conditioning on them should alter the empirical distribution of returns; if the information contained in such patterns has already been incorporated into returns, the conditional and unconditional distribution of returns should be close. Although this is a weaker test of the effectiveness of technical analysis—informativeness does not guarantee a profitable trading strategy—it is, nevertheless, a natural first step in a quantitative assessment of technical analysis.

To measure the distance between the two distributions, LMW2 use the Kolmogorov-Smirnov test.[43] which is designed to test the null hypothesis that two samples have the same distribution function, and is based on the empirical cumulative distribution functions of both samples. Under the null hypothesis, Smirnov (1939a, 1939b) has derived the limiting distribution of the statistic, and an approximate α-level test of the null hypothesis can be performed by computing the statistic and rejecting the null if it exceeds the upper 100α-th percentile for the null distribution (see Hollander and Wolfe, 1973, Table A.23, Csáki, 1984; and Press et al., 1986, Chapter 13.5; and Lo, Mamaysky, and Wang, 2000b).

trading strategies to the bootstrapped distribution, the authors find that for two of the six currencies in their sample (the yen and the Deutsche mark), trading strategies based on a head and shoulders pattern can lead to statistically significant profits. See, also, Neftci and Policano (1984), Pruitt and White (1988), and Brock, Lakonishok, and LeBaron (1992).

[43]LMW2 also compute chi-squared goodness-of-fit statistics but we omit them to conserve space.Note that the sampling distribution of the Kolmogorov-Smirnov statistic is derived under the assumption that returns are independently and identically distributed, which is not plausible for financial data. LMW2 attempt to address this problem by normalizing the returns of each security, i.e., by subtracting its mean and dividing by its standard deviation (see Section 7.3 below), but this does not eliminate the dependence or heterogeneity, and warrants further research.

7.3 Empirical Results

LMW2 apply the Kolmogorov-Smirnov test to the daily returns of individual NYSE/AMEX and Nasdaq stocks from 1962 to 1996 using data from the Center for Research in Securities Prices (CRSP). To ameliorate the effects of nonstationarities induced by changing market structure and institutions, they split the data into NYSE/AMEX stocks and Nasdaq stocks and into seven five-year periods: 1962 to 1966, 1967 to 1971, and so on. To obtain a broad cross-section of securities, in each five-year subperiod, they randomly select ten stocks from each of five market-capitalization quintiles (using mean market-capitalization over the subperiod), with the further restriction that at least 75 percent of the price observations must be non-missing during the subperiod.[44] This procedure yields a sample of 50 stocks for each subperiod across seven subperiods (note that they sample with replacement, hence there may be names in common across subperiods).

For each stock in each subperiod, LMW2 apply the procedure outlined in Section 7.1 to identify all occurrences of the 10 patterns they define mathematically according to the properties of the kernel estimator. For each pattern detected, they compute the one-day continuously compounded return three days after the pattern has completed. Therefore, for each stock, there are 10 sets of such conditional returns, each conditioned on one of the 10 patterns of Section 7.1.

For each stock, a sample of *unconditional* continuously compounded returns is constructed using non-overlapping intervals of length τ, and the empirical distribution function of these returns is compared with those of the conditional returns. To facilitate such comparisons, all returns are standardized—both conditional and unconditional—by subtracting means and dividing by standard deviations, hence:

$$X_{it} \quad = \quad \frac{R_{it} - \text{Mean}[R_{it}]}{\text{SD}[R_{it}]} \tag{49}$$

where the means and standard deviations are computed for each individual stock within each subperiod. Therefore, by construction, each normalized return series has zero mean and unit variance.

To increase the power of their goodness-of-fit tests, LMW2 combine the normalized returns of all 50 stocks within each subperiod; hence for each subperiod they have two

[44] If the first price observation of a stock is missing, they set it equal to the first non-missing price in the series. If the t-th price observation is missing, they set it equal to the first non-missing price prior to t.

samples—unconditional and conditional returns—from which two empirical distribution functions are computed and compared using the Kolmogorov-Smirnov test.

Finally, given the prominent role that volume plays in technical analysis, LMW2 also construct returns conditioned on increasing or decreasing volume. Specifically, for each stock in each subperiod, they compute its average share-turnover during the first and second halves of each subperiod, τ_1 and τ_2, respectively.[45] If $\tau_1 > 1.2 \times \tau_2$, they categorize this as a "decreasing volume" event; if $\tau_2 > 1.2 \times \tau_1$, they categorize this as an "increasing volume" event. If neither of these conditions holds, then neither event is considered to have occurred. Using these events, conditional returns can be constructed conditioned on two pieces of information: the occurrence of a technical pattern and the occurrence of increasing or decreasing volume. Therefore, the empirical distribution of unconditional returns can be compared with three conditional-return distributions: the distribution of returns conditioned on technical patterns, the distribution conditioned on technical patterns and increasing volume, and the distribution conditioned on technical patterns and decreasing volume.[46]

To develop some idea of the cross-sectional and time-series distributions of each of the 10 patterns, Figures 6 and 7 plot the occurrences of the patterns for the NYSE/AMEX and NASDAQ samples, respectively, where each symbol represents a pattern detected by the LMW2 algorithm. The vertical axis is divided into five quintiles, the horizontal axis is calendar time, and alternating symbols (diamonds and asterisks) represent distinct subperiods. These graphs show that there are many more patterns detected in the NYSE/AMEX sample than in the Nasdaq sample (Figure 6 is more densely populated than Figure 7). Also, for the NYSE/AMEX sample, the distribution of patterns is not clustered in time or among a subset of securities, but there seem to be more patterns in the first and last subperiods for the Nasdaq sample.

Table 9 contains the results of the Kolmogorov-Smirnov test of the equality of the conditional and unconditional return distributions for NYSE/AMEX and NASDAQ stocks from 1962 to 1996. Recall that conditional returns are defined as the one-day return starting three days following the conclusion of an occurrence of a pattern. The p-values are with respect to the asymptotic distribution of the Kolmogorov-Smirnov test statistics. The entries in the top

[45] For the Nasdaq stocks, τ_1 is the average turnover over the first third of the sample, and τ_2 is the average turnover over the final third of the sample.

[46] Of course, other conditioning variables can easily be incorporated into this procedure, though the "curse of dimensionality" imposes certain practical limits on the ability to estimate multivariate conditional distributions nonparametrically.

panel of Table 9 shows that for NYSE/AMEX stocks, five of the ten patterns—HS, BBOT, RTOP, RBOT, and DTOP—yield statistically significant test statistics, with p-values ranging from 0.000 for RBOT to 0.021 for DTOP patterns. However, for the other five patterns, the p-values range from 0.104 for IHS to 0.393 for DBOT, which implies an inability to distinguish between the conditional and unconditional distributions of normalized returns.

When LMW2 condition on declining volume trend as well as the occurrence of the patterns, the statistical significance declines for most patterns, but increases for TBOT. In contrast, conditioning on increasing volume trend yields an increase in the statistical significance of BTOP patterns. This difference may suggest an important role for volume trend in TBOT and BTOP patterns. The difference between the increasing and decreasing volume-trend conditional distributions is statistically insignificant for almost all the patterns (the sole exception is the TBOT pattern). This drop in statistical significance may be due to a lack of power of the Kolmogorov-Smirnov test given the relatively small sample sizes of these conditional returns.

The bottom panel of Table 9 reports corresponding results for the NASDAQ sample and in contrast to the NYSE/AMEX results, here all the patterns are statistically significant at the 5 percent level. This is especially significant because the NASDAQ sample exhibits far fewer patterns than the NYSE/AMEX sample (compare Figures 6 and 7), hence the Kolmogorov-Smirnov test is likely to have lower power in this case.

As with the NYSE/AMEX sample, volume trend seems to provide little incremental information for the NASDAQ sample except in one case: increasing volume and BTOP. And except for the TTOP pattern, the Kolmogorov-Smirnov test still cannot distinguish between the decreasing and increasing volume-trend conditional distributions, as the last pair of rows of Table 9 indicate.

When applied to many stocks over many time periods, LMW2's approach shows that certain technical patterns do provide incremental information, especially for NASDAQ stocks. While this does not necessarily imply that technical analysis can be used to generate "excess" trading profits, it does raise the possibility that technical analysis can add value to the investment process. Moreover, the evidence also suggests that volume trend provides incremental information in some cases. Although this hardly seems to be a controversial conclusion—that both prices and quantities contain incremental information for future returns—nevertheless, it comes from a rather surprising source that may contain other insights into the role of trading volume for economic activity.

8 Conclusion

Trading volume is an important aspect of the economic interactions of investors in financial markets. Both volume and prices are driven by underlying economic forces, and thus convey important information about the workings of the market. Although the literature on financial markets has focused almost exclusively on the behavior of returns based on simplifying assumptions about the market such as perfect competition, lack of frictions, and informational efficiency, we wish to develop a more realistic framework to understand the empirical characteristics of prices and volume.

In this paper, we hope to have made a contribution towards this goal. We first develop a dynamic equilibrium model for asset trading and pricing. The model qualitatively captures the most important motive for investors to participate in the market, namely, to achieve optimal allocations of wealth over time and across different risk profiles. We then explore the implications of the model for the behavior of volume and returns, particularly, the cross-sectional behavior of volume and the dynamic volume-return relations. We test these implications empirically and have found them to be generally consistent with the data. Fully realizing that our model merely provides a benchmark at best since it omits many important factors such as asymmetric information, market frictions, and other trading motives, we extend our model to include information asymmetry and transaction costs. We also go beyond the framework of our formal model and analyze the relation between price and volume in heuristic models of the market such as technical analysis. Our empirical analysis of these heuristic models finds some interesting connections between volume and price dynamics.

Our main approach in this paper has been to investigate study the behavior of price and volume using a structured equilibrium framework to motivate and direct our empirical analysis. While this has led to several interesting insights, there are many other directions to be explored. One important direction is to derive and test the implications of the model for identifying the specific risk factors that explain the cross-sectional variation in expected returns. We are currently pursuing this line of research in Lo and Wang (2000b). Another direction is to extend the framework to include other important factors, such as a richer set of trading motives, the actual trading mechanism, price impact, frictions, and other institutional aspects in our analysis. We hope to turn to these issues in the near future.

A Appendix

In this appendix, we give a proof for Theorem 1. We first solve the investors optimization problem under the stock prices in the form of (14) and then show that a and b can be chosen to clear the market.

Define $\theta_t \equiv (1; X_t; Y_t; Z_t)$ to be the state variable for an individual investor, say, i. For simplicity, we omit the superscript i for now. Then

$$d\theta_t = \alpha_\theta \theta_t dt + \sigma_\theta dB_t \tag{A.1}$$

where $\alpha_\theta \equiv \text{diag}\{0, \alpha_X, \alpha_Y, \alpha_Z\}$ is a diagonal matrix, $\sigma_\theta \equiv (0; \sigma_X; \sigma_Y; \sigma_Z)$. Given the price function in (14), the excess dollar return on the stocks can be written as

$$dQ_t = e_Q \theta_t dt + \sigma_Q dB_t \tag{A.2}$$

where $e_Q \equiv (ra, 0, 0, (r+\alpha_Z)b)$ and $\sigma_Q \equiv \sigma_D - b\sigma_Z$.

Let $J(W, \theta, t)$ denote the value function. We conjecture that it has the following form:

$$J(W, \theta, t) = -e^{-\rho t - r\gamma W - \frac{1}{2}\theta' v \theta} \tag{A.3}$$

The Bellman equation then takes the following form:

$$\begin{aligned}
0 &= \sup_{c,\, S} \quad -e^{-\rho t - \gamma c_t} + \mathrm{E}[dJ]/dt && \text{(A.4a)} \\
&= \sup_{c,\, S} \quad -e^{-\rho t - \gamma c_t} - J\Big[\rho + (r\gamma)(rW - c) - \tfrac{1}{2}\theta' m \theta \\
&\qquad\qquad +(r\gamma) S' e_Q \theta - \tfrac{1}{2}(r\gamma)^2 S' \sigma_{QQ} S - (r\gamma)^2 S' \sigma_{QN} \iota' \theta - (r\gamma) S' \sigma_{Q\theta} v \theta\Big] && \text{(A.4b)}
\end{aligned}$$

where

$$m \equiv (r\gamma)^2 \sigma_N^2 \iota\iota' + v\sigma_{\theta\theta} v + v\alpha_\theta + \alpha_\theta v + (r\gamma)(\iota \sigma_{N\theta} v + v \sigma_{\theta N} \iota'). \tag{A.5}$$

The first order condition for optimality gives

$$c = rW - \frac{1}{\gamma}\ln r + \frac{1}{2\gamma}\theta' v\theta \tag{A.6a}$$

$$S = \frac{1}{r\gamma}\left[e_Q - (r\gamma)\sigma_{QN}\iota' - \sigma_{Q\theta}v\right]\theta. \tag{A.6b}$$

Substituting into the Bellmen equation, we have

$$0 = v_{00} + \frac{r}{2}\theta' v\theta + \tfrac{1}{2}\theta' m\theta - \tfrac{1}{2}\theta\left(e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta}v\right)'\left(e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta}v\right)\theta \tag{A.7}$$

where $v_{00} \equiv r - \rho - r\ln r$. This then leads to the following equation for v:

$$0 = \bar{v} + \tfrac{1}{2}rv + \tfrac{1}{2}m - \tfrac{1}{2}\left(e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta}v\right)'\left(e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta}v\right) \tag{A.8}$$

where $\bar{v} \equiv v_{00}((1,0,0,0);(0,0,0,0);(0,0,0,0);(0,0,0,0))$.

We now consider market clearing. First,

$$S_t^i = \frac{1}{r\gamma}\left(\sigma_{QQ}\right)^{-1}\left[ra + (r+\alpha_Z)bZ_t - (r\gamma)\sigma_{QN}\iota'\theta_t - \sigma_{Q\theta}v\theta_t\right]. \tag{A.9}$$

Second, let $v \equiv (v_0; v_X; v_Y; v_Z)$. Since $\sigma_{Q\theta} = (0,0,0,\sigma_{QZ})$, we have $\sigma_{Q\theta}v = \sigma_{QZ}v_Z$. Thus,

$$\begin{aligned} S_t^i = \frac{1}{r\gamma}\left(\sigma_{QQ}\right)^{-1}\Big[&ra + (r+\alpha_Z)bZ_t - (r\gamma)\sigma_{QN}\left(X_t^i + Y_t^i + Z_t\right) \\ &-\sigma_{QZ}\left(v_{Z0} + v_{ZX}X_t^i + v_{ZY}Y_t^i + v_{ZZ}Z_t\right)\Big]. \end{aligned} \tag{A.10a}$$

Third, summing over the investors, we have

$$\iota = \sum_{i=1}^{I} S_t^i = \frac{1}{r\bar{\gamma}}\left(\sigma_{QQ}\right)^{-1}\left[ra + (r+\alpha_Z)bZ_t - (r\gamma)\sigma_{QN}Z_t - \sigma_{QZ}\left(v_{Z0} + v_{ZZ}Z_t\right)\right] \tag{A.11}$$

where $\bar{\gamma} = \gamma/I$. Thus, we have

$$a = \bar{\gamma}\left(\sigma_{QQ}\right)^{-1}\iota + (v_{Z0}/r)\sigma_{QZ} \tag{A.12a}$$

$$b = \frac{1}{r+\alpha_Z}\left(v_{ZZ}\sigma_{QZ} + r\gamma\sigma_{QN}\right). \tag{A.12b}$$

Substituting (A.12) (the equilibrium prices) into the expression for investors' asset demand gives us their equilibrium holdings:

$$S_t^i = (1/I)\iota - \left(X_t^i + Y_t^i\right)(\sigma_{QQ})^{-1}\sigma_{QN} - \frac{1}{r\gamma}\left(v_{ZX}X_t^i + v_{ZY}Y_t^i\right)(\sigma_{QQ})^{-1}\sigma_{QZ} \qquad \text{(A.13)}$$

which is the four-fund separation result in Theorem 1.

The remaining part of the proof is the existence of a solution to the system of algebraic equations defined by (A.8) and (A.12). The proof of the existence of a solution in the case of a single stock can be found in Huang and Wang (1997), which can be extended to the case of multiple stocks. In particular, equation (A.8) reduces to a Riccatti equation, which has a closed form solution (under certain parameter restrictions, see Huang and Wang (1997) for more details). The existence of a solution to (A.12) is then straightforward to establish.

References

Ajinkya B.B., and P.C. Jain, 1989, "The Behavior of Daily Stock Market Trading Volume", *Journal of Accounting and Economics* 11, 331–359.

Allen, F. and Karjalainen, R., 1999, "Using Genetic Algorithms to Find Technical Trading Rules", *Journal of Financial Economics* 51, 245–271.

Amihud, Y. and H. Mendelson, 1986a, "Asset Pricing and the Bid-Ask Spread", *Journal of Financial Economics* 17, 223–249.

Amihud, Y. and H. Mendelson, 1986b, "Liquidity And Stock Returns", *Financial Analysts Journal* 42, 43–48.

Andersen, T., 1996, "Return Volatility and Trading Volume: An Information Flow Interpretation", *Journal of Finance* 51, 169–204.

Antoniewicz, R.L., 1993, Relative Volume and Subsequent Stock Price Movements, working paper, Board of Governors of the Federal Reserve System.

Atkins, A. and E. Dyl, 1997, "Market Structure and Reported Trading Volume: NASDAQ versus the NYSE", *Journal of Financial Research* 20, 291-304.

Banz, R., 1981, "The Relation between Return and Market Value of Common Stocks", *Journal of Financial Economics* 9, 3–18.

Bamber, L., 1986, "The Information Content of Annual Earnings Releases: A Trading Volume Approach", *Journal of Accounting Research* 24, 40–56.

Black, F., 1976, "Studies of Stock Price Volatility Changes", in *Proceedings of the 1976 Meetings of the Business and Economic Statistics Section, American Statistical Association*, 177–181.

Black, F., M. Jensen, and M. Scholes, 1972, "The Capital Asset Pricing Model: Some Empirical Tests", *Studies in the Theory of Capital Markets* (M. Jensen ed.), Praeger Publishers.

Blume, L., D. Easley, and M. O'Hara, "Market Statistics and Technical Analysis: The Role of Volume", *Journal of Finance* 49, 153–181.

Brock, W., Lakonishok, J. and B. LeBaron, 1992, "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns", *Journal of Finance* 47, 1731–1764.

Brown, D. and R. Jennings, 1989, "On Technical Analysis", *Review of Financial Studies* 2, 527–551.

Brown, K., Van Harlow, W. and S. Tinic, 1993, "The Risk and Required Return of Common Stock Following Major Price Innovations", *Journal of Financial and Quantitative Analysis* 28, 101–116.

Campbell, J., Grossman S. and J. Wang, 1993, "Trading Volume and Serial Correlation in Stock Returns", *Quarterly Journal of Economics* 108, 905–939.

Campbell, J., A. Lo and C. MacKinlay, 1996, *The Econometrics of Financial Markets*, Princeton University Press.

Campbell, J. and A. Kyle, 1993, "Smart Money, Noise Trading, and Stock Price Behavior", *Review of Economic Studies* 60, 1–34.

Chamberlain, G., 1983, "Funds, Factors, and Diversification in Arbitrage Pricing Models", *Econometrica* 51, 1305–1323.

Chamberlain, G. and M. Rothschild, 1983, "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets", *Econometrica* 51, 1281–1304.

Chan, L., Jegadeesh, N. and J. Lakonishok, 1996, "Momentum Strategies", *Journal of Finance* 51, 1681–1713.

Chan, L. and J. Lakonishok, 1995, "The Behavior of Stock Prices Around Institutional Trades", *Journal of Finance* 50, 1147–74.

Chang, K. and C. Osler, 1994, "Evaluating Chart-Based Technical Analysis: The Head-and-Shoulders Pattern in Foreign Exchange Markets", working paper, Federal Reserve Bank of New York.

Constantinides, G.M., 1986, "Capital Market Equilibrium with Transaction Costs," *Journal of Political Economy* , Vol.94 (4), 842-862.

Cready, W.M., and Ramanan R., 1991, The Power of Tests Employing Log-Transformed Volume in Detecting Abnormal Trading, *Journal of Accounting and Economics* 14, 203-214.

Csáki, E., 1984, "Empirical Distribution Function", in P. Krishnaiah and P. Sen, eds., *Handbook of Statistics*, Volume 4. Amsterdam, The Netherlands: Elsevier Science Publishers.

Dhillon, U. and H. Johnson, 1991, "Changes in the Standard and Poor's 500 List", *Journal of Business* 64, 75–85.

Fama, E. and K. French, 1992, "The Cross-Section of Expected Stock Returns", *Journal of Finance* 47, 427–465.

Gallant, R., Rossi, P. and G. Tauchen, 1992, "Stock Prices and Volume", *Review of Financial Studies* 5, 199–242.

Goetzmann, W. and M. Garry, 1986, "Does Delisting From the S&P 500 Affect Stock Prices?", *Financial Analysts Journal* 42, 64–69.

Grundy, B. and S. Martin, 1998, "Understanding the Nature of the Risks and the Source of the Rewards to Momentum Investing", unpublished working paper, Wharton School, University of Pennsylvania.

Hamilton, W., 1922, *The Stock Market Barometer.* New York: John Wiley & Sons.

Hamilton, J., 1994, *Times Series Analysis.* Princeton, NJ: Princeton University Press.

Harris, L. and E. Gurel, 1986, "Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures," *Journal of Finance* 46, 815–829.

Heaton, John and Deborah J. Lucas, 1996, "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing," *Journal of Political Economy* , Vol.104 (3), 443-487.

He, H., and J. Wang, 1995, "Differential Information and Dynamic Behavior of Stock Trading Volume", *Review of Financial Studies* 8, 919–972.

Hollander, Myles and Douglas Wolfe, 1973, *Nonparametric Statistical Methods* (John Wiley & Sons, New York, NY).

Hu, S., 1997, "Trading Turnover and Expected Stock Returns: Does It Matter and Why?", working paper, National Taiwan University.

Huang, Chi-fu and Henri Pages, 1990, "Optimal consumption and portfolio policies with an infinite horizon: Existence and convergence," working paper, MIT.

Huang, Jennifer, and Jiang Wang, 1997, "Market Structure, Security Prices, and Informational Efficiency," *Macroeconomic Dynamics*, 1, 169-205.

Huang, Ming, 1998, "Liquidity Shocks and Equilibrium Liquidity Premia," unpublished working paper, Graduate School of Business, Stanford University.

Jacques, W., 1988, "The S&P 500 Membership Anomaly, or Would You Join This Club?", *Financial Analysts Journal* 44, 73–75.

Jain, P., 1987, "The Effect on Stock Price of Inclusion in or Exclusion from the S&P 500", *Financial Analysts Journal* 43, 58–65.

Jain, P. and G. Joh, 1988, "The Dependence between Hourly Prices and Trading Volume", *Journal of Financial and Quantitative Analysis* 23, 269–282.

Jegadeesh, N. and S. Titman, 1993, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency", *Journal of Finance* 48, 65–91.

Karpoff, J., 1987, "The Relation between Price Changes and Trading Volume: A Survey", *Journal of Financial and Quantitative Analysis* 22, 109–126.

Karpoff, J. and R. Walkling, 1988, "Short-Term Trading Around Ex-Dividend Days: Additional Evidence", *Journal of Financial Economics* 21, 291–298.

Karpoff, J. and R. Walkling, 1990, "Dividend Capture in NASDAQ Stocks", *Journal of Financial Economics* 28, 39–65.

Kwiatkowski, D., Phillips, P., Schmidt, P. and Y. Shin, 1992, "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?", *Journal of Econometrics* 54, 159–178.

Lakonishok, J. and S. Smidt, 1986, "Volume for Winners and Losers: Taxation and Other Motives for Stock Trading", *Journal of Finance* 41, 951–974.

Lakonishok, J. and T. Vermaelen, 1986, "Tax-Induced Trading Around Ex-Dividend Days", *Journal of Financial Economics* 16, 287–319.

Lamoureux, C. and J. Wansley, 1987, "Market Effects of Changes in the Standard & Poor's 500 Index", *Financial Review Journal* 22, 53–69.

LeBaron, B., 1992, "Persistence of the Dow Jones Index on Rising Volume", working paper, University of Wisconsin.

Lim, T., Lo, A., Wang, J. and P. Adamek, 1998, "Trading Volume and the MiniCRSP Database: An Introduction and User's Guide", MIT Laboratory for Financial Engineering Working Paper No. LFE–1038–98.

Llorente, G., R. Michaely, G. Saar and J. Wang, 2000, "Dynamic Volume-Return Relations for Individual Stocks", working paper, MIT.

Lo, A. and C. MacKinlay, 1988, "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test", *Review of Financial Studies* 1, 41–66.

Lo, A. and C. MacKinlay, 1997, "Maximizing Predictability in the Stock and Bond Markets", *Macroeconomic Dynamics* 1(1997), 102–134.

Lo, A. and C. MacKinlay, 1999, *A Non-Random Walk Down Wall Street*. Princeton, NJ: Princeton University Press.

Lo, A., H. Mamaysky and J. Wang, 2000a, "Asset Prices and Trading Volume under Fixed Transaction Costs", working paper, MIT.

Lo, A., H. Mamaysky and J. Wang, 2000b, "Foundations of Technical Analysis: Computational Algorithms, Statistical Inference and Empirical Implementation", *Journal of Finance* 55, 1705–1765.

Lo, A. and J. Wang, 2000a, Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory, *Review of Financial Studies* 13, 257–300.

Lo, A. and J. Wang, 2000b, "Trading Volume: Implications of an Intertemporal Capital Asset-Pricing Model", work in progress, MIT.

Lynch-Koski, J., 1996, "A Microstructure Analysis of Ex-Dividend Stock Price Behavior Before and After the 1984 and 1986 Tax Reform Acts", *Journal of Business* 69, 313–338.

Marsh, T. and R. Merton, 1987, "Dividend Behavior For The Aggregate Stock Market", *Journal of Business* 60, 1–40.

Merton, R., 1971, "Optimal Consumption and Portfolio Rules in a Continuous-Time Model", *Journal of Economic Theory* 3, 373–413.

Merton, R., 1973, "An Intertemporal Capital Asset Pricing Model", *Econometrica* 41, 867–887.

Merton, R., 1987, "A Simple Model of Capital Market Equilibrium with Incomplete Information", *Journal of Finance* 42, 483–510.

Michaely, R., 1991, "Ex-Dividend Day Stock Price Behavior: The Case of the 1986 Tax Reform Act", *Journal of Finance* 46, 845–860.

Michaely, R. and M. Murgia, 1995, "The Effect of Tax Heterogeneity on Prices and Volume Around the Ex-Dividend Day: Evidence from the Milan Stock Exchange", *Review of Financial Studies* 8, 369–399.

Michaely, R. and J. Vila, 1995, "Investors' Heterogeneity, Prices and Volume Around the Ex-Dividend Day", *Journal of Financial and Quantitative Analysis* 30, 171–198.

Michaely, R. and J. Vila, 1996, "Trading Volume with Private Valuation: Evidence from the Ex-Dividend Day", *Review of Financial Studies* 9, 471–509.

Michaely, R., J.-L. Vila and J. Wang, "A Model of Trading Volume with Tax-Induced Heterogeneous Valuation and Transaction Costs" *Journal of Financial Intermediation* 5, 340-371, 1996.

Morse, D., 1980, "Asymmetric Information in Securities Markets and Trading Volume", *Journal of Financial and Quantitative Analysis* 15, 1129–1148.

Muirhead, R., 1982, *Aspects of Multivariate Statistical Theory*. New York: John Wiley and Sons.

Neely, C., Weller, P. and R. Dittmar, 1997, "Is Technical Analysis in the Foreign Exchange Market Profitable? A Genetic Programming Approach", *Journal of Financial and Quantitative Analysis* 32, 405–426.

Neely, C. and P. Weller, 1998, "Technical Trading Rules in the European Monetary System", working paper, Federal Bank of St. Louis.

Neftci, S., 1991, "Naive Trading Rules in Financial Markets and Wiener-Kolmogorov Prediction Theory: A Study of Technical Analysis", *Journal of Business* 64, 549–571.

Neftci, S. and A. Policano, 1984, "Can Chartists Outperform the Market? Market Efficiency Tests for 'Technical Analyst' , *Journal of Future Markets* 4, 465–478.

Ohlson, J. and B. Rosenberg, 1976, "The Stationary Distribution of Returns and Portfolio Separation in Capital Markets: A Fundamental Contradiction," *Journal of Financial and Quantitative Analysis* .

Ohlson Osler, C. and K. Chang, 1995, "Head and Shoulders: Not Just a Flaky Pattern", Staff Report No. 4, Federal Reserve Bank of New York.

Press, W., Flannery, B., Teukolsky, S. and W. Vetterling, 1986, *Numerical Recipes: The Art of Scientific Computing*. Cambridge, UK: Cambridge University Press.

Pruitt, S. and J. Wei, 1989, "Institutional Ownership and Changes in the S&P 500", *Journal of Finance* 44, 509–513.

Pruitt, S. and R. White, 1988, "The CRISMA Trading System: Who Says Technical Analysis Can't Beat the Market?", *Journal of Portfolio Management* 14, 55–58.

Reinganum, M., 1992, "A Revival of the Small-Firm Effect", *Journal of Portfolio Management* 18, 55–62.

Richardson, G., Sefcik, S. and R. Thompson, 1986, "A Test of Dividend Irrelevance Using Volume Reaction to a Change in Dividend Policy", *Journal of Financial Economics* 17, 313–333.

Roll, R., 1984, "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market", *Journal of Finance* 39, 1127–1140.

Ross, S.A., 1989, "Discussion: Intertemporal Asset Pricing," *Theory of Valuation* (S. Bhattacharya and G. Constantinides, eds.) Rowman & Littlefield Publishers, Inc., 85-96.

Rouwenhorst, G., 1998, "International Momentum Strategies", *Journal of Finance* 53, 267–284.

Schroeder, M., 1998, "Optimal Portfolio Selection with Fixed Transaction Costs," working paper, Northwestern University.

Shleifer, A., 1986, "Do Demand Curves for Stocks Slope Down?", *Journal of Finance* 41, 579–590.

Smirnov, N., 1939a, Sur les écarts de la courbe de distribution empirique, *Rec. Math.* (*Mat. Sborn.*) 6, 3–26.

Smirnov, N., 1939b, On the estimation of the discrepancy between empirical curves of distribution for two independent samples, *Bulletin. Math. Univ. Moscow* 2, 3–14.

Stickel, S., 1991, "The Ex-Dividend Day Behavior of Nonconvertible Preferred Stock Returns and Trading Volume", *Journal of Financial and Quantitative Analysis* 26, 45–61.

Stickel, S. and R. Verrecchia, 1994, "Evidence that Volume Sustains Price Changes", *Financial Analysts Journal* (November-December), 57-67.

Tabell, A. and E. Tabell, 1964, "The Case for Technical Analysis", *Financial Analyst Journal* 20, 67–76.

Tkac, P., 1996, "A Trading Volume Benchmark: Theory and Evidence", working paper, Department of Finance and Business Economics, University of Notre Dame.

Treynor, J. and R. Ferguson, 1985, "In Defense of Technical Analysis", *Journal of Finance* 40, 757–773.

Vayanos, D., 1998, "Transaction Costs and Asset Prices: A Dynamic Equilibrium Model," *Review of Financial Studies* , Vol.11(1), 1-58.

Wang, J., 1994, "A Model of Competitive Stock Trading Volume", *Journal of Political Economy* 102, 127–168.

Woolridge, J and C. Ghosh, 1986, "Institutional Trading and Security Prices: The Case of Changes in the Composition of the S&P 500 Market Index", *Journal of Financial Research* 9, 13–24.

Wu, G.J. and C.S. Zhou, 2001, "A Model of Active Portfolio Management," working paper, University of Michigan.