

*Raja Velu<sup>1</sup>, Maxence Hardy<sup>2</sup> and Daniel Nehren<sup>3</sup>*

---

## ***Algorithmic Trading and Quantitative Strategies***

***(Version: November 18, 2017)***

<sup>1</sup>Department of Finance  
Whitman School of Management  
Syracuse University

<sup>2</sup>Linear Quantitative Research  
Global Equities  
J.P. Morgan - Hong Kong

<sup>3</sup>Author at Large



---

---

## *Contents*

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	From Trading Floors to Electronic Trading . . . . .	1
1.2	Liquidity and Market Making . . . . .	3
1.3	Algorithmic Trading: Some Basic Elements . . . . .	7
1.4	Low Frequency vs. High Frequency Trading Strategies . . . . .	12
1.5	Risk and Regulation in Quantitative Trading . . . . .	12
1.6	Trading in Other Asset Classes . . . . .	12
<b>I</b>	<b>Foundations: Basic Models and Empirics</b>	<b>17</b>
<b>2</b>	<b>Univariate Time Series Models</b>	<b>21</b>
2.1	Trades and Quotes Data and their Aggregation: From Point Processes to Discrete Time Series . . . . .	22
2.2	Trading Decisions as Short-Term Forecast Decisions . . . . .	24
2.3	Time Series Models for Aggregated Data: Modeling the Mean . . . . .	25
2.3.1	Stochastic Processes: Some Properties . . . . .	25
2.3.2	Some Basic Models and Descriptive Tools . . . . .	30
2.3.3	Specification and Estimation of ARIMA Models . . . . .	40
2.3.4	Testing for Nonstationary (Unit Root) in ARIMA Models . . . . .	50
2.3.5	Forecasting for ARIMA Processes . . . . .	52
2.4	Stylized Models for Asset Returns . . . . .	55
2.5	Time Series Models for Aggregated Data: Modeling the Variance . . . . .	59
2.6	Stylized Models for Variance of Asset Returns . . . . .	64
2.7	Other Topics: Bet Size etc. . . . .	65
2.8	Exercises (starred items have R-codes available) . . . . .	65
<b>3</b>	<b>Multivariate Time Series Models</b>	<b>71</b>
3.1	Multiple Time Series Modeling and Co-integration . . . . .	71
3.1.1	Introduction . . . . .	71
3.1.2	Co-integration, Co-movement and commonality in Multiple Time Series . . . . .	73
3.1.3	Multivariate Regression . . . . .	79
<b>4</b>	<b>Advanced Topics</b>	<b>85</b>
4.0.1	State-Space Modeling . . . . .	85
4.1	A Model for Volume-Volatility Relationship . . . . .	88
4.2	Regime Switching and Change-Point Models . . . . .	90

4.3	Models for Point Processes . . . . .	94
4.3.1	Stylized Models for High Frequency Financial Data . . . . .	95
4.3.2	Models for Multiple Assets: High-Frequency Context . . . . .	100
4.4	Time Aggregation and Loss of Information . . . . .	101
4.5	Realized Volatility and Econometric Models . . . . .	101
4.6	Analytics from Machine Learning Literature . . . . .	102
<b>II</b>	<b>Trading Algorithms</b>	<b>105</b>
<b>5</b>	<b>Statistical Trading Strategies and Back-Testing</b>	<b>107</b>
5.1	Introduction to Trading Strategies: Origin and History . . . . .	107
5.2	Trading Rules for Time Aggregated Data . . . . .	108
5.2.1	Filter Rules . . . . .	109
5.2.2	Moving Average Variants and Oscillators . . . . .	111
5.3	Patterns Discovery via Non-Parametric Smoothing Methods . . . . .	113
5.4	A Decomposition Algorithm . . . . .	115
5.5	Fair Value Models . . . . .	117
5.6	Back-Testing and Data Snooping: In-Sample and Out-of-Sample Performance Evaluation . . . . .	118
5.7	Pairs Trading . . . . .	120
5.7.1	Distance-Based Algorithms . . . . .	121
5.7.2	Co-Integration . . . . .	123
5.7.3	Some General Comments . . . . .	125
5.7.4	Practical Considerations . . . . .	127
5.8	Cross-Sectional Momentum Strategies . . . . .	129
5.9	Multiple Indicators and Boosting Methods . . . . .	133
5.10	Evaluation of Strategies: Various Measures . . . . .	135
5.11	Extraneous Signals: Trading Volume, Volatility, etc. . . . .	136
5.11.1	Filter Rules based on return and volume . . . . .	139
5.11.2	An illustrative example . . . . .	140
5.12	Market Segmentation and Inconsistent Pricing across Markets . . . . .	147
5.13	Supplements and Problems . . . . .	147
<b>6</b>	<b>News Analytics: From Market Attention and Sentiment to Trading</b>	<b>157</b>
6.1	Introduction to News Analytics: Behavioral Finance and Investor Cognitive Biases . . . . .	157
6.2	Market Micro Structure and Liquidity . . . . .	162
6.3	Automated News Analysis and Market Sentiment . . . . .	162
6.4	News Analytics and its Applications to Trading . . . . .	164
6.5	Discussion / Future of Social Media News . . . . .	173
<b>7</b>	<b>Active Portfolio Management and Dynamic Investment Strategies</b>	<b>175</b>
7.1	Introduction to Modern Portfolio Theory . . . . .	175
7.1.1	Mean-Variance Portfolio Theory . . . . .	176
7.1.2	Multifactor Models . . . . .	179
7.1.3	Tests Related to CAPM and APT . . . . .	180

7.1.4	An Illustrative Example . . . . .	182
7.1.5	Summary of Models and Implications for Investing . . . . .	184
7.2	Statistical Underpinnings . . . . .	187
7.2.1	Portfolio Allocation Using Regularization . . . . .	191
7.2.2	Portfolio Strategies: Some General Findings . . . . .	193
7.3	Dynamic Portfolio Selection . . . . .	195
7.4	Portfolio Tracking and Rebalancing . . . . .	197
7.5	Transaction Costs, Shorting and Liquidity Constraints . . . . .	199
7.6	Alphas, Betas, Information Ratios and Realized Returns . . . . .	202
7.7	Supplements and Problems . . . . .	202
<b>III</b>	<b>Execution Algorithms</b>	<b>209</b>
<b>8</b>	<b>Electronic Exchanges and Dynamics of Limit Order Books</b>	<b>211</b>
8.1	Limit Order Book (LOB): Basic Description and Characteristics . . . . .	214
8.2	Data from LOB and Relevant Questions . . . . .	216
8.3	Models for LOB dynamics, based on Level I and Level II data . . . . .	219
8.3.1	Application of a Hawkes process . . . . .	228
8.3.2	One-Dimensional Models . . . . .	233
8.3.3	An extension . . . . .	237
8.3.4	Two-Dimensional models . . . . .	237
8.3.5	Limitations and findings . . . . .	240
8.4	Models for Level II and Level III data . . . . .	241
8.5	Estimation of Hidden Orders . . . . .	241
8.6	Volume Imbalance and Other Indicators . . . . .	241
8.7	Hidden Markov Models (HMM) for dynamic market microstructure: Estimation of Parameter and States . . . . .	241
8.8	Appendix: Adaptive Particle Filters in HMM . . . . .	241
8.9	Supplements and Problems . . . . .	241
<b>9</b>	<b>Algorithmic Trading In Electronic Platforms</b>	<b>243</b>
9.1	Execution Algorithms: History and Review . . . . .	244
9.1.1	Bertsimas and Lo Model . . . . .	245
9.1.2	Almgren and Chriss Model . . . . .	246
9.1.3	Obizhaeva and Wang Model . . . . .	249
9.1.4	Easley, De Prado and O'Hara Model . . . . .	252
9.2	Multiple Exchanges: Smart Order Routing . . . . .	253
9.3	Market Impact Models: Permanent and Temporary Impacts . . . . .	258
9.4	Microstructure Noise Models for Multiple Stocks: Issues of Synchronizations, etc. . . . .	275
9.5	Simultaneous Execution Algorithms for Multiple Instruments . . . . .	275
9.6	A Stochastic Adaptive Control Approach to Optimal Execution . . . . .	275
9.7	Role of Market Makers in Limit Order Markets . . . . .	275
9.8	Open Issues . . . . .	275
9.9	Supplements and Problems . . . . .	275

<b>10 Algorithmic Trading: Some Potential Research Issues</b>	<b>277</b>
10.1 Extending the Algorithms to Derivative markets . . . . .	277
10.2 Developing a Trading Platforms for Simultaneous Execution of Assets . . . . .	277
10.3 Future of Algorithmic Trading . . . . .	277
<b>IV Informatics and Regulatory Issues</b>	<b>279</b>
<b>11 Informatics and Computation for Algorithmic Trading</b>	<b>281</b>
11.1 Model for electronic Trading Platforms . . . . .	281
11.1.1 Trading Rules . . . . .	281
11.1.2 Access to Order Book; Information Flow and the Infrastructure . . . . .	281
11.1.3 High Frequency vs. Low Frequency Set Up . . . . .	281
11.2 Current Practices and Alternative Trading Systems . . . . .	281
11.2.1 Computer Networks and Broadcasting . . . . .	281
11.2.2 Data Center and Co-location Service . . . . .	281
11.2.3 Dark Pools . . . . .	281
11.3 Developing Modern Trading Platforms . . . . .	281
<b>12 Risk Management and Regulatory Issues</b>	<b>283</b>
12.1 Operational Risk and Information Technology Risk . . . . .	283
12.2 Risk Management: Pre-trade and Post-trade . . . . .	283
12.3 Circuit Breakers . . . . .	283
12.4 Order-to-trade Ratio and Quote Stuffing . . . . .	283
12.5 Regulation of Parallel Processes . . . . .	283
12.6 Information Access and Fairness Issues . . . . .	283
<b>Bibliography</b>	<b>285</b>

# 1

---

## *Introduction*

---

### CONTENTS

1.1	From Trading Floors to Electronic Trading .....	1
1.2	Liquidity and Market Making .....	2
1.3	Algorithmic Trading: Some Basic Elements .....	7
1.4	Low Frequency vs. High Frequency Trading Strategies .....	12
1.5	Risk and Regulation in Quantitative Trading .....	12
1.6	Trading in Other Asset Classes .....	12

---

### **1.1 From Trading Floors to Electronic Trading**

In less than a decade, electronification of financial markets has profoundly changed the way equities are traded. In this short timeframe, equity trading evolved from primarily manual trading on a central exchange to primarily automated trading on multiple fragmented venues, driven by both advances in technology and by changes in the regulatory framework. This unprecedented transformation in market structure has given rise to new types of market participants and novel trading strategies that have become widespread, so the terms “High Frequency Trading” (HFT) and “Algorithmic Trading” have become part of household usage.

Secondary markets for trading equities serve two complementary purposes; price discovery and liquidity aggregation (O’Hara (2003) [216]). Buyers and sellers meet and agree on a price to exchange a security. When that transaction is made public it, in turn, informs other potential buyers and sellers of the most recent valuation of the security. Additionally, would-be buyers or sellers need to find matching opposite side interest in order to transact. In absence of enough selling interest, a large buyer would significantly displace the supply and demand equilibrium price and have to pay more to transact.

As such, when exchange price matching was still manual, a single, physical, centralised location (the Exchange) benefited from significant network externalities. The exchange trading floor was the physical location where buyers and sellers (or their intermediaries) met to negotiate the transactions. In addition to providing an aggregation of potential counterparts, the open outcry nature of the transaction rewarded physical presence on the floor and that was the only way to gather the most up to date information on securities prices.

## 2 Algorithmic Trading and Quantitative Strategies (Version: November 18, 2017)

Starting in the seventies in US, the implementation of electronic dissemination of quotes by NASDAQ removed the need for a physical presence at the exchange. The participants were granted the same level of real time access to price discovery irrespective of their locations. Once exchanges became fully automated allowing buy and sell orders to match without human intervention, traditional intermediaries gave place to automated market makers, leveraging technology to produce and display optimal bids and offers. These participants compete to provide liquidity to liquidity demanders by displaying quotes (bid to purchase or offer to sell) on the exchange for a certain number of shares at a given price. The spread between bid and ask compensates them for the inventory risk associated with providing liquidity.

The promulgation of *Regulation National Market System* (Reg NMS) by the *Securities and Exchange Commission* (SEC) in 2005 further accelerated the transformation of market structure in the US<sup>1</sup>. Through its two main rules, the Reg NMS was to promote best price execution for investors by encouraging competition among individual exchanges and individual orders. The Access Rule promoted non-discriminatory access to quotes displayed by Self-Regulatory Organization (SRO) trading centers as well as established a cap, limiting the fees that a trading center can charge for accessing a displayed quote. The Order Protection Rule required trading centers to obtain the best possible price for investors when such price was represented by an automated immediately accessible quote. In essence, aside from a series of exemptions, the Order Protection Rule mandated market participants to transact first at a price better or equal to the National Best Bid Offer (NBBO).

The latter rule, in particular, had a significant effect on the US market structure. Market makers, who had traditionally been competing for quotes on a given exchange, quickly became multi-venue market makers by leveraging technology to process real time quotes from all Reg NMS protected venues to determine where and at what price to provide their own quotations. Automated market makers therefore played a critical role in linking various trading venues and allowing for fragmentation to increase. The NYSE market share in NYSE-listed stocks was still above 80% in 2005 but that number quickly declined as Reg NMS facilitated competition among venues to reduce NYSE's share to 25% by 2010.

As fragmentation increased, speed and technology became major differentiating factors of success for market makers and other participants. The ability to process market data, analyze it, and submit one's orders faster than competing participants meant capturing fleeting opportunities or successfully avoiding adverse selection. To gain or maintain an edge, market participants heavily invested in technology, faster networks, and installed their servers in the same data centers as the venues they transacted on (co-location).

---

<sup>1</sup>equivalent regulation had a similar impact in Europe - see MiFID

## 1.2 Liquidity and Market Making

- a) **A definition of liquidity:** Financial markets are commonly described as carrying the function of efficiently directing the flow of savings and investments to the real economy to allow the production of goods and services. An important factor contributing to well-developed financial markets is their liquidity which enables investors to diversify their asset allocation and facilitate the transfers of securities at a reasonable transaction cost.

Properly describing “liquidity” often proves to be elusive as there is no commonly agreed upon definition of it. Black (1971) [28] proposes a relatively intuitive description of a liquid market: *“The market for a stock is liquid if the following conditions hold:*

- *There are always bid and ask prices for the investor who wants to buy or sell small amounts of stock immediately.*
- *The difference between the bid and ask prices (the spread) is always small.*
- *An investor who is buying or selling a large amount of stock, in the absence of special information, can expect to do so over a long period of time at a price not very different, on average, from the current market price.*
- *An investor can buy or sell a large block of stock immediately, but at a premium or discount that depends on the size of the block. The larger the block, the larger the premium or discount.”*

In other words, Black defines a liquid market as a continuous market having the characteristics of relatively tight spread, with enough depth on each side of the limit order book to accommodate instantaneous trading of small orders, and resilient enough to allow large orders to be traded slowly without significant impact on the price of the asset. This general definition remains particularly well suited to today’s modern electronic markets and can be used to by practitioners to assess the difference in liquidity between various markets when choosing where to deploy a strategy.

- b) **Different types of market participants: liquidity demanders (generally, AlgorithmicTraders, AT) and liquidity providers (generally, MarketMakers—MM):** Liquid financial markets carry their primary economic function by facilitating savings and investment flows as well as allowing investors to exchange securities in the secondary market. Economic models of financial markets attempt to classify market participants into different categories. These can be broadly delineated as:

**Informed Traders**, making trading decisions based on superior information that is not yet fully reflected in the asset price. That knowledge can be derived from either fundamental analysis or information not directly available nor known to other market participants.

**News Traders**, making trading decisions based on market news or announcements and trying to make profits by anticipating the market's response to a particular news or event. The electronification of news dissemination offers new opportunities for developing quantitative trading strategies by leveraging text-mining tools such as natural language processing to interpret, and trade on, machine readable news before it is fully reflected in the market prices. Chapter 4 will detail some of the advances in the field.

**Noise Traders** (as introduced by Kyle (1985) [182]), making trading decisions without particular information and at random times mainly for liquidity. They can be seen as adding liquidity to the market through additional volume transacted, but only have a temporary effect on price formation. Their presence in the market allows informed traders to not be immediately detected when they start transacting as market makers cannot normally distinguish the flow origin between these two types of participants. In a market without noise traders, being fully efficient, at equilibrium each trade would be revealing information that would instantly be incorporated in prices, hereby removing any profit opportunities.

**Market Makers**, providing liquidity to the market with the intent of collecting profits originating from trading frictions in the market place (bid ask spread). Risk-neutral market makers are exposed to adverse selection risk arising from the presence of informed traders in the marketplace, and therefore establish their trading decisions mostly based on their current inventory. As such, they are often considered in the literature to drive the determination of efficient prices by acting as the rational intermediaries.

Generalizing these concepts, market participants and trading strategies can be separated between liquidity providing and liquidity demanding. The former being essentially the domain of market makers whose level of activity, proxied by market depth, is proportional to the amount of noise trading and inversely proportional to the amount of informed trading (Kyle (1985) [182]). The latter being the domain of a variety of algorithmic trading users (mutual funds, hedge funds, asset managers,...).

Given the key role played by market makers in the liquidity of electronic markets, we want to elaborate a bit more on their activities.

- c) **The objectives of the modern Market Maker:** In present days, market making can broadly be separated into two main categories based on the trading characteristics. The first one is the provision of large liquidity—known as blocks—to institutional investors, and has traditionally been in the realm of sell side brokers acting as intermediaries and maintaining significant inventories. Such market makers usually transact through a non-continuous, negotiated, process based on their current inventory as well as assessment of the risk involved in liquidation of the position in the future. Larger or more volatile positions generally tend to come at a higher cost, reflecting the increased risk for the intermediary, but provide the end investor with a certain price and an immediate execution bearing no timing risk that is associated with execution over time. These transactions, because they

involve negotiation between two parties, still mostly happen in a manual fashion or over the phone and then get reported to an appropriate exchange for public dissemination.

The second category of market making involves the provision of quasi-continuous, immediately accessible quotes on an electronic venue. With the advent of electronic trading described in the previous section, market makers originally seated on the exchange floors have progressively been replaced by electronic liquidity providers (ELP) who leverage fast technology to disseminate timely quotes across multiple exchanges and develop automated quantitative strategies to manage their inventory and the associated risk. As such, most ELPs can be classified as high frequency traders. They derive their profits from three main sources: from liquidity rebates on exchanges that offer maker-taker fee structure, from spread earned when successfully buying on the bid and selling on the offer, and from short-term price move favorable to their inventory.

Hendershott, Brogaard and Riordan (2014) [152] find that HFT activity tends to be concentrated in large liquid stocks and postulate that this can be attributed to a combination of larger profit opportunities emanating from trades happening more often, and easier risk management due to larger liquidity that allow for easier exit of unfavorable positions at a reasonable cost.

- d) **Risk management:** In the existing literature on informed trading, it is observed that liquidity supplying risk-neutral market makers are adversely selected by informed traders suddenly moving prices against them. For example, a market marker buy quote tends to executed when large sellers are pushing the price down, resulting in even lower prices in the near term. This significant potential asymmetry of information at any point in time emphasizes the need for market makers to employ robust risk management techniques, particularly in the domain of inventory risk. For a market maker, risk management is generally accomplished by first adjusting market quotes upward or downward to increase the arrival rate of sellers or buyers and similarly adjusting the inventory in the desired direction. If biasing the quotes doesn't result in a successful inventory adjustment, the market makers generally employ limit orders to cross the spread.

Ho and Stoll (1981) [157] introduced a market making model in which the market makers objective is to maximize the profit while minimizing the probability of ruin by determining the optimal bid-ask spread to quote. The inventory held evolves through the arrival of bid and ask orders where arrival rate is taken to be a function of bid and ask prices. Their model also incorporates the relevant notion of spread size dependence on the market marker's time horizon. The longer the remaining time, there is a greater potential for adverse move risk for liquidity providers, and vice versa, which is consistent with observed spreads. In most markets, the spread is wider at the beginning of the day, narrowing toward the close. An additional reason annotated for the wider spread right after the beginning of the trading day is due to existence of potentially significant information asymmetry accumulated overnight. As market makers are directly exposed to that information asymmetry, they tend to quote wider

## 6 Algorithmic Trading and Quantitative Strategies (Version: November 18, 2017)

spreads while the price discovery process unfolds following the opening of continuous trading, and progressively tighten them as uncertainty about the fair price for the asset dissipates.

Understanding the dynamics of market makers inventory and risk management, and their effect on spreads, has direct implications for the practitioners who intend to deploy algorithmic trading strategies as the spread paid to enter and exit positions is a non-negligible source of cost that can erode the profitability of low alpha quantitative strategies.

Hendershott and Seasholes (2007) [156] confirms that market makers' inventories are negatively correlated with previous price changes and positively correlated with subsequent changes. This is consistent with market makers first acting as dampener of buying or selling pressure by bearing the risk of temporarily holding inventory in return for earning the spread and thus potential price appreciation from market reversal. This model easily links liquidity provision and asset prices dynamics.

In the context of market making, two important price behaviors are relevant to model: momentum and mean-reversion. Both of these price dynamics represent a departure from the pure random walk assumption for asset prices as well as the efficient market hypothesis developed by Eugene Fama that states that asset prices reflect all available information and consequently that past prices cannot predict future performance. Jegadeesh and Titman (1993) [168] identified the potential profitability of momentum in equities documenting the outperformance of strategies that suggest buying stocks that have performed well in the recent past while selling stocks that have performed poorly. CTA funds are among the first to have implemented momentum strategies in a systematic way, in the commodities futures space, both on single and multiple assets (such as baskets of commodities where momentum metrics can be applied either to the single assets or to the relative performance of assets within the basket).

The most naïve momentum strategies rely on simple time series analysis such as price return over a period of time to decide which asset go long (previous positive return) and which asset go short (previous negative return). Other commonly used momentum signals include simple moving averages (MA), moving averages crossovers (of a short-term MA with a longer term MA), exponential weighted moving averages (EWMA), as well as more sophisticated techniques leveraging machine learning tools. The key characteristic of momentum strategies is their flexibility and relative low complexity. They can be designed at various time intervals (from minutes to months), make use of cross-sectional indicators, be corrected for seasonal effects, and so on. Chapter 2 on Time Series will provide in-depth coverage of techniques that are commonly used by practitioners. In discrete time, the AR(1) process is typically used to model mean-reversion, and it will be described in more details in Chapter 2 as well.

In continuous time, one well known mean-reverting process is the Ornstein-Uhlenbeck (OU) stochastic process (Uhlenbeck and Ornstein (1930) [254]). It is described by the following stochastic differential equation:

$$dX_t = \theta(\mu - X_t) dt + \sigma dW_t \quad (1.1)$$

where  $W_t$  is a standard Brownian motion, and  $\theta, \sigma$  are positive constants. The value  $\mu$  is the long term mean of the process, the coefficient of  $dt$  is called the drift, and  $\sigma$  is the volatility. One can observe that the drift is negative for  $X_t > \mu$  and positive for  $X_t < \mu$ , making the process revert towards  $\mu$  over time.  $\theta$  is commonly known as the rate of mean reversion. While the OU process assumes the volatility  $\sigma$  is a constant term, more advanced models with varying volatility allow for behaviors closer to real asset price dynamics. With the increasing availability of ultra-high frequency data, the trading data arrives at a quasi-continuous time scale and the models such OU process have become easily testable.

Going back to Hendershott (2006) observation of a positive correlation of market makers inventory with subsequent prices changes, inventories can complement past returns when predicting future return. Since inventories are not publicly known, market participants use different proxies to infer their values throughout the day. Two commonly used proxies are trade imbalances (the net excess of buy or sell initiated trade volume) and spreads. Trade imbalance aims at signaling trades, either buy initiated or sell initiated, by comparing their price with the prevailing quote. Given market makers try to minimize their directional risk, they can only accommodate a limited amount of non-diversified inventory over a finite period of time. As such, spread sizes, and more particularly their sudden variation, have also been used as a proxy for detecting excess inventory from liquidity providers.

### 1.3 Algorithmic Trading: Some Basic Elements

The automation of trading and the resulting wealth of data have had a significant impact on the field of quantitative finance in general. It has allowed for a seamless integration of quantitative strategies used for portfolio construction with the executions based on algorithmic trading strategies. Progressively, portfolio managers realized that the alpha generation problem cannot be decoupled from the execution problem as transaction costs can have a material effect on the profitability of the overall strategy. For example, a fund manager back-testing a strategy might assume he or she will be able to obtain the VWAP (Volume-Weighted Average Price) of the day instead of the previous night close which is not a realistic benchmark. The initial algorithmic trading strategies were targeting average prices over the period of execution (either volume-weighted average for VWAP or time-weighted average for TWAP). The VWAP benchmark price is defined as,

$$\text{VWAP} = \frac{\sum_i V_i P_i}{\sum_i V_i} \quad (1.2)$$

where  $i$  represents each trade over the trading period,  $V_i$  represents the volume and  $P_i$  is the price.

Without having prior knowledge of  $P_i$  trajectory, a natural way to minimize slippage (executed price less arrival price) to VWAP is to try to achieve a constant pro-

portion of all  $V_i$  throughout the execution period. As a result, the first generation trading algorithms targeting VWAP had a natural incentive to break down parent orders in smaller child orders traded throughout the prescribed execution period, naively following the expected intraday distribution of market volume. The more granular data resulting from the automation of exchanges has allowed algorithmic traders to progressively build more sophisticated execution strategies. Data on trade and quote events (tick-by-tick data) are now easily accessible to researchers interested in studying microstructure behavior. The amount of available tick data now is very large and calls for techniques used for big data analysis and modeling. For instance, the constituents of the S&P500 alone record an average of 12 million trade events per day (the simplest level of information), and that number can grow significantly larger if one collects so called message data (containing all trade and quote events processed in the limit order book) published by the exchanges.

As such, algorithmic trading can be described as a broader field combining several areas such as mathematical modeling, computer science and statistical analysis. Quantitative strategies, such as the ones detailed in the subsequent chapters, call for robust and methodical research, a good understanding of economics fundamentals, while the amount of data available and the complexity of reproducing realistic trading conditions require a reliable and fast infrastructure for backtesting and calibrating models.

More simply, the practitioner working on developing quantitative algorithmic strategies goes through the following steps:

- **Data collection:** As mentioned previously, there is no dearth of available data. That said, the researcher needs to put particular emphasis on understanding the data structure as well as the methodology used to collect it (Which data feed is used? At what point is the data time stamped? What is the impact on recording the market performance of when there are activity bursts? How to order in the correct sequence, data coming from different origins or different asset classes in the correct time scale? etc.). High frequency data tends to be noisy, and one of the significant challenges faced in practice is to devise appropriate mechanisms to clean-up possibly corrupted data, to handle missing data, to tag outliers so that they can be isolated from the general modeling and to study how to use them for modeling extreme scenarios that may arise in the future. Failure to properly correct for errors in the data inevitably will lead to misspecified trading strategies that will not perform as expected.
- **Modeling:** The core of the subsequent chapters will cover in detail the models commonly used in existing quantitative strategies. These models cover the diverse nature of the trading data. Particular attention should be paid to the underlying assumptions of the model (characteristics of the distribution, independence, stationarity, etc.), as well as the robustness and stability of the calibration under model deviations. While violation of model assumptions is one of the most common pitfalls encountered in practice, model misspecification (whether due to missing variables or inclusion of wrong variables) can also lead to significant bias (see Rao (1973) [226]).

- **Backtesting:** Once the strategy is designed, the next step involves backtesting it, both on in and out of sample data. The objective is to determine the behavior and profitability of the strategy over time in order to fine tune the calibrated parameters. Backtesting can be carried out with different objectives, from simply maximizing returns to maximizing risk-adjusted returns metrics such as Sharpe ratio or Sortino ratio. However, a key factor in successful and realistic backtesting depends on the assumptions of the model. Among the relevant issues to consider for execution: the fill probability (for passive orders), the models for market impact that determines the actual transactions and the scalability of the strategy all require particular attention. Additionally, for high frequency strategies, the backtesting should account for bursts of market events and latency effects. Finally, while the goal of backtesting is to find the optimal model, the most common pitfall encountered by practitioners is overfitting with a large set of parameters that simply introduce noise rather than capturing the signal (see Easley, Lopez de Prado and O'Hara (2014) [94]).
- **Execution:** The final step involves deploying the strategy in production and continually monitoring its performance to determine if there is a need for potential adjustments. The field of Transaction Cost Analysis is relevant to both developers of algorithmic trading strategies and users of strategies developed by third parties such as sell-side brokers. We give some initial thoughts on this topic in the remainder of the chapter.

### Measuring algorithmic trading performance

Whether one builds a proprietary algorithmic trading strategy or uses the ones provided by a third party, the implementation requires the deployment of an appropriate performance measurement framework to assess the quality of decisions made at the time of trading and to explore potential avenues to improve future trades.

The main objective of algorithmic trading, whether it is a stand alone strategy or is part of a broader, longer term, trading strategy is to achieve the optimal balance between market impact (how much one moves the market with trader's own transactions) and timing risk (the risk that the traded instrument price moves away over time). These two quantities are respectively proportional and inversely proportional to the trading rate (how much of the current market volume is transacted by a given trader) and therefore a high level decision on strategy selection can be reduced to the question of speed, that is how fast to trade. Given the wide variety of trading strategies available, transaction cost analysis generally focuses on two questions:

- What is the optimal strategy to select?
- How does the strategy perform against a stated benchmark? What is the quality of the execution?

However, measuring the performance of a strategy is complex due to its multiple objectives (e.g. maximize liquidity sourcing while minimizing market impact), so, no single metric can capture all the dimensions. Here, we give a brief description of some commonly used metrics:

- **Slippage vs. arrival price:** This metric, also named Implementation Shortfall, was introduced by Almgren and Chriss (2000) [5] in a seminal paper. By measuring the deviation between the (arrival) price at the time of the investment decision (or at the initiation of the order) and the average price obtained by the investor, we can assess if the algorithm was good at capturing the desired alpha opportunity without generating excess market impact.
- **Slippage vs. period VWAP:** Comparing the average execution price with the market volume-weighted average price during the same trading period measures the ability of the algorithm to capture favorable price points during the execution window, rather than just simply following market volume.
- **Participation rate:** Comparing participation rates gives an indication of how much of the period volume is obtained by the investor. Generally, a higher participation rate, without higher market impact, would mean the order was completed faster, and thus incurring less exposure to timing risk. That said, a higher participation rate is not always desirable; for instance in the case of a favorable slippage versus arrival price (e.g. a stock going down over time while the investor is buying), trading too fast would prevent the investor from getting better prices. Studying the participation rate in conjunction with the patterns of momentum or mean-reversion at times of order entry can help assess if the rate of participation was appropriate.
- **Opportunity Cost:** In attempting to minimize price impact there is an obvious trade-off between lowering the trading rate and the potential for not being able to complete the order. The slower the trading rate, the more likely the market impact will be low, but a portion of the order is likely to remain unfilled, generating an opportunity cost for the investor. The common metrics for intraday opportunity cost measurement include performance versus close price as well as participation-weighted price (PWP) at different participation rates. For a strategy that spans over several days in splitting parent orders over many days, the opportunity cost of the unfilled quantity can be computed either with the last day close, or with the theoretical alpha left on the table.
- **Reversion:** An additional way to analyze market impact or short term opportunity cost is to study the stock price trajectory once the execution is over. A sustained high demand for liquidity results in a displacement of the Supply-Demand equilibrium, as liquidity providers need to manage their inventories, that may lead to an adverse price move (market impact). Empirical evidence (as mentioned in Gatheral (2010) [129]) suggests the temporary component of price impact reverts once the execution stops, and the impact function follows a power law decay function. Hence, studying reversion at a time window commensurates with the order duration informs the trader of the impact cost that might have been incurred due to a participation rate that is unadapted to the liquidity that the market was able to provide at the time of the transaction.
- **Ability to capture passive liquidity:** Another non-negligible source of transac-

tion cost that a trading strategy aims at reducing is the spread cost; thus the ability to trade on the passive side of the order book and capture the spread by providing liquidity should also be measured. The less the liquidity demand is, on the average, the lower the aggressive take rate would be as the algorithm has more time to source passive fills. While some strategies do not explicitly distinguish between providing or taking liquidity in their objective function (like VWAP), removing liquidity from the limit order book tends to generate higher market impact and results in information leakage particularly when a full level is depleted. The less liquid markets or securities are, the more the spread cost contributes to the total transaction cost. For instance, in low market capitalization stocks or emerging markets, bid-ask spreads greater than 50 bps are not uncommon and can significantly affect the profitability of higher turnover strategies. Furthermore, in markets where the pricing of exchange fees follows a maker-taker model, liquidity providers are entitled to receive a rebate (for instance, a cent per hundred shares traded) from the exchange where the transaction took place. This rebate can further improve the profitability of their strategy and encourages the focus on passive orders.

- **Order Routing Performance:** Fragmented markets, whether due to the existence of several lit venues or numerous dark pools, add an extra layer of complexity. In a fragmented market, the routing decisions become important to optimal liquidity sourcing. Sending a passive child order to a venue and resting it there for a period of time without getting any execution while transactions happen elsewhere in different venues might incur opportunity cost if the market moves away and the child order ends up being traded at a less favorable price later on. Similarly, if the algorithm sends an aggressive child order to capture liquidity on the far touch, routing decisions as well as the speed of the router will affect how much of the displayed liquidity at the time of decision can be captured (fill rate). Other market participants, particularly high frequency market makers, tend to display liquidity on multiple venues simultaneously and are likely to attempt to cancel all other outstanding child orders once they receive the execution confirmation from the first hit venue to avoid adverse selection, and may replace their quotes deeper in the book. Hence, the trading algorithm attempting to capture all visible liquidity on the far touch at a point in time might end up only in a partial execution and may have to trade the remainder at a less favorable price. Given the volume and complexity of events to process in a fragmented market environment and the sensitivity to latency, the routing decisions are often delegated to the Smart Order Router (SOR) rather than handled by the algorithmic strategy itself. Most advanced SOR handle real-time market data feeds from various venues and incorporate estimations of the distance separating them from the matching engines of these venues in order to optimize child order routing sequence. Some of the common metrics used to measure performance of liquidity sourcing include analyzing fill rate on aggressive limit orders (higher fill rate meaning a more efficient routing), the ratio of routed versus executed quantity, and average time to

fill on passive orders (less unsuccessful routing and shorter time to fill meaning less potential opportunity cost incurred by child orders).

The diversity of these metrics highlights the complexity of the task at hand when assessing the performance of an algorithmic trading strategy. Additionally, they have to be put in perspective with what is the state of the market at the time of execution. The outcome of a given trade is in most part determined by market conditions at the time of trading. Being a buyer “at 20%” of the volume in a market trending downward on large sellers volume is significantly easier than being a seller at 20 percent of the volume in similar conditions. As such, it is often necessary to normalize the data based on market conditions, or cluster executions by their most relevant features such as: market is significantly trending away or trending in, is there a low liquidity demand or a high liquidity demand, presence or not of user-imposed constraints such as minimum/maximum participation rate or limit price.

One point worth noting for the practitioner is that defining a priori the metrics to be used to assess the performance of the strategy can help inform the design of the algorithm itself, as well as the areas of focus for backtesting and in providing a framework for data collection a priori before deploying strategies in production.

---

## 1.4 Low Frequency vs. High Frequency Trading Strategies

---

## 1.5 Risk and Regulation in Quantitative Trading

---

## 1.6 Trading in Other Asset Classes

While the electronification of trading first took place in the equity space, recent years have seen a significant growth of electronic trading and market making in the fixed income world, opening new opportunities to quantitative trading strategies.

### a) Credit Derivatives:

When an entity, whether private or public, borrows money, the lender or debt holder bears some default risk until maturity. There is always a risk that the entity will not be able to repay the debt in full or per the agreed terms (coupons, maturity, ...). Credit derivatives allow debt holders to hedge all or part of that risk, as well as speculators to express their views on the creditworthiness of various entities.

A credit default swap (CDS) is a derivative contract allowing to buy or sell protection on a single reference entity. The protection buyer pays a fixed, running or upfront, premium in return for the right to receive a payment should the reference entity suffer

a credit event. Depending on the CDS contract, an eligible credit event might correspond to a payment default on the debt, a restructuring or a simple credit downgrade of the reference entity. In return for the premium, the writer of the CDS agrees to pay to the buyer  $(1 - \text{recovery rate})$  times the notional of the contract upon the realization of an eligible credit event. Here, the recovery rate represents the fraction of the face value of the debt that can be recovered following a credit event.

By entering a default swap, the buyer is essentially purchasing an insurance against the risk of default of a borrower. However, that insurance being paid by the CDS issuer bears counter-party risk (the ability of the issuer to pay the agreed amount in the event of a credit event affecting the borrower). Consequently, the valuation of a CDS depends, among other factors, on the default probability of the borrowing entity over the life of the contract, on the default probability of issuing counter-party, as well as the correlation between them.

The estimated CDS notional outstanding stands above \$10 trillions, after having peaked at about \$62 trillions at the end of 2007 (International Swaps and Derivatives Association (ISDA) [?]), and north of 1 million trades are recorded per week. In the aftermath of the Global Financial Crisis of 2008, trading liquidity shifted from single name CDS toward CDS Indices, which are essentially baskets of single name CDS. Roughly half of the outstanding notional originates from single reference entity contracts, while most of the other half emanates from credit indices (roughly 130,000 trades per week in 2016).

Over-the-counter (OTC) derivatives markets are considered dealer markets, as they tend to be traded through a network of private dealers who stand ready to provide liquidity in various instruments while maintaining a relatively neutral net risk exposure. This, traditionally, is conducted one on one, off exchanges. However, the introduction of the Dodd-Frank Act of 2010 brought significant transformation to the U.S. swap markets by mandating central clearing for standardized over-the-counter derivatives such as swaps, as well as forcing their execution on a swap execution facility (SEF).<sup>2</sup>

The term ‘swap execution facility’ defines a registered “trading platform in which multiple participants have the ability to execute or trade swaps by accepting bids and offers made by multiple participants” (Dodd-Frank Act, Section 733 [?]). Among the stated goals of mandating swaps execution on SEFs are promoting pre-trade transparency in the swaps market as well as facilitating the real-time publication of trading information such as price and volume to enhance price discovery. As such, the Dodd-Frank Act addressed for swap markets the three main challenges that had historically prevented a significant electronification of trading in fixed income markets: 1) lack of standardized instruments, 2) lack of centralized trading platforms, 3) lack of available transaction data.

SEF usually offer multiple trading protocols giving investors the flexibility to choose the most appropriate trading style for their needs. Existing trading protocols

---

<sup>2</sup>In Europe, the MiFID II directive adopted by the EU in 2014, with applicability on January 3rd, 2018, also introduces obligations for sufficiently liquid standardized derivatives contracts to be traded only on regulated platforms such as OTFs: Organized Trading Platform.

center around either an Order Book-like approach where all markets participants have the ability to execute available bids or offers from multiple participants as well as leave their own limit orders; or a Request-For-Quote (RFQ)-like approach where participants request a single or two-sided market from multiple dealers during a real-time auction.

A variety of credit Indices are fairly liquid and trading on SEFs, in particular U.S. indices (CDX(c)) and European indices (iTraxx(c)). The CDX indices are further broken out by the type of debt covered such as Investment Grade (IG) and High Yield (HY) for the most liquid ones. Due to their liquidity, both CDX IG and CDX HY are examples of credit indices that lends themselves well to electronic market making activities.

However, the CDS Index market presents a certain number of idiosyncrasies compared to equities markets when it comes to market making. While there is an order book available to all participants, most of the volume still gets transacted via RFQ mechanism for which market makers are not always allowed to see the quotes. Consequently, market makers only have partial information regarding the true position of the market when time comes to decide where to place their own orders. Using a mixture of historical trades and partial real-time information, market makers can reconstruct a theoretical mid price of the market. They can set then the bid and ask quotes at an appropriate distance from that mid price, accounting for the trade-off between their desire to obtain a fill and the risk associated with maintaining their inventory.

**b) Corporate Bonds:**

The corporate bond market is also an OTC market where market participants only have access to limited information prior to placing a trade. The main market participants can be categorized as pension funds and insurance companies which tend have longer investment horizons, hedge funds which tend to have more tactical trading allocations, and corporations treasury departments whose objectives and horizons can span a wide spectrum.

Compared to other assets, it is also interesting to note that most of the secondary market transactions happen in the first few months following issuance, and after that a significant portion of the amount of bonds issued is held to maturity. Similarly to other fixed income products, the key challenges for secondary trading of bonds and the electronification of its secondary market have historically been 1) lack of centralized market place, 2) lack of harmonized instrument characteristics and 3) lack of transaction data.

The corporate bonds secondary market remains dominated by one-to-one privately negotiated trades, but electronic platforms allow transacting via similar execution protocols as what traders can use for CDSs: limit order book type of executions as well as request-for-quotes (RFQ) on platforms such as Tradeweb, Bloomberg, MarketAccess and a variety of others.

In the U.S., for instance, under FINRA Rule 6730, Broker-dealer FINRA member firms have the obligation to report transactions in TRACE-eligible securities to the

TRACE<sup>3</sup> database as soon as practicable, but no later than within 15 minutes of the Time of Execution. FINRA Rule 6710 defines TRACE-eligible securities as USD denominated debt securities (whether issued by a U.S or foreign private issuer), and USD denominated debt issued or guaranteed by an Agency or Government-Sponsored Enterprise. Foreign sovereign, U.S. Treasuries and money market instruments are specifically excluded from the eligible list. The TRACE database then disseminates price, size<sup>4</sup>, time stamp and direction of the trade to the public.

While the rules require reporting in no more than 15 minutes, most reporting and public dissemination now happens within seconds or minutes of execution (with the exception of overnight trades being batch-reported the next morning), giving participants some transparency about the current market levels.

The combination of existing venues allowing for click-to-trade or RFQ protocols and trade events information facilitated the expansion of electronic trading for corporate bonds as well. A fifth to a quarter of the investment grade market is now transacted electronically, while on the high yield side, roughly 10% of the market is traded through RFQs.

Similarly to credit indices, the first challenge for dealers in corporate bond electronic market making is to infer the value of the current “fair mid” for the market following each transaction. Given the relatively infrequent updates observations even for bonds that are considered liquid, quantitative techniques employed tend to vary from the ones used in markets with higher frequency data available such as equities. Practitioners rely much more, for instance, on probabilistic state-space models (see further—Kalman Filter—Chapter 2 [?]) to estimate unobservable data.

This type of approach obviously becomes increasingly more challenging as one moves down the liquidity spectrum. It is not uncommon for illiquid bonds to not trade for months, in which case assessing a fair mid price cannot solely rely on prior transactions. Among possible solutions, modelers can use bonds of other maturities from the same issuers, or bonds from correlated issuers or comparable investment grades. Additionally, more liquid credit derivatives are also a potential sources of information that can be leveraged for better valuations.

---

<sup>3</sup>Trade Reporting And Compliance Engine

<sup>4</sup>The actual size disseminated back to the market is capped to prevent excessive information leakage: for investment-grade corporate bonds and agency debt securities, for any trade greater than \$5 million, the par value is displayed as \$5 MM+; for non-investment grade corporate bond, the displayed quantity is capped to \$1 million par value



---

---

## **Part I**

# **Foundations: Basic Models and Empirics**

---

---



In Part I, our goal is to provide a review of some basic and some advanced methodologies that are useful for developing trading algorithms. We begin with time series models (in Chapter 2) for univariate data and provide a broad discussion of autoregressive moving average models—from model identification, to estimation, inference and to model diagnostics. The stock price and return data do exhibit unique features and so we identify some stylized facts that have been confirmed; this work will greatly help to discern any anomalies and when they arise as then anomalies generally indicate deviation from efficient market hypothesis. Wherever possible we illustrate the methodology with examples and provide codes for computing and make the data accessible to the reader. This chapter is followed by methodologies for multiple time series data and the discussion in Chapter 3 is useful from pairs trading to portfolio optimization. The last chapter (Chapter 4) in Part I contains advanced topics such as theory of point processes as trading takes place on a continual basis and understanding the market macro-structure is essential for deciding about the entry to and exit from the market. This chapter also contains advanced topics such as machine learning.

A reader with strong statistics background can afford to skip some sections but it is recommended that the readers peruse these chapters as they contain discussion on topics that need further research work. In presenting the methodologies here, we kept the discussion lucid but immediately relevant to the main theme of this book—understanding the market behavior and developing effective trading strategies.



# 2

---

## *Univariate Time Series Models*

---

### CONTENTS

2.1	Trades and Quotes Data and their Aggregation: From Point Processes to Discrete Time Series .....	22
2.2	Trading Decisions as Short-Term Forecast Decisions .....	24
2.3	Time Series Models for Aggregated Data: Modeling the Mean .....	25
2.3.1	Stochastic Processes: Some Properties .....	25
2.3.2	Some Basic Models and Descriptive Tools .....	30
2.3.3	Specification and Estimation of ARIMA Models .....	39
2.3.4	Testing for Nonstationary (Unit Root) in ARIMA Models ...	50
2.3.5	Forecasting for ARIMA Processes .....	52
2.4	Stylized Models for Asset Returns .....	55
2.5	Time Series Models for Aggregated Data: Modeling the Variance ...	59
2.6	Stylized Models for Variance of Asset Returns .....	63
2.7	Other Topics: Bet Size etc. ....	65
2.8	Exercises (starred items have <i>R</i> -codes available) .....	65

With the advent of electronic trading a vast amount of data on orders is now recorded and is available to traders to make real-time decisions. The Trades and Quotes (TAQ) data on time-stamped sequence of trades with the updates on the price and depth of the best bid and ask quotes was used by researchers and traders initially. This was later expanded to the so-called Level II data, with time-stamped sequence of trades and quotes for the top 5 levels of the order book. Both TAQ and Level II data still do not provide the complete story of the buy and sell sides. With Level III data now available, we can obtain time-stamped sequence of all events that occur, with the exception of the information on hidden orders. Thus the data related to all trading activities is quite voluminous and is irregularly spaced. We take a broad view that for making strategic decisions to enter and exit the market, the trader needs to take an aggregate view of the trading activities, but that for the actual execution of the strategies the trader needs to understand the market micro-structure that relates to the actual process of trading. With the increased speed of decisions that are made nowadays the difference between the two steps may be blurred, but the aggregation process may help to look beyond trading fractions. In Section 2.1 we discuss the Trades and Quotes data and the aggregation issues. In Section 2.2 we define that the trading algorithms depend on prediction of short-term price movement.

The movement whether it is due to pure market frictions or due to information related to a stock is captured through a drift term that needs to be carefully monitored. Because time series models for discrete time units are well-studied and are still widely used by the traders in the form of aggregated price-bars, we focus on these methods in Sections 2–3 to 2.6. We broadly classify these methods into modeling the mean (return) and into modeling the variance (volatility). As these techniques are well-documented and can be found elsewhere, we only present a brief overview. The established facts about the stock returns and variances are reviewed along with the methodology as these will provide expected benchmarks; successful algorithms as one would notice, generally exploit the deviations from these benchmarks.

In Section 2.7, the time series models are augmented with other stock-related information such as volume. The flow of volume is taken to provide indirectly the flow of information about a stock. As trading now is taking a portfolio point of view, it is important to study multiple stocks together and discrete multiple time-series models provide a framework and these are presented in Section 2.9. Models that deal with the print-processes and the high-frequency analytical tools are covered in the rest of the chapter.

## **2.1 Trades and Quotes Data and their Aggregation: From Point Processes to Discrete Time Series**

In its simplest form, the trading activities of equity through an exchange that opens at 9:30 a.m. and closes at 4 p.m. can be described by a sequence of time stamps (“ticks”)  $t_0 < t_1 < \dots < t_n$  and the “marks”  $y_i$  at time  $t_i$ , in which  $t_0$  denotes 9:30 a.m. and after and  $t_n$  denotes the time of the last trade that occurs before or at 4 p.m. The marks  $y_i$  can be price, volume of an order placed on either the buy or sell side of the trade and in general can represent the characteristics of the order book at the time of  $i$ th activity (see Figure 2.1). The events with the marks associated with the ticks can be described mathematically by a marked point process. But our goal here is to first show how these data can be aggregated over regularly spaced time intervals and methods for linear homogeneous time series will be described to analyze the aggregated data and later some tools that are relevant for the analysis of point processes will be presented.

The typical level III data for CISCO on a single day in 2011 is given below in Table 2.1:

This outlay in Table 2.1 points to various issues that need to be addressed in processing this type of data. The order numbers are only unique during the lifetime of the order but can be reassigned at a later time. Several activities can take place at the same time. Hidden orders are revealed only when they are executed and as such do not have visible order numbers. Full cancellation quantity and price information should be traced back to the submission time. To summarize, at any given point in time using the data described in Table 2.1, limit order book that essentially records where the order is in the queue, can be constructed and thus providing the trade time and the associated

Table 2.1: CISCO–Trade Data

Timestamp	Order Number	Event	Quantity	Price	Exchange
⋮	⋮	⋮	⋮	⋮	⋮
34561915	4254917	D	0	0	J
34561915	4253917	B	200	20.55	J
⋮	⋮	⋮	⋮	⋮	⋮
34561923	13056048	S	200	20.58	Z
⋮	⋮	⋮	⋮	⋮	⋮
34573369	13255731	C	98	0	Q
⋮	⋮	⋮	⋮	⋮	⋮
34577331	6225085	E	300	0	K
⋮	⋮	⋮	⋮	⋮	⋮
34577338	6225085	F	0	0	K
⋮	⋮	⋮	⋮	⋮	⋮
34573379	2030057	S	100	20.77	K
⋮	⋮	⋮	⋮	⋮	⋮
34573382	NA	T	1700	20.37	P
⋮	⋮	⋮	⋮	⋮	⋮

\*Time-stamp is the time since midnight; Event: B: submission of LO on Buy side; S: submission of LO on sell side; T: full execution of a LO; F: full execution of a LO; E: partial execution of a LO; D: full cancellation of a LO; C: partial cancellation of a LO; Exchange: Q-NASDAQ; J-EDGE-A; Z-BATS-Z; K-EDGE-K; P-ARCA; etc

marks. These and other characteristics of the limit order book (superbook—if we collate this information from all exchanges) will be taken up in a later chapter, but the focus here is on constructing discrete time series data from this point process data.

To illustrate the aggregation methods, we will initially focus on price of the stock,  $p_t$ , and the associated volume  $V_t$ . Two types of aggregation methods are proposed in the literature. One when the time span  $T$  for the Exchange hours (9:30 a.m.–4:00 p.m.) is divided into ‘ $K$ ’ intervals so that the regularly spaced intervals are of size  $\Delta t = T/K$ . The other method will do aggregation when there is a change in a marker, such as price. Here we focus on the first method. Various summary measures within each period  $((i-1)\Delta t, i\Delta t)$  where  $i = 0, 1, \dots, K$  can be computed;

- number of transactions: ‘ $n_i$ ’
- Average price:  $\bar{p}_i = \sum p_{t_j} / n_i$
- Volume Weighted Average Price:  $\bar{p}_{wi} = \sum V_{t_j} p_{t_j} / \sum V_{t_j}$
- Average duration between transactions:  $\bar{d}_i = \sum_{j \in ((i-1)\Delta t, i\Delta t)} (t_j - t_{j-1}) / n_i$
- Mid-Price: (Best bid + Best ask)/2

It is possible that for infrequently traded stocks, some time intervals may not contain any trading activity. For heavily traded stocks, the duration can be very small and

thus may not provide any consequential information. We will address ways to analyze high-frequency data later in this chapter. To begin with, we discuss the analysis for data aggregated over fixed time intervals. Depending on the size of the interval, the data can be called low frequency or high frequency but the time series methods that we discussed here apply to *all* aggregated data.

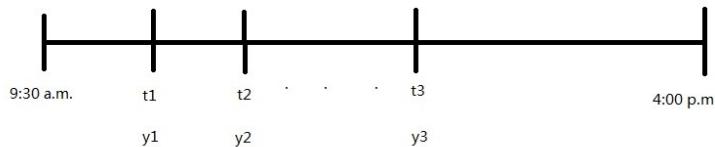


Figure 2.1: Trading Activities

## 2.2 Trading Decisions as Short-Term Forecast Decisions

Most trading decisions generally involve the following considerations

- How much to trade?
- When to trade? ie when to enter and when to exit?
- How to trade? How fast? What is the objective? What are the benchmarks?

Answer to these questions involve in some way predicting the market movement and in particular the future (short-term and long-term) price of assets. In the short term it is possible that prices have a tendency to ‘trend’ or exhibit ‘momentum’ but in the long run, prices have a tendency to ‘mean-revert’ as the information about the asset gets incorporated in the price.

Let  $p_{it}$  denote the price of the  $i$ th asset at time  $t$  and let  $p_t = (p_{1t}, p_{2t}, \dots, p_{nt})$  denote the price vector for ‘ $n$ ’ assets. Let  $y_{it}$  denote a vector of characteristics, e.g. volume of transactions, price volatility, number of transactions and the intensity of trading as captured by the average duration between transactions, etc of the  $i$ th asset at time  $t$ . These quantities are aggregated from high-frequency data of type given in Table 2.1 and illustrated in Figure 2.1 to be used as the characteristics associated with the price changes. In addition, we can also consider factors  $f_t = (f_{1t}, f_{2t}, \dots, f_{nt})$  that may include market and industry factors as well as asset characteristics such as market capitalization, book-to-market ratio etc. The trading rules can be broadly grouped as follows:

A. Statistical Arbitrage Rules:  $E(p_{i,t+1} | p_{i,t}, p_{i,t-1}, \dots, y_{i,t}, y_{i,t-1}, \dots)$

- Predicting the price of  $i$ th stock at  $t+1$  based on the past trading information; this is sometimes labelled as time series momentum.

B. Momentum:  $E(p_{t+1} | p_t, p_{t-1}, \dots, y_t, y_{t-1}, \dots)$ 

- Predicting the cross-sectional momentum of a subset of stocks based on their past trading characteristics. This is not only useful for portfolio formation and rebalancing, also for pairs trading which is based on tracking two stocks simultaneously so that their divergence can be exploited.

C. Fair Value:  $E(p_{t+1} | p_t, p_{t-1}, \dots, V_t, V_{t-1}, \dots, f_t, f_{t-1}, \dots)$ 

- Predicting the price using all relevant quantities. Many of these factors can be at a more macro level than the time scale considered for the price prediction, but nevertheless could be useful, as explained in Chapter 3.

Thus the price (or alternatively, return) and volatility (or squared return) prediction can be formulated as a time series prediction problem and the common tools such as autoregressive and ARCH models can be used.

Predicting stock market behavior has been studied extensively starting in the early 20th century. Cowles (1933, 1944) [70, 71] evaluated the performance of the stock market forecasters. Analyzing the performance of several financial services that made some 7500 recommendations on common stocks over a period, January 1, 1925 to July 1, 1952, it was pointed out the record was worse than the performance of average common stock by 1.43% annually. Thus ‘buy-and-hold’ strategy out-performed the stock picking strategies of those days. In fact, a much harsher assessment is made: “A review of the various statistical tests, applied to the records for this period, of these 24 forecasters, indicates that the most successful records are little, if any, better than what might be expected to result from pure chance. There is some evidence, on the other hand, to indicate that the least successful records are worse than what could reasonably be attributed to chance.” Many of these comments are still valid and have stood the test of the time. Yet, we will demonstrate in this book that there are opportunities for statistical arbitrage. The key is to have access to relevant information and use the appropriate techniques to exploit this in a timely manner.

## 2.3 Time Series Models for Aggregated Data: Modeling the Mean

In this section we present a broad overview of the time series models for data aggregated over discrete time intervals.

### 2.3.1 Stochastic Processes: Some Properties

Formally, a discrete time series or stochastic process  $Y_1, Y_2, \dots, Y_T$  is a sequence of random variables (r.v.’s) defined on a probability space, and possessing a joint probability distribution, with  $Y_t$  denoting the r.v. corresponding to the value of the

series at time  $t$ . A particular sequence of observations of the stochastic process  $\{Y_t, t = 1, \dots, T\}$  is known as a realization of the process. In the analysis of time series we are interested in aspects of the joint probability distribution of collections of random variables such as  $Y_1, Y_2, \dots, Y_T$ . For such a collection of r.v.'s, the joint distribution function is defined by

$$F_{12\dots T}(y_1, y_2, \dots, y_T) = \Pr[Y_1 \leq y_1, Y_2 \leq y_2, \dots, Y_T \leq y_T] \quad (2.1)$$

In general, time series analysis is concerned with determining the properties and identifying the probability structure which generated the observed time series. A general feature of the approach to this analysis is that we do not attempt to study the joint distribution of  $Y_1, Y_2, \dots, Y_T$  directly, since this may be too complicated for large  $T$ , but we study and attempt to describe the probabilistic mechanism which generates the process sequentially through time, and from this to derive the conditional distribution of future observations for purposes of prediction.

Because the means, variances, and covariances are useful summary descriptions of the joint probability distribution of the stochastic process  $\{Y_t, t = 1, \dots, T\}$ , we will first consider these quantities. Thus, for any stochastic process  $\{Y_t\}$ , we define the following functions of  $t$ :

$$\begin{aligned} \text{mean, } \mu(t) &= E(Y_t) \\ \text{variance, } \sigma^2(t) &= \text{Var}(Y_t) = E[(Y_t - \mu(t))^2] \\ \text{autocovariance, } \gamma(t, s) &= \text{Cov}(Y_t, Y_s) = E[(Y_t - \mu(t))(Y_s - \mu(s))] \\ \text{autocorrelation, } \rho(t, s) &= \text{Corr}(Y_t, Y_s) = \frac{\gamma(t, s)}{\sigma(t)\sigma(s)} \end{aligned} \quad (2.2)$$

Obviously, it is not possible to estimate the many unknown parameters (means, variances, and covariances) under such a general specification from a single realization of  $T$  observations. In addition, such generality in the specification of the mean and covariance structure offers no information concerning the prediction of future observations of the process. Hence we must impose some additional structure on the joint distribution of the process in order to substantially reduce the number of unknown parameters, as well as to enable meaningful predictions into the future to be possible. The concept of stationarity of a process is useful in this regard and also serves as a realistic assumption for many types of time series. Stationarity is motivated by the fact that for many time series in practice, segments of the series at different points in time behave similarly, or exhibit homogeneous behavior over time.

### Stationarity

A time series  $\{Y_t\}$  is said to be stationary if for every integer  $m$  and any finite set of  $m$  time points  $t_1, t_2, \dots, t_m$ , the set of variables  $Y_{t_1}, Y_{t_2}, \dots, Y_{t_m}$  must depend only on the distance between the times  $t_1, t_2, \dots, t_m$ , rather than on their actual values. So a stationary process  $\{Y_t\}$  tends to behave in a homogeneous manner as it moves through time. The means and variances of the  $Y_t$  are the same for all  $t$ , that is,  $E(Y_t) = \mu$  and

$\text{Var}(Y_t) = \sigma^2$  are constant, for all  $t$ . So we may express the autocovariance function as

$$\gamma(s) = \text{Cov}(Y_t, Y_{t+s}) = E[(Y_t - \mu)(Y_{t+s} - \mu)], \text{ for all } s = 0, \pm 1, \dots \quad (2.3)$$

Note that by this notation,  $\gamma(0) = E[(Y_t - \mu)^2] = \text{Var}(Y_t)$ . Also, the *autocorrelation function* of the stationary process may be expressed

$$\rho(s) = \text{Corr}(Y_t, Y_{t+s}) = \frac{\text{Cov}(Y_t, Y_{t+s})}{(\text{Cov}(Y_t) \text{Cov}(Y_t))^{1/2}} = \frac{\gamma(s)}{\gamma(0)}, s = 0, \pm 1, \dots \quad (2.4)$$

$\rho(s)$  will be referred to as the autocorrelation of the process at lag  $s$ . As we will find, the autocorrelation function (which we will abbreviate as ACF)  $\rho(s) = \gamma(s)/\gamma(0)$  of a stationary process  $\{Y_t\}$  is a very important tool in describing the characteristics of the process, since it is a convenient summary of the correlation structure of the process over all time lags.

### Examples of Stationary Stochastic Processes

**Example 1 (White Noise).** Let  $\dots, \epsilon_{-1}, \epsilon_0, \epsilon_1, \dots$  be a sequence of independent random variables defined on the discrete time points  $\dots, -1, 0, 1, \dots$ , with mean  $E(\epsilon_t) = 0$  and variance  $E(\epsilon_t^2) = \sigma^2$ , for all  $t$ . Set  $Y_t = \mu + \epsilon_t$ ,  $t = 0, \pm 1, \dots$ . Then  $E(Y_t) = \mu$  for all  $t$ , and since independence implies that the random variables are uncorrelated, we have  $\text{Cov}(Y_t, Y_{t+s}) = \text{Var}(Y_t) = \sigma^2$  if  $s = 0$  and  $\text{Cov}(Y_t, Y_{t+s}) = 0$  if  $s \neq 0$ . Thus the process is stationary. Such a process is referred to as purely random process or a *white noise process*, and will be found to be the foundation for the construction of many other processes of interest.

**Example 2 (Moving Average).** Let  $\{\epsilon_t\}$  be independent r.v.'s as in Example 1, and define a new process  $\{Y_t\}$  by

$$Y_t = \mu + \epsilon_t + \epsilon_{t-1}, \quad t = 0, \pm 1, \dots$$

where  $\mu$  is a constant. Then  $E(Y_t) = \mu$ , for all  $t$ , and

$$\text{Cov}(Y_t, Y_{t+s}) = \gamma(s) = \begin{cases} 2\sigma^2 & \text{if } s = 0 \\ \sigma^2 & \text{if } s = 1 \\ 0 & \text{if } s > 1 \end{cases}$$

which depends only on the time lag  $s$  and not on  $t$ . Hence the process  $\{Y_t\}$  is stationary with ACF  $\rho(s) = \gamma(s)/\gamma(0)$  such that  $\rho(0) = 1$ ,  $\rho(1) = 1/2$  and  $\rho(s) = 0$  for  $|s| > 1$ .

Figure 2.2 (a) shows a time series plot of a computer simulated series of 100 values from a Gaussian white noise process  $\{\epsilon_t\}$  with  $\sigma^2 = 1$ , while Figure 2.2 (b) shows a plot of the corresponding values of the series  $Y_t = 1.0 + \epsilon_t + \epsilon_{t-1}$  using these same  $\{\epsilon_t\}$ . Note the relatively much smoother behavior over time of the data in (b) than in (a), due to the positive autocorrelation at lag one of the process in (b) which tends

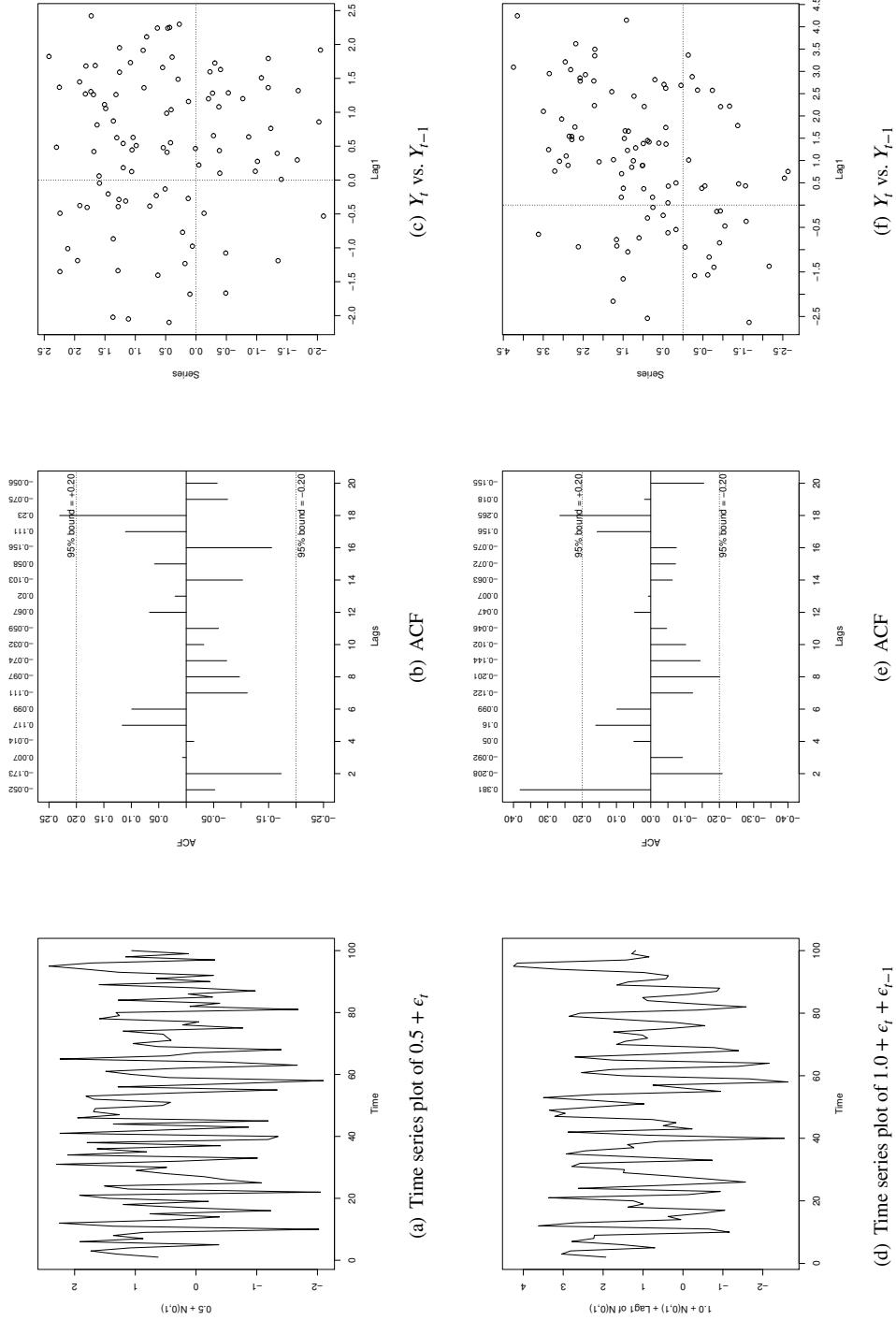


Figure 2.2: Processes & Diagnostics from  $Y_t = 0.5 + \epsilon_t$  and from  $Y_t = 1.0 + \epsilon_t + \epsilon_{t-1}$

to make successive values of the series behave more similarly to each other. Similar to the last example, it is easy to show that the process  $\{Y_t\}$  defined by the relation  $Y_t = \mu + \epsilon_t - \epsilon_{t-1}$  is stationary with autocorrelation function  $\rho(0) = 1$ ,  $\rho(1) = -1/2$ , and  $\rho(s) = 0$  for  $|s| > 1$ . A plot of a simulated series from this process is given in Figure 2.2 (d) where the much “rougher” or less smooth behavior over time relative to the behavior in (a) and (b) due to the addition of extra noise to the model. These figures form the basis of model building and the subsequent diagnostic checks on the fitness of models.

### Examples of Nonstationary Stochastic Processes

Many time series that occur in finance do not exhibit stationary behavior. For example, many stock price series show a trend or change in the mean level over time, reflecting information flow about the stock or investors’ behavior. Series also exhibit seasonal behavior, either in the form of nonstationary or deterministic seasonal patterns, both of which need to be distinguished from purely nondeterministic stationary stochastic behavior. Other departures from stationarity include changes in variability of the process over time (nonconstant variance or volatility regimes), abrupt shifts in the mean level, and changes in the autocorrelation structure of the process over time which is a feature of the series which would typically be more difficult to detect. An important practical matter in the analysis of time series data involves methods of transforming a possibly nonstationary time series into a stationary series, and accounting for and modeling of the nonstationary aspects of the original series. Many of the methods that are found useful in practice involve considering successive changes or differences over time in the original series to remove the nonstationary features of the series, especially changes in the mean level, or using linear regression techniques to help account for the nonstationary mean behavior. We will illustrate this with a few simple examples.

**Example 3 (Random Walk with a Drift).** Let  $\{\epsilon_t, t = 0, 1, 2, \dots\}$  be a sequence of independent random variables with mean 0 and variance  $\sigma^2$ , and define a new process  $\{Y_t\}$  recursively by

$$Y_t = Y_{t-1} + \delta + \epsilon_t, \quad t = 0, 1, 2, \dots, \quad Y_0 = 0,$$

where  $\delta$  is a constant. By successive substitutions in the above relations for  $Y_t$ , we find that  $\{Y_t\}$  can also be expressed as

$$Y_t = \delta t + \sum_{j=1}^t \epsilon_j = \delta t + \epsilon_t + \epsilon_{t-1} + \dots + \epsilon_1, \quad t = 1, 2, \dots.$$

Hence we see that  $E(Y_t) = \delta t$  and  $\text{Var}(Y_t) = \text{Var}(\sum_{j=1}^t \epsilon_j) = \sum_{j=1}^t \text{Var}(\epsilon_j) = t \sigma^2$ , so that the process  $\{Y_t\}$  is not stationary and exhibits change in both mean and variance. Moreover, for any  $s > 0$ , we find that

$$\text{Cov}(Y_t, Y_{t+s}) = \text{Cov}\left(\sum_{j=1}^t \epsilon_j, \sum_{i=1}^{t+s} \epsilon_i\right) = \sum_{j=1}^t \text{Var}(\epsilon_j) = t \sigma^2,$$

so that  $\text{Corr}(Y_t, Y_{t+s}) = t/\sqrt{t(t+s)} = 1/\sqrt{1+s/t}$ . Note that this correlation is close to one for large  $t$  and relatively small  $s$ , so that nearby values  $Y_t$  and  $Y_{t+s}$  of the process  $\{Y_t\}$  are highly positively correlated. However, note that the time series of successive changes or *first differences* of the process  $\{Y_t\}$ , defined by  $Z_t = Y_t - Y_{t-1}$ , does form a stationary process, since  $Z_t = Y_t - Y_{t-1} = \delta + \epsilon_t$ , a white noise series as in Example 1. Such a process  $\{Y_t\}$  whose first differences form a white noise series, is called a random walk process. Time series plots of simulated data from random walk exhibit behavior with little tendency to vary around a fixed constant mean level. The parameter  $\delta$  in the random walk model is referred to as the drift in the random walk process. Note the trend behavior in the process with  $\delta > 0$ .

Of course, a more general class of models is obtained by supposing that a non-stationary process  $\{Y_t\}$  is such that its first differences form a stationary process, but not necessarily white noise. It has been found from practical experience that this type of model is quite useful in producing adequate representations for actual observed financial series in many cases.

### 2.3.2 Some Basic Models and Descriptive Tools

Suppose we have a sample realization of  $T$  observations  $Y_1, Y_2, \dots, Y_T$  from a stationary process  $\{Y_t\}$  which has a mean  $\mu = E(Y_t)$ , autocovariance function (ACF)  $\gamma(s) = \text{Cov}(Y_t, Y_{t+s})$  and autocorrelation function  $\rho(s) = \text{Corr}(Y_t, Y_{t+s}) = \gamma(s)/\gamma(0)$ . Since the theoretical ACF  $\rho(s)$  is such a basic quantity which describes the correlation properties of the process, we are naturally interested in obtaining estimates of  $\rho(s)$  from the sample data.

First, a natural estimator of  $\mu$  is the sample mean  $\bar{Y} = (1/T) \sum_{t=1}^T Y_t$ . The most common estimator used for  $\gamma(s)$  is the sample autocovariance function defined by

$$\hat{\gamma}(s) = \frac{1}{T} \sum_{t=1}^{T-s} (Y_t - \bar{Y})(Y_{t+s} - \bar{Y}), \quad s = 0, 1, \dots$$

Usually one only computes  $\hat{\gamma}(s)$  for lags  $s$  much smaller than  $T$ , say,  $0 \leq s \leq T/4$ , since  $\hat{\gamma}(s)$  for large lags are typically not very informative or of much interest. Note that  $\hat{\gamma}(s)$  has the form of an ordinary sample covariance between the pairs of variables  $(Y_t, Y_{t+s})$ , except for the use of the common sample mean  $\bar{Y}$  for both variables and the divisor  $T$  in place of  $T - s$ . Note also that  $\hat{\gamma}(0) = T^{-1} \sum_{t=1}^T (Y_t - \bar{Y})^2$  represents the sample variance of the data. The estimator of  $\rho(s)$  follows directly from  $\hat{\gamma}(s)$  and is given by

$$\hat{\rho}(s) = \frac{\hat{\gamma}(s)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{T-s} (Y_t - \bar{Y})(Y_{t+s} - \bar{Y})}{\sum_{t=1}^T (Y_t - \bar{Y})^2}, \quad s = 0, 1, \dots \quad (2.5)$$

which is called the sample ACF. The sample ACF will also sometimes be denoted as  $\hat{\rho}(s) \equiv r(s)$ .

We now briefly discuss a few properties of the sampling distributions of the above estimators, with particular emphasis on the sample ACF  $r(s)$ . The mean of  $\bar{Y}$  is easily seen to be  $E(\bar{Y}) = (1/T) \sum_{t=1}^T E(Y_t) = \mu$ , so that  $\bar{Y}$  is an unbiased estimator of  $\mu$ . The variance of  $\bar{Y}$  is

$$\begin{aligned}\text{Var}(\bar{Y}) &= \frac{1}{T^2} \text{Var}\left(\sum_{t=1}^T Y_t\right) \\ &= \frac{1}{T^2} \left[ \sum_{t=1}^T \text{Var}(Y_t) + 2 \sum_{t=1}^{T-1} \sum_{u=1}^{T-t} \text{Cov}(Y_t, Y_{t+u}) \right] \\ &= \frac{1}{T^2} \left[ T\gamma(0) + 2 \sum_{t=1}^{T-1} \sum_{u=1}^{T-t} \gamma(u) \right] \\ &= \frac{\gamma(0)}{T} \left[ 1 + 2 \sum_{u=1}^{T-1} \frac{T-u}{T} \rho(u) \right]\end{aligned}\tag{2.6}$$

Hence if  $\rho(u) \rightarrow 0$  as  $u \rightarrow \infty$ , then so that  $\text{Var}(\bar{Y}) \rightarrow 0$  as  $T \rightarrow \infty$ . This implies that  $\bar{Y}$  converges in probability to  $\mu$  and from Equation 2.6  $\text{Var}(\bar{Y}) = \gamma(0)/T$  occurs when  $\rho(u) = 0$  for all  $u \neq 0$ ; that is, when the  $Y_t$  are mutually uncorrelated. Under some mild conditions it can also be shown that  $\bar{Y}$  is asymptotically normally distributed as  $T \rightarrow \infty$ .

The sampling properties of the sample ACF  $r(s)$  are quite complicated and exact results are not available except in a few special cases. We indicate one result from Anderson and Walker (1964) [13] based on large sample theory. Under linearity of the time series, the joint distribution of  $T^{1/2}(r(1) - \rho(1)), \dots, T^{1/2}(r(k) - \rho(k))$ ,  $k$  fixed, tends to a multivariate normal  $N(0, W)$  distribution as  $T \rightarrow \infty$  where  $W$  is a  $k \times k$  matrix with  $(i, j)$ th element  $w_{ij}$  given by

$$\begin{aligned}w_{ij} &= \sum_{v=-\infty}^{\infty} [\rho(v+i)\rho(v+j) + \rho(v-i)\rho(v+j) - 2\rho(v)\rho(j)\rho(v+i) \\ &\quad - 2\rho(v)\rho(i)\rho(v+j) + 2\rho(v)^2\rho(i)\rho(j)], i, j = 1, \dots, k.\end{aligned}\tag{2.7}$$

Hence for large  $T$  we have the approximations  $E(r(s)) \approx \rho(s)$ ,  $\text{Var}(r(s)) \approx w_{ss}/T$ ,  $\text{Cov}(r(s), r(u)) \approx w_{su}/T$ , and the  $r(s)$  are approximately normally distributed. The Equation 2.7 is also called Bartlett's formula and dates back to 1946.

There are a few special cases of Equation 2.7 worth noting. In the purely random or white noise case where the  $Y_t$  are independent r.v.'s, we then have  $\rho(v) = 0$  for all  $v \neq 0$  and so the approximations reduce to  $E(r(s)) \approx 0$ ,  $\text{Var}(r(s)) \approx 1/T$ ,  $\text{Cov}(r(s), r(u)) \approx 0$ , or more exactly,  $\text{Var}(r(s)) \approx (T-s)/T^2$ . Hence we can use these results to check for "non-randomness" of a series of comparing the sample values  $r(s)$  at various lags  $s = 1, \dots, k$  with the approximate 95% limits  $\pm 2\sqrt{(T-s)/T^2} \approx 2/\sqrt{T}$  for the normal distribution to see whether the  $r(s)$  generally fall within these limits. We will use this test repeatedly as a diagnostic tool for model adequacy.

For a second case, suppose it is assumed that the theoretical ACF satisfies  $\rho(v) = 0$  for all  $|v| > q$ . Then, for any  $s > q$  the expression for the approximate variance of  $r(s)$  reduces to

$$\text{Var}(r(s)) \approx \frac{1}{T} \sum_{v=-q}^q \rho(v)^2 = \frac{1}{T} \left[ 1 + 2 \sum_{v=1}^q \rho(v)^2 \right], \quad s > q \quad (2.8)$$

and

$$\text{Cov}(r(s), r(u)) \approx \frac{1}{T} \sum_{v=-q}^q \rho(v) \rho(v - s + u), \quad s, u > q.$$

As one final special case, suppose the process has ACF of the simple form (which results from an autoregressive model of order one (AR(1)), defined later)  $\rho(v) = \phi^{|v|}$ ,  $v = 0, \pm 1, \dots$ , with  $|\phi| < 1$ . Then

$$\text{Var}(r(s)) \approx \frac{1}{T} \left[ \frac{(1 + \phi^2)(1 - \phi^{2s})}{1 - \phi^2} - 2s\phi^{2s} \right],$$

and in particular  $\text{Var}(r(1)) \approx \frac{1}{T}(1 - \phi^2)$ .

The plot of ACF for a white noise sequence for a large number of lags may have a few isolated lags lie outside the limits,  $\pm \frac{2}{\sqrt{T}}$ . But characteristic behavior of nonstationary series is that the sample ACF values  $r(s)$  are slow to “dampen out” as the lag  $s$  increases.

### Linear Models for Stationary Time Series

A stochastic process  $\{Y_t\}$  will be called a *linear process* if it can be represented as the output of a (one-sided) absolutely summable linear filter with white noise input process  $\{\epsilon_t\}$ ; that is,

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \quad (2.9)$$

where the  $\epsilon_t$  are independent with 0 mean and variance  $\sigma_\epsilon^2$ , and  $\sum_{j=0}^{-\infty} |\psi_j| < \infty$ . This process may also be referred to as an infinite moving average process.

We now introduce, for notational convenience, the backward shift operator  $B$ , defined by the property that  $B Y_t = Y_{t-1}$  for any time series  $\{Y_t\}$ . Hence, on successive applications, we obtain  $B^j Y_t = Y_{t-j}$  and the infinite moving average (2.9) may be expressed as

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j B^j \epsilon_t = \mu + \psi(B) \epsilon_t,$$

where  $\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \dots$  is referred to as the infinite moving average operator. Since the input  $\{\epsilon_t\}$  is a stationary process, it follows that  $\{Y_t\}$  in (2.9) forms

a stationary process, with mean  $E(Y_t) = \mu$  and autocovariance function

$$\begin{aligned}\gamma(s) &= \text{Cov}(Y_t, Y_{t+s}) = E \left[ \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \psi_j \psi_k \epsilon_{t-j} \epsilon_{t+s-k} \right] \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \psi_j \psi_k \gamma_e(j - k + s) \\ &= \sigma_e^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+s},\end{aligned}\tag{2.10}$$

since  $\gamma_e(s) = \sigma_e^2$  if  $s = 0$  and  $\gamma_e(s) = 0$  if  $s \neq 0$ , so that  $\gamma_e(j - k + s) = 0$  when  $k \neq j = s$ .

The following theorem indicates the generality of the class of linear processes in stationary theory, in the sense that every purely non-deterministic stationary process may be expressed in an infinite moving average representation as in (2.9), although the  $\epsilon_t$  need not be independent but merely uncorrelated.

**Theorem 2.1 (Wold's Theorem).** Let  $Y_t, t = 0, \pm 1, \dots$ , be a purely non-deterministic weakly stationary stochastic process with mean  $\mu$ . Then  $Y_t$  may be expressed as

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j},$$

with  $\psi_0 = 1$ ,  $\sum_{j=0}^{\infty} \psi_j^2 < \infty$  where the  $\epsilon_t$  are uncorrelated random variables with  $E(\epsilon_t) = 0$  and  $\text{Var}(\epsilon_t) = \sigma_e^2$  for all  $t$ . The r.v.  $\epsilon_t$  is called the innovation or random shock at time  $t$ .

The general linear representation model given by (2.9) although useful for studying the properties of time series models but is not directly useful in practice to represent a stationary process, since it requires the determination of an infinite number of unknown parameters  $\psi_1, \psi_2, \dots$ , from a finite record of data. Hence we now consider finite parameter models of the same type which will be useful representations. These models essentially express the infinite number of coefficients  $\psi_j$  in terms of a finite number of parameters, and the particular classes of finite parameter models considered are sufficiently general to well approximate the model for any process of the form (2.9).

### Finite Moving Average Processes

A direct way to obtain finite parameter models from the general form (2.9) is simply to restrict the  $\psi_j$  to be zero beyond some lag  $q$ . Thus (with a change of notation), a stationary process  $\{Y_t\}$  is said to be a *moving average process* of order  $q$ , which is denoted as  $\text{MA}(q)$ , if it satisfies

$$Y_t = \mu + \epsilon_t - \sum_{j=1}^q \theta_j \epsilon_{t-j},\tag{2.11}$$

where the  $\epsilon_t$  are independent with mean 0 and variance  $\sigma^2$ . Using the backward shift operator notation  $B$ , the MA( $q$ ) model can be expressed as

$$\begin{aligned} Y_t &= \mu + \epsilon_t - \sum_{j=1}^q \theta_j B^j \epsilon_t \\ &= \mu + (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \epsilon_t = \mu + \theta(B) \epsilon_t, \end{aligned}$$

where  $\theta(B) = 1 - \sum_{j=1}^q \theta_j B^j$ . An MA( $q$ ) process is always stationary, by Theorem 2.1, since  $\sum_{j=0}^{\infty} |\psi_j| = 1 + \sum_{j=1}^q |\theta_j|$  is always finite. The mean of the process is  $\mu = E(Y_t)$ , and the autocovariance function is (from (2.10))

$$\begin{aligned} \gamma(s) &= \text{Cov}(Y_t, Y_{t+s}) = \sigma^2(-\theta_s + \theta_1 \theta_{s+1} + \cdots + \theta_{q-s} \theta_q) \\ &= \sigma^2 \sum_{j=0}^{q-s} \theta_j \theta_{j+s}, \quad s = 0, 1, 2, \dots, q, \end{aligned}$$

and  $\gamma(s) = 0$  for  $s > q$ , where  $\theta_0$  is defined to be  $-1$ . So in particular,  $\gamma(0) = \text{Var}(Y_t) = \sigma^2(1 + \sum_{j=1}^q \theta_j^2)$ . Hence the autocorrelation function of an MA( $q$ ) process (2.11) is

$$\rho(s) = \frac{-\theta_s + \sum_{j=1}^{q-s} \theta_j \theta_{j+s}}{1 + \sum_{j=1}^q \theta_j^2}, \quad s = 0, 1, \dots, q \quad (2.12)$$

and  $\rho(s) = 0$  for  $s > q$ . The prominent feature of the ACF of an MA( $q$ ) process is that it equals zero or “cuts off” after a finite number of lags,  $q$ . Thus in one sense the “memory” of an MA( $q$ ) process is  $q$  periods long. In practice, useful MA( $q$ ) models are those for which the value of  $q$  is typically quite small, such as  $q = 1, 2$ , or  $3$ . For financial time series, usually  $q = 1$  will suffice for modeling, which we will discuss below.

**Example 4 (Moving Average Model of Order 1).** The simplest example of a moving average process is that of order one, MA(1), given by

$$Y_t = \mu + \epsilon_t - \theta \epsilon_{t-1} = \mu + (1 - \theta B) \epsilon_t.$$

Its autocovariances are  $\gamma(0) = \sigma^2(1 + \theta^2)$ ,  $\gamma(1) = -\theta\sigma^2$ , and  $\gamma(s) = 0$  for  $s > 1$ , and hence the ACF is  $\rho(1) = -\theta/(1+\theta^2)$ ,  $\rho(s) = 0$ ,  $s > 1$ . It is easy to show that the largest possible value of  $|\rho(1)| = |\theta|/(1 + \theta^2)$  for an MA(1) process is  $\rho(1) = \pm 0.5$  which occurs when  $\theta = \pm 1$ . This illustrates that in general there are further restrictions on the autocorrelation function  $\rho(s)$  of an MA( $q$ ) process in addition to  $\rho(s) = 0$ ,  $s > q$ . These restrictions are essentially the result of the requirement that the autocorrelation function  $\rho(s)$  of the MA( $q$ ) process be positive semidefinite.

**Example 5 (Seasonal Moving Average).** For time series which may exhibit seasonal correlation, such as certain financial time series with week-day effect, the following type of special seasonal MA model is useful. Let  $S$  denote the number of discrete

time periods per seasonal period (for example,  $S = 4$  for quarterly data of an annual seasonal nature), and define the process  $\{W_t\}$  by

$$W_t = \mu + \epsilon_t - \tau_S \epsilon_{t-S}$$

Then we find that  $\gamma(0) = \sigma^2(1 + \tau_S^2)$ ,  $\gamma(S) = -\tau_S/(1 + \tau_S^2)$ , and  $\rho(j) = 0$  for all  $j \neq 0, \pm S$ . Hence this process exhibits nonzero autocorrelation only at the seasonal lag  $S$ . This model is known as a seasonal MA model of seasonal order 1 (and period  $S$ ), and its generalizations are useful when modeling seasonal differences  $W_t = Y_t - Y_{t-S}$  of an original seasonal time series  $\{Y_t\}$ . In the context of financial data, this could be used to model the so called Monday and Friday effects.

**Example 6 (Autoregressive Model of Order 1).** Consider the process (2.9) with coefficients given by  $\psi_j = \phi^j$ ,  $j = 0, 1, \dots$ , for some  $\phi$  with  $|\phi| < 1$ . Then the process  $Y_t = \mu + \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$  is stationary, since  $\sum_{j=0}^{\infty} |\psi_j| = \sum_{j=0}^{\infty} |\phi|^j = 1/(1 - |\phi|)$  is absolutely convergent for  $|\phi| < 1$ . From (2.10), the autocovariance function of  $\{Y_t\}$  is

$$\gamma(s) = \sigma_{\epsilon}^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+s} = \sigma_{\epsilon}^2 \phi^s \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_{\epsilon}^2 \phi^s}{1 - \phi^2}, \quad s \geq 0.$$

In particular,  $\gamma(0) = \text{Var}(Y_t) = \sigma_{\epsilon}^2/(1 - \phi^2)$ . Hence the autocorrelation function of  $\{Y_t\}$  is  $\rho(s) = \gamma(s)/\gamma(0) = \phi^s$ ,  $s \geq 0$  which declines exponentially (geometrically) as the lag  $s$  increases.

Note that the process  $\{Y_t\}$  defined above actually satisfies the relation  $Y_t = \phi Y_{t-1} + \delta + \epsilon_t$ ,  $t = \dots, -1, 0, 1, \dots$ , with  $\delta = \mu(1 - \phi)$ . This process is called a first-order autoregressive process, and is a special case of the general class of autoregressive processes which is discussed below. Also note that if the value of  $\phi$  in the above is taken as  $\phi = 1$ , then the arguments for stationarity of  $\{Y_t\}$  no longer hold, since  $\sum_{j=0}^{\infty} |\psi_j| = 1 + 1 + \dots$  does not converge. We see that in this case the process is actually a (nonstationary) random walk satisfying the relation  $Y_t = Y_{t-1} + \epsilon_t$ , a common representation of the behavior of the stock prices under the efficient market hypothesis.

### General Order Autoregressive Processes

Consider the general order AR( $p$ ) process  $\{Y_t\}$  defined by

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \delta + \epsilon_t \quad (2.13)$$

or,  $(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)Y_t = \delta + \epsilon_t$ . When  $p = 1$ ,  $Y_t = \phi Y_{t-1} + \delta + \epsilon_t$ , the ACF,  $\rho(v) = \phi^{|v|}$  takes the simple form, as observed earlier.

**Theorem 2.2.** If the roots of  $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p = 0$  are greater than one in absolute value (equivalently roots of  $m^p - \phi_1 m^{p-1} - \dots - \phi_p = 0$  satisfy  $|m_i| < 1$ ), then  $\{Y_t\}$  defined by (2.13) is a stationary AR( $p$ ) process and has a convergent one-sided infinite moving average representation as

$$Y_t = \phi(B)^{-1} \delta + \phi(B)^{-1} \epsilon_t = \mu + \Psi(B) \epsilon_t = \mu + \sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j} \quad (2.14)$$

where  $\mu = E(Y_t) = \delta/(1 - \phi_1 - \dots - \phi_p)$ ,  $\Psi(B) = \sum_{j=0}^{\infty} \Psi_j B^j$ .

The autocovariance  $\gamma(s)$  for the AR( $p$ ) process satisfy the *Yule-Walker equations* given by

$$\begin{aligned}\gamma(s) &= \text{Cov}(Y_t, Y_{t-s}) = \text{Cov}(\phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \delta + \epsilon_t, Y_{t-s}) \\ &= \phi_1 \gamma(s-1) + \phi_2 \gamma(s-2) + \dots + \phi_p \gamma(s-p), \quad s = 1, 2, \dots,\end{aligned}$$

noting again that  $\text{Cov}(\epsilon_t, Y_{t-s}) = 0$ , for  $s > 0$ . Dividing by  $\gamma(0)$ , the ACF  $\rho(s)$  also satisfies the relations

$$\rho(s) = \phi_1 \rho(s-1) + \phi_2 \rho(s-2) + \dots + \phi_p \rho(s-p), \quad s = 1, 2, \dots. \quad (2.15)$$

The Yule-Walker equations (2.15) for  $s = 1, 2, \dots, p$  are particularly useful for determining the autoregressive parameters  $\phi_1, \dots, \phi_p$  in terms of the autocorrelations. Note that these equations can be expressed in matrix form as  $P_p \phi = \rho$ , where

$$= \begin{bmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(p-1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(p-2) \\ \vdots & & \ddots & & \vdots \\ \rho(p-1) & \cdots & & \rho(1) & 1 \end{bmatrix}, \quad \phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}, \quad \rho = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(p) \end{bmatrix} \quad (2.16)$$

These equations can be used to solve for  $\phi_1, \dots, \phi_p$  in terms of  $\rho(1), \dots, \rho(p)$ , with solution  $\phi = P_p^{-1} \rho$ . Note also, that the variance of the process,  $\gamma(0) = \text{Var}(Y_t)$ , and  $\sigma^2 = \text{Var}(\epsilon)$ , are related by

$$\gamma(0) = \text{Cov}(Y_t, \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \delta + \epsilon_t) = \phi_1 \gamma(1) + \dots + \phi_p \gamma(p) + \sigma^2,$$

since  $\text{Cov}(Y_t, \epsilon_t) = \sigma^2$ . Hence  $\sigma^2$  can be determined in terms of  $\gamma(0)$  and the autocorrelations  $\rho(s)$  through the Yule-Walker equations and use of

$$\sigma^2 = \gamma(0) - (\phi_1 \gamma(1) + \dots + \phi_p \gamma(p)) = \gamma(0)(1 - \phi_1 \rho(1) - \dots - \phi_p \rho(p)). \quad (2.17)$$

Also note that a measure of the strength of association or “predictability” of  $Y_t$  based on its past values  $Y_{t-1}, \dots, Y_{t-p}$  can be given by the (squared) multiple correlation coefficients of  $Y_t$  with  $(Y_{t-1}, \dots, Y_{t-p})$ , which is  $R^2 = (\gamma(0) - \sigma^2)/\gamma(0) = 1 - (\sigma^2/\gamma(0)) = \phi_1 \rho(1) + \dots + \phi_p \rho(p)$ . Since  $\gamma(0)$  represents the “total” variance of  $Y_t$ , and  $\sigma^2$  may be interpreted as the “unexplained” or error variance (i.e., that portion of the variance not explained by the past values), the quantity  $R^2$  has the usual interpretation as in linear regression of representing the proportion of the total variance of the variable  $Y_t$  that is explained by (auto)regression on the past values  $Y_{t-1}, \dots, Y_{t-p}$ .

### Partial Autocorrelation Function

When an AR model is being fit to observed time series data, the sample version of the Yule-Walker equations (2.15), (2.16), in which the values  $\rho(j)$  are replaced by the

sample ACF values  $r(j)$ , can be solved to obtain estimates  $\hat{\phi}_j$  of the autoregressive parameters. However, initially it will not be known which order  $p$  is appropriate to fit to the data. The sample partial autocorrelation function (PACF) will be seen to be useful for determination of the appropriate order of an AR process as how the autocorrelation function (ACF) is useful for the determination of the appropriate order of an MA process. First, we discuss the theoretical PACF in general.

Suppose  $\{Y_t\}$  is a stationary process, not necessarily an AR process, with ACF  $\rho(j)$ . For general  $k \geq 1$ , consider the first  $k$  Yule-Walker equations for the ACF associated with an AR( $k$ ) process.

$$\rho(j) = \phi_1\rho(j-1) + \phi_2\rho(j-2) + \cdots + \phi_k\rho(j-k), \quad j = 1, 2, \dots, k \quad (2.18)$$

and let  $\phi_{1k}, \phi_{2k}, \dots, \phi_{kk}$  denote the solution for  $\phi_1, \phi_2, \dots, \phi_k$  to these equations. Given the ACF  $\rho(j)$ , (2.18) can be solved for each value  $k = 1, 2, \dots$ , and the quantity  $\phi_{kk}$ , regarded as a function of the lag  $k$ , is called the (theoretical) *partial autocorrelation function* (PACF) of the process  $\{Y_t\}$ . The values  $\phi_{1k}, \phi_{2k}, \dots, \phi_{kk}$  which are the solution to (2.18) are the regression coefficients in the regression of  $Y_t$  on  $Y_{t-1}, \dots, Y_{t-k}$ ; that is, values of coefficients  $b_1, \dots, b_k$  which minimize  $E[(Y_t - b_0 - \sum_{i=1}^k b_i Y_{t-i})^2]$ .

If  $\{Y_t\}$  is truly an AR process of order  $p$ , the  $\phi_{kk}$  will generally be nonzero for  $k \leq p$ , but  $\phi_{kk}$  will always be zero for  $k > p$ . This is so since the ACF  $\rho(j)$  actually satisfies the Yule-Walker equations (2.18) for  $k = p$ , and hence for any  $k > p$  the solution to (2.18) must be  $\phi_{kk} = \phi_1, \dots, \phi_{pk} = \phi_p, \phi_{p+1,k} = 0, \dots, \phi_{kk} = 0$ , where  $\phi_1, \dots, \phi_p$  are the true AR coefficients. Thus, the PACF  $\phi_{kk}$  of an AR( $p$ ) process “cuts off” (is zero) after lag  $p$ , and this property serves to distinguish (identify) an AR( $p$ ) process.

The quantity  $\phi_{kk}$  defined above is called the *partial autocorrelation* at lag  $k$ , since it is actually equal to the partial correlation between the r.v.’s  $Y_t$  and  $Y_{t-k}$  adjusted for the intermediate variables  $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$ , and  $\phi_{kk}$  measures the correlation between  $Y_t$  and  $Y_{t-k}$  after adjusting for the effects of  $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$  (or the correlation between  $Y_t$  and  $Y_{t-k}$  not accounted for by  $Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}$ ).

Based on a sample  $Y_1, \dots, Y_T$  from a process  $\{Y_t\}$ , the sample PACF value at lag  $k$ ,  $\hat{\phi}_{kk}$ , is obtained as the solution to the sample Yule-Walker equations of order  $k$ ,

$$r(j) = \hat{\phi}_{tk}r(j-1) + \cdots + \hat{\phi}_{kk}r(j-k), \quad j = 1, \dots, k, \quad (2.19)$$

for each  $k = 1, 2, \dots$ . These are the same form as the equations (2.18) but with sample ACF values  $r(j)$  used in place of the theoretical ACF values  $\rho(j)$ . In matrix notation similar to that used in (2.16), the solution is  $\hat{\phi} = R_k^{-1}r$ . Now under the assumption that the process  $\{Y_t\}$  is an AR of order  $p$ , the estimated PACF values  $\hat{\phi}_{kk}$  for lags  $k > p$  are approximately independently and normally distributed for large  $T$ , with mean  $E(\hat{\phi}_{kk}) = 0$  and St.Dev.  $\hat{\phi}_{kk} = 1/\sqrt{T}$  for  $k > p$ . These facts concerning the sampling properties of the sample PACF  $\hat{\phi}_{kk}$  can be used to assess whether the coefficients in estimated AR models may be treated as close to zero after some lag  $p$  (e.g. if  $|\hat{\phi}_{kk}| < 2/\sqrt{T}$  for  $k > p$ ), and hence can be useful in the selection of an appropriate

order  $p$  for fitting an AR model to sample data.

### Linear Models for Nonstationary Time Series

Often, in practice, series  $\{Y_t\}$  will be encountered which are nonstationary. One type of nonstationary series that occurs commonly, are series that exhibit some homogeneous behavior over time in the sense that, except for local level or perhaps local level and local trend, one segment of the series behaves much like any other part of the series. In cases where the series  $\{Y_t\}$  exhibits such homogeneous nonstationarity, it may be found that the first difference of the series,  $(1 - B)Y_t = Y_t - Y_{t-1}$ , is a stationary series. For seasonal nonstationary time series that exhibit homogeneous behavior apart from a seasonal mean level or trend, with seasonal period  $S$ , the seasonal difference of the series,  $(1 - B^S)Y_t = Y_t - Y_{t-S}$ , may be stationary. More generally, a useful class of models for this type of homogeneous nonstationary time series  $\{Y_t\}$  is obtained by assuming that the  $d$ th difference,  $W_t = (1 - B)^d Y_t$ , is a stationary series, and  $W_t$  can be represented by an ARMA( $p, q$ ) model. The most common case is  $d = 1$ , for financial time series.

### Autoregressive Integrated Moving Average Processes

We consider models for  $\{Y_t\}$  such that  $W_t = (1 - B)^d Y_t$  is a stationary ARMA( $p, q$ ) process. The process  $\{Y_t\}$  is then said to be an *autoregressive integrated moving average process* of order  $(p, d, q)$ , denoted as ARIMA( $p, d, q$ ). Since the  $d$ th differences  $W_t$  form an ARMA( $p, q$ ) process, they satisfy  $\phi(B)W_t = \delta + \theta(B)\varepsilon_t$ . So the ARIMA process  $Y_t$  is generated by  $\phi(B)(1 - B)^d Y_t = \delta + \theta(B)\varepsilon_t$ , or

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 - B)^d Y_t = \delta + (1 - \theta_1 B - \dots - \theta_q B^q)\varepsilon_t, \quad (2.20)$$

where  $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$  and  $\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$ .

We first briefly comment on the use of the term “integrated” to describe such a process as given above. Consider the case  $d = 1$  so that the process  $W_t = Y_t - Y_{t-1}$  is a stationary ARMA( $p, q$ ) process. Then the stationary series  $W_t$  must be summed or “integrated” to obtain the process  $\{Y_t\}$ ; that is,

$$Y_t = W_t + Y_{t-1} = W_t + W_{t-1} + Y_{t-2} = \dots = W_t + W_{t-1} + \dots + W_1 + Y_0,$$

or  $Y_t = W_t + W_{t-1} + W_{t-2} + \dots$ . Using the differencing operator notation, we have that  $W_t = (1 - B)Y_t$  yields  $Y_t = (1 - B)^{-1}W_t = (1 + B + B^2 + \dots)W_t = \sum_{j=0}^{\infty} W_{t-j}$ . In fact, the summation must be truncated at some finite time in the past with a known initial value. Hence, the process  $\{Y_t\}$  is formed by “integrating” a stationary ARMA( $p, q$ ) process  $\{W_t\}$ , and so  $Y_t$  is called an *integrated process* (of order  $d = 1$ ), more precisely, an autoregressive integrated moving average process, ARIMA( $p, 1, q$ ) in this case with  $d = 1$ . More generally, with  $d > 1$ , the stationary process  $W_t = (1 - B)^d Y_t$  must be summed  $d$  times to obtain the process  $Y_t$ . Note that the random walk process  $Y_t = Y_{t-1} + \delta + \varepsilon_t$  is the simplest example of an integrated process, and is also the most commonly observed. Since then the first differences  $W_t = Y_t - Y_{t-1} = \delta + \varepsilon_t$  form a (stationary) white noise process, i.e., the process  $\{Y_t\}$  is ARIMA(0,1,0).

The differencing operation is useful in eliminating “random changes in level” or “random trends” in a (homogeneous) nonstationary series  $\{Y_t\}$ , thus transforming the series to a stationary series  $W_t = Y_t - Y_{t-1}$ . Differencing is a more general method of removing trend behavior in a series than to fit a deterministic trend function to the series by regression methods. Differencing is appropriate to apply to processes which contain a stochastic trend component (not a deterministic trend component), and stochastic trend behavior is often more likely to be present in financial time series than deterministic trend behavior.

Some observations are in order. First, if a series does not require differencing, over differencing will lead to unnecessarily more complicated models. Second, the differencing can sometimes wipe out all the information leaving series with white noise behavior. Third, there are identifiability issues in ARMA models; it is possible to model the same series with various ARMA structures, but we will follow the principle of parsimony in modeling.

**Example 7 (Autoregressive Integrated Moving Average Model).** Consider a process  $Y_t$  that may be thought to be composed of a trend component  $U_t$  and a noise component  $N_t$ ,  $Y_t = U_t + N_t$ , where we assume that the (random) trend component follows a random walk model (possibly with drift),  $(1 - B)U_t = \delta + a_t$ , and we assume that  $N_t$  is a stationary process, to be specific assume that  $N_t$  is an AR(1) process  $(1 - \phi B)N_t = b_t$ , where  $\{a_t\}$  and  $\{b_t\}$  are independent white noise processes with zero means and variances  $\sigma_a^2$  and  $\sigma_b^2$ , respectively. Then, applying the first differencing operator, we have  $(1 - B)Y_t = (1 - B)U_t + (1 - B)N_t = \delta + a_t + (1 - B)N_t$ , and hence

$$\begin{aligned}(1 - \phi B)(1 - B)Y_t &= (1 - \phi B)\delta + (1 - \phi B)a_t + (1 - \phi B)(1 - B)N_t \\ &= (1 - \phi)\delta + (1 - \phi B)a_t + (1 - B)b_t.\end{aligned}\quad (2.21)$$

It can be shown that  $Z_t = (1 - \phi B)a_t + (1 - B)b_t$  is the sum of two independent MA(1) processes and thus an MA(1) process, so that  $Y_t$  is an ARIMA(1,1,1) process. Thus ARIMA  $(p, d, q)$  can arise naturally when several stochastic elements with different behaviors are added together.

This general observation concerning the sum of two ARMA processes gives rise to the following result.

**Theorem 2.3 (Aggregation).** If  $X_t$  is an ARMA( $p_1, q_1$ ) process and  $Y_t$  is an ARMA( $p_2, q_2$ ) process, with  $X_t$  and  $Y_t$  independent process, then  $Z_t = X_t + Y_t$  follows an ARMA( $p, q$ ) model, where  $p \leq p_1 + p_2$  and  $q \leq \max(p_1 + q_2, p_2 + q_1)$ .

These types of models have potential applications in studying aggregate market behavior of several series. Also where the observed series may be viewed as the sum of the true process of interest plus observational error or noise (as in the classical “signal-plus-noise” model), and in the use of structural component models where an observed time series is represented as the sum of unobservable components that correspond to factors such as trend, seasonality, stationary variations, and so on. We will provide financial time series examples later in the chapter.

### 2.3.3 Specification and Estimation of ARIMA Models

In previous sections, theoretical linear models for stochastic processes and properties implied by these models were discussed. We now consider the problem of identifying appropriate models and fitting them to time series data. The following four steps highlight the basic stages in the proposed model building and testing procedure. Assume that a general class of models has been postulated for consideration for the data, such as the generating class of (linear) ARIMA time series models. Then the four stages that apply for any model building are:

1. Model Specification or Identification—specify or identify specific models to be entertained as appropriate, based on preliminary examination of certain fundamental statistical features of the data, such as features of the sample ACF and PACF for time series data.
2. Parameter Estimation—for the specified model or models, estimate parameters in the tentatively entertained models efficiently, usually by the methods of maximum likelihood (ML) or least squares (LS).
3. Model Checking—perform diagnostic checks on the adequacy of the estimated model, usually by examination of various features of the residuals from the fitted model.
4. Model Validation—confirm that the model is appropriate for out-of-sample prediction. When multiple models are tried out on a single set of data, it may result in data snooping bias. By keeping the data for building the model and the data for validating the model separate, we avoid the snooping bias.

If the estimated model is assessed as being adequate, it can then be used for forecasting or other purposes of interest. Otherwise, one would return to step one and specify an alternate model or models for further consideration. We now discuss some aspects of the model specification procedures in some detail, and point out the key features.

#### Model Specification

Given the time series data  $Y_1, \dots, Y_T$ , basic time series plotting of the series is a fundamental step in the initial model building, along with other data plots that might seem informative. We will then consider the sample ACF and PACF of the original series  $Y_t$ , and if necessary also the sample ACFs and PACFs of certain transformed series derived from the series  $Y_t$ , such as first and possibly second differences, seasonal differences, logarithms or other instantaneous transformation of the original series, residuals from regression to remove deterministic seasonal component or linear trend, and so on. That is, if the original time series appears to be nonstationary we consider transformations, such as differencing or residuals from regression methods, of the original series which will yield a *stationary* series. More formal procedures to ‘test’ for certain unit-root type nonstationarity can also be considered, and

will be discussed later in the context of estimation for AR models. While dealing with financial time series the following considerations are usually made: for example,  $r_t = \text{return} = \ln(P_t) - \ln(P_{t-1})$ , where  $P_t$  is the price of the stock, the differencing of the log price is naturally considered and for volume  $V_t$ , because of its size is usually logged  $v_t = \log(V_t)$  to avoid heteroscedasticity.

The sample ACF and PACF of the stationary series are then examined and can be compared against features of the theoretical ACFs and PACFs of various types of ARMA models to find an appropriate model for the observed series on the basis of good correspondence in features between sample and theoretical correlation functions. Of course, the sample ACF values that are defined in (2.5) are only *estimates* of the theoretical ACF values  $\rho(j)$  and hence are subject to sampling errors. Thus, we must recognize that the sample ACF of an observed series will never correspond in all details to an underlying theoretical ACF, and we need to look for correspondence in broad features. To properly interpret the sample ACF [ $\hat{\rho}(j)$ ], we should recall the basic sampling properties of the estimates  $\hat{\rho}(j)$  mentioned earlier. In particular, under the assumption that  $\rho(j) = 0$  for all  $j > q$ , we have  $E[\hat{\rho}(j)] = 0$  and  $\text{St.Dev.}[\hat{\rho}(j)] = \frac{1}{\sqrt{T}}[1 + 2 \sum_{i=1}^q \hat{\rho}(i)^2]^{1/2}$  for  $j > q$  and the  $\hat{\rho}(j)$  are approximately normally distributed, for moderate and large  $T$ , a result given in (2.7).

Similarly, we need to know about the sampling properties of the sample PACF  $\{\hat{\phi}_{kk}\}$ , where  $\hat{\phi}_{kk}$  is obtained as the solution, for the  $k$ th coefficients, to the sample Yule-Walker equations of order  $k$  in (2.18). It has been proven that under the assumption that the process  $\{Y_t\}$  is an AR of order  $p$ , then the estimated PACFs for lags  $p+1$  and higher are approximately independently and normally distributed, with  $E[\hat{\phi}_{kk}] = 0$  and  $\text{St.Dev.}[\hat{\phi}_{kk}] = 1/\sqrt{T}$  for  $k > p$ .

These facts may be used to assess whether the last coefficient in an estimated AR model is essentially zero. In particular, the limits  $\pm 2 \text{ St.Dev.}[\hat{\phi}_{kk}] = \pm 2/\sqrt{T}$  can typically be used to assess the ‘significance’ of sample PACF values  $\hat{\phi}_{kk}$ ; if the sample PACF possesses the feature that the sample values  $[\hat{\phi}_{kk}]$  are all relatively small (close to zero) compared to  $\pm 2/\sqrt{T}$  for all  $k > p$ , where  $p$  is a ‘low’ order (such as  $p = 1, 2$ , or 3), then an AR model of orders  $p$  may be suggested as appropriate.

We review some general characteristics of the sample and theoretical ACF and PACF, which might be useful to keep in mind in the initial model identification step.

1. The sample ACF of a unit-root nonstationary process will generally fail to dampen out sufficiently fast as the lag increases. The process may be nonstationary due to deterministic or stochastic trend in mean, and we might consider first differences of the series or residuals from the regression on a linear trend. The process may also exhibit seasonal nonstationary behavior, in which case seasonal components can be considered. Thus the order of differencing depends on the series being trend or seasonal nonstationary.
2. The sample ACF [ $\hat{\rho}(j)$ ] of a typical stationary process should dampen out sufficiently fast; that is,  $\hat{\rho}(j) \rightarrow 0$  at a sufficiently fast rate as  $j$  increases (subject to sampling error), since the corresponding theoretical ACF  $\rho(j)$  of stationary

ARMA processes approach zero exponentially fast. We also briefly mention the general behavior of the (stationary) MA( $q$ ), AR( $p$ ), and ARMA( $p, q$ ) processes.

- MA( $q$ ) process has ACF[ $\rho(j)$ ] that will ‘cut off’ after  $q$  lags; that is,  $\rho(j) = 0$  for  $j > q$ , so this will tend to be the behavior of the sample ACF  $\hat{\rho}(j)$  from an MA( $q$ ) process (subject to sampling variability). For example, the MA(1) process has  $\rho(j) = 0$  for  $j > 1$ .
  - AR( $p$ ) process has ACF[ $\rho(j)$ ] which can have exponential decay or damped sinusoidal behavior, or a mixture of both, so this will tend to be the behavior of the sample ACF. For example, the AR(1) process has  $\rho(j) = \phi\rho(j - 1)$ , or  $\rho(j) = \phi^j$ , for all  $j \geq 1$ .
  - ARMA( $p, q$ ) process has ACF [ $\rho(j)$ ] which can have an irregular pattern for the first  $q$  lags and then similar behavior to a corresponding AR( $p$ ) process for lags than  $q$ . For example, the ARMA(1,1) has  $\rho(j) = \phi\rho(j - 1)$ , or  $\rho(j) = \phi^{j-1}\rho(1)$ , for  $j \geq 2$ , which is similar behavior to the AR(1) after lag 1.
3. The sample PACF  $\hat{\phi}_{kk}$  can be useful to identify the order  $p$  of a finite order AR( $p$ ) process, since for an AR( $p$ ) process we know that the theoretical PACF  $\phi_{kk}$  ‘cuts off’ after lag  $p$ ; that is,  $\phi_{kk} = 0$  for all  $k > p$ . For the sample PACF  $\hat{\phi}_{kk}$  of an AR( $p$ ), we have  $E[\hat{\phi}_{kk}] = 0$  and  $\text{St.Dev.}[\hat{\phi}_{kk}] = 1/\sqrt{T}$ , for  $k > p$ . The sample PACF may also be helpful in combination with the sample ACF to suggest an MA( $q$ ) or ARMA( $p, q$ ) model, since for these the PACF  $\phi_{kk}$  will not become zero after some distinct (small) lag, but instead its behavior will be dominated by exponential decay or damped sinusoids or a mixture of both.

Although the sample autocorrelation and partial autocorrelation functions are extremely useful in model identification, there are sometimes cases involving mixed ARMA models where they will not provide unambiguous results. There has been considerable interest in developing additional tools for use at the model identification stage. Interested readers can refer to standard texts in time series such as Shumway and Stoffer (2011)[241].

*Use of model selection criteria.* Another approach to model selection is the use of information criteria such as AIC proposed by Akaike (1974) [3] or the BIC of Schwarz (1978) [237]. In the implementation of this approach, a range of potential models is estimated and for each, a criterion such as AIC or BIC (normalized by sample size  $T$ ), given by

$$\begin{aligned} \text{AIC}_{p,q} &= \frac{-2 \log(\text{maximized likelihood}) + 2n^*}{T} \approx \log \hat{\sigma}_\epsilon^2 + \frac{n^* 2}{T} \\ \text{BIC}_{p,q} &= \log \hat{\sigma}_\epsilon^2 + \frac{n^* \log T}{T} \end{aligned} \quad (2.22)$$

is evaluated, where  $\hat{\sigma}_\epsilon^2$ , and  $n^* = p + q + 1$  denote the number of parameters estimated in the model, including constant term. In the above criteria, the first term essentially corresponds to minus  $2/T$  times the log of the maximized likelihood, while

the second term is a “penalty factor” for inclusion of additional parameters in the model. In the information criteria approach, models which yield a minimum value for the criterion are to be preferred, as these criteria penalize for overfitting.

### Estimation of Model Parameters

#### Method of Moments

Once the orders  $p, d, q$ , of an ARIMA model (2.20), rewritten as

$$\phi(B)(1 - B)^d Y_t = \theta_0 + \theta(B)\varepsilon_t \quad (2.23)$$

for the series  $Y_1, \dots, Y_T$  have been tentatively identified, it is often helpful to obtain preliminary (initial) estimates of the parameters  $\phi_1, \dots, \phi_p, \theta_0, \dots, \theta_q$  and  $\sigma^2$  in the model. These may be obtained by the *method of moments* procedure, which is based on the sample ACF  $\hat{\rho}(j), j = 1, \dots, p + q$ . The method of moments estimates have the appeal of being relatively simple to compute, although these initial estimates are not statistically efficient (except for a pure AR); they can usefully serve as initial values to obtain more efficient estimates using iterative numerical procedures, namely maximum likelihood.

Assuming that  $W_t = (1 - B)^d Y_t$  follows an ARMA( $p, q$ ) model, we know that the ACF  $\rho(j)$  of  $W_t$  satisfies the “generalized” Yule-Walker equations in (2.15). Hence we can obtain initial estimates for the AR parameters  $\phi_1, \dots, \phi_p$  by solving the sample version of these equations for  $j = q + 1, \dots, q + p$ ,

$$\hat{\rho}(j) - \sum_{i=1}^p \hat{\phi}_i \hat{\rho}(j-i) = 0, \quad j = q + 1, \dots, q + p. \quad (2.24)$$

Note that these equations could also be useful in identifying the orders  $(p, q)$  of an ARMA model. Since the mean  $\mu_w = E(W_t)$  of the stationary process  $W_t$  in (2.23) is  $\mu_w = \theta_0/(1 - \phi_1 - \dots - \phi_p)$ , we estimate  $\mu_w$  by sample mean  $\hat{\mu}_w = \bar{W}$  and hence the constant term  $\theta_0$  by  $\hat{\theta}_0 = (1 - \hat{\phi}_1 - \dots - \hat{\phi}_p)\hat{\mu}_w$ .

Initial estimates of the MA parameters could then be based on the autocovariances of the ‘derived’ MA( $q$ ) process,  $\bar{W}_t = \phi(B)W_t = \theta_0 + \theta(B)\varepsilon_t$ . The autocovariances of  $\bar{W}_t$  are related to those of  $W_t$  by

$$\gamma(s) = \text{Cov}(\bar{W}_t, \bar{W}_{t+s}) = \sum_{i=0}^p \sum_{j=0}^p \phi_i \phi_j \gamma(s+i-j), \quad s = 0, 1, \dots \quad (2.25)$$

with  $\phi_0 = -1$ , where  $\gamma(j) = \text{Cov}(W_t, W_{t+j})$ . But since  $\bar{W}_t = \theta_0 + \theta(B)\varepsilon_t$  also satisfies the MA( $q$ ) model

$$\gamma_*(s) = \sigma_\varepsilon^2(-\theta_s + \theta_1 \theta_{s+1} + \dots + \theta_{q-s} \theta_q), \quad s = 1, 2, \dots, q, \quad (2.26)$$

with  $\gamma_*(0) = \sigma_\varepsilon^2(1 + \theta_1^2 + \dots + \theta_q^2)$ . Hence, to obtain initial estimates of  $\theta_1, \dots, \theta_q$  and

$\sigma_\varepsilon^2$  we first form the estimated autocovariances  $\hat{\gamma}_*(s)$  of  $\bar{W}_t$  based on (2.31) using the initial  $\hat{\phi}_i$  and the sample autocovariance  $\hat{\gamma}(j)$  of  $W_t$ . We then substitute the  $\hat{\gamma}_*(s)$  for  $\gamma_*(s)$  in the equations (2.32), and solve for parameters  $\theta_1, \dots, \theta_q$  and  $\sigma_\varepsilon^2$ . The resulting equations are nonlinear in the  $\theta_i$  and hence must be solved by an iterative numerical procedure (except for the MA(1) case where an explicit solution is available).

An alternative, but essentially equivalent way to obtain initial estimates for the MA parameters  $\theta_1, \dots, \theta_q$  and  $\sigma_\varepsilon^2$  is to consider the first  $q+1$  autocovariance equations for the ARMA( $p, q$ ) model,

$$\gamma(j) = \phi_1\gamma(j-1) + \dots + \phi_p\gamma(j-p) - \sigma_\varepsilon^2(\theta_j + \theta_{j+1}\Psi_1 + \dots + \theta_q\Psi_{q-j}), \quad j = 0, 1, \dots, q$$

where  $\theta_0 = -1$ . Then substitute the initial estimates  $\hat{\phi}_1, \dots, \hat{\phi}_p$  into these equations, along with the sample autocovariance  $\hat{\gamma}(0), \dots, \hat{\gamma}(p)$  of the series  $W_t$ , and solve for the initial estimates of  $\theta_1, \dots, \theta_q$  and  $\sigma_\varepsilon^2$ . We now illustrate these methods with examples of a few simple models.

**Example 8 (AR(1) Model).** In this case, the only equation needed for the AR parameter  $\phi_1$  is the first-order Yule-Walker equation, which gives the estimate  $\hat{\phi}_1 = \hat{\rho}(1)$ , and then  $\gamma(0) = (1 - \hat{\phi}_1^2)\gamma(0) \equiv \sigma_\varepsilon^2$ , so that  $\hat{\sigma}_\varepsilon^2 = (1 - \hat{\phi}_1^2)\hat{\gamma}(0)$  is the corresponding estimate of  $\hat{\sigma}_\varepsilon^2$ .

**Example 9 (ARMA(1,1) Model).** The equation for  $\phi_1$  comes from  $\rho(2) = \phi_1\rho(1)$ , so that the moment estimate is  $\hat{\phi}_1 = \hat{\rho}(2)/\hat{\rho}(1)$ . Then, in one approach, we can form  $\hat{\gamma}(0) = (1 + \hat{\phi}_1^2)\hat{\gamma}(0) - 2\hat{\phi}_1\hat{\gamma}(1)$ ,  $\hat{\gamma}(1) = (1 + \hat{\phi}_1^2)\hat{\gamma}(1) - \hat{\phi}_1\hat{\gamma}(0) - (-\hat{\phi}_1\hat{\gamma}(2))$ , and solve the equations,  $\hat{\gamma}(0) = \sigma_\varepsilon^2(1 + \theta_1^2)$ ,  $\hat{\gamma}(1) = -\sigma_\varepsilon^2\theta_1$ . We have  $\hat{\rho}(1) \equiv \hat{\gamma}(1)/\hat{\gamma}(0) = -\hat{\theta}_1/(1 + \hat{\theta}_1^2)$ , which leads to the solution to a quadratic equation as  $\hat{\theta}_1 = (-1 + \sqrt{1 - 4\hat{\rho}(1)^2})/(2\hat{\rho}(1))$  (provided  $|\hat{\rho}(1)| < 0.5$ , in which case we get an invertible value with  $|\hat{\theta}_1| < 1$ , and then  $\hat{\sigma}_\varepsilon^2 = \hat{\gamma}(0)/(1 + \hat{\theta}_1^2)$ .

In the alternate approach, we consider directly the sample versions of the first two autocovariance equations for the ARMA(1,1) process,

$$\hat{\gamma}(0) - \hat{\phi}_1\hat{\gamma}(1) = \sigma_\varepsilon^2[1 - \theta_1(\hat{\phi}_1 - \theta_1)]; \quad \hat{\gamma}(1) - \hat{\phi}_1\hat{\gamma}(0) = -\sigma_\varepsilon^2\theta_1$$

We can eliminate  $\sigma_\varepsilon^2$  from the above two equations and obtain the initial estimate  $\hat{\theta}_1$  by solving a quadratic equation. Then the initial estimate of  $\sigma_\varepsilon^2$  is  $\hat{\sigma}_\varepsilon^2 = [\hat{\gamma}(0) - \hat{\phi}_1\hat{\gamma}(1)]/[1 - \hat{\theta}_1(\hat{\phi}_1 - \hat{\theta}_1)]$ . This same procedure would specialize to the MA(1) model, with the simplification that the first step of estimation of  $\phi_1$  is eliminated and we use simply  $\hat{\gamma}(0) = \sigma_\varepsilon^2(1 + \theta_1^2)$  and  $\hat{\gamma}(1) = -\sigma_\varepsilon^2\theta_1$ .

**Example 10 (AR( $p$ ) Model).** For the AR( $p$ ) model, the method of moments estimates of the  $\phi_1$  parameters come directly from solving the system of the first  $p$  sample Yule-Walker equations in [25].

In matrix notation, these equations can be written as  $R_p\hat{\phi} = R_p^{-1}r_p$ . These estimates are generally known as the *Yule-Walker estimates* for the AR( $p$ ) model. The corresponding estimate of  $\sigma_\varepsilon^2$  is based on equation  $\sigma_\varepsilon^2 = \gamma(0)[1 - \phi_1\rho(1) - \dots - \phi_p\rho(p)]$ , so that the estimate is  $\hat{\sigma}_\varepsilon^2 = \hat{\gamma}(0)[1 - \hat{\phi}_1\hat{\rho}(1) - \dots - \hat{\phi}_p\hat{\rho}(p)]$ , where  $\hat{\gamma}(0)$  is the sample variance of the series  $W_t$ .

The method of moments was discussed as a method for preliminary parameter estimation and also in connection with model specification. Now we will focus on more efficient estimation methods, and the general approach will be the use of Gaussian maximum likelihood (ML) methods, and the closely related method of least squares estimation. This discussion is made brief as there are excellent books such as Shumway and Stoffer (2011) [241] on these topics. For ease of presentation, we will first examine ML estimation for pure autoregressive models, and extend to the general ARMA model in later sections because certain estimation features are relatively simple for the AR model compared to the general ARMA model. It must be noted that the least-squares estimator does not guarantee that the roots of the characteristic function lie outside the unit circle, but the Yule-Walker estimator does.

### Conditional Likelihood Estimation

We consider parameter estimation for the AR( $p$ ) process with mean  $\mu$ , or constant term  $\delta$ .

$$Y_t = \delta + \sum_{i=1}^p \phi_i Y_{t-i} + \varepsilon_t, \text{ or } Y_t - \mu = \sum_{i=1}^p \phi_i (Y_{t-i} - \mu) + \varepsilon_t, \quad (2.27)$$

where  $\delta = \mu(1 - \phi_1 - \dots - \phi_p)$ , and we assume the  $\varepsilon_t$  are i.i.d. normal  $N(0, \sigma^2)$ . Given a sample realization of  $T$  observations  $Y_1, Y_2, \dots, Y_T$ , the conditional likelihood function (i.e., the conditional p.d.f of  $Y_{p+1}, \dots, Y_T$ , given the first  $p$  observations  $Y_1, \dots, Y_p$ , or say given  $Y_p = (Y_1, \dots, Y_p)'$ ) is given by

$$f(Y_{p+1}, \dots, Y_T | Y_p; \mu, \phi, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(T-p)/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{t=p+1}^T (Y_t - \mu - \phi_1(Y_{t-1} - \mu) - \dots - \phi_p(Y_{t-p} - \mu))^2 \right]$$

where  $\phi$  denotes  $(\phi_1, \dots, \phi_p)$ . The *conditional maximum likelihood estimates* (CMLE) of  $\mu$ ,  $\phi$  and  $\sigma^2$  are values of the parameters that maximize the conditional likelihood function.

First consider the AR(1) case. Then the conditional (on knowing  $y_1$ ) sum of squares function to be minimized is

$$S_*(\mu, \phi) = \sum_{t=2}^T (Y_t - \mu - \phi(Y_{t-1} - \mu))^2 = \sum_{t=2}^T (Y_t - \delta - \phi Y_{t-1})^2$$

To minimize, we find the derivatives with respect to  $\mu$  and  $\phi$ . Taking  $\bar{Y}_{(0)} = \frac{1}{T-1} \sum_{t=2}^T Y_t$  and  $\bar{Y}_{(1)} = \frac{1}{T-1} \sum_{t=2}^T Y_{t-1}$ , the least squares estimates (LSE) of  $\phi$  and  $\delta = \mu(1 - \phi)$  are given by

$$\hat{\phi} = \frac{\sum_{t=2}^T (Y_t - \bar{Y}_{(0)})(Y_{t-1} - \bar{Y}_{(1)})}{\sum_{t=2}^T (Y_{t-1} - \bar{Y}_{(1)}^2)}, \quad \hat{\delta} = \bar{Y}_{(0)} - \hat{\phi} \bar{Y}_{(1)}$$

Now since  $\bar{Y}_{(1)} \approx \bar{Y}_{(0)} \approx \bar{Y}$ , where  $\bar{Y} = (1/T) \sum_{t=1}^T Y_t$  is the overall sample mean, we have approximately that  $\hat{\mu} \approx \bar{Y}$ , which is an intuitively appealing estimator. Then using  $\hat{\mu} \approx \bar{Y}$  we obtain

$$\hat{\phi} = \frac{\sum_{t=2}^T (Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=2}^T (Y_{t-1} - \bar{Y})^2} \quad (2.28)$$

This estimator  $\hat{\phi}$  is exactly the estimator one would obtain if the AR equation (adjusted for the overall sample mean  $\bar{Y}$ )  $Y_t - \bar{Y} = \phi(Y_{t-1} - \bar{Y}) + \varepsilon_t, t = 2, \dots, T$ , is treated as an ordinary regression with  $Y_{t-1} - \bar{Y}$  as the “independent” variable. A further approximation that is noted from [50] is  $\hat{\phi} = \hat{\rho}(1)$ , which is also an appealing estimator since  $\rho(1) = \phi$  for the AR(1) process. This estimator of  $\hat{\phi}$  is called the Yule-Walker estimator since it is the solution to the first order sample Yule-Walker equation for  $p = 1$ .

Higher order AR( $p$ ) processes may also be estimated by conditional least squares in a straight forward manner. The estimates are simply obtained by ordinary least squares regression of the “dependent” variable  $Y_t$  on the “independent” variables  $Y_{t-1}, \dots, Y_{t-p}$  and a constant term. This can be obtained equivalently by ordinary least squares regression in terms of the “mean-adjusted” variables  $Y_{t-i} - \bar{Y}$ , where  $\bar{Y} = 1/(T-p) \sum_{t=p+1}^T Y_t$ . Similar to the “regression” method one fits the data to the model  $Y_t - \bar{Y} = \phi_1(Y_{t-1} - \bar{Y}) + \dots + \phi_p(Y_{t-p} - \bar{Y}) + \varepsilon_t, t = p+1, \dots, T$ , treating this as if it were an ordinary linear regression with “dependent” variable  $Y_t - \bar{Y}$  and “independent” variables  $Y_{t-i} - \bar{Y}, i = 1, \dots, p$ . Except for the approximation  $\hat{\mu} = \bar{Y}$ , this approach will produce the conditional least squares estimates. A second approximate method consists of solving the first  $p$  sample Yule-Walker equations for  $\hat{\phi}_1, \dots, \hat{\phi}_p$ . In matrix notation these sample Yule-Walker equations are  $R_p \hat{\phi} = r$  with solution  $\hat{\phi} = R_p^{-1} r$ , which is the Yule-Walker estimator of  $\phi$ .

For moderately large sample size  $T$ , both conditional least squares estimation (“regression” methods) and Yule-Walker estimation will give approximately the same estimated values of  $\phi_1, \dots, \phi_p$ . This is because the sample sums of squares and cross-products that are involved in both estimation methods differ only by “end-effects” involving the treatment of the first and/or last few observations in the series. That is, in the “regression” method the estimator is  $\hat{\phi} = (X'X)^{-1} X'Y = C_*^{-1} c_*$ , where

$$X = \begin{bmatrix} Y_p - \bar{Y} & Y_{p-1} - \bar{Y} & \dots & Y_1 - \bar{Y} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{T-1} - \bar{Y} & Y_{T-2} - \bar{Y} & \dots & Y_{T-p} - \bar{Y} \end{bmatrix}, \quad Y = \begin{bmatrix} Y_{p+1} - \bar{Y} \\ \vdots \\ Y_T - \bar{Y} \end{bmatrix}.$$

with  $C_* = (1/T)X'X$  and  $c_* = (1/T)X'Y$ . Now, for example, the diagonal elements of  $C_*$  are equal to  $1/T$  times  $\sum_{t=p+1}^T (Y_{t-1} - \bar{Y})^2, \sum_{t=p+1}^T (Y_{t-2} - \bar{Y})^2, \dots, \sum_{t=p+1}^T (Y_{t-p} - \bar{Y})^2$ , which are all nearly equal to  $c(0) = (1/T) \sum_{t=1}^T (Y_t - \bar{Y})^2$ , and similar for the cross-product terms. Also, the elements of  $c_*$  are equal to  $1/T$  times  $\sum_{t=p+1}^T (Y_{t-1} - \bar{Y}) (Y_t - \bar{Y})$ .

$\bar{Y})(Y_t - \bar{Y}), \sum_{t=p+1}^T (Y_{t-2} - \bar{Y})(Y_t - \bar{Y}), \dots, \sum_{t=p+1}^T (Y_{t-p} - \bar{Y})(Y_t - \bar{Y})$ , which are nearly equal to  $c(1), c(2), \dots, c(p)$ , respectively. Hence,  $C_*$  and  $c_*$  differ from  $C = c(0)R$  and  $c = c(0)r = [c(1), c(2), \dots, c(p)]'$  by only small “end-effects” which should be negligible for large  $T$ . Therefore, the LSE  $\hat{\phi} = C_*^{-1}c_*$  and the Yule-Walker estimator  $\hat{\phi} = C^{-1}c = R^{-1}r$  will be very similar. Generally, LS estimators are preferred over YW estimators since they tend to have better “finite-sample” properties, especially for AR models that are nearly nonstationary (i.e., have roots of  $\phi(B)$  close to the unit circle).

The (conditional) least squares estimator of  $\sigma^2 = Var(\varepsilon_t)$  is given by

$$\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^T \hat{\varepsilon}_t^2 = \frac{1}{T-p} \sum_{t=p+1}^T (Y_t - \hat{\phi} - \hat{\phi}_1 Y_{t-1} - \dots - \hat{\phi}_p Y_{t-p})^2$$

From standard results concerning the residual sum of squares from least squares linear regression.

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{T-p} (Y - X\hat{\phi})'(Y - X\hat{\phi}) = \frac{1}{T-p} [Y'Y - \hat{\phi}'X'X\hat{\phi}] \\ &= \frac{1}{T-p} \left[ \sum_{t=p+1}^T (Y_t - \bar{Y})^2 - Tc_*'C_*^{-1}c_* \right] \approx c(0) - c'C^{-1}c, \end{aligned}$$

so we also have the approximation  $\hat{\sigma}^2 = c(0)(1 - \hat{\phi}'r) = c(0)[1 - \sum_{i=1}^p \hat{\phi}_i r(i)]$ . Note that this form of  $\hat{\sigma}^2$  is the same as the relation for  $\sigma^2$  in terms of theoretical autocorrelations,  $\sigma^2 = \gamma(0)[1 - \sum_{i=1}^p \phi_i \rho(i)]$ .

Now for large sample size  $T$ , much of the classical linear regression sampling theory can be applied to the estimation in the AR( $p$ ) situation. Thus, for the LS estimator and for the Yule-Walker estimator (which are asymptotically equivalent to each other as  $T \rightarrow \infty$ ), it can be proven that the distribution of  $\sqrt{T}(\hat{\phi} - \phi)$  converges as  $T \rightarrow \infty$  to the multivariate normal distribution  $N(0, \sigma^2 \Gamma_p^{-1})$ , a multivariate normal with mean 0 and covariance matrix  $\sigma^2 \Gamma_p^{-1}$ , where  $\Gamma_p \equiv \gamma(0)P_p$ , where  $P_p$  is defined in (2.16). Thus, for large  $T$ , we can use the approximation that  $\hat{\phi}$  is distributed as  $N(\phi, (\sigma^2/T)\Gamma_p^{-1})$ . The approximate covariance matrix of  $\hat{\phi}$  is estimated by  $\widehat{Cov}(\hat{\phi}) = (\hat{\sigma}^2/T * c(0))R^{-1} = (\hat{\sigma}^2/T)C^{-1} = \hat{\sigma}^2(X'X)^{-1}$ . For a simple example, in the AR(1) case, the LS or Yule-Walker estimator  $\hat{\phi} = r(1) = \hat{\rho}(1)$  has an approximate normal distribution with mean  $\phi = \rho(1)$  and variance  $Var(\hat{\phi}) = (\sigma^2/T)[\gamma(0)]^{-1} = (1/T)(\sigma^2/\gamma(0)) = (1/T)(1 - \phi^2)$ , since  $\gamma(0) = \sigma^2/(1 - \phi^2)$  for an AR(1) process, and  $Var(\hat{\phi})$  is estimated by  $(1 - \hat{\phi}^2)/T$ .

Maximum likelihood and least squares estimation are much more computationally difficult for a mixed ARMA( $p, q$ ) model,  $Y_t - \sum_{i=1}^p \phi_i Y_{t-i} = \delta + \varepsilon_t - \sum_{i=1}^q \theta_i \varepsilon_{t-i}$ , than for a pure AR( $p$ ) model because the log-likelihood and the sum of squares are more complicated (non-quadratic) functions of the unknown parameters  $\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \delta$ , and  $\sigma^2$ . The main difference between ML and the conditional likelihood method is how the initial values are treated. Under ML, their distributions

are incorporated in the estimation but under conditional likelihood, they are taken to be given or fixed. We will be brief on this topic.

For computational convenience, in the conditional ML (or conditional LS) estimation approach, for a sample  $Y_1, Y_2, \dots, Y_T$ , one conditions on the first  $p$  observations  $Y_1, Y_2, \dots, Y_p$  and also on  $q$  “starting values”  $\varepsilon_p^0, \varepsilon_{p-1}^0, \dots, \varepsilon_{p+1-q}^0$  for the needed initial  $\varepsilon_t$ ’s. Given these values, the remaining  $\varepsilon_t$ ’s can be computed recursively as

$$\varepsilon_t^0 = Y_t - \sum_{i=1}^p \phi_i Y_{t-i} - \delta + \sum_{i=1}^q \theta_i \varepsilon_{t-i}^0, \quad t = p+1, \dots, T$$

and the conditional log-likelihood function is given by

$$l_*(\phi, \theta, \delta, \sigma^2) = -\frac{T-p}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} S_*(\phi, \theta, \delta), \quad (2.29)$$

where  $S_* = S_*(\phi, \theta, \delta) = \sum_{t=p+1}^T \varepsilon_t^{02}$  is the condition sum of squares function. The conditional MLE or conditional LSE of  $\phi, \theta$  and  $\delta$  are the values which minimizes  $S_*$ , and the corresponding estimate of  $\sigma^2$  is  $\hat{\sigma}^2 = [1/(T-p)] \sum_{t=p+1}^T \varepsilon_t^{02} = [1/(T-p)] S_*(\hat{\phi}, \hat{\theta}, \hat{\delta})$ , where the residual  $\hat{\varepsilon}_t^0$  are computed from the LSE’s  $\hat{\phi}_i, \hat{\theta}_i$ , and  $\delta$  and the starting values. In the conditional approach, the most common procedure is to set  $\varepsilon_p^0 = 0, \dots, \varepsilon_{p+1-q}^0 = 0$ , that is, use zero starting values for the initial  $\varepsilon_t$ ’s. A nonlinear least squares numerical procedure such as Gauss-Newton or Newton-Raphson is used to find the values which minimize  $S_*$ . These methods are iterative and are based on a linear approximation of a nonlinear model. Interested readers should refer to the classical texts in this area.

We briefly present the results with the two most applicable models in finance.

**Example 11 (MA(1) Model).** Let  $Y_t = \varepsilon_t - \theta \varepsilon_{t-1}$ , write  $Z_t = -\partial \varepsilon_t / \partial \theta$  that satisfies the relation  $\partial \varepsilon_t / \partial \theta = \varepsilon_t + \theta \partial \varepsilon_{t-1} / \partial \theta$ , or  $Z_t = \theta Z_{t-1} - \varepsilon_t, t = 1, 2, \dots, T$ . This can be computed recursively from the starting values  $\varepsilon_0^0 = 0, Z_0 = 0$  and the  $\varepsilon_t^0$ . Then we have  $\hat{\theta} = \theta_0 + \sum_{t=1}^T \bar{Z}_t \bar{\varepsilon}_t^0 / \sum_{t=1}^T \bar{Z}_t^2$ , with  $\text{Var}(\hat{\theta}) = \hat{\sigma}^2 / \sum_{t=1}^T \bar{Z}_t^2$ . Note that in theory, the stochastic process  $\{Z_t\}$  follows an AR(1) model,  $(1 - \theta B)Z_t = -\varepsilon_{t-1}$ , so  $V = E(Z_t^2) = \sigma^2 / (1 - \theta^2)$ , and we have  $\text{Var}(\hat{\theta}) = (\sigma^2 / T)V^{-1} = (1 - \theta^2) / T$ .

**Example 12 (MA(1,1) Model).** With  $Y_t - \phi Y_{t-1} = \varepsilon_t - \theta \varepsilon_{t-1}$ , the derivatives  $Z_{1t} = -\partial \varepsilon_t / \partial \phi$  and  $Z_{2t} = -\partial \varepsilon_t / \partial \theta$  are determined recursively from the relations

$$\frac{\partial \varepsilon_t}{\partial \phi} = -Y_{t-1} + \theta \frac{\partial \varepsilon_{t-1}}{\partial \phi}, \quad \frac{\partial \varepsilon_t}{\partial \theta} = \varepsilon_{t-1} + \theta \frac{\partial \varepsilon_{t-1}}{\partial \theta},$$

or  $Z_{1t} = \theta Z_{1,t-1} + Y_{t-1}, Z_{2t} = \theta Z_{2,t-1} - \varepsilon_{t-1}, t = 2, \dots, T$ , with  $Z_{11} = Z_{21} = 0$ . Notice that the stochastic processes  $\{Z_{1t}\}$  and  $\{Z_{2t}\}$  satisfy  $(1 - \theta B)Z_{1t} = Y_{t-1} = (1 - \phi B)^{-1}(1 - \theta B)\varepsilon_{t-1}$ , or  $(1 - \phi B)Z_{1t} = \varepsilon_{t-1}$ , and  $(1 - \theta B)Z_{2t} = -\varepsilon_{t-1}$ , so that  $\{Z_{1t}\}$  and  $\{Z_{2t}\}$  are both AR(1) processes with related white noise inputs. Hence we find that  $E(Z_{1t}^2) = \sigma^2 / (1 - \phi^2)$ ,  $E(Z_{2t}^2) = \sigma^2 / (1 - \theta^2)$ , and  $E(Z_{1t} Z_{2t}) =$

$-E \left( \sum_{t=0}^{\infty} \phi^t \varepsilon_{t-1-i} \cdot \sum_{j=0}^{\infty} \theta^j \varepsilon_{t-1-j} \right) = -\sigma^2 \sum_{i=0}^{\infty} \phi^i \theta^i = -\sigma^2 / (1 - \phi \theta)$ . Thus the large sample approximate covariance matrix of the MLE  $\hat{\beta} = (\hat{\phi}, \hat{\theta})'$  is

$$\text{Cov}(\hat{\phi}, \hat{\theta}) \approx \frac{\sigma^2}{T} V^{-1} = \frac{1}{T} \begin{pmatrix} \frac{1}{1-\phi^2} & \frac{-1}{1-\phi\theta} \\ \frac{-1}{1-\phi\theta} & \frac{1}{1-\theta^2} \end{pmatrix}$$

Note that when  $\phi = \theta$  the asymptotic variances of  $\hat{\phi}$  and  $\hat{\theta}$  become infinite. This is because in this case the common factor  $(1 - \phi B) = (1 - \theta B)$  cancels from both sides of the ARMA(1,1) model equation, which reduces to  $Y_t = \varepsilon_t$ , a white noise process. Thus there are no “true” values for the parameter  $\phi$  and  $\theta$  when  $\phi = \theta$ , their common value is actually arbitrary, and cannot be estimated. This is an example of what is known as parameter redundancy (or parameter non-identifiability), and can arise when one attempts to “overfit” an ARMA model using too large a value for both  $p$  and  $q$ .

## Model Diagnostics

As several models can be entertained from the specification tools such as ACF and PACF, one quick way to check if all the informative dependencies are captured is to use the model residuals and examine if these residuals reflect the features that are assumed. One key feature is the assumption of independence that can be tested through ACF. Another feature is the constancy in the variance which again can be examined by the plot of squared residuals versus time scale and the time dependence by the ACF of the squared residuals. Thus residuals or functions of the residuals play an important role in model diagnostics.

## Model Validation

The debate on the evaluation of a model for its consistency with the data at hand or its forecasting power is a long-standing one. The risk of overfitting a model purely by data exploration will be exposed through poor performance in forecasting. There is a chance to fit the features of the data that may be unique but may not be present in the future. We will follow in general the principle of parsimony; that is develop models with a few parameters that can adequately represent the data.

Ideally we should have two random samples from the same population; a training set and a test set. Because of the chronological nature of the time series, usually the training set precedes in time the test set. Thus if  $y_t, t = 1, 2, \dots, n_1, n_1 + 1, \dots, n_1 + n_2$  and if the model is built on ‘ $n_1$ ’ observations and the last ‘ $n_2$ ’ observations are used for model validation, it is usual to compare the prediction mean square error (PMSE),

$$\text{PMSE} = \sum_{t=1}^{n_2} \frac{(y_{n_1+t} - \hat{y}_{n_1+t})^2}{n_2} \quad (2.30)$$

with the MSE of the fitted model. While other forms of cross-validation methods are applicable to non-dependent data, for time series data, to apportion the data into data sets with non-overlapping behavior requires knowing in the first place what the appropriate model is. So we suggest simply using the out of sample method.

### 2.3.4 Testing for Nonstationary (Unit Root) in ARIMA Models

As noted, initially the decision concerning the need for differencing is based, informally, on features of the time series plot of  $Y_t$  and of its sample autocorrelation function (eg. failure of the  $\rho(k)$  to dampen out sufficiently quickly). This can be evaluated further based on efficient estimation of model parameters as discussed earlier. However, it must be borne in mind that distribution theory for estimates of parameters differs markedly between stationary and nonstationary models. This has led to an interest in more formal inference procedures concerning the appropriateness of a differencing operator (or a unit root in the AR operator) and distribution theory for parameter estimates in unit root nonstationary time series models.

Consider first the simple AR(1) model  $Y_t = \phi Y_{t-1} + \varepsilon_t, t = 1, 2, \dots, T, Y_0 = 0$ , and consider testing for a random walk model; that is,  $\phi = 1$ . The (conditional) least squares (LS) estimator of  $\phi$  is given by

$$\hat{\phi} = \frac{\sum_{t=2}^T Y_{t-1} Y_t}{\sum_{t=2}^T Y_{t-1}^2} = \phi + \frac{\sum_{t=2}^T Y_{t-1} \varepsilon_t}{\sum_{t=2}^T Y_{t-1}^2} \quad (2.31)$$

In the stationary case with  $|\phi| < 1$ ,  $\{Y_t\}$  has the representation  $Y_t = \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}$  with constant variance  $\text{Var}(Y_t) = \gamma(0) = \sigma^2/(1 - \phi^2)$ . Hence, as indicated earlier,  $T^{1/2}(\hat{\phi} - \phi) = T^{-1/2} \sum_{t=2}^T Y_{t-1} \varepsilon_t / T^{-1} \sum_{t=2}^T Y_{t-1}^2$  has an approximate normal distribution with zero mean and variance  $(1 - \phi^2)$ , since  $T^{-1/2} \sum_{t=2}^T Y_{t-1} \varepsilon_t \xrightarrow{D} N(0, \sigma^2 \gamma(0))$  and  $T^{-1} \sum_{t=2}^T Y_{t-1}^2 \xrightarrow{P} \gamma(0)$  as  $T \rightarrow \infty$ . However, when  $\phi = 1$ , so  $Y_t = \sum_{j=0}^{t-1} \varepsilon_{t-j} + Y_0$  in the (truncated) integrated form, with  $\text{Var}(Y_t) = E(Y_t^2) = t\sigma^2$ , it can be shown that  $T(\hat{\phi} - 1) = T^{-1} \sum_{t=2}^T Y_{t-1} \varepsilon_t / T^{-2} \sum_{t=2}^T Y_{t-1}^2 = O_p(1)$ , bounded in probability as  $T \rightarrow \infty$ , with both the numerator and denominator possessing non-degenerate and nonnormal limiting distributions. Hence, in the nonstationary case the estimator  $\hat{\phi}$  approaches its true value  $\phi = 1$  with increasing sample size  $T$  at a faster rate than in the stationary case.

A representation for the limiting distribution of  $T(\hat{\phi} - 1)$  has been given by Dickey and Fuller (1979) [84], such that

$$T(\hat{\phi} - 1) \xrightarrow{D} \frac{1}{2} (\Lambda^2 - 1) / \Gamma \quad (2.32)$$

where  $(\Gamma, \Lambda) = (\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2, \sum_{i=1}^{\infty} 2^{1/2} \gamma_i Z_i)$ , with  $\gamma_i = 2(-1)^{i+1}/[(2i-1)\pi]$ , and the  $Z_i$  are i.i.d.  $N(0, 1)$  distributed r.v.'s. Tables for the percentiles of the limiting distribution of  $T(\hat{\phi} - 1)$  have been given by Fuller (1996) [125]. Also, the “Studentized”

statistic

$$\hat{\tau} = \frac{\hat{\phi} - 1}{s_\epsilon \left( \sum_{t=2}^T Y_{t-1}^2 \right)^{-1/2}} \quad (2.33)$$

where  $s_\epsilon^2 = (T-2)^{-1} (\sum_{t=2}^T Y_t^2 - \hat{\phi} \sum_{t=2}^T Y_{t-1} Y_t)$  is the residual mean square, has been considered. The limiting distribution of the statistic  $\hat{\tau}$  has been derived, and tables of the percentiles of this distribution under  $\phi = 1$  are given in Fuller (1996) [125]. The test rejects  $\phi = 1$  when  $\hat{\tau}$  is “too negative.”

For higher order AR( $p+1$ ) model  $Y_t = \sum_{j=1}^{p+1} \varphi_j Y_{t-j} + \epsilon_t$ , it is seen that the model can be expressed in an equivalent form as  $W_t = (\rho - 1)Y_{t-1} + \sum_{j=1}^p \phi_j W_{t-j} + \epsilon_t$ , where  $W_t = Y_t - Y_{t-1}$ ,  $\rho - 1 = -\varphi(1) = \sum_{j=1}^{p+1} \varphi_j - 1$ , and  $\phi_j = -\sum_{i=j+1}^{p+1} \varphi_i$ . Hence, the existence of a unit root in the AR operator  $\varphi(B)$  is equivalent to  $\rho = \sum_{j=1}^{p+1} \varphi_j = 1$ . So, based on this last form of the model, let  $(\hat{\rho} - 1, \hat{\phi}_1, \dots, \hat{\phi}_p)$  denote the usual least squares regression estimates for the model,

$$W_t = (\rho - 1)Y_{t-1} + \sum_{j=1}^p \phi_j W_{t-j} + \epsilon_t \quad (2.34)$$

Then, under the unit root model where  $\rho = 1$  and  $\phi(B)$  is stationary, it has been shown by Fuller (1996, Th. 10.1.1 and Cor. 10.1.1.1) [125] that  $(\hat{\rho} - 1)/[s_\epsilon(\sum_{t=p+2}^T Y_{t-1}^2)^{-1/2}]$  has the same limiting distribution as the statistic  $\hat{\tau}$  in (2.33) for the AR(1) case, while  $(T-p-1)(\hat{\rho} - 1)c$ , where  $c = \sum_{j=0}^\infty \psi_j$ , with  $\psi(B) = \phi^{-1}(B)$ , has approximately the same distribution as the statistic  $T(\hat{\phi} - 1)$  for the AR(1) case. Also, it follows that the statistic, denoted as  $\hat{\tau}$ , formed by dividing  $(\hat{\rho} - 1)$  by its “usual estimated standard error” from the least squares regression will be asymptotically equivalent to the statistic  $(\hat{\rho} - 1)/[s_\epsilon(\sum_{t=p+2}^T Y_{t-1}^2)^{-1/2}]$ , and hence will have the same limiting distribution as the statistic  $\hat{\tau}$  for the AR(1) case.

The test statistic  $\hat{\tau}$  formed in the above manner can be used to test the hypothesis that  $\rho = 1$  in the AR( $p+1$ ) model. The above methods extend to the case where a constant term is included in the least squares regression estimation, with the statistic analogous to  $\hat{\tau}$  denoted as  $\hat{\tau}_\mu$ . Thus, for example, in the AR(1) model  $Y_t = \phi Y_{t-1} + \theta_0 + \epsilon_t$ , one obtains the LSE,

$$\hat{\phi}_\mu = \frac{\sum_{t=2}^T (Y_{t-1} - \bar{Y}_{(t)})(Y_t - \bar{Y}_{(0)})}{\sum_{t=2}^T (Y_{t-1} - \bar{Y}_{(1)})^2} \quad (2.35)$$

where  $\bar{Y}_{(i)} = (T-1)^{-1} \sum_{t=2}^T Y_{t-i}$ ,  $i = 0, 1$ . The corresponding test statistic for  $\phi = 1$  in the AR(1) case is

$$\hat{\tau}_\mu = (\hat{\phi}_\mu - 1)/[s_\epsilon \left( \sum_{t=2}^T (Y_{t-1} - \bar{Y}_{(1)})^2 \right)^{-1/2}] \quad (2.36)$$

and percentiles of the distribution of  $\hat{\tau}_\mu$  and when  $\phi = 1$  are given in Table 2.2.

These test procedures and other similar ones have also been extended for testing for unit roots in univariate ARIMA models have been given by Dickey, Bell, and Miller (1986) [83].

**Example 13 (To difference or not).** To illustrate here, we consider briefly the quarterly series of U.S. Commercial Paper interest rates for 1953–1970 in Figure 2.3. The series is identified as ARIMA(2,1,0), and unconditional least squares estimation gave

$$W_t = (1 - B)Y_t = 0.739W_{t-1} - 0.424W_{t-2}.$$

When conditional LS estimation of an AR(3) model for  $Y_t$ , as represented in (2.34), was performed, the results were  $W_t = -0.0147Y_{t-1} + 0.741W_{t-1} - 0.399W_{t-2} + \hat{\delta} + \varepsilon_t$ , and the estimate  $\hat{\rho} - 1 = -0.0147$  had estimated standard error of 0.0246. Clearly the test statistic  $\hat{\tau} = -0.60$  is not significant, and there is no cause to doubt the need for first differencing of the series  $Y_t$  (i.e., that there is a unit root for  $Y_t$ ).

Table 2.2: Approximate percentiles

	Probability of a Smaller Value							
	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
$\hat{\tau}$	-2.58	-2.23	-1.95	-1.62	0.89	1.28	1.62	2.00
$\hat{\tau}_\mu$	-3.43	-3.12	-2.86	-2.57	-0.44	-0.07	0.23	0.60

The use of the unit root test statistics will become apparent when we discuss pairs trading based on co-integration (see Chapter 3). A precondition for forming co-integrated pairs is that each series must be non-stationary on its own.

### 2.3.5 Forecasting for ARIMA Processes

Given a realization of ARIMA( $p, d, q$ ) process through time  $t$ ,  $\{Y_s, s \leq t\}$ , we consider the problem of forecasting of future values  $Y_{t+l}, l = 1, 2, \dots$ . For this purpose, it is assumed that the model for  $\{Y_t\}$  is known exactly including the values of the model parameters. Although, in practice, the model must be specified and the parameters estimated from available sample data, errors due to estimation of parameters will not have too much effect on forecast properties for sufficiently large sample sizes. Also, the practical problem of future values based on the finite past sample data  $Y_1, \dots, Y_T$  will be considered briefly, but results based on the infinite past data assumption usually provide an adequate approximation to the finite past data prediction problem.

#### Minimum Mean Squared Error Prediction

Before presenting details of forecasting for ARIMA processes, we review some basic principles concerning prediction in a more general context. Let  $Y$  be a r.v. and let  $X = (X_1, \dots, X_p)'$  be an  $p$ -dimensional random vector related to  $Y$ , and consider

the problem of predicting (estimating) the unknown value of  $Y$  by some function of  $X$ , say  $\hat{Y} = g(x)$ . The mean squared error (MSE) of a predictor  $\hat{Y}$  is  $E[(Y - \hat{Y})^2]$ . We refer to the minimum mean squared error (minimum MSE) predictor of  $Y$  as that function  $\hat{Y} = g^*(X)$  such that, among all possible functions of  $X$ ,  $\hat{Y}$  minimizes the MSE  $E[(Y - \hat{Y})^2]$ . Then it is well-known that the *minimum mean squared error* predictor is given by  $\hat{Y} = E(Y|X)$ , the conditional expectation of  $Y$  given  $X$ , with prediction error  $e = Y - \hat{Y} = Y - E(Y|X)$ . One can also restrict prediction to consider only linear functions of  $X$ , and hence consider the minimum MSE *linear* predictor as the linear function  $\hat{Y}^* = a + b'X$  which minimizes the prediction MSE among all linear functions of  $X$ . It is well-known that the minimum MSE linear predictor is

$$\hat{Y}^* = \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (X - \mu_x) \quad (2.37)$$

with the prediction error  $e^* = Y - \hat{Y}^*$  having mean zero and prediction MSE (variance)

$$\text{Var}(e^*) = \text{Var}(Y - \hat{Y}^*) = \sigma_y^2 - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \quad (2.38)$$

where  $\mu_y = E(Y)$ ,  $\mu_x = E(X)$ ,  $\sigma_y^2 = \text{Var}(Y)$ ,  $\Sigma_{yx} = \text{Cov}(Y, X)$ , and  $\Sigma_{xx} = \text{Cov}(X)$ . Note  $\Sigma_{yx}$  is a  $1 \times p$  row vector and  $\Sigma_{xx}$  is a  $p \times p$  variance-covariance matrix. Moreover, if the prediction error  $e^* = Y - \hat{Y}^*$  of the best linear predictor is such that  $E(e^*|X) = 0$  (e.g., if  $e^*$  is independent of  $X$ ), then  $\hat{Y}^*$  is also the minimum MSE predictor, i.e.  $\hat{Y}^* = \hat{Y} = E(Y|X)$  is a linear function of  $X$  and the prediction error  $e = e^* = Y - \hat{Y}$  has variance  $\text{Var}(e^*)$  as given above in (2.38).

### Forecasting for ARIMA Processes and Properties of Forecast Errors

For forecasting in the time series setting, we assume that the process  $\{Y_t\}$  follows an ARIMA( $p, d, q$ ) model,  $\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t$ , and we assume the white noise series  $\varepsilon_t$  are mutually independent random variables. We are interested in forecasting the future value  $Y_{t+l}$  based on observations  $Y_t, Y_{t-1}, \dots$ . From the result in previous section, the minimum MSE forecast of  $Y_{t+l}$  based on  $Y_t, Y_{t-1}, \dots$ , which we will denote as  $\hat{Y}_t(l)$ , is such that  $\hat{Y}_t(l) = E(Y_{t+l} | Y_t, Y_{t-1}, \dots)$ . The prediction  $\hat{Y}_t(l)$  is called the lead  $l$  or  $l$ -step ahead forecast of  $Y_{t+l}$ ,  $l$  is the lead time, and  $t$  is the forecast origin.

To obtain a representation for  $\hat{Y}_t(l)$ , recall the ARIMA process has the “infinite” MA form  $Y_t = \psi(B)\varepsilon_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$ , and hence a future value  $Y_{t+l}$  at time  $t+l$ , relative to the current time or “forecast origin”  $t$ , can be expressed as

$$Y_{t+l} = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t+l-i} = \sum_{i=0}^{l-1} \psi_i \varepsilon_{t+l-i} + \sum_{i=l}^{\infty} \psi_i \varepsilon_{t+l-i}. \quad (2.39)$$

The information contained in the past history of the  $Y_t$ 's,  $\{Y_s, s \leq t\}$ , is the same as that contained in the past random shocks  $\varepsilon_t$ 's (because the  $Y_t$ 's are generated by the  $\varepsilon_t$ 's). Also,  $\varepsilon_{t+h}$ , for  $h > 0$  is independent of present and past values  $Y_t, Y_{t-1}, \dots$ , so that  $E(\varepsilon_{t+h} | Y_t, Y_{t-1}, \dots) = 0, h > 0$ . Thus

$$\hat{Y}_t(l) = E(Y_{t+l} | Y_t, Y_{t-1}, \dots) = E(Y_{t+l} | \varepsilon_t, \varepsilon_{t-1}, \dots) = \sum_{i=l}^{\infty} \psi_i \varepsilon_{t+l-i}, \quad (2.40)$$

using the additional property that  $E(\varepsilon_{t+l-i} | \varepsilon_t, \varepsilon_{t-1}, \dots) = \varepsilon_{t+l-i}$  if  $i \geq l$ . The  $l$ -step ahead prediction error is given by

$$e_t(l) = Y_{t+l} - \hat{Y}_t(l) = \sum_{i=0}^{l-1} \psi_i \varepsilon_{t+l-i}.$$

So we have  $E[e_t(l)] = 0$  and the mean squared error or variance of the  $l$ -step prediction error is

$$\sigma^2(l) = \text{Var}(e_t(l)) = E[e_t^2(l)] = \text{Var}\left(\sum_{i=0}^{l-1} \psi_i \varepsilon_{t+l-i}\right) = \sigma^2 \sum_{i=0}^{l-1} \psi_i^2, \quad (2.41)$$

with  $\text{St.Dev.}(e_t(l)) = \sigma(l) = \sigma(1 + \psi_1^2 + \dots + \psi_{l-1}^2)^{1/2}$ . Note, in particular, that for  $l = 1$  step we have  $e_t(1) = Y_{t+1} - \hat{Y}_t(1) = \varepsilon_{t+1}$  with error variance  $\sigma^2 = \text{Var}(\varepsilon_{t+1})$ , so that the white noise series  $\varepsilon_t$  can also be interpreted as the one-step ahead forecast errors for the process. Under the normality assumption that the  $\varepsilon_t$  are normally distributed as i.i.d.  $N(0, \sigma^2)$ , the  $l$ -step ahead forecast errors  $e_t(l) = Y_{t+l} - \hat{Y}_t(l) = \sum_{i=0}^{l-1} \psi_i \varepsilon_{t+l-i}$  will also be normally distributed with mean 0 and variance  $\sigma^2(l) = \sigma^2 \sum_{i=0}^{l-1} \psi_i^2$ . Hence, the conditional distribution of  $Y_{t+l}$ , given  $Y_t, Y_{t-1}, \dots$ , is normal with mean  $\hat{Y}_t(l)$  and variance  $\sigma^2(l)$ .

For forecasting purposes, the variance  $\sigma^2(l)$  of the  $l$ -step forecast error provides a measure of the degree of accuracy of the point forecast  $\hat{Y}_t(l)$ . Along with the point forecast  $\hat{Y}_t(l)$ , we need to provide prediction intervals for the future observations  $Y_{t+l}$  so that the degree of uncertainty of the point forecasts is properly reflected. From the preceding paragraph, it follows that the standardized variable  $Z_t(l) = e_t(l)/\sigma(l) = (Y_{t+l} - \hat{Y}_t(l))/\sigma(l)$  is distributed as standard normal  $N(0, 1)$ , which is also the conditional distribution of this variable, given  $Y_t, Y_{t-1}, \dots$ . Then, for example, from  $0.90 = R[-1.65 < Z_t(l) < 1.65] = P[\hat{Y}_t(l) - 1.65 \sigma(l) < Y_{t+l} < \hat{Y}_t(l) + 1.65 \sigma(l)]$ , it follows that a 90% prediction interval (interval forecast) for  $Y_{t+l}$  is given by  $\hat{Y}_t(l) \pm 1.65 \sigma(l)$ , where  $\sigma(l) = \sigma(\sum_{i=0}^{l-1} \psi_i^2)^{1/2}$ . Prediction intervals for other desired levels of confidence,  $100(1 - \alpha)\%$ , can be constructed similarly. Of course, in practice, estimates of  $\sigma^2$  and of the weights  $\psi_i$  based on the estimates  $\hat{\phi}_i$  and  $\hat{\theta}_i$  of parameters in the ARIMA model obtained from the sample data will actually be used in estimating the standard deviation  $\sigma(l)$  of the forecast errors and in forming the prediction intervals.

### Eventual Forecast Function for ARIMA Processes

It can be shown that the forecast function  $\hat{Y}_t(l)$ , considered as a function of the lead time  $l$  with the origin  $t$  fixed, satisfies the linear difference equation

$$\hat{Y}_t(l) = \sum_{i=1}^{p+d} \phi_i \hat{Y}_t(l-i) + \delta, \quad \text{for } l > q. \quad (2.42)$$

We can express this relation as  $\phi^*(B) \hat{Y}_t(l) = \delta$ , for  $l > q$ , where the backshift operator

$B$  operates on the lead time  $l$ , with origin  $t$  fixed. Hence for  $l > q - p - d$ , the function  $\hat{Y}_t(l)$ , which will be referred to as the *eventual forecast function*, has the features of solution to the difference equation of order  $p + d$ . So, in general, as a function of  $l$ ,  $\hat{Y}_t(l)$  can consist of a mixture of polynomial terms, exponential decaying terms, and damped (and undamped) sinusoidal terms. The exact form of  $\hat{Y}_t(l)$  is determined by the roots of the polynomial associated with the generalized AR operator  $\phi^*(B) = \phi(B)(1 - B)^d$ .

**Example 14 (IMA(1,1) Model).** Let  $Y_t - Y_{t-1} = \delta + \varepsilon_t - \theta\varepsilon_{t-1}$ . Then, from (2.39)–(2.40), the forecasts are  $\hat{Y}_t(1) = Y_t + \delta - \theta\varepsilon_t$ ,  $\hat{Y}_t(2) = Y_t + \delta = Y_t - \theta\varepsilon_t + 2\delta$ , and in general,

$$\hat{Y}_t(l) = \hat{Y}_t(l-1) + \delta = Y_t - \theta\varepsilon_t + \delta(l) = \hat{Y}_t(1) + \delta(l-1), \quad l = 1, 2, \dots$$

So the “eventual” forecast function  $\hat{Y}_t(l)$  is a straight line with a deterministic (fixed) slope  $\delta$ , and an “adaptive” intercept  $\hat{Y}_t(1) - \delta = Y_t - \theta\varepsilon_t$  which is random and is adaptive to the process values which occur through time  $t$ . The variance of the  $l$ -step ahead forecast error  $e_t(l)$  is  $\sigma^2(l) = \sigma^2(1 + (l-1)(1-\theta)^2)$ ,  $l \geq 1$ . Hence, we see that the values  $\sigma^2(l)$  increase without bound as the lead time  $l$  increases, which is a general feature for (nonstationary) ARIMA processes with  $d > 0$ . Also notice from the infinite AR form of the model (with  $\delta = 0$ ) with AR coefficients  $\pi_i = (1-\theta)\theta^{i-1}$ ,  $i \geq 1$ ,  $Y_{t+1} = (1-\theta)\sum_{i=1}^{\infty} \theta^{i-1} Y_{t+1-i} + \varepsilon_{t+1}$ , we have the representation  $\hat{Y}_t(1) = (1-\theta)\sum_{i=1}^{\infty} \theta^{i-1} Y_{t+1-i}$ . This provides the exponential smoothing, or exponentially weighted moving average (EWMA), form of the one-step forecast  $\hat{Y}_t(1)$  in the IMA(1,1). This model with a short term drift is of immense interest in modeling the stock prices.

## 2.4 Stylized Models for Asset Returns

It is generally known that the securities market is extremely efficient in reflecting information about stocks: When new information arrives, it gets incorporated in the price without delay. Thus the efficient market hypothesis, is widely accepted by financial economists. It is implied that neither technical analysts that study the past prices in an attempt to predict future prices nor fundamental analysts of financial information related to company earnings and asset values would carry any advantage over the returns obtained from a randomly selected portfolio of individual stocks.

The efficient market hypothesis is usually associated with the idea of the “random walk”, which implies that all future price changes represent random departures from past prices. As stated in Malkiel (2012) [203]: “The logic of the random walk is that if the flow of information is unimpeded and information is immediately reflected in stock prices, then tomorrow’s price change will reflect only tomorrow’s news and will be independent of the price changes today. But news by definition is unpredictable, and, thus resulting price changes must be unpredictable and random.”

The random walk (RW) model (Example ??) without the drift term can be stated as:

$$p_t = p_{t-1} + \varepsilon_t \quad (2.43)$$

where  $\varepsilon_t \sim N(0, \sigma^2)$  i.i.d and  $p_t = \ln(P_t)$ . Note as observed earlier this model is a particular case of AR(1) model if the constant term is assumed to be zero and the slope,  $\phi$ , is assumed to be one. Thus, the RW model is a non-stationary model and considering  $\varepsilon_t = p_t - p_{t-1} = r_t$  is the differencing of the series,  $p_t$ , makes the series  $\varepsilon_t$  stationary. Note  $r_t \approx \frac{P_t - P_{t-1}}{P_{t-1}}$  and the returns are purely random and are unpredictable. The decision concerning the need for differencing is based, informally, on the features of time series plot of  $p_t$ , its sample autocorrelation function; that is, its failure to dampen out sufficiently quickly.

**Testing for a RW model:** Consider the AR(1) model  $p_t = \phi p_{t-1} + \varepsilon_t$ ,  $t = 1, \dots, T$  and  $p_0$  is fixed; we want to test  $\phi = 1$ . The details of this test are given in Section 2.3.4, along with the table of the percentiles of the distribution of the test statistics. While the results for testing for non-stationarity for higher order autoregressive models were discussed earlier, the price series once differenced does not generally exhibit autocorrelation structure that requires modeling beyond AR(1).

It will become obvious in subsequent discussions, that in order to exploit any pattern in the returns, one has to look beyond linear time dependencies. We observe a series of established stylized facts regarding price (return).

**Stylized fact 1: Absence of Autocorrelation in returns:** Linear autocorrelations are often insignificant for a lower frequency data.

To illustrate the main ideas, we consider the exchange rate between USD and GBP; these are daily closing rates from Jan 4, 1999 to Sept 16, 2013. The time series plot of the data in (Figure 2.3) clearly indicates that the price is non-stationary and it is confirmed by the never dying autocorrelation function in (Figure 2.3). The LS estimate of  $\hat{\phi} = 0.9996$  and  $s_e = 0.00958$  and  $\sum_{t=2}^T p_{t-1}^2 = 10431$ , provides a value of  $\hat{\tau} = -0.001$  from (2.33) which confirms that the series is a RW. The differenced series,  $r_t = \ln(p_t) - \ln(p_{t-1})$  is plotted in (Figure 2.3) and its autocorrelation function is given in (Figure 2.3). Thus it would appear that return series is a white noise process.

Many industry analysts tend to look at not just the closing price but also the so called price bars that consist of open, high, low and close prices. Select averages over these four quantities are monitored to study the stock momentum. Two such quantities are pivot, the average of high, low and close prices; and mid-range, the average of high and low prices. The plot of the pivot for the exchange rate is overlaid on the closing price in Figure 2.4. It is not easy to distinguish between the two series from this plot, but an important property of the (aggregated) pivot series is worth noting here. The autocorrelation of the returns dated on pivot series at lag 1 is quite significant (Figure 2.4). This happens mainly due to ‘aggregation’ of data as observed by Working (1960) [262] and by Daniels (1966) [78]. Working (1960) [262] shows that if the average is computed over ‘ $m$ ’ terms in each time segment (in this illustration, a day) the first order correlation is  $\frac{(m^2-1)}{2(2m^2+1)}$ . When  $m = 3$ , we can expect the

lag 1 autocorrelation to be about 0.21. Daniels (1966) [78] observes that we could expect lag 1 autocorrelation in the order of 0.4 for the mid-range series. Thus, the main point of this discussion is that the aggregation of prices (which is the basis of several low-frequency analysis) induces the autocorrelation, but it has nothing to do with reflecting market efficiency.

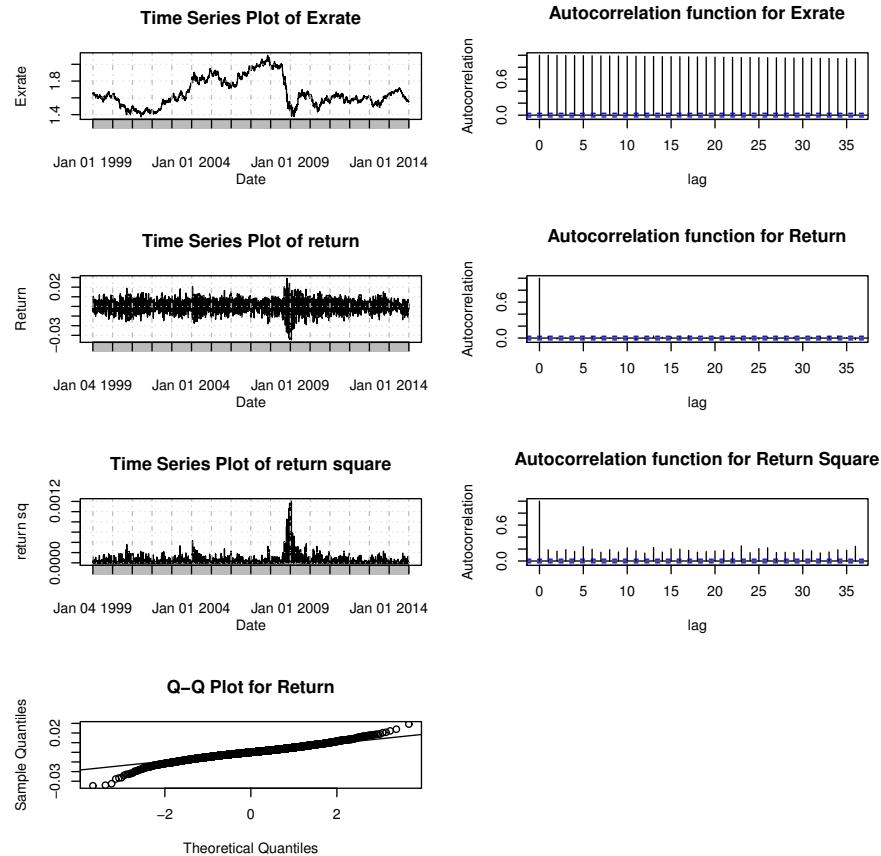


Figure 2.3: Exchange rate related plots.

An assumption of the random walk model that is often made is that the errors (returns) follow a normal distribution. But in practice, the errors (or returns) tend to have heavier tails than the normal. We suggest using the quantile-quantile plot which is quite sensitive to even modest deviations from normality and also a test based on the properties of normal distribution. An omnibus test based on the skewness ( $= 0$  for normal) and the kurtosis ( $= 3$  for normal) is Jarque-Bera test:

$$JB = n \left( \frac{\widehat{SK}^2}{6} + \frac{(\widehat{K} - 3)^2}{24} \right) \sim \chi^2_2 \quad (2.44)$$

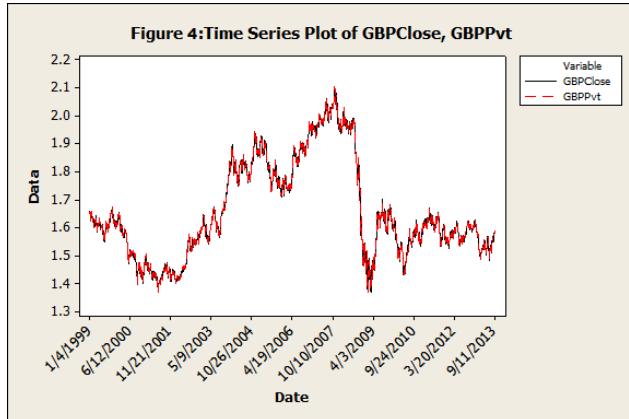


Figure 2.4: Times series plot of GBPclose, GBPPvt.

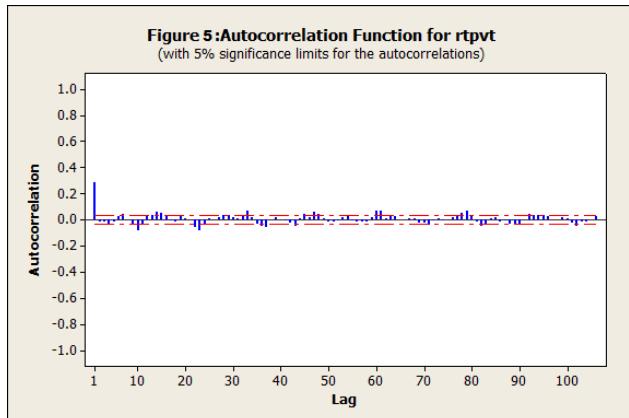


Figure 2.5: Autocorrelation function for rtpvt

where  $SK(\text{ewness}) = E\left(\frac{X-\mu}{\sigma}\right)^3$ ,  $K(\text{urtosis}) = E\left(\frac{X-\mu}{\sigma}\right)^4$ . The corresponding sample numbers are substituted in JB (see Jarque and Bera (1980)[166]). The probability (Q-Q) plot of  $r_t = \ln(p_t) - \ln(p_{t-1})$  as given in Figure 3 clearly indicates that the distribution has somewhat heavy tails. For this data  $\widehat{SK} = -0.31$  and  $\widehat{K} = 3.67$  resulting in  $JB = 129.7$  with  $n = 3236$ , thus rejecting the normal distribution.

**Stylized fact 2:** Heavy Tails: The returns are likely to display a heavy tailed distribu-

tion such as a power-law or Pareto.

**Stylized fact 3:** Asymmetry: When prices fall, they tend to fall faster than they tend to rise, thus drawdowns tend to be larger than upward rises.

These stylized facts are particularly important for risk management of strategies, and how they could be incorporated in trading decisions is still a subject of much research and will be discussed later.

## 2.5 Time Series Models for Aggregated Data: Modeling the Variance

In the ARMA( $p, q$ ) model  $\phi(B)Y_t = \delta + \theta(B)\varepsilon_t$  for series  $\{Y_t\}$ , when the errors  $\varepsilon_t$  are *independent* r.v.'s (with the usual assumptions of zero mean and constant variance  $\sigma^2$ ), an implication is that the *conditional* variance of  $\varepsilon_t$  given the past information is a constant not depending on the past. This, in turn, implies the same feature for  $l$ -step ahead forecast errors  $e_t(l) = \sum_{i=0}^{l-1} \psi_i \varepsilon_{t+l-i}$ . However, in some settings, particularly for data from finance, the variability of errors may exhibit dependence on the past variability. Modeling the variance, which is essential for studying the risk-return relationship is an important topic in finance.

### Autoregressive Conditional Heteroscedastic (ARCH) Models

The autoregressive conditional heteroscedastic (ARCH) models were originally proposed by Engle (1982) [99], Engle and Kraft (1983) [102], to allow for the conditional error variance in an ARMA process to depend on the past squared innovations, rather than be a constant as in ARMA models with independent errors. For example, an AR( $p$ ) process with ARCH( $q$ ) model errors is given as  $Y_t = \sum_{i=1}^p \phi_i Y_{t-i} + \delta + \varepsilon_t$ , where,

$$\begin{aligned} E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) &= 0 \text{ and} \\ h_t = \text{Var}(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) &= E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) = \omega_0 + \sum_{i=1}^q \omega_i \varepsilon_{t-i}^2 \end{aligned} \quad (2.45)$$

with  $\omega_0 > 0$  and  $\omega_i \geq 0$  for  $i = 1, \dots, q$ . In the generalized ARCH (GARCH) model, introduced by Bollerslev (1986) [32], it is assumed that

$$h_t = E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) = \omega_0 + \sum_{i=1}^q \omega_i \varepsilon_{t-i}^2 + \sum_{i=1}^k \beta_i h_{t-i}, \quad (2.46)$$

where  $\beta_i \geq 0$  for all  $i = 1, \dots, k$ . Much subsequent research on and applications of the ARCH and GARCH models have occurred since the original research papers.

Let us briefly discuss some basic results and implications of the ARCH and GARCH models. The errors  $\varepsilon_t$  in the model have zero mean, since  $E(\varepsilon_t) = E[E(\varepsilon_t | \varepsilon_{t-1}, \dots)] = 0$ , and they are serially uncorrelated; that is,  $E(\varepsilon_t \varepsilon_{t-j}) = 0$  for  $j \neq 0$  (since for  $j > 0$ , for example,  $E(\varepsilon_t \varepsilon_{t-j}) = E[E(\varepsilon_t \varepsilon_{t-j} | \varepsilon_{t-1}, \dots)] = E[\varepsilon_{t-j} E(\varepsilon_t | \varepsilon_{t-1}, \dots)] = 0$ ). But the  $\varepsilon_t$  are not mutually independent r.v.'s since they are inter-related through their conditional variances (i.e., their second moment). We will also assume that the  $\varepsilon_t$  have equal unconditional variances,  $\text{Var}(\varepsilon_t) = \sigma^2$ , for all  $t$ , so they are weakly stationary. Consider, then, the simple case of the first-order ARCH or ARCH(1) model,

$$h_t = E(\varepsilon_t^2 | \varepsilon_{t-1}, \dots) = \omega_0 + \omega_1 \varepsilon_{t-1}^2, \quad (2.47)$$

Assuming that  $\omega_1 < 1$  ( $\omega_0$  and  $\omega_1 \geq 0$  is always assumed), the *unconditional* variance of  $\varepsilon_t$  is  $\sigma^2 = \text{Var}(\varepsilon_t) = \frac{\omega_0}{1-\omega_1}$ , since  $\sigma^2 = E(\varepsilon_t^2) = E[E(\varepsilon_t^2 | \varepsilon_{t-1}, \dots)] = E[\omega_0 + \omega_1 \varepsilon_{t-1}^2] = \omega_0 + \omega_1 \sigma^2$ . Therefore, the conditional variance of  $\varepsilon_t$  can be written as

$$h_t = \omega_0 + \omega_1 \varepsilon_{t-1}^2 = \sigma^2(1 - \omega_1) + \omega_1 \varepsilon_{t-1}^2 = \sigma^2 + \omega_1 (\varepsilon_{t-1}^2 - \sigma^2), \quad (2.48)$$

or in the form of deviation from the unconditional variance as  $h_t - \sigma^2 = \omega_1 (\varepsilon_{t-1}^2 - \sigma^2)$ . So the conditional variance will be above the unconditional variance whenever  $\varepsilon_{t-1}^2$  exceeds its unconditional variance  $\sigma^2$ . If  $\varepsilon_t$  were conditionally normally distributed, given the past, the fourth unconditional moment  $E(\varepsilon_t^4)$  will exceed  $3\sigma^4$  (the value in the normal distribution), so that the marginal distribution of  $\varepsilon_t$  will exhibit fatter tails than the normal distribution. Conditional variances of multi-step ahead forecast errors can also be established to depend on the past squared errors. These basic results indicated for the ARCH(1) model also tend to hold for the higher order ARCH( $q$ ) models. A basic impact of the ARCH errors in a process is in assessment of the accuracy of forecast, since the forecast errors will have conditional variances (given the past) that depend on the past. This allows for formation of correct and more informative (conditional) prediction intervals that would be obtained under the usual assumption that conditional error variances were constant independent of the past.

Now consider briefly the popular GARCH(1,1) model,  $h_t = E(\varepsilon_t^2 | \varepsilon_{t-1}, \dots) = \omega_0 + \omega_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}$ . Similar to the result for the ARCH(1) model, assuming that  $\omega_1 + \beta_1 < 1$ , it is readily shown that the unconditional variance of  $\varepsilon_t$  is equal to  $\sigma^2 = \text{Var}(\varepsilon_t) = \omega_0/[1 - (\omega_1 + \beta_1)]$ . Let  $v_t = \varepsilon_t^2 - h_t$ , so that the r.v.'s  $v_t$  have zero mean, and they are serially uncorrelated since

$$\begin{aligned} E((\varepsilon_t^2 - h_t)(\varepsilon_{t-j}^2 - h_{t-j})) &= E[E\{(\varepsilon_t^2 - h_t)(\varepsilon_{t-j}^2 - h_{t-j}) | \varepsilon_{t-1}, \dots\}] \\ &= E[(\varepsilon_{t-j}^2 - h_{t-j})E\{(\varepsilon_t^2 - h_t) | \varepsilon_{t-1}, \dots\}] = 0. \end{aligned}$$

Then, since  $h_t = \varepsilon_t^2 - v_t$ , note that the GARCH(1,1) model can be rearranged as  $\varepsilon_t^2 - v_t = \omega_0 + \omega_1 \varepsilon_{t-1}^2 + \beta_1 (\varepsilon_{t-1}^2 - v_{t-1})$ , or

$$\varepsilon_t^2 = \omega_0 + (\omega_1 + \beta_1) \varepsilon_{t-1}^2 + v_t - \beta_1 v_{t-1}, \quad (2.49)$$

This form reveals that the process of squared errors,  $\{\varepsilon_t^2\}$ , follows an ARMA(1,1) model with uncorrelated innovations  $v_t$  (however, the  $v_t$  are conditionally heteroscedastic and more strongly are Martingale differences). Although this process looks like a linear process, it is not, since  $\{v_t\}$  are uncorrelated but dependent). This fact motivates the use of the ACF and PACF of the squares  $\varepsilon_t^2$  for model specification and for basic preliminary checking for the presence of ARCH/GARCH effects in the errors  $\varepsilon_t$ . In practice, a starting point would be examination of the sample ACF and PACF of the squared residuals  $\hat{\varepsilon}_t^2$ , where  $\hat{\varepsilon}_t$  are residuals obtained from fitting of a usual ARMA( $p, q$ ) model (e.g., McLeod and Li, 1983 [205]). The correspondence between ARCH and ARMA is further noted via the condition,  $0 < \sum_{i=1}^q \omega_i < 1$ , that implies the roots of its characteristic equation lie outside the unit circle (like a causal ARMA).

For maximum likelihood (ML) estimation of models with ARCH or GARCH errors, assuming that  $\varepsilon_t$  is conditionally normally distributed as  $N(0, h_t)$ , we have the log-likelihood function (apart from some initial conditions) given by  $\log(L) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^T \log(h_t) - \frac{1}{2} \sum_{t=1}^T \varepsilon_t^2/h_t$ . Maximum likelihood estimation procedures, properties of the estimators, and Lagrange multiplier (LM) and likelihood ratio (LR) procedures for testing for the presence of ARCH effects in the conditional variances of the  $\varepsilon_t$  have been considered by Engle (1982) [99] and Weiss (1984) [258], among others.

As there are other excellent sources to learn about ARCH and GARCH models (see Tsay (2010) [252], Chapter 3), we have limited our discussion to some basic details.

**Preliminary Testing for ARCH Effect:** The test is the omnibus test similar to Ljung-Box test that examines the presence of autocorrelation up to certain number of lags, which is equivalent to testing,  $H_0 : \alpha_1 = \alpha_2 \cdots = \alpha_q = 0$  in the model

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2 + a_t \quad (2.50)$$

where  $a_t$  is white noise series; note that  $\{a_t\}$  does not need to follow a normal distribution. The usual  $F$ -test used in regression can be applied here; Let  $R^2$  be the coefficient of determination that result from (??) when  $\hat{\varepsilon}_t$ 's are used; then

$$F = \frac{R^2/q}{(1 - R^2)/(T - 2q - 1)} \sim \chi^2/q \quad (2.51)$$

We use  $\chi^2$  instead of  $F$  due to the fact that  $\varepsilon_t$ 's are estimated quantities. More informally, one could use the ACF of  $\hat{\varepsilon}_t^2$  and if any of the autocorrelations are out of the normal bounds,  $\pm 2/\sqrt{T}$ , we may examine further for ARCH effects.

**GARCH vs ARCH:** While ARCH models care for the time dependence in volatilities and have elegant properties such as having excessive kurtosis (matching the empirical finding related to asset returns), they tend to overpredict the volatility and treat both negative and positive,  $\varepsilon_t$ 's symmetrically. To overcome some of these issues and as

well as to come up with a more parsimonious representation (as in AR vs ARMA considerations) GARCH models stated in (2.49) are proposed.

**IGARCH:** The variance can persist for extended periods and can change at different time spans, thus leading to non-stationary behavior. In the GARCH(1,1) model, it is possible that  $w_1 + \beta_1 = 1$ , so that IGARCH(1,1) (Integrated GARCH) can be written as

$$\varepsilon_t = \sqrt{h_t} \cdot a_t, \quad h_t = w_0 + w_1 h_{t-1} + (1 - w_1) \varepsilon_{t-1}^2 \quad (2.52)$$

If  $h_t$  is proxied by  $\varepsilon_t^2$ , the above model leads to  $\varepsilon_t^2 - \varepsilon_{t-1}^2 = w_0 + w_1(\varepsilon_{t-1}^2 - \varepsilon_{t-2}^2)$ , which is similar in form to non-stationary autoregressive model. In practice because of level shifts in volatility, IGARCH appears to be a natural model to fit. Because the IGARCH(1,1) is similar to ARIMA(0,1,1), the model can be taken as an exponential smoothing model for the  $\varepsilon_t^2$  series; By iterative substitutions, we can show that

$$h_t = (1 - w_1)(\varepsilon_{t-1}^2 + w_1 \varepsilon_{t-2}^2 + w_1^2 \varepsilon_{t-3}^2 \dots) \quad (2.53)$$

where  $w_1$  can be taken as a discount factor.

**Prediction Performance of GARCH:** The predictive power of GARCH based on ex-post forecast evaluation criteria tend to suggest that the model provide poor forecasts, although the in-sample parameter estimates turn out to be highly significant. As Andersen and Bollerslev (1998) [10] note, in the model (2.49), the latent volatility,  $h_t$ , evolves over time. The approach to evaluate the performance of the model for  $h_t$  via  $\hat{\varepsilon}_t^2$  is not appropriate. While  $\hat{\varepsilon}_t^2$  provides an unbiased estimate of  $h_t$ , it is a noisy measurement due to idiosyncratic error terms associated with it; these error terms maybe due to the result of compounding frictions in the trading process. This will be clarified when we discuss the analysis of high-frequency data.

**Time-Varying ARCH Processes (tvARCH):** The underlying assumptions of ARCH models is stationarity and with changing pace of economic conditions the assumption of stationarity is not appropriate for modeling financial returns over long intervals. We may obtain a better fit by relaxing the assumption of stationarity. Dahlhaus and Subba Rao (2006) [76] generalize the ARCH model with time-varying parameters:

$$\varepsilon_t = \sqrt{h_t} \cdot a_t, \quad h_t = w_0(t) + \sum_{j=1}^{\infty} w_j(t) \varepsilon_{t-j}^2 \quad (2.54)$$

which  $a_t$ 's are i.i.d. with mean zero and variance, one. By rescaling the parameters to unit intervals, the tvARCH process can be approximated by a stationary ARCH process. A broad class of models resulting from (??) can be stated as:

$$\varepsilon_{t,N} = \sqrt{h_{t,N}} \cdot a_t, \quad h_{t,N} = w_0 \left( \frac{t}{N} \right) + \sum_{j=1}^p w_j \left( \frac{t}{N} \right) \varepsilon_{t-j,N}^2 \quad (2.55)$$

for  $t = 1, 2, \dots, N$ . This model captures the slow decay of the sample autocorrelations

in squared returns that is commonly observed in financial data which is attributed to the long memory of the underlying process. But tvARCH( $p$ ) is a non-stationary process that captures the property of long memory.

Fryzlewicz, Sapatinas and Subba Rao (2008) [124] propose a kernel normalized-least squares estimator which is easy to compute and is shown to have good performance properties. Rewriting (2.55) as

$$\varepsilon_{t,N}^2 = w_0\left(\frac{t}{N}\right) + \sum_{j=1}^p w_j\left(\frac{t}{N}\right) \varepsilon_{t-j,N}^2 + (a_t^2 - 1)h_{t,N}^2 \quad (2.56)$$

in the autoregressive form, the least squares criterion with the weight function,  $k(u_0, \chi'_{k-1,N})$ , where  $\chi'_{k-1,N} = (1, \varepsilon_{k-1,N}^2, \dots, \varepsilon_{k-p,N}^2)$  is

$$L_{t_0,N}(\alpha) = \sum_{k=p+1}^N \frac{1}{b_N} w\left(\frac{t_0-k}{b_N}\right) \frac{\left(\varepsilon_{k,N}^2 - \alpha_0 - \sum_{j=1}^p \alpha_j \varepsilon_{k-j,N}^2\right)^2}{k(u_0, \chi_{k-1,N})^2} \quad (2.57)$$

Minimizing (2.57) yields,  $\hat{a}_{t_0,N} = R_{t_0,N}^{-1} r_{t_0,N}$  where

$$R_{t_0,N} = \sum_{k=p+1}^N \frac{1}{b_N} ; w\left(\frac{t_0-k}{b_N}\right) \frac{\chi_{k-1,N} \chi'_{k-1,N}}{k(u_0, \chi_{k-1,N})^2} \quad (2.58)$$

and

$$r_{t_0,N} = \sum_{k=p+1}^N \frac{1}{b_N} w\left(\frac{t_0-k}{b_N}\right) \frac{\varepsilon_{k,N}^2 \chi_{k-1,N}}{k(u_0, \chi_{k-1,N})^2}$$

The kernel  $w$  is a function defined on  $[-\frac{1}{2}, \frac{1}{2}]$  which is of bounded variation. The choice of weight function suggested is as follows: it is motivated by the differing variances at various values of  $k$ . Estimate,

$$\hat{\mu}_{t_0,N} = \sum_{k=1}^N \frac{1}{b_N} w\left(\frac{t_0-k}{b_N}\right) \varepsilon_{k,N}^2 \quad (2.59)$$

and

$$s_{k-1,N} = \sum_{j=1}^p \varepsilon_{k-j,N}^2$$

and use  $k_{t_0,N}(s_{k-1}, N) = \hat{\mu}_{t_0,N} + s_{k-1,N}$  as the weight function. Thus the estimation is a two-stage scheme. For the bandwidth selection, the cross-validation method based on a subsample of the observations is suggested.

The methodology presented here is useful because of the changing nature of financial volatility. Real data applications will be discussed in Chapter 3.

## 2.6 Stylized Models for Variance of Asset Returns

Volatility has many important implications in finance. It is one of the most commonly used risk measure and plays a vital role in asset allocation. The estimate of volatility of an asset is obtained using the prices of stock or options or both. Three different measures that are normally studied are stated below:

- Volatility is the conditional standard deviation of daily or low frequency returns.
- Implied volatility is derived from options price under some assumed relationship between the options and the underlying stock prices.
- Realized volatility is an estimate of daily volatility using high frequency intraday returns.

In this section we will mainly focus on the first item and the others will be discussed in a later section on high frequency data.

Consider  $r_t = \ln(p_t) - \ln(p_{t-1})$ , return of an asset. We observed that ' $r_t$ ' exhibits no serial correlation. This does not imply that the series ' $r_t$ ' consists of independent observations. The plot of  $r_t^2$  given in Figure ?? clearly indicates that volatility tends to cluster over certain time spans. The autocorrelations in Figure ?? confirms that there is some time dependence for the volatility. In fact, there is some long range dependence that spans up to sixty days.

**Stylized fact 4:** Volatility clustering: High volatility events tend to cluster in time and this can be measured through positive correlations that exist over several lags.

To formally model the volatility, recall the conditional mean and the conditional variance of the return,  $r_t$ , are:  $\mu_t = E(r_t | F_{t-1})$ ,  $\sigma_T^2 = \text{Var}(r_t | F_{t-1}) = E((r_t - \mu_t)^2 | F_{t-1})$ .

We have observed for the exchange rate series that  $\mu_t \approx \text{constant}$  and is generally close to zero. But  $\sigma_t^2$  clusters around certain time spans and modeling variation in  $\sigma_t^2$  is typically done through ARCH/GARCH models and also through stochastic volatility models. The stochastic volatility models have the advantage of modeling jumps in prices that are also often used in derivative pricing. To begin with, we can test for the heteroscedasticity using the test in (2.51) or one of the following which are all equivalent asymptotically:

- **Ljung-Box Omnibus Test:**

$$Q_k = T(T+2) \cdot \sum_{j=1}^k \frac{(\hat{e}^{(j)})^2}{T-j} \sim \chi^2_{k-m} \quad (2.60)$$

where  $\hat{e}^{(j)}$  is the  $j$ th lag autocorrelation of  $r_t^2$ ;  $m$  is the number of independent parameters. Here  $H_0 : \rho_1 = \rho_2 = \dots = \rho_k = 0$ .

- **Lagrange Multiplier Test:** Regress  $r_t^2$  on  $r_{t-1}^2, \dots, r_{t-q}^2$  and obtain  $R^2$  the coefficient of determination; Test the  $H_0$ : Slope coefficients are all zero, by

$$T \cdot R^2 \sim \chi_q^2 \quad (2.61)$$

These tests were carried out on the exchange rate data; Table 2.3 has the result of Ljung-Box Test:

Table 2.3: Ljung-Box Chi-Square Statistics for Exchange rate data

h	12	24	36	48
$Q_h$	73.2	225.7	463.4	575.7
df	6	18	30	42
p-value	0.000	0.000	0.000	0.000

The Lagrange Multiplier test with  $q = 5$  also gives  $R^2 = 0.065$ , that  $T \cdot R^2 = 243.6$  and compared with  $\chi_5^2$  table values has  $p$ -value  $\approx 0$ . These results clearly confirm that there is some time dependence in the volatility.

Among the conditional heteroscedasticity models ARCH/GARCH that were discussed in the last section, GARCH is simple to use and also results in more parsimonious modeling. Therefore, we present only the estimated GARCH model here. For exchange rate data, the estimated GARCH model is  $\hat{\sigma}_t^2 = 2.61 \times 10^{-7} + 0.037r_{t-1}^2 + 0.95\hat{\sigma}_{t-1}^2$ . Observe that the coefficients  $r_{t-1}^2$  and  $\hat{\sigma}_{t-1}^2$  add up to unity approximately and thus indicating the volatility is best modeled through IGARCH, which seems to hold generally for many asset volatilities.

## 2.7 Other Topics: Bet Size etc.

## 2.8 Exercises (starred items have R-codes available)

1. Consider the exchange rate daily data from December 4, 2006 to November 5, 2010 (Rupee versus Dollar, Pound and Euro), Rates.csv.
  - (a) Compute the sample average, standard deviation and the first order autocorrelation of daily returns over the entire sample period. Test if the mean and the first order autocorrelation are significantly different from zero using the tests proposed in Section 2.3.2.
  - (b) Plot histograms of the returns over the entire sample period; Does the distribution look normal? Test it through Jarque-Bera test in (??).
  - (c) Aggregate the data at the weekly level; do (a) and (b) on the aggregated data. Compare the result with the result for the daily level.

2. For the returns of all three time series in problem 1, construct
  - (a) ARMA models for the returns; identify the model structure via ACF and PACF.
  - (b) GARCH models for the squared returns; compare the model coefficients for the three series. Comment on the difference if any.
  - (c) Is there a co-movement among the three exchange-rate series? To make the plots on a comparable scale, convert the starting points of the series unity. Does the co-movement vary over different time regimes? (Back up your claim with solid analysis.) Identify the transition states and speculate how you can exploit this for trading decisions.
3. Exchange Rates: The file `Exchange_Rates.csv` contains exchange rates between US dollar and twenty-five major currencies. The daily data spans from Jan 3, 2000 to April 10, 2012. Use this data to perform the following tasks: after computing the returns,  $r_t = \log(p_t) - \log(p_{t-1})$  for each series.
  - (a) Plot histograms of the returns and test if the distributions are normal, via Q-Q plots.
  - (b) Compute the auto-correlation function and identify which lags are significant. What is the tentative ARMA model?
  - (c) Is there any ARCH effect? Why?
  - (d) Fit GARCH(1,1,) and IGARCH(1,1) models using both normal and t-distributions for the innovations. Which volatility model appears to be the best?
4. Exchange rates: Consider the original series,  $p_t$ , for the duration starting June 11, 2003. Answer the following after dividing each series by  $p_1$ , so that the starting points are the same.
  - (a) Identify for each series, if there are any regime changes. Test if these changes are statistically significant.
  - (b) Identify how many clusters are there and if the membership changes under different regimes.
  - (c) Are the series co-integrated? Use Johansen's rank test (??) to identify the number of co-integrating relationships. Interpret the coefficients of the co-integrating vectors.
  - (d) Compare the results in (c) with (b). What is the connection between clustering and co-integration? Briefly discuss.
5. Consider daily price of Apple stock from `Apple.csv`. The data can be obtained from Yahoo Finance and have 7 columns (namely, Date, Open, High, Low, Close, Volume, Adj Close). We focus on the adjusted closing price in the last column. Note the data needs to be reversed from the past to the recent.

- (a) Compute the daily log returns. Is there any serial correlation in the daily log returns? Use the test for white noise as outlined in the text.
  - (b) Consider the pivot quantities on the average of high and low price and the average of high, low and close prices. Compute the returns based on the pivot log prices and test for serial correlation. Compare this result with the finding in (a).
  - (c) Consider the log price series of AAPL stock. Is the log price series unit-root nonstationary? Perform a unit-root (Dickey-Fuller) test to answer the question and present your conclusion.
6. Consider daily price of Apple stock again.
- (a) Compute various measures of variance computed from the entries of the price bars. Comment on their correlation with log volume.
  - (b) Use the ARIMA modeling to come up with a parsimonious model for log volume. Comment on the model accuracy by setting aside a validation data set.
7. The *Roll* model for trade prices discussed in the text can be more specifically stated as,

$$p_t^* = p_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$

and  $p_t = p_t^* + cs_t$ , where  $s_t \in \{+1, -1\}$

Here  $p_t^*$  is the “true” value of a security and  $p_t$  is the observed trade price, which differ because of the bid-offer spread:  $c$  is the half-spread and  $s_t$  is the direction of the  $t$ th trade.  $s_t$  and  $\epsilon$  are serially independent and independent of each other.

- (a) Calculate the serial correlation of the observed prices  $p_t$ . Construct an MA(1) model with the same autocorrelation structure. Of the two such models, take the invertible one.
- (b) For the invertible model, construct the associated AR model.
- (c) Use the model to show how to estimate the bid-ask spread and apply the results on the Apple data used in Problem (5) and (6).

**Questions 8–10 refer to Level III data described below:**

The dataset consists of (CSV) file that contains the entire trading session, including early and late hours from 6:00 a.m. to 8:00 p.m. EST. It contains intraday depth book activity several major tickers from up to 5 major national exchanges: Nasdaq, Direct Edge, NYSE, ARCA, and BATS (exact number of exchanges depends on the historical period). All messages are consolidated into one file ordered by timestamp. This dataset allows to build a full national depth book (super book) at any moment intraday.

For the analysis of Problems 8–10, we consider the data for CISCO.

Table 2.4: Column Format

Column	Description
Timestamp	Number of milliseconds after the midnight.
Ticker	Equity symbol (up to 8 characters)
Order	Unique order ID.
T	Message type. Allowed Values: <ul style="list-style-type: none"> <li>• “B”–Add buy order</li> <li>• “S”–Add sell order</li> <li>• “E”–Execute outstanding order in part</li> <li>• “F”–Execute outstanding order in full</li> <li>• “D”–Delete outstanding order in full</li> <li>• “X”–Bulk volume for the cross event</li> <li>• “T”–Execute non-displayed order</li> </ul>
Shares	Order quantity, available for the “B”, “S”, “E”, “X”, “C”, “T” messages. Zero for “F” and “D” messages.
Price	Order price, available for the “B”, “S”, “X”, and “T” order messages. Zero for cancellation and executions. The last 4 digits are decimal digits. The decimal portion is padded on the right with zeros. The decimal point is implied by position; it does not appear inside the price field. Divide by 10000 to convert into currency value.
MPID	Market Participant ID associated with the transaction (4 characters)
MCID	U Market Center Code (originating exchange–1 character)

8. You need to aggregate the data into 5-min intervals before answering the following.
- (a) Let  $x_t$  denote the number of trades in the  $t$ th 5-minute interval. Ignore the time gaps between trading days. Plot the time series and its ACF. Determine if there are intraday period patterns in the series.
  - (b) Using the last transaction in the  $t$ th 5-minute interval as the stock price in that interval, plot the time series  $y_t$  of 5-minute returns during the period and the corresponding ACF.
  - (c) Consider the bivariate time series  $(x_t, y_t)$ . How does  $y_t$  vary with  $x_t$ ? Are there intraday periodic patterns in  $(x_t, y_t)$ ?
  - (d) Plot the durations of the transaction times for Mondays and Fridays. Is there any difference in the patterns?
  - (e) Fit GARCH models for the durations and interpret the coefficients.
  - (f) Is there any particular exchange that offers better price?

9. Use the transaction data in Problem 8. There are seventy-eight five-minute intervals in a trading day. Let  $d_i$  be the average of all log durations for the  $i$ th 5-minute interval across all trading days. Let  $t_j$  be tick time of a trade during the  $i$ th 5-minute interval, define an adjusted duration as  $\Delta t_j^* = \Delta t_j / \exp(d_i)$  where  $\Delta t_j = t_j - t_{j-1}$ .
  - (a) Is there a diurnal pattern in the adjusted duration series?
  - (b) Build EACD and WACD models for the adjusted duration and compare them. Now aggregate the data over the 5-minute intervals; Let  $r_t$  be the return and  $V_t$  is the volume of transactions in the  $t$ th interval.
  - (c) Is there any relationship between the volume  $V_t$  and volatility  $r_t^2$ ? If there is develop a model that will capture the dependence of volatility on volume. Let  $\bar{p}_t$  be the average price in the  $t$ th 5-minute interval.
10. Use the CISCO execution data only; aggregate at various time intervals: 1 min, 2 min, ..., 15 min. For these levels record
  - (a) Average price, VWAP, realized volatility and aggregate volume.
  - (b) By relating the returns and volatility estimates based on average price to realized volatility, identify the optimal duration for aggregation that will cancel out the market friction but pick up the true information. (You may want to refer to Bandi and Russell (2006) [20]).



# 3

## Multivariate Time Series Models

### CONTENTS

3.1	Multiple Time Series Modeling and Co-integration .....	71
3.1.1	Introduction .....	71
3.1.2	Co-integration, Co-movement and commonality in Multiple Time Series .....	73
3.1.3	Multivariate Regression .....	79

### 3.1 Multiple Time Series Modeling and Co-integration

#### 3.1.1 Introduction

Multiple time series analysis is concerned with modeling and estimation of dynamic relationships among  $m$  related time series  $y_{1t}, \dots, y_{mt}$ , based on observations on these series over  $T$  equally spaced time points  $t = 1, \dots, T$ , and also between these series and potential exogenous time series variables  $x_{1t}, \dots, x_{nt}$ , observed over the same time period. We shall explore the use of these techniques in statistical arbitrage in leveraging multiple markets in Chapter 3. We first introduce a general model for multiple time series modeling, but will specialize to vector autoregressive (VAR) models for more detailed investigation. As some concepts are similar to those of the univariate models discussed in Section 2.3, the presentation will be brief.

Let  $Y_t = (y_{1t}, \dots, y_{mt})'$  be an  $m \times 1$  multiple time series vector of response variables and let  $X_t = (x_{1t}, \dots, x_{nt})'$  be an  $n \times 1$  vector of input or predictor time series variables. Let  $\epsilon_t$  denote an  $m \times 1$  white noise vector of errors, independently distributed over time with  $E(\epsilon_t) = 0$  and  $\text{Cov}(\epsilon_t) = \Sigma_{\epsilon\epsilon}$ , a positive-definite matrix. We consider the multivariate time series model

$$Y_t = \sum_{s=0}^p C_s X_{t-s} + \epsilon_t \quad (3.1)$$

where the  $C_s$  are  $m \times n$  matrices of unknown parameters. In the general setting, the ‘input’ vectors  $X_t$  could include past (lagged) values of the response series  $Y_t$ . In the context of trading  $Y_t$  could denote the prices of related assets,  $X_{t-s}$  could represent the past values of  $Y_t$  and the values of volume, volatility, market, industry factors, etc. Note that the model (3.1) can be written in the form of a multivariate regression

model. An important issue that arises in the multiple series modeling is as follows: even for moderate values of the dimensions  $m$  and  $n$ , the number of parameters that must be estimated in model (3.1), when no constraints are imposed on the matrices  $C_s$  can become quite large. Because of the potential complexity of these models, it is often useful to consider canonical analyses or other dimension reduction procedures such as reduced-rank regression methods described in Reinsel and Velu (1998) [230]. This may also lead to more efficient models due to the reduction in the number of unknown parameters.

The class of models stated in (3.1) is rather flexible and provides many options for modeling multivariate time series data. Specifically consider the VAR( $p$ ) model,

$$Y_t = \Phi_0 + \sum_{j=1}^p \Phi_j Y_{t-j} + \epsilon_t \quad (3.2)$$

with ' $p$ ' lags. As in Section 2.3, we can define the mean, variance, and autocovariance functions for stationary series as follows; in addition we have the cross-correlation matrix that captures the leading relationships among all ' $m$ ' series.

Mean:	$\mu = E(Y_t)$
Variance:	$\Gamma(0) = E[(Y_t - \mu)(Y_t - \mu)']$
Autocovariance:	$\Gamma(l) = E[(Y_t - \mu)(Y_{t-l} - \mu)']$
Cross-correlation:	$\rho(l) = D^{-1}\Gamma(l)D^{-1}$

Here  $D$  is the diagonal matrix of size  $m \times m$ , consisting of the ' $m$ ' standard deviations taken from the variance terms, from the diagonal of  $\Gamma(0)$ . It is easy to observe,  $\Gamma(l) = \Gamma(-l)'$  and  $\rho(l) = \rho(-l)'$ . The linear dependence among all the variables in terms of leads-lags is fully captured by examining  $\rho_{ij}(l)$ . There are excellent sources to study about modeling multiple time series (see Tsay (2010) [252] and Reinsel (2002) [229]). We focus mainly on the use of Vector AR( $p$ ) (VAR( $p$ )) model in studying the relationships among stocks or markets; we elaborate on them below.

Using the back-shift operator  $B$ , the model (3.2), VAR( $p$ ) can be written as,  $\Phi(B)Y_t = \Phi_0 + \epsilon_t$ , where  $\Phi(B) = I - \Phi_1 B - \dots - \Phi_p B^p$  is a matrix polynomial. It can be shown for the process  $Y_t$  to be stationary, the roots of  $|\Phi(\lambda) - I| = 0$  should all lie outside the unit circle. If  $Y_t$  is stationary, then  $\mu = [\Phi(1)]^{-1}\Phi_0$  where  $\Phi(1) = I - \Phi_1 - \dots - \Phi_p$ . An easy consequence of this form is the result of the Yule-Walker equations, which are useful to identify,  $\Phi$ 's, the autoregressive coefficients from the moment conditions:

$$\begin{aligned} \Gamma(l) &= \Phi_1 \Gamma(l-1) + \dots + \Phi_p \Gamma(l-p) \\ \rho(l) &= \Phi_1^* \rho(l-1) + \dots + \Phi_p^* \rho(l-p) \end{aligned} \quad (3.3)$$

Here  $\Phi_i^* = D^{-\frac{1}{2}}\Phi_i D^{-\frac{1}{2}}$ . A general approach to building a VAR model is to sequentially increase the lag value, ' $p$ ' and check if the added lag is significant. The lag coefficients are estimated from (3.3) with their sample counter-parts and they provide good initial estimates. The VAR( $p$ ) model building at the  $i$ th stage is fitting the

model,  $Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \dots + \Phi_i Y_{t-i} + \epsilon_t$  via least-squares method and estimate the residual covariance matrix via

$$\hat{\Sigma}_i = \left( \frac{1}{T - 2i - 1} \right) \sum_{t=i+1}^T \hat{\epsilon}_t^{(i)} \hat{\epsilon}_t^{(i)'} , \quad i \geq 0 \quad (3.4)$$

where  $\hat{\epsilon}_t^{(i)} = Y_t - \hat{\Phi}_0^{(i)} - \hat{\Phi}_1^{(i)} Y_{t-1} - \dots - \hat{\Phi}_i^{(i)} Y_{t-i}$  are the residuals. The significance of added lag ' $i$ ' is tested with  $H_0 : \Phi_i = 0$  versus  $H_a : \Phi_i \neq 0$ , using the likelihood-ratio criterion:

$$M(i) = -(T - m - i - \frac{3}{2}) \ln \left( \frac{|\hat{\Sigma}_i|}{|\hat{\Sigma}_{i-1}|} \right) \quad (3.5)$$

This is distributed asymptotically as a Chi-squared distribution with  $m^2$  degrees of freedom. Alternatively, the AIC or BIC criteria that penalizes for over parameterization can be used;

$$\begin{aligned} AIC(i) &= \ln(|\hat{\Sigma}_i|) + \frac{2m^2 i}{T} \\ BIC(i) &= \ln(|\hat{\Sigma}_i|) + \frac{m^2 i \ln(T)}{T} \end{aligned} \quad (3.6)$$

Here  $\tilde{\Sigma}_i$  is the maximum likelihood estimate of  $\Sigma_i$  where the divisor in (3.4) is  $T$  instead of  $T - 2i - 1$ .

Two important properties of VAR( $p$ ) model models are worth noting and will be used later. First modeling several variables jointly usually results in lower order vector model as the marginal models for the individual variables can result in a higher order of autoregressive moving average terms that may be hard to interpret.

**Result 1:** If  $Y_t$  is a  $m$ -dimensional series following VAR( $p$ ), the components of  $Y_t$  follow ARMA with maximal orders  $[mp, (m-1)p]$ . Second in many practical situations a certain linear combination of  $Y_t$  is important to study, particularly in dimension reduction aspect whether using the principal components or canonical variables. The linear combination can still exhibit time series behavior.

**Result 2:** Let  $Y_t \sim \text{VAR}(p)$ , then  $Z_t = l' Y_t$ , where  $l$  is a  $m \times 1$  known vector, follows ARMA with maximum orders  $[mp, (m-1)p]$ .

### 3.1.2 Co-integration, Co-movement and commonality in Multiple Time Series

The dynamic relationships among components of the vector time series,  $Y_t$  can be studied in various ways. While the estimation and inference aspects are similar to the univariate series discussed in Section 2.3, because of the sheer dimension of  $Y_t$  there are more challenges in the precise estimation of the model coefficients but there are also more opportunities to study the linkages among the components of

$Y_t$ . The number of unknown elements in (3.2) is  $m^2(p + 1)$  which can be quite large even for modest values of ‘ $m$ ’ and ‘ $p$ ’; therefore considerations of the dimension reduction aspects in modeling (3.2) have been given by Box and Tiao (1977) [39] and Brillinger (1981) [45] among others. The traditional tool used in multivariate analysis for dimension reduction, principal components method, extracting linear combinations that account for the most variance is based on the contemporaneous covariance matrix, and thus it ignores the time dependence. But Result 2, stated in the previous section indicates that the linear combination can exhibit time dependence but we expect the optimal linear combination to exhibit other desirable features, which are described here. Identifying the common features among the series have been focused on by several authors; for example Engle and Kozicki (1993) [104], Vahid and Engle (1993) [255] among others. The series if they are stationary, are studied for their co-movement and if they are non-stationary, they are studied for their co-integration. A method that can capture both aspects can be stated as reduced-rank autoregressive modeling procedure which we briefly describe below. The basic VAR( $p$ ) model in (3.6) without the constant term can be written as

$$Y_t = CY_{t-1}^* + \epsilon_t \quad (3.7)$$

where  $C = [\Phi_1, \dots, \Phi_p]$  and  $Y_{t-1}^* = (Y_{t-1}, Y_{t-2}, \dots, Y_{t-p})'$ . Observe that ‘ $C$ ’ matrix is of dimension  $m \times mp$ . We impose a condition that

$$\text{Rank}(C) = r < m, \quad (3.8)$$

a condition that is often met in modeling large scale data. This has two implications; one the ‘ $C$ ’ matrix can be written as

$$C = A \cdot B \quad (3.9)$$

where  $A$  is  $m \times r$  and  $B$  is  $r \times n$  and are matrices of full rank. The other is that, there are  $m - r$  constraints on ‘ $C$ ’ matrix,

$$l_i' \cdot C = 0, \quad i = 1, \dots, (m - r) \quad (3.10)$$

The first one recovers, via  $B \cdot Y_{t-1}^*$ , the most relevant information from the past  $Y_{t-1}$  that relates to  $Y_t$ . The second implication reveals how certain linear combinations,  $l_i' \cdot Y_t$  are not related to the past  $Y_{t-1}^*$  and can be taken as white noise series. We are assuming that the process ‘ $Y_t$ ’ is stationary; that is  $\det(\Phi(z)) \neq 0$  for all complex numbers  $|z| \leq 1$ . Estimation of  $A$ ,  $B$ , and  $l$ ’s all follow from the theory of reduced-rank regression (see Reinsel and Velu (1998), Ch 2)[230].

For the stationary vector AR( $p$ ) process  $\{Y_t\}$  above, note the covariance matrix at lag  $j$  as defined earlier,  $\Gamma(j) = E(Y_{t-j}Y_t')$ , and set  $\Gamma_* = [\Gamma(1)', \dots, \Gamma(p)']' = E(Y_{t-1}^*Y_t')$ , and  $\Gamma_p = E(Y_{t-1}^*Y_{t-1}^{**})$  as the  $mp \times mp$  matrix which has  $\Gamma(i - j)$  in the  $(i, j)$ th block. Under the reduced rank assumption, for any choice of positive definite matrix  $\Gamma$ , it can be shown that the population coefficient matrices  $A$  and  $B$  in the reduced rank AR model can be expressed as

$$\begin{aligned} A &= \Gamma^{-1/2}[V_1, \dots, V_r] = \Gamma^{-1/2}V = [\alpha_1, \dots, \alpha_r], \\ B &= V'\Gamma^{1/2}\Gamma_*'\Gamma_p^{-1} = [\beta_1, \dots, \beta_r]', \end{aligned} \quad (3.11)$$

where the  $V_j$  are (normalized) eigenvectors of the matrix  $\Gamma^{1/2} \Gamma_p' \Gamma_p^{-1} \Gamma_* \Gamma^{1/2}$  associated with the  $r$  largest eigenvalues  $\lambda_j^2$ . This also implies that  $A$  and  $B$  in (3.11) satisfy the normalizations  $B\Gamma_p B' = \Lambda^2$  and  $A'\Gamma A = I_r$ , where  $\Lambda^2 = \text{diag}(\lambda_1^2, \dots, \lambda_r^2)$ .

Based on a sample  $Y_1, \dots, Y_T$ , let  $\hat{\Gamma}_p = \frac{1}{T} \sum_t Y_{t-1}^* Y_{t-1}'$  and  $\hat{\Gamma}_* = \frac{1}{T} \sum_t Y_{t-1}^* Y_t'$ . Then the (conditional) ML estimates of  $A$  and  $B$  corresponding to the population quantities in (3.11) are given by

$$\begin{aligned}\hat{A} &= \tilde{\Sigma}_{\epsilon\epsilon}^{1/2} [\hat{V}_1, \dots, \hat{V}_r] = \tilde{\Sigma}_{\epsilon\epsilon}^{1/2} \hat{V} = [\hat{\alpha}_1, \dots, \hat{\alpha}_r], \\ \hat{B} &= \hat{V}' \tilde{\Sigma}_{\epsilon\epsilon}^{-1/2} \hat{\Gamma}_* \hat{\Gamma}_p^{-1} = [\hat{\beta}_1, \dots, \hat{\beta}_r]',\end{aligned}\quad (3.12)$$

where  $\hat{V} = [\hat{V}_1, \dots, \hat{V}_r]$  and  $\hat{V}_j$  is the (normalized) eigenvector of  $\tilde{\Sigma}_{\epsilon\epsilon}^{-1/2} \hat{\Gamma}_* \hat{\Gamma}_p^{-1} \hat{\Gamma}_* \tilde{\Sigma}_{\epsilon\epsilon}^{-1/2}$  corresponding to the  $j$ th largest eigenvalue  $\hat{\lambda}_j^2$ , and  $\tilde{\Sigma}_{\epsilon\epsilon} = (1/T)[\mathbf{Y}\mathbf{Y}' - \mathbf{Y}\mathbf{Y}_{-1}^* (\mathbf{Y}_{-1}^* \mathbf{Y}_{-1}')^{-1} \mathbf{Y}_{-1}^* \mathbf{Y}']$ . Here,  $\mathbf{Y} = [Y_1, \dots, Y_T]$  and  $\mathbf{Y}_{-1}^* = [Y_0^*, \dots, Y_{T-1}^*]$  are the  $m \times T$  and  $mp \times T$  data matrices containing the observations on  $Y_t$  and  $Y_{t-1}^*$ , respectively, for  $t = 1, \dots, T$ .

The asymptotic theory holds for the LR testing for the rank of the coefficient matrix  $C$  in the reduced rank autoregressive model. That is, under the null hypothesis  $H_0 : \text{rank}(C) \leq r$ , the LR test statistic  $\mathcal{M} = [T - mp + (mp - m - 1)/2] \sum_{j=r+1}^m \log(1 + \hat{\lambda}_j^2)$  has an asymptotic  $\chi^2_{(m-r)(mp-r)}$  distribution. It can be shown that  $\hat{\lambda}_j^2 = \frac{\hat{\rho}_j^2}{1 - \hat{\rho}_j^2}$  where  $\hat{\rho}_j^2$  are the canonical correlations between  $Y_t$  and  $Y_{t-1}^*$  and the LR test above can be alternatively stated in terms of the canonical correlations.

The description of canonical correlations can be found in a standard multivariate text (see Anderson (1963)). It is based on the concept of maximizing the correlations between two linear combinations of  $Y_t$  and  $Y_{t-1}^*$ . The linear vectors are related to the eigenvectors of  $R_* = \hat{\Gamma}_{(0)}^{-1/2} \hat{\Gamma}_* \hat{\Gamma}_p^{-1} \Gamma_* \Gamma_{(0)}^{-1/2}$  that correspond to the ' $r$ ' largest eigenvalues. Because of the fact that  $\tilde{\Sigma}_{\epsilon\epsilon} = \hat{\Gamma}_{(0)} - \hat{\Gamma}_* \hat{\Gamma}_p^{-1} \Gamma_*$ , the eigenvalues  $\hat{\lambda}_j^2$  and  $\hat{\rho}_j^2$  - the canonical correlations are related as stated above.

We briefly address one additional issue unique to time series analysis. This is the aspect of nonstationarity of the vector time series process  $\{Y_t\}$  and the relation of nonstationarity to the canonical correlation analysis. Thus, in the comovement and the cointegration analysis, the canonical correlations as descriptive tools play an important role.

One interesting result in the work of Box and Tiao (1977) [39] is the development of the correspondence between nonstationarity in the VAR(1) process, as reflected by the roots of  $\det\{I - \Phi(z)\} = 0$  (equivalently, the eigenvalues of  $\Phi$ ) and the (squared) canonical correlations between  $Y_t$  and  $Y_{t-1}^*$  in the canonical analysis, as reflected in the eigenvalues of the matrix  $R_* = \Gamma(0)^{-1/2} \Gamma(1)' \Gamma(0)^{-1} \Gamma(1) \Gamma(0)^{-1/2}$ . To develop the correspondence more easily, note that we can write the matrix  $R_*$  as  $R_* = \Gamma(0)^{-1/2} \Phi \Gamma(0) \Phi' \Gamma(0)^{-1/2}$ , since  $\Phi = \Gamma(1)' \Gamma(0)^{-1}$ . The specific result to be established is that the number of eigenvalues of  $\Phi$  approaching the unit circle (the nonstationary boundary) is the same as the number of canonical correlations between

$Y_t$  and  $Y_{t-1}$  that approach unity. This correspondence was observed by Box and Tiao (1977) [39]. Velu, Wichern, and Reinsel (1987) [257] gave an alternative more direct proof of the result.

The main implication of the result in Box and Tiao (1977) [39] is that the existence of  $d$  canonical correlations  $\rho_j$  close to one in the vector AR(1) process implies that the associated canonical components will be (close to) nonstationary series, and these may reflect the nonstationary trending or dynamic growth behavior in the multivariate system, while the remaining  $r = m - d$  canonical variates possess stationary behavior. This approach to investigation of the nonstationary aspects of the vector AR process has more recently been explored by Bossaerts (1988) [34] and Bewley and Yang (1995) [26].

A particularly important instance of nonstationarity in the vector AR process is the case where the only roots of  $\det\{\Phi(z)\} = 0$  on the unit circle are unity. This corresponds typically to the situation where the nonstationary vector process  $Y_t$  becomes stationary upon first differencing, that is,  $W_t = (1 - L)Y_t$  is stationary. For the present, we will establish a connection between the number of unit eigenvalues of  $\Phi$  in the vector AR(1) process and the number of zero canonical correlations in a certain canonical analysis. For this, we first note that the vector AR(1) model,  $Y_t = \Phi Y_{t-1} + \epsilon_t$ , can be expressed in an equivalent form as

$$W_t \equiv Y_t - Y_{t-1} = -(I - \Phi)Y_{t-1} + \epsilon_t, \quad (3.13)$$

which is referred to as an error-correction form. Notice that if  $\Phi$  has eigenvalues approaching unity then  $I - \Phi$  will have eigenvalues approaching zero (and will become of reduced rank). The methodology of interest for exploring the number of unit roots in the AR(1) model is motivated by the model from (3.13), and consists of a canonical correlation analysis between  $W_t = Y_t - Y_{t-1}$  and  $Y_{t-1} (\equiv X_t)$ . Under stationarity of the process  $Y_t$ , note that the covariance matrix between  $W_t$  and  $Y_{t-1}$  is equal to  $\Sigma_{wx} = -(I - \Phi)\Gamma(0)$ . Thus, the squared canonical correlations between  $W_t$  and  $Y_{t-1}$  are the eigenvalues of

$$\Sigma_{ww}^{-1}\Sigma_{wx}\Sigma_{xx}^{-1}\Sigma_{xw} = \Sigma_{ww}^{-1}(I - \Phi)\Gamma(0)(I - \Phi)', \quad (3.14)$$

where  $\Sigma_{ww} = \text{Cov}(W_t) = 2\Gamma(0) - \Phi\Gamma(0) - \Gamma(0)\Phi' = (I - \Phi)\Gamma(0) + \Gamma(0)(I - \Phi)'$ . We now present the basic result.

**Result:** Suppose  $Y_t$  follows the stationary vector AR(1) model, with error-correction form given in (3.13), where  $\text{Cov}(\epsilon_t) = \Sigma_{ee}$  is assumed to be positive definite and fixed. With respect to variation of the matrix  $\Phi$ , the condition that  $d \leq m$  of the eigenvalues of  $\Sigma_{ww}^{-1}(I - \Phi)\Gamma(0)(I - \Phi)'$  (i.e., squared canonical correlations between  $W_t$  and  $Y_{t-1}$ ) tend to zero implies that at least  $d$  of the eigenvalues of  $\Phi$  approach unity. (For a proof see Reinsel and Velu (1998), p 138-139) [230]

Thus for a vector process,  $Y_t$ , the possibility exists for each component series  $y_{it}$  to be nonstationary with its first difference  $(1 - B)y_{it}$  stationary (in which case  $y_{it}$  is said to be integrated of order one), but such that certain linear combinations  $z_{it} = b_i'Y_t$  of  $Y_t$  will be stationary. In such circumstances, the process  $Y_t$  is said to be *cointegrated*

with cointegrating vectors  $b_i$  (e.g., Engle and Granger (1987) [101]). An interpretation of cointegrated vector series  $Y_t$ , particularly related to economics, is that the individual components  $y_{it}$  share some common nonstationary components or “common trends” and, hence, they tend to have certain similar movements in their long-term behavior. A related interpretation is that the component series  $y_{it}$ , although they may exhibit nonstationary behavior, satisfy (approximately) a long-run equilibrium relation  $b'_i Y_t \approx 0$  such that the process  $z_{it} = b'_i Y_t$ , which represents the deviations from the equilibrium, exhibits stable behavior and so is a stationary process. This fact is exploited in ‘pairs trading’ or more broadly in trading multiple stocks and can be very useful in the context of portfolio rebalancing.

A specific nonstationary AR structure for which cointegration occurs is the model where  $\det\{\Phi(z)\} = 0$  has  $d < m$  roots equal to one and all other roots are greater than one in absolute value, and also the matrix  $\Phi(1) = I - \Phi_1 - \dots - \Phi_p$  has rank equal to  $r = m - d$ . Then, for such a process, it can be established that  $r$  linearly independent vectors  $b_i$  exist such that  $b'_i Y_t$  is stationary, and  $Y_t$  is said to have cointegrating rank  $r$ . Properties of nonstationary cointegrated systems have been investigated by Engle and Granger (1987) [101]. We will illustrate the usefulness of reduced rank estimation methods in the analysis of such models.

The VAR( $p$ ) model,  $\Phi(B)Y_t = Y_t - \sum_{j=1}^p \Phi_j Y_{t-j} = \epsilon_t$ , can be represented in the *error-correction form* (Engle and Granger (1987) [101]) as  $\Phi^*(B)(1-B)Y_t = -\Phi(1)Y_{t-1} + \epsilon_t$ ; that is,

$$W_t = CY_{t-1} + \sum_{j=1}^{p-1} \Phi_j^* W_{t-j} + \epsilon_t, \quad (3.15)$$

where  $W_t = (1-B)Y_t = Y_t - Y_{t-1}$ ,  $\Phi^*(B) = I - \sum_{j=1}^{p-1} \Phi_j^* B^j$ , with  $\Phi_j^* = -\sum_{i=j+1}^p \Phi_i$ , and  $C = -\Phi(1) = -(I - \sum_{j=1}^p \Phi_j)$ .

The form (3.15) is obtained from direct matrix algebra, since  $\Phi(B)$  can always be re-expressed as  $\Phi(B) = \Phi^*(B)(1-B) + \Phi(1)B$  with  $\Phi^*(B)$  as defined. For example, in the VAR(2) model that is most likely to apply in the asset prices movements, we have  $\Phi(B)\Phi^*(B)(1-B) + \Phi(1)B$ , where  $\Phi^*(B) = I + \Phi_2 B \equiv I - \Phi_1^* B$  with  $\Phi_1^* = -\Phi_2$ . Under the assumptions on  $\Phi(B)$ , it can also be shown that  $\Phi^*(B)$  is a stationary AR operator having all roots of  $\det\{\Phi^*(z)\} = 0$  outside the unit circle. The error-correction form (3.15) is particularly useful because the number of unit roots in the AR operator  $\Phi(B)$  can conveniently be incorporated through the “error-correction” term  $CY_{t-1}$ , so that the nature of nonstationarity of the model is conveniently concentrated in the behavior of the single coefficient matrix  $C$  in this form.

When maximum likelihood estimation of the parameter matrix  $C$  in (3.15) is considered subject to the reduced-rank restriction that  $\text{rank}(C) = r$ , it is convenient to express  $C$  as  $C = AB$ , where  $A$  and  $B$  are  $m \times r$  and  $r \times m$  full rank matrices, respectively.

Specifically, note that the model can be written as

$$W_t = ABY_{t-1} + \sum_{j=1}^{p-1} \Phi_j^* W_{t-j} + \epsilon_t \equiv ABY_{t-1} + DW_{t-1}^* + \epsilon_t, \quad (3.16)$$

where  $D = [\Phi_1^*, \dots, \Phi_{p-1}^*]$  and  $W_{t-1}^* = (W'_{t-1}, \dots, W'_{t-p+1})'$ .

When the  $\text{rank}(C) = r$ , the Gaussian reduced-rank estimator in model (3.15) can also be obtained explicitly through the *partial canonical correlation analysis*. This approach was presented by Johansen (1988, 1991) [173] [174]. To describe the results, let  $\tilde{W}_t$  and  $\tilde{Y}_{t-1}$  denote the “adjusted” or residual vectors from the least squares regressions of  $W_t$  and  $Y_{t-1}$ , respectively, on the lagged values  $W_{t-1}^* = (W'_{t-1}, \dots, W'_{t-p+1})'$ , and let

$$S_{\tilde{w}\tilde{w}} = \frac{1}{T} \sum_{t=1}^T \tilde{W}_t \tilde{W}'_t, \quad S_{\tilde{w}\tilde{y}} = \frac{1}{T} \sum_{t=1}^T \tilde{W}_t \tilde{Y}'_{t-1}, \quad S_{\tilde{y}\tilde{y}} = \frac{1}{T} \sum_{t=1}^T \tilde{Y}_{t-1} \tilde{Y}'_{t-1}.$$

Then, the sample partial canonical correlations  $\hat{\rho}_i$  between  $W_t$  and  $Y_{t-1}$ , given  $W_{t-1}, \dots, W_{t-p+1}$ , and the corresponding vectors  $\hat{V}_i^*$  are the solutions to

$$(\hat{\rho}_i^2 I_m - S_{\tilde{w}\tilde{w}}^{-1/2} S_{\tilde{w}\tilde{y}} S_{\tilde{y}\tilde{y}}^{-1} S_{\tilde{y}\tilde{w}} S_{\tilde{w}\tilde{w}}^{-1/2}) \hat{V}_i^* = 0, \quad i = 1, \dots, m, \quad (3.17)$$

The *reduced-rank Gaussian estimator* of  $C = AB$  can then be obtained explicitly as  $\hat{C} = S_{\tilde{w}\tilde{w}}^{1/2} \hat{V}_* \hat{V}'_* S_{\tilde{w}\tilde{w}}^{-1/2} \tilde{C}$ , where  $\tilde{C} = S_{\tilde{w}\tilde{y}} S_{\tilde{y}\tilde{y}}^{-1}$  is the full rank LS estimator, and  $\hat{V}_* = [\hat{V}_1^*, \dots, \hat{V}_r^*]$  are the (normalized) vectors corresponding to the  $r$  largest partial canonical correlations  $\hat{\rho}_i$ ,  $i = 1, \dots, r$ . This form of the estimator provides the reduced-rank factorization as  $\hat{C} = (S_{\tilde{w}\tilde{w}}^{1/2} \hat{V}_*)(\hat{V}'_* S_{\tilde{w}\tilde{w}}^{-1/2} \tilde{C}) \equiv \hat{A} \hat{B}$ , with  $\hat{A} = S_{\tilde{w}\tilde{w}}^{1/2} \hat{V}_*$  satisfying the normalization  $\hat{A}' S_{\tilde{w}\tilde{w}}^{-1} \hat{A} = I_r$ . The Gaussian estimator for the other parameters  $\Phi_1^*, \dots, \Phi_{p-1}^*$  can be obtained by ordinary least squares regression of  $W_t - \hat{C} Y_{t-1}$  on the lagged variables  $W_{t-1}, \dots, W_{t-p+1}$ .

The dual implications, (3.9) and (3.10) of reduced-rank structure come into play in modeling VAR as well. Note from the earlier discussion that the expression  $\hat{B} = \hat{V}'_* S_{\tilde{w}\tilde{w}}^{-1/2} \tilde{C}$  shows that the estimated cointegrating vectors (rows of  $\hat{B}$ ) are obtained from the vectors that determine the first  $r$  canonical variates of  $Y_{t-1}$  in the partial canonical correlation analysis. In addition, it is seen from the form of the reduced-rank estimator  $\hat{C} = S_{\tilde{w}\tilde{w}}^{1/2} \hat{V}_* \hat{V}'_* S_{\tilde{w}\tilde{w}}^{-1/2} \tilde{C}$  given above that the Gaussian estimates of the “common trends” components  $Z_{1t} = Q'_1 Y_t$  can be obtained with  $\hat{Q}_1 = S_{\tilde{w}\tilde{w}}^{-1/2} [\hat{V}_m^*, \dots, \hat{V}_{r+1}^*]$ , so that the rows of  $\hat{Q}'_1$  are the vectors  $\hat{V}_i^* S_{\tilde{w}\tilde{w}}^{-1/2}$ ,  $i = r+1, \dots, m$ , corresponding to the  $d$  smallest partial canonical correlations, with the property that  $\hat{Q}'_1 \hat{C} = 0$  since  $\hat{Q}'_1 \hat{A} = [\hat{V}_m^*, \dots, \hat{V}_{r+1}^*]' \hat{V}_* = 0$ . This method of estimation of the common “trend” or “long-memory” components was explored by Gonzalo and Granger (1995) [137].

Within the context of model (3.15), it is necessary to specify or determine the rank  $r$  of cointegration or the number  $d$  of unit roots in the model. Thus, it is also of interest to test the hypothesis  $H_0 : \text{rank}(C) \leq r$ , which is equivalent to the hypothesis that the number of unit roots in the AR model is greater than or equal to  $d$  ( $d = m-r$ ), against the general alternative that  $\text{rank}(C) = m$ . The *likelihood ratio test* for this hypothesis is considered by Johansen (1988) [173] and the test statistic is given as  $-T \log(U) = -T \log(|S| / |S_0|)$ , where  $S = \sum_{t=1}^T \hat{\epsilon}_t \hat{\epsilon}'_t$  denotes the residual sum of squares matrix in the full or unconstrained model, while  $S_0$  is the residual sum of

squares matrix obtained under the reduced-rank restriction on  $C$  that  $\text{rank}(C) = r$ . It follows from the work of Anderson (1951) [11] in the multivariate linear model with reduced-rank structure that the LR statistic can be expressed equivalently as

$$-T \log(U) = -T \sum_{i=r+1}^m \log(1 - \hat{\rho}_i^2) \quad (3.18)$$

where the  $\hat{\rho}_i$  are the  $d = m-r$  smallest sample partial canonical correlations between  $W_t = Y_t - Y_{t-1}$  and  $Y_{t-1}$ , given  $W_{t-1}, \dots, W_{t-p+1}$ . The limiting distribution for the likelihood ratio statistic does not take standard form and has been derived by Johansen (1988) [173], and by Reinsel and Ahn (1992) [228]. The asymptotic distribution of the likelihood ratio test statistic under  $H_0$  depends only on the number of unit roots,  $d$ , and not on any other parameters or on the order  $p$  of the VAR model.

Critical values of the asymptotic distribution of (3.18) have been obtained by simulation by Johansen (1988) [173] and Reinsel and Ahn (1992) [228] and can be used in the test of  $H_0$ . It is known that inclusion of a constant term in the estimation of the nonstationary AR model affects the limiting distribution of the estimators and test statistics. Testing for this and other model variations can be found elsewhere in the literature (see Johansen (1991) [174], Engle and Granger (1987) [101], etc.).

### 3.1.3 Multivariate Regression

We want to briefly discuss the general multivariate regression model, its estimation and testing as these tools come handy to test the CAPM and APT models. Consider the general multivariate linear model,

$$Y_t = CX_t + \epsilon_t \quad (3.19)$$

where  $Y_t = (Y_{1t}, \dots, Y_{mt})'$  is an  $m \times 1$  vector of responsive variables and  $X_k = (X_{1t}, \dots, X_{nt})'$  is an  $n \times 1$  vector of predictors,  $C$  is an  $m \times n$  coefficient matrix and  $\epsilon_t = (\epsilon_{1t}, \dots, \epsilon_{mt})'$  is the  $m \times 1$  vector of random errors. We assume  $E(\epsilon_t) = 0$  and  $\text{Cov}(\epsilon_t) = \Sigma_{\epsilon\epsilon}$  is a  $m \times m$  positive definite covariance matrix. The errors are independent over 't', the time index. Thus if we arrange the data and error as matrices,

$$Y = [Y_1, \dots, Y_T], X = [X_1, \dots, X_T] \text{ and } \epsilon = [\epsilon_1, \dots, \epsilon_T] \quad (3.20)$$

and if  $e = \text{vec}(\epsilon')$ ,

$$E(e) = 0, \text{Cov}(e) = \Sigma_{\epsilon\epsilon} \otimes I_T \quad (3.21)$$

where the symbol  $\otimes$  signifies the Kronecker's product.

The unknown parameters in model (3.19) are the elements of the regression coefficient matrix  $C$  and the error covariance matrix  $\Sigma_{\epsilon\epsilon}$ . These can be estimated by the method of least squares, or equivalently, the method of maximum likelihood under the assumption of normality of the  $\epsilon_k$ . Note we can rewrite the model (3.19) in terms of the complete data matrices  $\mathbf{Y}$  and  $\mathbf{X}$  as

$$\mathbf{Y} = C\mathbf{X} + \epsilon \quad (3.22)$$

The least squares estimates are obtained by minimizing

$$\mathbf{e}'\mathbf{e} = \text{tr}(\mathbf{e}\mathbf{e}') = \text{tr}[(\mathbf{Y} - \mathbf{C}\mathbf{X})(\mathbf{Y} - \mathbf{C}\mathbf{X})'] = \text{tr} \left[ \sum_{t=1}^T \epsilon_t \epsilon_t' \right] \quad (3.23)$$

where  $\text{tr}(A)$  denotes the trace of a square matrix  $A$ , that is, the sum of the diagonal elements of  $A$ . This yields a unique solution for  $\mathbf{C}$  as

$$\tilde{\mathbf{C}} = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1} = \left( \frac{1}{T} \mathbf{Y}\mathbf{X}' \right) \left( \frac{1}{T} \mathbf{X}\mathbf{X}' \right)^{-1} \quad (3.24)$$

Hence, it can be seen from the form of (3.24) the rows of the matrix  $\mathbf{C}$  are estimated by the least squares regression of each response variable on the predictor variables and therefore the covariances among the response variables do not enter into the estimation. The multivariate regression is essentially a series of multiple regressions.

For maximum likelihood (ML) estimation, the criterion that must be minimized to estimate  $\mathbf{C}$  is

$$\text{tr}(\Sigma_{\epsilon\epsilon}^{-1}) = \text{tr} \left[ \Sigma_{\epsilon\epsilon}^{-1/2} (\mathbf{Y} - \mathbf{C}\mathbf{X})(\mathbf{Y} - \mathbf{C}\mathbf{X})' \Sigma_{\epsilon\epsilon}^{-1/2} \right] \quad (3.25)$$

which yields the same solution for  $\mathbf{C}$  as given in (3.24). The ML estimator of the error covariance matrix  $\Sigma_{\epsilon\epsilon}$  is obtained from the least squares residuals  $\hat{\mathbf{e}} = \mathbf{Y} - \tilde{\mathbf{C}}\mathbf{X}$  as  $\tilde{\Sigma}_{\epsilon\epsilon} = \frac{1}{T}(\mathbf{Y} - \tilde{\mathbf{C}}\mathbf{X})(\mathbf{Y} - \tilde{\mathbf{C}}\mathbf{X})' \equiv \frac{1}{T}S$ , where  $S = \sum_{t=1}^T \hat{\epsilon}_t \hat{\epsilon}_t'$  is the error sum of squares matrix, and  $\hat{\epsilon}_t = Y_t - \tilde{\mathbf{C}}\mathbf{X}_t$ ,  $t = 1, \dots, T$ , are the least squares residual vectors. An unbiased estimator of  $\Sigma_{\epsilon\epsilon}$  is obtained as

$$\bar{\Sigma}_{\epsilon\epsilon} = \frac{1}{T-n} \hat{\epsilon} \hat{\epsilon}' = \frac{1}{T-n} \sum_{t=1}^T \hat{\epsilon}_t \hat{\epsilon}_t'. \quad (3.26)$$

### Inference Properties:

Under the assumption that the matrix of predictor variables  $\mathbf{X}$  is fixed and non-stochastic, properties of the least squares estimator  $\tilde{\mathbf{C}}$  in (3.24) can readily be derived. If  $X_t$  is a stochastic vector, the LS estimator  $\tilde{\mathbf{C}}$  can be viewed as conditional on fixed observed values of the  $X_t$ . First,  $\tilde{\mathbf{C}}$  is an unbiased estimator of  $\mathbf{C}$ ,  $E(\tilde{\mathbf{C}}) = \mathbf{C}$  and  $\text{Cov}(\tilde{\mathbf{C}}_{(i)}, \tilde{\mathbf{C}}_{(j)}) = \sigma_{ij}(\mathbf{X}\mathbf{X}')^{-1}$  for  $i, j = 1, \dots, m$ , where  $\sigma_{ij}$  denotes the  $(i, j)$ th element of the covariance matrix  $\Sigma_{\epsilon\epsilon}$ , since  $\text{Cov}(\mathbf{Y}_{(i)}, \mathbf{Y}_{(j)}) = \sigma_{ij}I_T$ . The distributional properties of  $\tilde{\mathbf{C}}$  follow easily from multivariate normality of the error terms  $\epsilon_t$ . Specifically, we consider the distribution of  $\text{vec}(\tilde{\mathbf{C}}')$ , where the “vec” operation transforms an  $m \times n$  matrix into an  $mn$ -dimensional column vector by stacking the columns of the matrix below each other.

**Result 3.1.** For the model (3.19) under the normality assumption on the  $\epsilon_k$ , the least squares estimator  $\tilde{\mathbf{C}} = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}$  is the same as the maximum likelihood estimator of  $\mathbf{C}$ , and the distribution of the least squares estimator  $\tilde{\mathbf{C}}$  is such that

$$\text{vec}(\tilde{\mathbf{C}}') \sim N(\text{vec}(\mathbf{C}'), \Sigma_{\epsilon\epsilon} \otimes (\mathbf{X}\mathbf{X}')^{-1}). \quad (3.27)$$

Note, in particular, that this result implies that the  $j$ th row of  $\tilde{C}$ ,  $\tilde{C}_{(j)}'$ , which is the vector of least squares estimates of regression coefficients for the  $j$ th response variable, has the distribution  $\tilde{C}_{(j)} \sim N(C_{(j)}, \sigma_{jj}(\mathbf{X}\mathbf{X}')^{-1})$ . The inference on the elements of the matrix  $C$  can be made using the result in (3.27). In practice, because  $\Sigma_{ee}$  is unknown, a reasonable estimator such as the unbiased estimator in (3.26) is substituted for the covariance matrix in (3.27).

We shall also indicate the likelihood ratio procedure for testing simple linear hypotheses regarding the regression coefficient matrix. Consider  $\mathbf{X}$  partitioned as  $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2']$  and corresponding  $C = [C_1, C_2]$ , so that the model (3.22) is written as  $\mathbf{Y} = \tilde{C}_1\mathbf{X}_1 + C_2\mathbf{X}_2 + \epsilon$ , where  $C_1$  is  $m \times n_1$  and  $C_2$  is  $m \times n_2$  with  $n_1 + n_2 = n$ . It may be of interest to test the null hypothesis  $H_0 : C_2 = 0$  against the alternative  $C_2 \neq 0$ . The null hypothesis implies that the predictor variables  $\mathbf{X}_2$  do not have any (additional) influence on the response variables  $\mathbf{Y}$ , given the impact of the variables  $\mathbf{X}_1$ . Using the likelihood ratio (LR) testing approach, it is easy to see that the LR test statistic is  $\lambda = U^{T/2}$ , where  $U = |S|/|S_1|$ ,  $S = (\mathbf{Y} - \tilde{C}\mathbf{X})(\mathbf{Y} - \tilde{C}\mathbf{X})'$  and  $S_1 = (\mathbf{Y} - \tilde{C}_1\mathbf{X}_1)(\mathbf{Y} - \tilde{C}_1\mathbf{X}_1)'$ . The matrix  $S$  is the residual sum of squares matrix from maximum likelihood (i.e., least squares) fitting of the full model, while  $S_1$  is the residual sum of squares matrix obtained from fitting the reduced model with  $C_2 = 0$  and  $\tilde{C}_1 = \mathbf{Y}\mathbf{X}_1'(\mathbf{X}_1\mathbf{X}_1')^{-1}$ . This LR statistic result follows directly by noting that the value of the multivariate normal likelihood function evaluated at the maximum likelihood estimates  $\tilde{C}$  and  $\tilde{\Sigma}_{ee} = (1/T)S$  is equal to a constant times  $|S|^{-T/2}$ , because  $\text{tr}(\tilde{\Sigma}_{ee}^{-1}\hat{\epsilon}\hat{\epsilon}') = \text{tr}(T\hat{\Sigma}_{ee}^{-1}\hat{\Sigma}_{ee}) = Tm$  is a constant when evaluated at the ML estimate. Thus, the likelihood ratio for testing  $H_0 : C_2 = 0$  is  $\lambda = |S_1|^{-T/2}/|S|^{-T/2} = U^{T/2}$ . It has been shown [see Anderson (1984, Chap. 8)] [12] that for moderate and large sample size  $T$ , the test statistic

$$\mathcal{M} = -[T - n + (n_2 - m - 1)/2]\log(U) \quad (3.28)$$

is approximately distributed as Chi-squared with  $n_2m$  degrees of freedom ( $\chi^2_{n_2m}$ ); the hypothesis is rejected when  $\mathcal{M}$  is greater than a constant determined by the  $\chi^2_{n_2m}$  distribution.

Finally, we briefly consider testing the more general null hypotheses of the form  $H_0 : F_1CG_2 = 0$  where  $F_1$  and  $G_2$  are known full-rank matrices of appropriate dimensions  $m_1 \times m$  and  $n \times n_2$ , respectively. In the hypothesis  $F_1CG_2 = 0$ , the matrix  $G_2$  allows for restrictions among the coefficients in  $C$  across (subsets of) the predictor variables, whereas  $F_1$  provides for restrictions in the coefficients of  $C$  across the different response variables. For instance, the hypothesis  $C_2 = 0$  considered earlier is represented in this form with  $G_2 = [0', I_{n_2}']'$  and  $F_1 = I_m$ , and the hypothesis of the form  $F_1CG_2 = 0$  with the same matrix  $G_2$  and  $F_1 = [I_{m_1}, 0]$  would postulate that the predictor variables  $\mathbf{X}_2$  do not enter into the linear regression relationship *only* for the first  $m_1 < m$  components of the response variables  $\mathbf{Y}$ . Using the LR approach, the likelihood ratio is given by  $\lambda = U^{T/2}$  where  $U = |S|/|S_1|$  and  $S_1 = (\mathbf{Y} - \hat{C}\mathbf{X})(\mathbf{Y} - \hat{C}\mathbf{X})'$  with  $\hat{C}$  denoting the ML estimator  $C$  obtained subject to the constraint that  $F_1CG_2 = 0$ . In fact, using Lagrange multiplier methods, it can be

shown that the constrained ML estimator under  $F_1 C G_2 = 0$  can be expressed as

$$\hat{C} = \hat{C} - S F'_1 (F_1 S F'_1)^{-1} F_1 \hat{C} G_2 [G'_2 (\mathbf{X} \mathbf{X}')^{-1} G_2]^{-1} G'_2 (\mathbf{X} \mathbf{X}')^{-1} \quad (3.29)$$

From (3.29) we therefore have

$$\begin{aligned} \mathbf{Y} - \tilde{C} \mathbf{X} &= (\mathbf{Y} - \tilde{C} \mathbf{X}) \\ &\quad + S F'_1 (F_1 S F'_1)^{-1} F_1 \tilde{C} G_2 [G'_2 (\mathbf{X} \mathbf{X}')^{-1} G_2]^{-1} (\mathbf{X} \mathbf{X}')^{-1} \mathbf{X} \end{aligned} \quad (3.30)$$

where the two terms on the right are orthogonal. We thus readily find that

$$\begin{aligned} S_1 &= S + S F'_1 (F_1 S F'_1)^{-1} (F_1 \tilde{C} G_2) [G'_2 (\mathbf{X} \mathbf{X}')^{-1} G_2]^{-1} \\ &\quad \times (F_1 \tilde{C} G_2)' (F_1 S F'_1)^{-1} F_1 S \equiv S + H, \end{aligned} \quad (3.31)$$

and  $U = 1/|I + S^{-1} H|$  so that the LR test can be based on the eigenvalues of  $S^{-1} H$ .

The nonzero eigenvalues of the  $m \times m$  matrix  $S^{-1} H$  are the same as those of the  $m_1 \times m_1$  matrix  $S_*^{-1} H_*$ , where  $S_* = F_1 S F'_1$  and  $H_* = F_1 H F'_1 = F_1 \tilde{C} G_2 [G'_2 (\mathbf{X} \mathbf{X}')^{-1} G_2]^{-1} (F_1 \tilde{C} G_2)'$ , so that  $U^{-1} = |I + S_*^{-1} H_*| = \prod_{i=1}^{m_1} (1 + \hat{\lambda}_i^2)$  where the  $\hat{\lambda}_i^2$  are the eigenvalues of  $S_*^{-1} H_*$ . Analogous to (3.28), the test statistic used is  $\mathcal{M} = -[T - n + \frac{1}{2}(n_2 - m_1 - 1)] \log(U)$  which has an approximate  $\chi^2_{n_2 m_1}$  distribution under  $H_0$ . Somewhat more accurate approximation of the null distribution of  $\log(U)$  is available based on the  $F$ -distribution, and in some cases (notably when  $n_2 = 1$  or 2, or  $m_1 = 1$  or 2) exact  $F$ -distribution results for  $U$  are available [see Anderson 1984, Chap. 8] [12]. Values relating to the (exact) distribution of the statistic  $U$  for various values of  $m_1$ ,  $n_2$ , and  $T - n$  have been tabulated by Schatzoff (1966) [236], Pillai and Gupta (1969) [225], and Lee (1972) [190]. However, for most of the applications the large sample  $\chi^2_{n_2 m_1}$  approximation for the distribution of the LR statistic  $\mathcal{M}$  is normally used.

Concerning the exact distribution theory associated with the test procedure, under normality of the  $\epsilon_k$ , it is established (e.g., Srivastava and Khatri, 1979, Chap. 6) [243] that  $S_*$  has the  $W(F_1 \Sigma_{\epsilon\epsilon} F'_1, T - n)$  distribution, the Wishart distribution with matrix  $F_1 \Sigma_{\epsilon\epsilon} F'_1$  and  $T - n$  degrees of freedom, and  $H_*$  is distributed as Wishart under  $H_0$  with matrix  $F_1 \Sigma_{\epsilon\epsilon} F'_1$  and  $n_2$  degrees of freedom. The sum of squares matrices  $S_*$  and  $H_*$  are also independently distributed, which follows from the well known fact that the residual sum of squares matrix  $S$  and the least squares estimator  $\tilde{C}$  are independent, and  $H_*$  is a function of  $\tilde{C}$  alone. Thus, in particular, in the special case  $m_1 = 1$  we see that  $S_*^{-1} H_* = H_*/S_*$  is distributed as  $n_2/(T - n)$  times the  $F$ -distribution with  $n_2$  and  $T - n$  degrees of freedom.

### Prediction:

Another problem of interest in regard to the multivariate linear regression model (3.19) is confidence regions for the mean responses corresponding to a fixed value  $X_0$  of the predictor variables. The mean vector of responses at  $X_0$  is  $C X_0$  and its estimate is  $\tilde{C} X_0$ , where  $\tilde{C} = \mathbf{Y} \mathbf{X}' (\mathbf{X} \mathbf{X}')^{-1}$  is the LS estimator of  $C$ . Under the assumption of normality of the  $\epsilon_k$ , noting that  $\tilde{C} X_0 = \text{vec}(X'_0 \tilde{C}') = (I_m \otimes X'_0) \text{vec}(\tilde{C}')$ ,

we see that  $\tilde{C}X_0$  is distributed as multivariate normal  $N(CX_0, X'_0(\mathbf{XX}')^{-1}X_0\Sigma_{\epsilon\epsilon})$  from Result 3.1. Since  $\tilde{C}$  is distributed independently of  $\bar{\Sigma}_{\epsilon\epsilon} = \frac{1}{T-n}S$  with  $(T-n)\bar{\Sigma}_{\epsilon\epsilon}$  distributed as Wishart with matrix  $\Sigma_{\epsilon\epsilon}$  and  $T-n$  degrees of freedom, it follows that

$$\mathcal{T}^2 = (\tilde{C}X_0 - CX_0)' \bar{\Sigma}_{\epsilon\epsilon}^{-1} (\tilde{C}X_0 - CX_0) / \{X'_0(\mathbf{XX}')^{-1}X_0\} \quad (3.32)$$

has the Hotelling's  $T^2$ -distribution with  $T-n$  degrees of freedom (Anderson, 1984, Chap. 5) [12]. Thus,  $\{(T-n-m+1)/m(T-n)\}\mathcal{T}^2$  has the  $F$ -distribution with  $m$  and  $T-n-m+1$  degrees of freedom, so that a  $100(1-\alpha)\%$  confidence ellipsoid for the true mean response  $CX_0$  at  $X_0$  is given by

$$\begin{aligned} & (\tilde{C}X_0 - CX_0)' \bar{\Sigma}_{\epsilon\epsilon}^{-1} (\tilde{C}X_0 - CX_0) \\ & \leq \{X'_0(\mathbf{XX}')^{-1}X_0\} \left[ \frac{m(T-n)}{T-n-m+1} F_{m,T-n-m+1}(\alpha) \right] \end{aligned} \quad (3.33)$$

where  $F_{m,T-n-m+1}(\alpha)$  is the upper  $100\alpha$ th percentile of the  $F$ -distribution with  $m$  and  $T-n-m+1$  degrees of freedom.

A related problem of interest is the prediction of values of a new response vector  $Y_0 = CX_0 + \epsilon_0$  corresponding to the predictor variables  $X_0$ , where it is assumed that  $Y_0$  is independent of the values  $Y_1, \dots, Y_T$ . The predictor of  $Y_0$  is given by  $\hat{Y}_0 = \tilde{C}X_0$ , and it follows similarly to the above result (3.32) that a  $100(1-\alpha)\%$  prediction ellipsoid for  $Y_0$  is

$$\begin{aligned} & (Y_0 - \tilde{C}X_0)' \bar{\Sigma}_{\epsilon\epsilon}^{-1} (Y_0 - \tilde{C}X_0) \\ & \leq \{1 + X'_0(\mathbf{XX}')^{-1}X_0\} \left[ \frac{m(T-n)}{T-n-m+1} F_{m,T-n-m+1}(\alpha) \right] \end{aligned} \quad (3.34)$$



# 4

## *Advanced Topics*

### CONTENTS

4.0.1	State-Space Modeling .....	85
4.1	A Model for Volume-Volatility Relationship .....	88
4.2	Regime Switching and Change-Point Models .....	90
4.3	Models for Point Processes .....	93
4.3.1	Stylized Models for High Frequency Financial Data .....	95
4.3.2	Models for Multiple Assets: High-Frequency Context .....	100
4.4	Time Aggregation and Loss of Information .....	101
4.5	Realized Volatility and Econometric Models .....	101
4.6	Analytics from Machine Learning Literature .....	102

### 4.0.1 State-Space Modeling

There has been much recent interest in the representation of ARIMA models in the state-space form, for purpose of forecasting, as well as for model specification and maximum likelihood estimation of parameters. This approach helps to update the model coefficients as and when new data arrives.

The state-space models were initially developed by control systems engineers to measure a signal contaminated by noise. The signal at time “ $t$ ” is taken to be a linear combination of variables, called state variables that form the so called state vector at time,  $t$ . The key property of the state vector is that it contains the minimum set of information from past and present data but the future behavior of the system is independent of the past value and depends only on the present values. Thus the state vector evolves according to the Markov property. The state equation is stated as,

$$Z_t = \Phi_t Z_{t-1} + a_t \quad (4.1)$$

and observation equation

$$Y_t = H_t Z_t + N_t \quad (4.2)$$

where it is assumed that  $a_t$  and  $N_t$  are independent white-noise processes;  $a_t$  is a vector white noise with covariance matrix  $\Sigma_a$  and  $N_t$  has variance  $\sigma_N^2$ . The matrix  $\Phi_t$  in (4.1) is an  $r \times r$  transition matrix and  $H_t$  in (4.2) is a  $l \times r$  vector, which are allowed to vary in time.

We will consider the state-space form of an ARIMA( $p, d, q$ ) process  $\phi(B)Y_t = \theta(B)\varepsilon_t$ , define the forecasts  $\hat{Y}_t(j) = E(Y_{t+j} | Y_t, Y_{t-1}, \dots)$  for  $j = 0, 1, \dots, r$ , with  $r = \max(p+d, q+1)$ , and  $\hat{Y}_t(0) = Y_t$ .

From the updating equations  $\hat{Y}_{t+1}(l) = \hat{Y}_t(l+1) + \psi_l \varepsilon_{t+1} = \hat{Y}_t(l+1) + \psi_l [Y_{t+1} - \hat{Y}_t(1)]$  for the general ARIMA model, we have that  $\hat{Y}_t(j-1) = \hat{Y}_{t-1}(j) + \psi_{j-1} \varepsilon_t$ ,  $j = 1, 2, \dots, r-1$ . Also for  $j = r > q$ , observe that  $\hat{Y}_t(j-1) = \hat{Y}_{t-1}(j) + \psi_{j-1} \varepsilon_t = \sum_{i=1}^{p+d} \phi_i \hat{Y}_{t-1}(j-i) + \psi_{j-1} \varepsilon_t$ . So we define the “state” vector at time  $t$ ,  $Z_t$ , with  $r$  components as  $Z_t = (Y_t, \hat{Y}_t(1), \dots, \hat{Y}_t(r-1))'$ . Then from the above relations we find that  $Z_t$  satisfies the equations

$$Z_t = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & & \ddots & 1 \\ \phi_r & \phi_{r-1} & \cdots & \cdots & \phi_1 \end{bmatrix} Z_{t-1} + \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \psi_{r-1} \end{bmatrix} \varepsilon_t$$

where  $\phi_i = 0$  if  $i > p+d$ . So we have

$$Z_t = \Phi Z_{t-1} + \Psi \varepsilon_t \quad (4.3)$$

together with the observation equation

$$Z_t = Y_t + N_t = [1, 0, \dots, 0] Z_t + N_t = H Z_t + N_t \quad (4.4)$$

where the additional noise  $N_t$  would be present only if the ARIMA process  $\{Y_t\}$  is observed subject to additional noise; otherwise we simply have  $Z_t \equiv Y_t = H Z_t$ . The above two equations constitute what is known as a state-space representation of the ARIMA model. We note that there are many other constructions of the state vector  $Z_t$  that will give rise to state-space equations of the general form of (4.1)–(4.2); that is, the state-space form of an ARIMA model is not unique.

Observe, in applications the matrices  $\Phi_t$  and  $H_t$  are constant matrices,  $\Phi \equiv \Phi_t$  and  $H \equiv H_t$  for all  $t$ , that do not depend on  $t$ , as in the state-space form (4.3)–(4.4) of the ARIMA model. In this case the system or model is said to be time-invariant. A nice feature of expressing a model in state-space form is that an updating procedure can be readily used when a new observation becomes available to revise the current state vector and to produce forecasts. These forecasts in general tend to fare better.

For the state-space model, define the finite sample estimates of the state vector  $Z_{t+1}$  based on observations  $Y_t, \dots, Y_l$  over the finite past time period, as

$$\hat{Z}_{t+l|t} = E[Z_{t+l}|Y_t, \dots, Y_l], \text{ with } V_{t+l|t} = E[(Z_{t+l} - \hat{Z}_{t+l|t})(Z_{t+l} - \hat{Z}_{t+l|t})'].$$

A convenient computational procedure, known as the *Kalman filter equations*, is then available to obtain the current estimate  $\hat{Z}_{t|t}$ , in particular. It is known that, starting from some appropriate initial values  $Z_0 \equiv \hat{Z}_{0|0}$  and  $V_0 \equiv V_{0|0}$ , the optimal filtered estimate,  $\hat{Z}_{t|t}$ , is given through the following recursive relations:

$$\hat{Z}_{t|t} = \hat{Z}_{t|t-1} + K_t(Y_t - H_t \hat{Z}_{t|t-1}) \quad (4.5)$$

where  $K_t = V_{t|t-1} H'_t [H_t V_{t|t-1} H'_t + \sigma_N^2]^{-1}$  with  $\hat{Z}_{t|t-1} = \Phi_t \hat{Z}_{t-1|t-1}$ ,  $V_{t|t-1} =$

$\Phi_t V_{t|t-1} \Phi'_t + \Sigma_a$  and  $V_{t|t} = [I - K_t H_t] V_{t|t-1} = V_{t|t-1} - V_{t|t-1} H'_t [H_t V_{t|t-1} H'_t + \sigma_N^2]^{-1} H_t V_{t|t-1}$  for  $t = 1, 2, \dots$ .

In (4.2), the quantity  $\varepsilon_{t|t-1} = Y_t - H_t \hat{Z}_{t|t-1} \equiv Y_t - \hat{Y}_{t|t-1}$  is called the (finite sample) innovation at time  $t$ , because it is the new information provided by the measurement  $Y_t$  which was not available from the previous observed (finite) history of the system. The factor  $K_t$  is called the “Kalman gain” matrix. The filtering procedure in (4.5) has the recursive “prediction-correction” or “updating” form, and the validity of these equations as representing the minimum mean square error predictor can readily be verified through the principles of “updating”. For example, verification of (4.5) follows from the principle, for linear prediction, that

$$\begin{aligned} E[Z_t | Y_t, \dots, Y_1] &= E[Z_t | Y_t - \hat{Y}_{t|t-1}, Y_{t-1}, \dots, Y_1] \\ &= E[Z_t | Y_{t-1}, \dots, Y_1] + E[Z_t | Y_t - \hat{Y}_{t|t-1}], \end{aligned}$$

since  $\varepsilon_{t|t-1} = Y_t - \hat{Y}_{t|t-1}$  is independent of  $Y_{t-1}, \dots, Y_1$ . From (4.5) it is seen that the estimate of  $Z_t$  based on observations through time  $t$  equals the prediction of  $Z_t$  from observations through time  $t-1$  updated by the factor  $K_t$  times the innovation  $\varepsilon_{t|t-1}$ . The quantity  $K_t$  can be interpreted as the regression coefficients of  $Z_t$  on the innovation  $\varepsilon_{t|t-1}$ , with  $\text{Var}(\varepsilon_{t|t-1}) = H_t V_{t|t-1} H'_t + \sigma_N^2$  and  $\text{Cov}(Z_t, \varepsilon_{t|t-1}) = V_{t|t-1} H'_t$  following directly from (4.4) since  $\varepsilon_{t|t-1} = H_t(Z_t - \hat{Z}_{t|t-1}) + N_t$ . Thus, the general *updating relation* is  $\hat{Z}_{t|t} = \hat{Z}_{t|t-1} + \text{Cov}(Z_t, \varepsilon_{t|t-1})\{\text{Var}(\varepsilon_{t|t-1})\}^{-1}\varepsilon_{t|t-1}$ , where  $\varepsilon_{t|t-1} = Y_t - \hat{Y}_{t|t-1}$ , and the relation in  $V_t$  is the usual updating of the error covariance matrix to account for the new information available from the innovation  $\varepsilon_{t|t-1}$ , while the *prediction relations* (4.5) follow directly from (4.3). In general, forecasts of future state values are available directly as  $\hat{Z}_{t+l|t} = \Phi_{t+l} \hat{Z}_{t+l-1|t}$  for  $l = 1, 2, \dots$  with the covariance matrix of the forecast errors generated recursively essentially through (2.43) as  $V_{t+l|t} = \Phi_{t+l} V_{t+l-1|t} \Phi'_{t+l} + \Sigma_a$ . Finally, forecasts of future observations  $Y_{t+l}$  are then available as  $\hat{Y}_{t+l|t} = H_{t+l} \hat{Z}_{t+l|t}$ , since  $Y_{t+l|t} = H_{t+l} Z_{t+l} + N_{t+l}$ , with forecast error variance  $v_{t+l|t} = E[(Y_{t+l} - \hat{Y}_{t+l|t})^2] = H_{t+l} V_{t+l|t} H'_{t+l} + \sigma_N^2$ .

For ARIMA models, with state-space representation (4.3)–(4.4) and  $Y_t = H Z_t$  with  $H = [1, 0, \dots, 0]$ , this Kalman filtering procedure constitutes an alternate method to obtain exact finite sample forecasts, based on data  $Y_t, Y_{t-1}, \dots, Y_1$ , for future values in the ARIMA process, subject to specification of appropriate initial conditions to use in the terms in (4.5). For stationary zero-mean processes  $Y_t$ , the appropriate initial values are  $\hat{Z}_{0|0} = 0$ , a vector of zeros, and  $V_{0|0} = \text{Cov}(Z_0) \equiv V_*$ , the covariance matrix of  $Z_0$  which can easily be determined under stationarity through the definition of  $Z_t$ . Specifically, since the state vector  $Z_t$  then follows the stationary vector AR(1) model  $Z_t = \Phi Z_{t-1} + \psi \varepsilon_t$ , its covariance matrix  $V_* = \text{Cov}(Z_t)$  satisfies  $V_* = \Phi V_* \Phi' + \sigma_\varepsilon^2 \Psi \Psi'$ , which can be readily solved for  $V_*$ . For nonstationary ARIMA processes, additional assumptions need to be specified. The “steady-state” values of the Kalman filtering procedure  $l$ -step ahead forecasts  $\hat{Y}_{t+l|t}$  and their forecast error variances  $v_{t+l|t}$ , which are rapidly approached as  $t$  increases, will be identical to the expressions given previously,  $\hat{Y}_t(l)$  and  $\sigma^2(l) = \sigma^2(1 + \sum_{i=1}^{l-1} \psi_i^2)$ .

In particular, for the ARIMA process in state-space form, we can obtain the exact

(finite sample) one-step ahead forecasts  $\hat{Y}_{t-1}(1) = E[Y_t | Y_{t-1}, \dots, Y_1] = H\hat{Z}_{t|t-1}$ , and their error variances  $v_{t|t-1} = HV_{t|t-1}H'$ , conveniently through the Kalman filtering equation (4.5). This can be particularly useful for evaluation of the Gaussian likelihood function, based on  $T$  observations  $Y_1, \dots, Y_T$  from the ARIMA process, applied to the problem of maximum likelihood estimation of model parameters.

This state-space approach to exact likelihood function determination has also been shown to be quite useful in dealing with estimation problems for ARMA models when some values  $Y_t$  of the series are not observed, that is, there are missing values among the sequence of  $Y_t$  values.

An alternative set-up, alternative to (4.4) and (4.5) is to assume that the deviations in the observation and transition equations are related. With the observation equation stated in (4.2), the transition equation is modified as

$$Z_t = \Phi Z_{t-1} + \alpha N_t \quad (4.6)$$

This model studied in Ord, Koehler, and Snyder (1997) [218] with a single source of error ( $N_t$ ) is shown to be closely related to various exponential smoothing procedures. There are other formulations of state-space models, such as structural models by Harvey (1989) [146] and dynamic linear models of West and Harrison (1997) [259] that are found to be useful for model building.

## 4.1 A Model for Volume-Volatility Relationship

While the price information via returns and the volatility can be modeled through methods described earlier, it has become important to incorporate other relevant trading data to augment the trading strategies. To that effect, we want to seek some guidance from economic theory how that information is linked to price movements.

In the market microstructure theory, it is assumed that price movements occur primarily due to the arrival of new information and this information is incorporated efficiently into market price. Other variables such as trading volume, bid-ask spread and market liquidity are observed to be related to the volatility of the returns. Early empirical studies have documented a positive relationship between daily trading volume and volatility (See Clark (1973) [63]). Epps and Epps (1976) [105] and Tauchen and Pitts (1983) [249] assume that the volume and price are driven by the underlying ‘latent’ information ( $I$ ) and provide a theoretical framework to incorporate this information. Assuming that there are fixed number of traders who trade in a day and the number of daily equilibria,  $I$  is random because the number of new pieces of information is random, the return and the volume are written as a bivariate normal mixture model with the same mixing variable  $I$ . Conditional on  $I$ :

$$\begin{aligned} r_t &= \sigma_1 \sqrt{I_t} Z_{1t} \\ V_t &= \mu_V I_t + \sigma_2 \sqrt{I_t} Z_{2t} \end{aligned} \quad (4.7)$$

where  $Z_{1t}$  and  $Z_{2t}$  are standard normal variables and  $Z_{1t}$ ,  $Z_{2t}$ , and  $I_t$  are mutually independent. Thus the volatility-volume relationship,

$$\begin{aligned}\text{Cov}(r_t^2, V_t) &= E(r_t^2 V_t) - E(r_t^2) E(V_t) \\ &= \sigma_1^2 \mu_V \text{Var}(I_t) > 0\end{aligned}\tag{4.8}$$

is positive due to the variance in  $I_t$ . If there is no new information or if there is no variation in the mixing variable,  $\text{Var}(I_t) = 0$  and thus the relationship vanishes. The theory of arrival rates suggest a Poisson distribution and based on empirical evidence, the lognormal distribution is taken to be a candidate for distribution of mixing variable,  $I_t$ . The parameters of the model in (4.7) are estimated through maximum likelihood. Gallant, Rossi and Tauchen (1992) [126] using a semi-parametric estimate of the joint density of  $r_t$  and  $V_t$  conduct a comprehensive study of NYSE data from 1925 to 1987. The following summarizes their findings:

#### **Stylized Facts 5 (Volume–Volatility Relationship):**

- Positive correlation between conditional volatility and volume.
- Large price movements are followed by high volume.
- Conditioning on lagged volume substantially attenuates the leverage (which is an asymmetry in the conditional variance of current price change against past price change) effect.
- After conditioning on lagged value, there is a positive risk-return relation.

Andersen (1996) [9] modifies the model (4.7) by integrating the microstructure setting of Glosten and Milgrom (1985) [136] with the stochastic volatility, that is built on weaker conditions on the information arrival process. While the first equation in (4.7) remains the same, the volume  $V_t$  has informed and noise components,  $V_t = IV_t + NV_t$ . Noise trading component,  $NV_t$  is taken as a time homogeneous Poisson process,  $P_0(m_0)$ . Therefore, the systematic variation in trading is mainly due to informed volume. Then  $IV_t|I_t \sim \text{Poisson}(I_t \mu)$  and thus

$$V_t|I_t \sim \text{Poisson}(m_0 + I_t m_1)\tag{4.9}$$

It is easy to see  $\text{Cov}(r_t^2, V_t) = \sigma^2 m_1 \text{Var}(I_t) > 0$  under this setting as well. These models clearly indicate that the intraday return volatility and volume processes jointly contain predictable elements.

There are other studies that focus on the trading volume and returns relationship and we mention only a select few here. Campbell, Grossman, and Wang (1993) [50] demonstrate for individual large stocks and stock indices the first-order daily auto-correlation in the returns tend to decline with volume. The authors develop a theoretical model where the economy has two assets, risk-free asset and a stock and there are two types of investors, one with constant risk aversion and the other with risk aversion changing over time. Under this set-up, it is implied that a decline in a stock

price on a high-volume day is more likely than a decline on a low-volume day to be associated with an increase in the expected stock return.

Gervais, Kaniel, and Mingelgrin (2001) [131] investigate the role of trading activity in providing information about future prices. It is shown that periods of extremely high (low) volume tend to be followed by positive (negative) excess returns. The formation period for identifying extreme trading volume is a day or a week, but the effect lasts at least twenty days and holds across all stock sizes. To test if this information can be used profitably, the authors construct a trading strategy, by sending buy (sell) limit orders at the existing bid (ask) price, at the end of the formation period. If the orders are not taken, they are converted into market orders at the closing. The strategy is shown to result in profits, in particular with the small-medium firm stocks, after adjusting for the transaction costs. The model used for the empirical framework is the vector autoregressive model discussed later, with an addition of a set of exogenous control variables:

$$Y_t = \Phi_0 + \sum_{j=1}^p \Phi_j Y_{t-j} + \sum_{l=1}^L B_l X_{t-l} + \epsilon_t \quad (4.10)$$

The components of  $Y_t$  include stock or market related variables such as detrended log of stock turnover, the stock return and the value weighted market return. The control variables are market volatility based on the daily return standard deviation and the dispersion which is the cross-sectional standard deviation of the security returns. The impulse response functions are used to aggregate the overall relationship among the endogenous variables. Statman, Thorley, and Vorkink (2006) [245] show that the trading volume is dependent on past returns over many months and it is argued that this may be due to the overconfidence of the investors.

These and other studies generally confirm the information content of the volume and turnover of stocks traded; strategies to exploit these, especially in a high frequency context, will be examined in Chapter 3. We will discuss how intra-day flow of volume can be related to price changes.

In algorithmic trading a key ingredient of many strategies is forecast of intra-day volume. Typically a parent order is split into several child orders and the timing of the submission of child orders could depend on the volume forecast for an interval of time that is considered for trading. Brownless, Cipollini and Gallo (2011) [48] provide prediction model for intra-day volume. This will be described in detail in the next chapter.

---

## 4.2 Regime Switching and Change-Point Models

It has been long noted in finance that there are two kinds of recurrences in stock prices: underreaction and overreaction to a series of good or bad news. The securities that have a good record for an extended period tend to become overpriced and have

low returns subsequently. Barberis, Shleifer and Vishny (1998) [21] present a model where investor believes that the returns can arise from one of two regimes although returns follow random walk: mean-reverting or trending. The transition probabilities are taken to be fixed and the regime is more likely to stay the same rather than to switch. But the investor's beliefs are updated as and when the returns data are observed. It is presumed that in many instances, we may not discern these shifts directly whether or when they occur but instead can draw probabilistic inference based on the observed behavior posthumously. Hamilton (2016) [142] reviews models for regime-switching applied in several contexts in macroeconomics, building upon the elegant model introduced in Hamilton (1989) [140]. Here in this section, we will cover only studies that are relevant to finance applications, particularly relevant to trading that exploit anomalies.

Consider the model,

$$y_t = \mu_{s_t} + \epsilon_t, \quad s_t = 1, 2 \quad (4.11)$$

where  $y_t$  is the observed variable, such as asset return and  $s_t$  represents two distinct regimes and  $\epsilon_t$ 's are i.i.d.  $\sim N(0, \sigma^2)$ . Let  $\{F_{t-1}\}$  be the information set available as of ' $t - 1$ '. The transition between the regimes is taken to be Markovian,

$$\text{Prob}(s_t = j \mid s_{t-1} = i, F_{t-1}) = p_{ij}, \quad i, j = 1, 2 \quad (4.12)$$

thus leading to an AR(1) model for  $\mu_{s_t}$  as

$$\mu_{s_t} = \phi_0 + \phi_1 \mu_{s_{t-1}} + a_t \quad (4.13)$$

where  $a_t$  by definition can take four possible values depending upon  $s_t$  and  $s_{t-1}$ . Here  $\phi_0 = p_{21}\mu_1 + p_{12}\mu_2$  and  $\phi_1 = p_{11} - p_{21}$ . Note the model for  $y_t$  in (4.11) is the sum of an AR(1) process and white noise, then from Theorem 2.3,  $y_t \sim \text{ARMA}(1,1)$ ; but because of discrete nature of ' $a_t$ ', (4.11) is a non-linear process. Note that the unknown parameters are  $(\mu_1, \mu_2, \sigma, p_{11}, p_{22})' = \lambda$  and they can be estimated via maximum likelihood as follows: note  $y_t | (s_t = j, F_{t-1}) \sim N(\mu_j, \sigma^2)$ . The predictive density of  $y_t$  is given as,

$$f(y_t \mid F_{t-1}) = \sum_{i=1}^2 \text{Prob}(s_t = i \mid F_{t-1}) f(y_t \mid s_t = i, F_{t-1}) \quad (4.14)$$

and estimate of  $\lambda$  is obtained by maximizing the likelihood function  $L(\lambda) = \sum_{t=1}^T \log f(y_t \mid F_{t-1}, \lambda)$ . Two useful quantities are:

Predicted regime:

$$\text{Prob}(s_t = j \mid F_{t-1}) = p_{1j} \text{Prob}(s_{t-1} = 1 \mid F_{t-1}) + p_{2j} \text{Prob}(s_{t-1} = 2 \mid F_{t-1}) \quad (4.15)$$

Optimal forecast:

$$E(y_t \mid F_{t-1}) = \sum_{i=1}^2 (\phi_0 + \phi_1 \mu_i) \text{Prob}(s_{t-1} = i \mid F_{t-1}). \quad (4.16)$$

The forecast for ‘ $k$ ’ period ahead can be stated as:

$$E(y_{t+k} | F_{t-1}) = \mu + \phi_1^k \sum_{i=1}^2 (\mu_i - \mu) \text{Prob}(s_t = i | F_{t-1}) \quad (4.17)$$

where  $\mu = \phi_0 / (1 - \phi_1)$ .

The basic model has been extended to cover multiple regimes and to vector processes. Interested readers should refer to Hamilton (2016) [142] and references therein. Ang and Timmermann (2012) [14] discuss a model that accounts for changes not only in the mean but also in variances and autocorrelations:

$$y_t = \mu_{s_t} + \phi_{s_t} y_{t-1} + \epsilon_t \quad (4.18)$$

where  $\epsilon_t$ ’s are independent  $\sim N(0, \sigma_{s_t}^2)$ . Using the data on excess S&P500 returns, FX returns, etc., it is identified that there are volatility regimes, but no level shifts.

Hamilton (1990) [141] proposes using EM (Expectation-Maximization) algorithm (Dempster, Laird and Rubin (1977) [82]) to obtain the maximum likelihood estimate, instead of recursive filter approach (see Hamilton (1989) [140] to optimize the loglihood of (4.14), updated with  $y_t$ . Note by Bayes’ Theorem

$$\text{Prob}(s_t = j | F_t) = \frac{\text{Prob}(s_t = j | F_{t-1}) f(y_t | s_t = j, F_{t-1})}{f(y_t | F_{t-1})} \quad (4.19)$$

To get an estimate of  $p_{ij}$  in step ‘ $l + 1$ ’, use

$$\hat{p}_{ij}^{(l+1)} = \frac{\sum_{t=1}^{T-1} \text{Prob}(s_t = i, s_{t+1} = j | F_t, \hat{\lambda}^{(l)})}{\sum_{t=1}^{T-1} \text{Prob}(s_t = i | F_t, \hat{\lambda}^{(l)})} \quad (4.20)$$

and with these estimates obtain the ML estimates of the rest of the parameters ( $\mu_1, \mu_2$ , and  $\sigma$ ) by maximizing the likelihood function. More details can be found in Hamilton (1990) [141].

There is a vast literature on Bayesian approach to this problem, see for example, Chib (1998) [60] and subsequent publications that followed. An interesting application is given in Pastor and Stambaugh (2001) [222]. Lai and Xing (2013) [184] study a general model, somewhat similar to (4.18):

$$\text{ARX-GARCH: } y_t = \beta_t' x_t + v_t \sqrt{h_t} \epsilon_t \quad (4.21)$$

where  $h_t \sim \text{GARCH}(p, q)$  and  $\beta_t$  and  $v_t$  are piecewise linear. Using the weakly returns of S&P500 index, AR(1)-GARCH(1,1) model is attempted and it is identified that 1998–2003 had higher volatility. Figure 4.1 provides the plot of closing price, returns, volatility and volume data.

In most of the applications of the above regime-switching models, particularly in finance, the plots of relevant quantities clearly reveal the pattern and these models provide a way to confirm the level and size shifts. We want to conclude this section with some simple ideas borrowed from the area of statistical process control. Recall

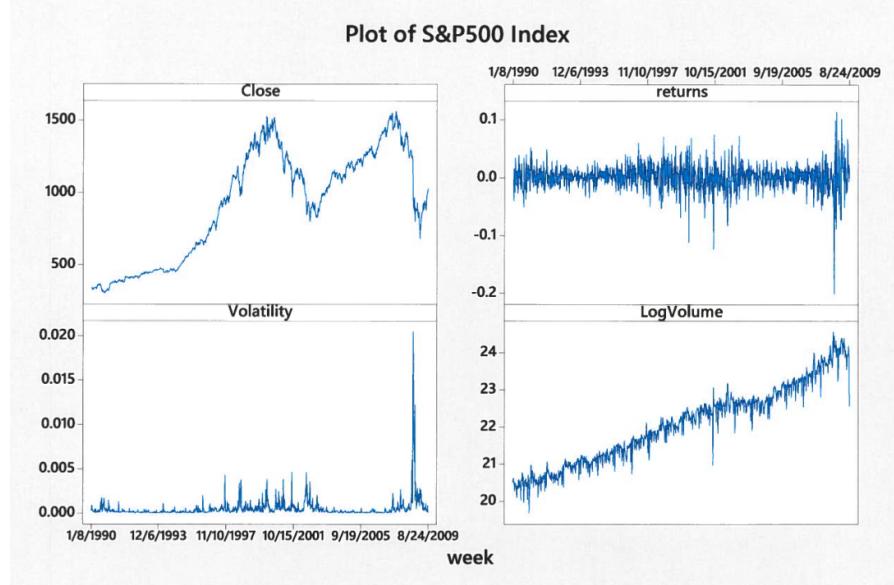


Figure 4.1: Plot of S&amp;P 500 Index.

that if the market is efficient, the return  $r_t = \ln P_t - \ln P_{t-1} \sim \text{IID } (\mu_t, \sigma^2)$ , but  $\mu_t = \delta$  when  $t \geq t_0$  and zero elsewhere. The cumulative sum chart (CUSUM) to detect the positive and negative mean shift are based on recursive calculations:

$$\begin{aligned} c_t^+ &= \max(0, c_{t-1}^+ + (r_t - k)) \\ c_t^- &= \min(0, c_{t-1}^- - (r_t - k)) \end{aligned} \quad (4.22)$$

where  $k = \delta/2$ . The chart signals as soon as  $c_t^+$  reaches ' $h$ '—a decision point set by the user. Jiang, Shu and Apley (2008) [171] modify (4.22) with exponentially weighted moving average (EWMA) type of estimates that is adaptive in nature. It can be stated as,

$$s_t^+ = \max \left( 0, s_{t-1}^+ + w(\hat{\delta}_t) \left( r_t - \frac{\hat{\delta}_t}{2} \right) \right) \quad (4.23)$$

where  $w(\hat{\delta}_t)$  is a weight function; a simple weight function  $w(\hat{\delta}_t) = \hat{\delta}_t$  is suggested. The  $\hat{\delta}_t$  is updated as

$$\hat{\delta}_t = (1 - \lambda) \hat{\delta}_{t-1} + \lambda r_t \quad (4.24)$$

The adaptive CUSUM chart is effective at detecting a range of mean shift sizes. Similar results for the variance need to be worked out.

In practical applications of these regime-shifting models, it is not easy to discern between a single outlier and persistent level shift. Valid measures that can distinguish these two situations on a timely basis could be very valuable to a trader.

### 4.3 Models for Point Processes

In many fields of study, observations occur in a continuum, space or time in the form of point events. The continuum can be multi-dimensional, but our focus is on the one-dimensional time scale with points distributed irregularly along the time scale. The main interest lies in estimating the mean rate of occurrence of events or more broadly on the patterns of occurrence. There are excellent monographs on this topic: Cox and Lewis (1966) [72], Daley and Vere-Jones (2003) [77], etc. But the interest in financial applications was revived by the seminal paper by Engle and Russell (1998) [103]. Financial market microstructure theories as discussed in Kyle (1985) [182], Admati and Pfleiderer (1988) [1] and Easley and O’ Hara (1992) [92] suggest that the frequency and the timing of transactions, that include posting, canceling and executing an order, carry information about the state of the market. The transactions generally tend to cluster during certain times of the day and the change in the mean rate of occurrences may suggest a new information flow about a stock.

To illustrate, recall Figure ??, on trading activities in Section 2.1. If we denote the observed intervals between successive activities as durations by  $d_1, d_2, \dots, d_r$ , the time of occurrences are obtained by forming the cumulative sums of the  $d$ ’s,  $t_1 = d_1, t_2 = t_1 + d_2, \dots, t_r = t_{r-1} + d_r$ . Cox and Lewis (1966) [72] suggest two ways to present this type of data graphically. One method is based on cumulative numbers of events that have occurred at or before ‘ $t$ ’ against ‘ $t$ ’. The slope of the line between any two points is the average number of events per unit for that period. One way to standardize the plot would be to approximate the graph by a line, ‘ $at$ ’ when ‘ $a$ ’ is the slope of the graph indicating the average rate of occurrence for the entire duration. The second method calls for dividing the time scale into equally spaced time intervals and count the number of events in each interval; this also can be modified by fixing a certain number of events and spacing on the time it takes for this number to occur. In the context of stock price data, this could mean simply recording not when the trade occurs but when the price changes. The advantage of the second plot is that the local fluctuations are readily observed and the advantage of the first plot is that it enables us to see systematic changes in the rate of occurrence.

The baseline distribution for the durations is exponential resulting from Poisson arrivals for a specified interval that assume a constant rate. It is suggested to plot  $\log(1 - \frac{i}{n_0+1})$  against  $d_{(i)}$  where  $d_{(i)}$  results from the ordered durations,  $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(n_0)}$  calculated between successive  $n_0 + 1$  occurrences. Departures from linearity would indicate that the exponential distribution may not hold. In addition it is also suggested to plot  $d_{i+1}$  against  $d_i$  to check on the dependance of durations. Cox and Lewis (1966, p14) [72] state that, “It is not clear how correlation in such data should be measured and tested; in particular it is not obvious that the ordinary correlation coefficient will be the most suitable measure with such data”. Engle and Russell (1998) [103] show how the dependence in the duration data can be explicitly modeled in the context of high frequency transaction data. We illustrate this with the trading data for Microsoft (MSFT) for the month of January in the year 2013. This graph (Fig-

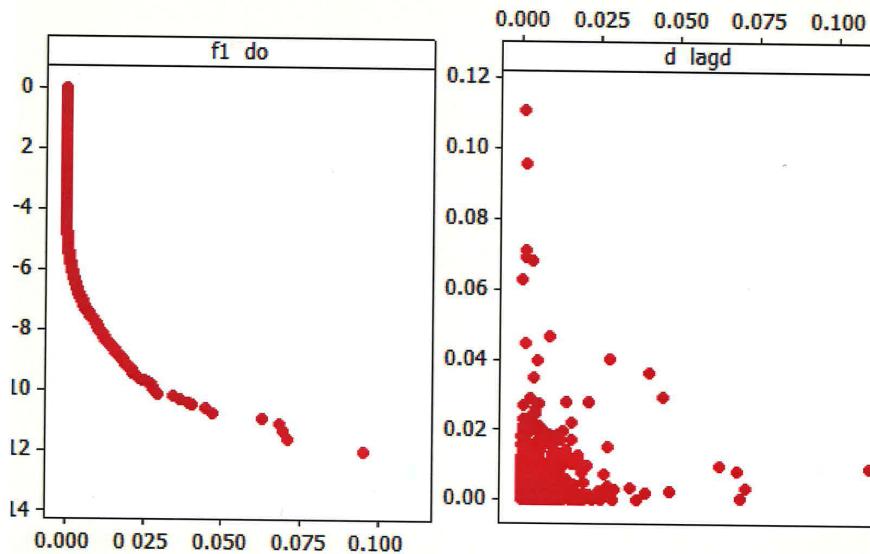
**Figures: Empirical Survivor Function and Serial Correlation plots**

Figure 4.2: Empirical Survivor Function and Serial Correlation plots.

ure 4.2) clearly indicates that the durations do not follow an exponential distribution and do exhibit some dependence. This is confirmed by the autocorrelation function for durations (Figure 4.3).

Cox and Lewis (1966) [72] also observe a few issues that can arise with the point-process type data. It is possible that two or more events recorded as happening at the same time. This may be due to latency issues but if they occur due to genuine coincidences it is better to analyze the number of events attached to each occurrence time as a variate separately. Further complications are that there may be events, such as price and volume that may be interdependent and events that occur with related series that may be useful for modeling and predicting the future occurrence times.

#### 4.3.1 Stylized Models for High Frequency Financial Data

There has been a great deal of interest in studying market micro-structure to better understand the trading mechanisms and the process of price formation. With the availability of Trades and Quotes (TAQ) data that contains all equity transactions on major exchanges, it has now became possible to better understand the market micro-structure. The extensive details from the order books over multiple exchanges have provided massive amount of data. Analyzing these data will be taken up in later chapters as well. Here we want to present some unique characteristics of the high frequency

**Figure7:Autocorrelation Function for duration**  
 (with 5% significance limits for the autocorrelations)

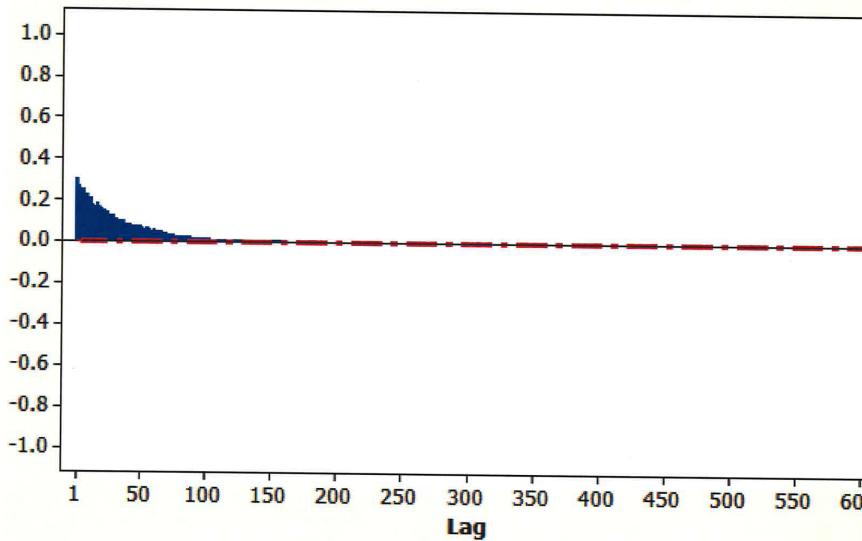


Figure 4.3: Autocorrelation Function for duration.

data and present most commonly used models in practice. To reiterate, the transactions (trades, quotes, bids, etc) may occur at any point in time (Figure 4.4) during the exchange hours as given below. High-frequency data refer to the ‘tick’  $t_i$  data which

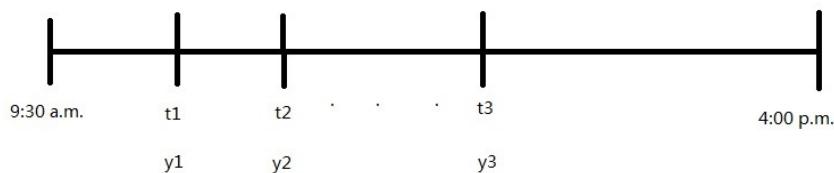


Figure 4.4: Exchange hours.

contains, in addition to the exact time of even occurrence, ‘ $y_i$ ’ called marks that may refer to all other elements of the limit order book such as traded price or quote etc. The traditional time series methods that we discussed in the context of low frequency data analysis are not applicable here as the ticks can occur at any point in time when the exchange are open. Standard econometric techniques are for data in discrete time intervals. Aggregating the high-frequency data to some fixed time interval will not capture the advantages of having access to detailed transaction data. Even for some

heavily traded stocks, if the intervals are chosen to be short, there may be many intervals with no data and if the intervals are long, the microstructure features will be lost. Also certain key features such as the imbalance between bid side and ask side of the limit order book and when and how the market orders cross the spread are not easy to aggregate in a fixed time interval, however short or long it may be. Moreover, the timings of transactions provide valuable information for trading. Some noted features of high frequency data are:

- **Nonsynchronicity:** Different stocks have different trade intensities and even for a single stock, the intensity can vary over a day. For the aggregated low frequency (daily) data, thus, we cannot assume that daily returns occur in equally-spaced time series.
- **Multiple transactions with the same time stamp:** It is possible that in periods of heavy trading especially in the opening and closing times of the exchange, each tick may contain multiple transactions. For the analysis of this type of occurrence, simple aggregate summary measures to more elaborate measures such as the variation in prices in addition to average prices are suggested.
- **Multiple Exchanges:** In the US market there are at least sixteen known lit exchanges; due to latency issues there could be time delays in recording the submission times of orders that get dispatched at the same time. Also with the rule (NBBO) on getting best price anywhere in the market, an aggregated view of the market given a fixed time interval can miss capturing the dependency over exchanges.
- **Intra-day Periodicity:** Generally, it is observed for stocks transaction activities are higher near the open and the close than in the middle of the data. Thus volatility is higher, immediately after the opening and before the closing of the market resulting in U-shape pattern.
- **Temporal Dependence:** High frequency data generally exhibit some dependence. The dependence is due to
  - price discoveries
  - bid-ask bounce
  - Execution clustering of orders

Under aggregation with low frequency data generally the dependence tends to decrease.

- Volatility, volume, spread are higher near the open and close.
- Time between trades, shorter near open and close.

**Stylized fact 6: Negative autocorrelation in returns:** Non-synchronous trading and bid-ask bounce both can introduce negative lag-1 autocorrelation in the returns. See Roll (1984) [?] and Lo and Mackinlay (1990) [198].

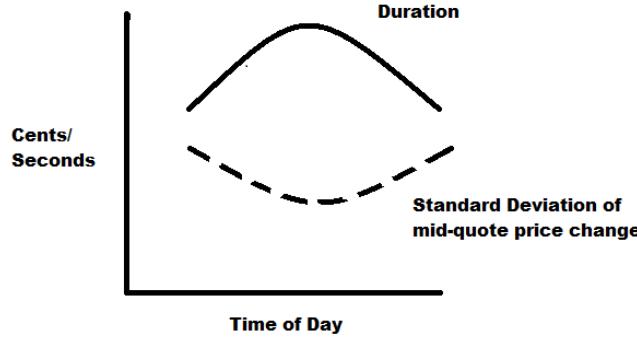


Figure 4.5: Typical Diurnal patterns.

We want to briefly discuss the Roll model; on a typical day of trading stocks exhibit price changes numerous times. These changes signify merely the market friction, the friction between the demand and supply side and they are not necessarily due to true information about the stock. Assume that the true price (unobservable) follows a random walk model (??) as,

$$p_t^* = p_{t-1}^* + \epsilon_t \quad (4.25)$$

and the observed price,

$$p_t = p_t^* + u_t \quad (4.26)$$

with ' $\epsilon_t$ ' denoting the true information about the stock and ' $u_t$ ' denoting the market friction. But the observed return,

$$r_t = p_t - p_{t-1} = \epsilon_t + u_t - u_{t-1} \quad (4.27)$$

inducing the negative autocorrelation of lag 1 in the return due to the carry over term, ' $u_{t-1}$ '. The term on the right-hand side because of zero autocorrelation beyond lag-1 can be written as MA(1) model:

$$r_t = a_t - \theta a_{t-1} \quad (4.28)$$

The value of ' $\theta$ ' generally tends to be small, but nevertheless significant.

*Commonly used duration models:* The timing aspect of transaction was discussed in Easley and O'Hara (1992) [92] and a rigorous set of tools was developed by Engle and Russell (1998) [103], as an alternative to fixed interval analysis. Treating the ticks as random variables that follow a point process, the intensity of transactions in an interval of length,  $\Delta$ , is defined as

$$\lambda(t) = \lim_{\Delta \rightarrow 0} \frac{E[N(t + \Delta) - N(t)|F_t]}{\Delta} \quad (4.29)$$

where  $N(t)$  denotes the number of ticks until ‘t’ and  $F_t$  is the information available until that time. The commonly used model for such arrival rates is Poisson that implies that the time between the two successive ticks is independent of pairwise ticks and follows an exponential distribution. But the duration data,  $d_i = t_i - t_{i-1}$  exhibit some dependence, thus implying that time moves faster sometimes, faster than the clock time. The intensity rate,  $\lambda(t)$  is not a constant, which is a typical characteristic of the exponential distribution. We briefly outline how the durations are modeled. For detailed discussion, see Tsay (2010) [252] or Lai and Xing (2008, Section 11.2) [183]. The autoregressive conditional duration (ACD) model is defined as follows:

$$\begin{aligned} d_i &= t_i - t_{i-1} = \psi_i \varepsilon_i, \text{ and} \\ \psi_i &= \alpha + \sum_{j=1}^p \alpha_j d_{t-j} + \sum_{v=1}^q \beta_v \psi_{i-v} \end{aligned} \quad (4.30)$$

where  $\varepsilon_i$  are i.i.d. with  $E(\varepsilon_i) = 1$ ; If  $\varepsilon_i$ 's are assumed to follow exponential the models (??) is called E(xponential)ACD. Two alternative models WACD and GGACD are based on the following specification for the errors:

$$\begin{aligned} \text{Weibull: } h(x) &= \frac{\alpha}{\beta^\alpha} x^{\alpha-1} \exp \left\{ -\left(\frac{x}{\beta}\right)^\alpha \right\} \\ \text{Generalized Gamma: } y &= \lambda \frac{x}{\beta} \end{aligned}$$

where  $\lambda = \frac{\sqrt{K}}{\sqrt{K+\frac{1}{\alpha}}}$ . The hazard function of Generalized Gamma is quite flexible and may fit better various patterns. The stationary conditions would require that the roots of  $\alpha(B) = 1 - \sum_{j=1}^q (\alpha_j + \beta_j) B^j$  are outside the unit circle when  $g = \max(p, q)$  and  $B$  is the back-shift operator. It is clear that the model (4.30) is similar to GARCH( $p, q$ ) model given in (??). The estimation is typically done through conditional maximum likelihood and all inferences are asymptotic.

An alternative more direct ARMA type model for log durations is suggested by Ghysels, Gourieroux and Jasiak (2004) [133]:

$$\ln(d_i) = \omega + \sum_{j=1}^p \alpha_j \ln(d_{t-j}) + \sum_{j=1}^v \beta_j \varepsilon_{i-j} + \varepsilon_i \quad (4.31)$$

and  $\varepsilon_i = \sqrt{h_i^v} u_i$ ,  $h_i^v = \omega^v + \sum_{j=1}^{p^v} \alpha_j^v \varepsilon_{i-j}^2 + \sum_{j=1}^{q^v} \beta_j^v h_{i-j}^v$  where  $u_i \sim N(0, 1)$  i.i.d.; here the conditional mean and variance of durations are assumed to be separable. The duration volatility that is modeled using the second equation in (??) is interpreted as ‘liquidity’ risk.

**Other Duration Related Models:** So far the discussion has been around modeling the durations, but the information set,  $F_t$  has information on ‘marks’, the information associated with the past ticks, such as the number of units transacted, price etc.

McCulloch and Tsay (2009) [] suggest the following model:

$$\ln d_i = \beta_0 + \beta_1 \ln(d_{t-1}) + \beta_2 s_{i-1} + \sigma \varepsilon_i \quad (4.32)$$

where  $s_i$  is the size of the  $i$ th price change measured in ticks; other relevant variables can be easily added to this model.

#### 4.3.2 Models for Multiple Assets: High-Frequency Context

There is a considerable interest in extending the univariate duration models to multiple stocks or to multiple types of arrival processes. Because investment managers generally consider a portfolio of stocks rather than a single one, it is important for them to follow the transaction processes of several stocks simultaneously. Because of non-synchronous trading relating two stocks with different trading intensities on a common time scale is somewhat difficult. Engle and Lunde (2003) [98] establish the following model for a bivariate point process for trades and quotes: De-

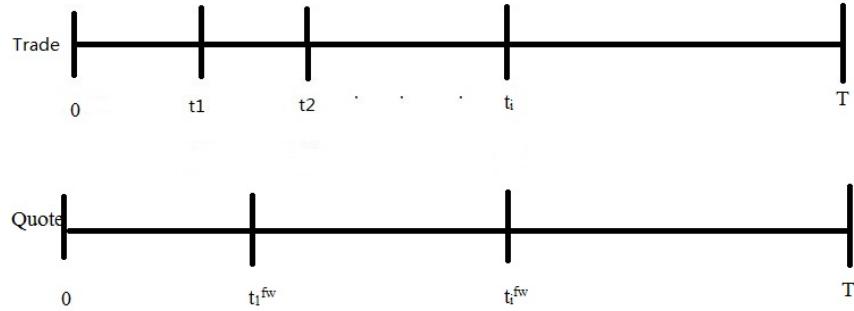


Figure 4.6: Trades and Quotes.

fine  $x_i = t_i - t_{i-1}$  as trade duration and  $y_i = t_i^{fw} - t_{i-1}^{fw}$  as the forward quote duration; our goal is to model  $(x_i, y_i)$  as a bivariate duration process but the underlying timings are not synchronized. Thus define, if  $y_i > x_i$ , then

$$\tilde{y}_i = (1 - d_i)y_i + d_i x_i \quad (4.33)$$

where  $d_i$  is indicator variable,  $d_i = I_{\{y_i > x_i\}}$ . Model  $(x_i, \tilde{y}_i)$  as censored process. The equation (98) provides a mechanism in a way to align the two processes. The joint distribution is given by

$$p(x_i, \tilde{y}_i | F_{i-1}, \omega) = g(x_i | F_{i-1}, \omega_1) f(\tilde{y}_i | x_i, F_{i-1}, \omega_2) \quad (4.34)$$

The first term is the trade density and the second term is the quote density. The  $\psi_i$  terms as in (??) are modeled as EACD,  $\psi_i = f(\psi_{i-1}, x_{i-1}, z_{i-1})$  and quote density is also modeled as EACD after adjusting for censoring:

$$f(y_i | \cdot) = [h(y_i | \cdot)]^{1-d_i} \times S(x_i | \cdot)^{d_i} \quad (4.35)$$

The estimation of the parameters are done through quasi maximum likelihood. Using data for eight stocks, Engle and Lund (2003) [98] observe that high trade arrival rates, large volume per trade, wide bid-ask spreads, all predict more rapid price revisions. The model given in Engle and Lund (2003) [98] can be conceptually applied to any two point processes. Extending this concept to more than two processes could be of interest to the traders who may want to monitor the trading exchanges simultaneously.

#### 4.4 Time Aggregation and Loss of Information

#### 4.5 Realized Volatility and Econometric Models

Estimating the volatility using the high-frequency data has received a great deal of attention lately. In the low frequency (daily) level some estimators based on select prices were presented. Intra-day dynamics of volatility is of interest to traders and that is not fully captured by the stock price indicators sampled during the day. If higher frequency data are used to estimate volatility of lower frequency data, it is important to know the model for the return at the lower frequency. To illustrate this, we consider data observed at two time scales; although they are at low frequency level, the concept can be easily extended to high-frequency context. If  $r_{t,i}$  is the  $i$ th day return in  $t$ th month, the  $t$ th month return assuming there 'n' trading days,  $r_t^m = \sum_{i=1}^n r_{t,i}$ . Note that  $\sigma_m^2 = \text{Var}(r_t^m | F_{t-1}) = \sum_{i=1}^n \text{Var}(r_{t,i} | F_{t-1}) + 2 \sum_{i < j} \text{Cov}(r_{t-i}, r_{t-j} | F_{t-1})$ . If  $r_{t,i}$  is white noise sequence, then

$$\hat{\sigma}_m^2 = \frac{n}{n-1} \sum_{i=1}^n (r_{t,i} - \bar{r}_t)^2 \quad (4.36)$$

But we had observed that in the high frequency data returns do exhibit some serial correlation and so adjusting (??) for serial correlation is important to get a more accurate estimate of volatility. A simpler estimate of  $\sigma_m^2$  is the so called realized volatility ( $RV_t$ )

$$RV_t = \sum_{i=1}^n r_{t,i}^2 \quad (4.37)$$

But in the estimation of intra-day volatility, the number of sampling intervals, 'n' and hence, ' $\Delta$ ', the interval size can affect the estimate. If  $\Delta \rightarrow 0$ , the value  $RV_t$  in the  $\Delta$ -interval goes to infinity. Thus the optimal choice of ' $\Delta$ ' is crucial and is reviewed below, for a judicious choice.

Bandi and Russell (2006) [20] argue that the observed price is the sum of efficient price and a friction price (see (??)) that is induced by bid-ask bounce, price discreteness etc. Thus the observed variance such as (??) is the sum of variance of the efficient returns and the variance of micro-structure noise. As they indicate, both are useful:

“The variance of the efficient return process is a crucial ingredient in the practice and theory of asset valuation and risk management. The variance of microstructure noise components reflects the market structure and the price setting behavior of market participants and thereby contains information about the market fine-grain dynamics.” Both components can be estimated using high frequency data but sampled at different frequencies. Data sampled at low frequency are used to estimate the efficient return variable and the data of high frequency are used to estimate the microstructure noise variance.

On any given day, assume that the observed price, at time ‘ $i$ ’ is as in (??)

$$p_i = p_i^* + u_i, \quad i = 1, 2, \dots, n \quad (4.38)$$

where  $p_i^*$  is the efficient price and  $u_i$  is the of microstructure noise. Dividing a trading day into  $M$  sub-periods, where  $\delta = 1/M$  is the size of the sub-period, write the observed returns as

$$r_{j,i} = r_{j,i}^* + \varepsilon_{j,i}, \quad j = 1, 2, \dots, M \quad (4.39)$$

Assume that the frictions,  $u$ ’s are i.i.d. mean zero and variance  $\sigma_u^2$ . Because  $\varepsilon_{j,i} = u_{(i-1)+j\delta} - u(i-1) + (j-1)\delta$ ,  $\text{Var}(\varepsilon_{j,i}) = 2\sigma_\eta^2$ . It has been noted already that the return series exhibit negative first order autocovariance and  $\varepsilon$ ’s can be taken as MA(1) structure for the  $\eta$ ’s. Thus, the returns are computed at higher frequency at which the new information arrives. Thus  $\sum_{j=1}^M r_{j,i}^2 / M$  can be used to estimate noise variance. For the estimation of efficient price variance, we need to consider an optimal division of a day into ‘ $M$ ’ number of subperiods. The optimal sampling frequency over ‘ $n$

day is chosen as  $\delta^* = 1/M^*$  with  $M^* = \left[ \frac{\hat{Q}}{\hat{\alpha}} \right]^{\frac{1}{3}}$ , where  $\hat{\alpha} = \sum_{i=1}^n \sum_{j=1}^M \bar{r}_{j,i}^2 / (nM)$

and  $\hat{Q}_i = \frac{M}{3} \sum_{j=1}^M \hat{r}_{j,i}^4$ . Based on the analysis of a sample of S&P 100 stocks mid quotes for the months of February 2002, Bandi and Russell (2006) [20] conclude that 5-min frequency is optimal and the 15-min interval may provide the variance estimates ‘excessively volatile’. The economic benefit of optimal sampling is that high frequency traders can employ a realistic variance forecasts in formulating the entry and exit strategies. Additional references on this topic include, Aït-Sahalia, Mykland and Zhang (2005) [2] and Barndorff-Nielsen and Shepard (2002) [22].

## 4.6 Analytics from Machine Learning Literature

The technical trading rules that are discussed in Chapter 3 involve a relatively small information set to carry out prediction. It treats each of the  $n$  time series of asset returns as autonomous. As pointed out by Malkiel (2012) [203], these rules can be easily implemented by most market participants if such opportunities should emerge, and market efficiency would rule them out as winning strategies. More powerful prediction methods using a very large set of potential predictors and yet capable

of avoiding overfitting are needed to take advantage of transient opportunities. We give a overview here of recent advances in high-dimensional regression and classification and in machine learning and computer science to handle different types of data (including text mining).



---

---

## **Part II**

# **Trading Algorithms**

---

---



# 5

---

## *Statistical Trading Strategies and Back-Testing*

---

### CONTENTS

5.1	Introduction to Trading Strategies: Origin and History .....	107
5.2	Trading Rules for Time Aggregated Data .....	108
5.2.1	Filter Rules .....	109
5.2.2	Moving Average Variants and Oscillators .....	111
5.3	Patterns Discovery via Non-Parametric Smoothing Methods .....	113
5.4	A Decomposition Algorithm .....	115
5.5	Fair Value Models .....	117
5.6	Back-Testing and Data Snooping: In-Sample and Out-of-Sample Performance Evaluation .....	118
5.7	Pairs Trading .....	120
5.7.1	Distance-Based Algorithms .....	121
5.7.2	Co-Integration .....	123
5.7.3	Some General Comments .....	124
5.7.4	Practical Considerations .....	126
5.8	Cross-Sectional Momentum Strategies .....	129
5.9	Multiple Indicators and Boosting Methods .....	133
5.10	Evaluation of Strategies: Various Measures .....	135
5.11	Extraneous Signals: Trading Volume, Volatility, etc. ....	136
5.11.1	Filter Rules based on return and volume .....	139
5.11.2	An illustrative example .....	140
5.12	Market Segmentation and Inconsistent Pricing across Markets .....	146
5.13	Supplements and Problems .....	147

---

### 5.1 Introduction to Trading Strategies: Origin and History

**Statistical arbitrage opportunity and strategies:** Trading strategies are central to the notion of statistical arbitrage. Strategies that claim to have better performance than the market date back to the inception of trading in finance. As we had already noted, there are a large number of studies that empirically conclude that the behavior of stock prices appear to contradict efficient market hypothesis. Broadly there are two types of strategies, one the momentum strategy that buys well-performing stocks and sells poor-performing stocks (see Jegadeesh and Titman (1993) [170]) and the other,

contrarian strategy that advocates buying past losers and selling past winners. The average excess return of some strategies relative to a standard capital asset pricing model is shown to be around 12%. Other strategies, such as buying value and selling glamour stocks (see Lakonishok et al. (1994) [186]) use other company fundamental data such as price earnings ratios, dividends, book-to-market values, cash flows and sales growth.

Hogan, Jarrow, Teo and Warachka (2004) [158] define ‘a statistical arbitrage’ as a long horizon trading opportunity that generates a riskless profit. Four conditions that are to be met are: it is a self-financing strategy with a zero initial cost, in the limit has positive expected discounted profits, a probability of a loss converging to zero and a time-averaged variance converging to zero; the last condition implies that an arbitrage opportunity eventually produces riskless profit. Citing anomalies to reject market efficiency has been criticized on the ground that such tests depend upon a specified model for equilibrium returns. Hogan et al. (2004) [158] argue that statistical arbitrage can be defined without any reference to an equilibrium model and rejects market efficiency. They provide a statistical test to determine if a trading strategy constitutes statistical arbitrage based on short horizon incremental profits. In this chapter we will cover some of these trading strategies and finally we will return to the proposed statistical test.

## 5.2 Trading Rules for Time Aggregated Data

Trading Rules that are broadly labeled as technical analysis involve the prediction of future short-term price movements from an inductive analysis of past movements using either quantitative techniques, such as moving averages or qualitative methods such as recognition of patterns from visual inspection of a time-series plot of the asset price data or a combination of both via semi-parametric procedures such as kernel smoothing. In foreign exchange market, as Menkhoff and Taylor (2002) [207] observe the widespread use of these techniques is puzzling, “since technical analysis eschews scrutiny of economic fundamentals and relies on information on past exchange rate movements that, according to the weakest notion of market efficiency, should already be embedded in the current exchange rate, making its use either unprofitable or implying that any positive returns that are generated are accompanied by an unacceptable risk exposure.” In general there are economic fundamentals that can explain long term movements, there is still no fundamental-based exchange rate model that is capable of forecasting short-term movements.

Technical analysis (also known as Chartist analysis) is a set of techniques for obtaining forecasts based on past history of prices along with the history of trading volumes. Clearly for the technical analysis to be useful, price movements should follow some regular, recurring patterns and these patterns must hold for long enough periods to sustain the transaction costs and false signals so that the investment results in a profit. What constitutes a useful forecast from an investor point of view could be

if the past history of prices can be used to increase expected gains. But as Fama and Blume (1966) [107] point out, “In a random walk market, with either zero or positive drift, no mechanical trading rule applied to an individual security would consistently outperform a policy of simply buying and holding the security”. The past movements could be useful only when there is some degree of dependence in successive price changes. The serial correlation measures may not capture the complex dependence that may exist. In what follows we briefly describe some commonly used trading rules and comment on their noted performance.

### 5.2.1 Filter Rules

Alexander (1961, 1964) [] studied the filter technique that is intuitive and is easy to implement; the basic idea is summarized below:

**x% Filter:** If a daily closing price moves up at least x% from its past low, buy and hold, until its price moves down at least x% from a subsequent high, sell and go short. Maintain the short position until price rises at least x% above a subsequent low, buy. Moves less than x% in either direction are ignored.

The general logic behind this rule is that if the price moved up x%, it is likely to move up more than x% further before it moves down by x%. Thus the random walk model is not likely to be upheld by the data. As observed by Mandelbrot (1963) [], because of the price jumps that occur at the time of transaction, the purchase price and sale price will not exactly match as required by the rule.

The initial application of this rule on daily data for S&P industrials showed that the trading rules with 5, 6 and 8% filters generated larger gross profits than the buy and hold strategy. This study did not account for the transaction cost. But a later study that incorporated a commission of 2% for each round-trip, only the larger filter ~ 46% rule beat the buy and hold strategy and all the other filters adjusting for the transaction costs did not perform better than simply buying and holding.

Fama and Blume (1966) [107] after applying Alexander’s filter technique for each of the individual security in Dow Jones Industrial Average for the daily data during a five year period, 1957–1962, concluded, “When commissions are taken into account, the larger profits under the filter technique are those of the brokers. When commissions are omitted, the returns from the filter techniques are, of course, greatly improved but are still not as large as the returns from simply buying and holding.” It is argued that under random walk model, the average return on long positions should be approximately equal to the buy and hold average returns. This may not be true if the drift occurs at the time of entry or exit; the filter rule can be stated under RW Model as follows; let the return,  $r_t = \ln P_t - \ln P_{t-1} = \mu_t + a_t$ , when  $a_t \sim N(0, \sigma_P^2)$  and ‘ $\mu_t$ ’ is the random drift term and  $\sigma_P^2$  is the volatility; as for the filter rule, enter if  $r_t > r$  and suppose this occurs at time  $t_0$ . If the drift term is approximately picked up at time ‘ $t_0$ ’, so that  $\mu_t = 0$ , for  $t < t_0$  but  $\mu_t = \mu > 0$  for  $t \geq t_0$ , the returns from the filter should fare better. The lasting of the drift is to be carefully monitored and using the exit strategy when  $r_t < r$  will guarantee any undue losses.

When the filters are small it has been observed that there is evidence of positive dependence in returns and it may be useful to predict only short term dependence. This may signal opportunities for more frequent transactions but the cost of transactions as mentioned earlier can eat away the profits.

Sweeney (1998) [248] develops a test for risk-adjusted filters profits. The stocks considered in Fama and Blume (1966) [107] that are “winners” in one period tend to persist as winners in later periods. Thus focusing only on winners rather than entire set of stocks gives better results. Assume that the investor uses the filters rule each day to decide to enter or exit based solely as past information. Consider the span of  $N$  days and with a filter strategy, the asset is held on  $N_{in}$  days and is not held on  $N_{out}$  days. At the end of  $N$  days, the filter’s performance is evaluated. The average excess rate of buy and hold is

$$\bar{R}_{BH} = \frac{1}{N} \sum_{t=1}^N (r_t - r_f)$$

where  $r_t$  is the stock’s rate of return in period ‘ $t$ ’ and  $r_f$  is the risk free rate. Assume that ‘ $f$ ’ is the proportion of days, the asset is not held; then the average excess rate of the filter is

$$\bar{R}_F = \frac{1-f}{N_{in}} \sum_{t \in I} (r_t - r_f)$$

Assume that under any of the asset pricing models (CAPM, APT etc.) expected excess rate of return equals a constant risk premium,  $PR_j$ . Note  $E(\bar{R}_{BH}) = PR$  and  $E(\bar{R}_F) = (1-f)PR$ ; Use the adjusted statistic

$$x = \bar{R}_F - (1-f)\bar{R}_{BH} \quad (5.1)$$

to judge the filter rule; observe  $E(x) = 0$  and  $\sigma_x = \sigma[\frac{f(1-f)}{N}]^{1/2}$ , where ‘ $\sigma$ ’ is the standard deviation of excess return  $r_t - r_f$ . Fama and Blume (1966) [107] use

$$X = \bar{R}_{FI} - \bar{R}_{BH} \quad (5.2)$$

where  $\bar{R}_{FI} = \frac{1}{N_{in}} \sum_{t \in I} (r_t - r_F)$ ; Note  $E(X) = 0$  and  $\sigma_x = \sigma_X$ .

The main advantage of  $X$  is that it measures profits for each day in the duration of ‘ $N$ ’ days and  $x$  measures profits per day for those days when the stock is held. The above results hold for the case of time varying risk premia as well; Suppose

$$r_f - r_f = PR + a_t$$

where  $a_t$  is stochastic; thus

$$x = \frac{1}{N_{in}} \sum_{t \in I} a_t - \frac{1}{N} \sum_{t=1}^N a_t + (\bar{PR}_{N_{in}} - \bar{PR}_N) \quad (5.3)$$

If the investor has no insight into changes in risk premia, then  $E(\bar{PR}_{N_{in}} - \bar{PR}_N) = 0$ ;

If the investor has no insight about the stochastic,  $a_t$ , the filter rule may be forecasting the risk premia.

It has been pointed out that Fama and Blume measure is biased against finding successful filter rules. The statistics X and x are meant for long strategies where the investor has the funds in the risk free asset when not holding. If we define,

$$\bar{R}_{F'} = \frac{1}{N} \left( \sum_{t \in I} (r_t - r_f) - \sum_{t \in O} (r_t - r_f) \right)$$

and  $E(\bar{R}_{F'} - \bar{R}_{BH}) = (1 - 2f)PR - PR = -2fPR$ , a proportion of constant risk premium, thus resulting in a bias.

### 5.2.2 Moving Average Variants and Oscillators

The moving average trading rules are generally based on comparing the current price with the current moving average. Consider the prices,  $P_t, P_{t-1}, \dots, P_{t-m+1}$  with a moving window of width ‘m’. Various versions of moving average are available in the literature;

$$(SMA) \text{ Simple Moving Average: } \bar{P}_t^{(m)} = \frac{1}{m} \sum_{i=0}^{m-1} P_{t-i} \quad (5.4)$$

$$(EWA) \text{ Exponentially Weighted Average: } \bar{P}_t^{(m)} = \left( \frac{1}{\sum_{i=0}^{m-1} \lambda^{i+1}} \right) \sum_{i=0}^{m-1} \lambda^{i+1} P_{t-i}$$

$$(BWA) \text{ Bollinger Weighted Moving Average: } \hat{P}_t^{(m)} = \left( \frac{1}{\sum_{i=1}^m} \right) \sum_{i=0}^{m-1} (m-i) P_{t-i}$$

The indicators of the above kind will tend to be profitable in markets that exhibit definite trends and so they are generally known as “trend-following” or sometimes, called as “momentum” indicators. Some simple rules that are normally used in practice are:

**SMA Rule:** If  $\frac{\bar{P}_t^{(m)}}{P_t} > U$  or  $< L$ , trade;

**EWA Rule:** If  $\frac{\bar{P}_t^{(m)}(\lambda)}{P_t} > U$  or  $< L$ , trade

The two types of potential strategies are broadly known as “Contrarian” vs “Momentum” and they differ mainly on when the entry is carried out. The contrarian in a way looks for local minimum to enter the market and the momentum follower looks for a local maximum to enter.

While the above rules are easy to implement and one may observe that they do not take into account the variance or the volatility of the price (return) process. The widely used Bollinger band, conceptually similar to monitoring the quality of a manufacturing process, is based on the weighted moving average and the variance. Define

$\hat{\sigma}_t^{(m)} = \{ \frac{1}{m-1} \sum_{i=0}^{m-1} (P_{t-i} - \bar{P}_t^{(m)})^2 \}^{1/2}$  and the rule can be stated as follows:

**Bollinger Rule:** If  $P_t > P_t^+ = \hat{P}_t^{(m)} + 2\hat{\sigma}_t^{(m)}$ , the upper band, then sell; buy if  $P_t < P_t^- = \hat{P}_t^{(m)} - 2\hat{\sigma}_t^{(m)}$ , the lower band, then buy; No action is taken in-between.

Broadly the key qualities such as moving average, weighted or exponentially weighted average etc. correspond to the running-mean smoother discussed in Lai and Xing (2009, Section 7.2.1) [183]. Instead of the running-mean, some technical rules use the running maximum (i. e  $\max_{0 \leq i \leq m} P_{t-i}$ ) and the running minimum (i. e  $\min_{0 \leq i \leq m} P_{t-i}$ ) in the above rules.

**Moving Average Oscillator:** The use of the technical rules by the foreign exchange professionals is well documented (See Menkhoff and Taylor (2007) [207]). The so called moving average oscillator rule requires computing two moving averages of short and long time spans. Let  $\bar{P}_t^{(m)}$  and  $\bar{P}_t^{(n)}$  be the short and long averages, where  $m < n$ . The rule would signal a break in trend when the longer moving average is crossed by a shorter moving average (or spot rate). The upward trend is signaled when short moving average intersect from below the long average. Conversely a downward break in trend is signaled if the crossing occurs from above. Precisely the rule is:

**Oscillator Rule:**

Enter if  $\bar{P}_{t-1}^{(m)} < \bar{P}_{t-1}^{(n)}$  and  $\bar{P}_t^{(m)} > \bar{P}_t^{(n)}$

Exit if  $\bar{P}_{t-1}^{(m)} > \bar{P}_{t-1}^{(n)}$  and  $\bar{P}_t^{(m)} < \bar{P}_t^{(n)}$

**RSI Oscillator:** Another oscillator that is designed to measure the strength of price movement in a particular direction is the relative strength indicator (RSI). The logic being that if the changes are too rapid in one direction, a correction in the opposite direction is likely to occur soon. Thus these types of rules are called ‘reversal’ or ‘overbought/oversold’ indicators. To define RSI, let  $U_t = \sum_{i=1}^m I(P_{t-i} - P_{t-i-1} > 0) |P_{t-i} - P_{t-i-1}|$  and  $D_t = \sum_{i=1}^m I(P_{t-i} - P_{t-i-1} < 0) |S_{t-i} - S_{t-i-1}|$  where  $I(\cdot)$  is an indicator function. Thus  $U_t$  and  $D_t$  aggregate the positive and negative side of the price changes. Now let  $RS_t = \frac{U_t}{D_t}$  and define

$$RSI_t = 100 \times \frac{U_t}{U_t + D_t} = 100 - 100 \left( \frac{1}{1 + RS_t} \right) \quad (5.5)$$

By construction, RSI is normalized to lie between 0 and 100.

**Reversal Rule:**

**If  $RSI_t > 70$ , it is overbought signaling a downward correction and hence exit or short sell.**

**If  $RSI_t < 30$ , it is oversold signaling an upward correction and hence enter.**

These rules that appear to be somewhat ad hoc can all be related back to the random walk model (see Chapter 2) with a drift:

$$p_t = \mu + p_{t-1} + a_t \quad (5.6)$$

where  $a_t \sim N(0, \sigma^2)$ ,  $\sigma^2$  is the volatility of the stock and note  $r_t = p_t - p_{t-1}$ ; thus  $r_t \sim N(\mu, \sigma^2)$ . If there is any autocorrelation in ' $r_t$ ', and if  $\bar{r}_n$  is the average based on the last ' $n$ ' observations, then  $\bar{r}_n \sim N\left(\mu, \sum_{|h|<n} \left(1 - \frac{|h|}{n}\right) \frac{\gamma(h)}{n}\right)$ , where  $\gamma(h)$  is the autocovariance function of lag ' $h$ ', with  $\gamma(0) = \sigma^2$ . When  $\gamma(h) = 0$  for  $h > 0$ , that  $P(\bar{r}_n > 0) = P(\bar{a}_n > -\mu) = P\left(\frac{\bar{a}_n}{\sigma} > \frac{-\mu}{\sigma}\right) = 1 - \Phi\left(-\frac{\mu}{\sigma}\right)$  where  $\Phi$  is the c.d.f. of the standard normal distribution. Thus we can evaluate  $RS_t$  by the odds ratio,  $\frac{1-\Phi\left(-\frac{\mu}{\sigma}\right)}{\Phi\left(-\frac{\mu}{\sigma}\right)}$ . In fact the reversal rule stated above implies for  $RSI_t$  to be greater than 70,  $RS_t > 2.33$ . All the rules stated in this section can be written in the form of the c.d.f. of the underlying distribution of the return. This model-based interpretation of the rules can help us in accommodating other factors such as volume more formally in the model.

### 5.3 Patterns Discovery via Non-Parametric Smoothing Methods

Charting has been a part of financial practice for many decades; but because of its highly subjective nature of interpretation of geometric shapes in price series, reading charts is “often in the eyes of the beholder”. Lo, Mamaysky and Wang (LMW) (2000) [196] propose a systematic approach to pattern recognition using nonparametric kernel regression. The general goal of this analysis is to identify regularities in the price series by extracting nonlinear patterns. The smoothing estimators that can perform the signal extraction by averaging out the noise are ideally suited for this task.

Assume that prices follow,

$$P_t = m(x_t) + a_t \quad (5.7)$$

where  $m(x_t)$  is an arbitrary function of the state variable,  $X_t$ . For most practical purposes, assume  $X_t = t$ . Kernel regression, orthogonal series expansion, projection

pursuit, neural networks are all examples of smoothing methods that are applied to estimate  $\hat{m}(x_t)$ :

$$\hat{m}(x) = \frac{1}{T} \sum_{t=1}^T w_t(x) P_t \quad (5.8)$$

where the weights  $\{w_t(x)\}$  are large for those  $P_t$ 's paired with  $x_t$ 's near  $x$  and are small for those away from  $x$ . The weight function in (5.8) is defined for a neighborhood of width ' $h$ ' around ' $x$ ' can be defined as

$$\hat{m}_h(x) = \frac{1}{T} \sum_{t=1}^T w_{t,h}(x) P_t = \frac{\sum_{t=1}^T K_h(x - x_t) P_t}{\sum_{t=1}^T K_h(x - x_t)} \quad (5.9)$$

where  $K(x)$  is a probability density function and is also called as 'Kernel', and  $\hat{m}_h(x)$  is called Nadaraya-Watson kernel estimator. A popular estimator is the density of a standard normal distribution,  $K(t) = \frac{e^{-t^2/2}}{\sqrt{2\pi}}$ .

Selecting the appropriate bandwidth, ' $h$ ' is important as a small ' $h$ ' may not smooth much and a large ' $h$ ' may result in too smooth a graph. The choice of ' $h$ ' is made through cross-validation, by minimizing

$$CV(h) = \frac{1}{T} \sum_{t=1}^T (P_t - \hat{m}_h(t))^2 \quad (5.10)$$

where

$$\hat{m}_{h,t} = \frac{1}{T} \sum_{\tau \neq t}^T w_{\tau,h} P_\tau \quad (5.11)$$

Thus the bandwidth choice depends on how well the estimator fits each observation,  $P_t$ , when it is not used in the construction of a Kernel estimator. LMW(2000) [196] observe that generally the bandwidth choice made through (5.10) when applied to price series, tend to be larger and they suggest using  $0.3 \times h$ .

To automate the detection of charting patterns, first the pattern is defined in terms of its geometric properties, then  $\hat{m}(x)$  is estimated via (5.9), to match with it. LMW(2000) [196] define ten different patterns based on local extremes. The window size is  $(t, t + l + d - 1)$ , where  $l$  and  $d$  are fixed parameters, so that a pattern is completed within  $l + d$  trading days.

Within each window, estimate a kernel regression using prices in that window:

$$\hat{m}_h(\tau) = \frac{\sum_{s=t}^{t+l+d-1} K_h(\tau - s) P_s}{\sum_{s=t}^{t+l+d-1} K_h(\tau - s)}, \quad t = 1, \dots, T - l - d - 1 \quad (5.12)$$

and the local extremes are identified through sign changes in  $\hat{m}_h(\tau)$ .

To test if the kernel regression yields better prediction, LMW(2000) [196] compare the unconditioned distribution of returns to the conditioned distribution, conditioned on the existence of a pattern. These comparisons are made through a goodness-of-fit test and Kolmogorov - Smirnov test. It is generally found that the Kernel regression method yields somewhat better predictions. Jegadeesh (2000) [167] in his

discussion of LMW(2000) [196] questions if the chartists' belief that selected patterns of the past tend to repeat as there is a great deal of noted subjectivity in the interpretation of charts. It is further noted that the practitioners tend to use the visual patterns in conjunction with other information such as macro level events that exert large influence compared to stock specific moves, to time their trades.

## 5.4 A Decomposition Algorithm

The rules that were discussed in previous sections are mostly trend following and use the past price or returns data. The predictability of returns generally tends to be weak. But the direction of change and volatility of returns tend to exhibit somewhat stronger dependence over time. Anatolyev and Gospodinov (2010) [8] exploit this and decompose the returns into a product of sign and absolute value components whose joint distribution is obtained through a model for absolute value of the return and a binary choice model for direction of the return and a copula for their interaction. Precisely the returns, ' $r_t$ ', is written as

$$r_t = c + |r_t - c| \operatorname{sgn}(r_t - c) = c + |r_t - c|(2 \cdot I(r_t > c) - 1) \quad (5.13)$$

where  $I(r_t > c)$  is an indicator function. The values of  $c$  are determined by the user; it could be a multiple of the standard deviation of ' $r_t$ ' or it could represent the transaction cost or the risk free rate. This model is conceptually similar to Rydberg and Shephard (2003) [233] and potential usefulness of the decomposition is also mentioned in Anatolev and Gerko (2005) [7].

An advantage of the decomposition can be illustrated as follows. Assume for convenience,  $c = 0$ . From the model for  $|r_t|$  and  $I(r_t > 0)$ , write  $E_{t-1}(|r_t|) = \alpha_{|r|} + \beta_{|r|} RV_{t-1}$  and  $Pr_{t-1}(r_t > 0) = \alpha_{|I|} + \beta_{|I|} RV_{t-1}$ , where  $RV$  denotes the realized volatility and  $E_{t-1}(r_t) = -E_{t-1}(|r_t|) + 2E_{t-1}(|r_t| \cdot I(r_t > 0)) = \alpha_r + \beta_r RV_{t-1} + \sigma_r RV_{t-1}^2$ .

Thus, the decomposition in (5.13) results in accommodating non-linearity and the model can include other asset related variables. Hence any dynamic modeling that captures the hidden non-linearities is likely to improve the prediction of  $r_t$ . With the random walk model for price,  $r_t$  is taken to be white noise, but the above formulation helps to capture any time dependence. We describe below some useful formulations.

The marginal distributions of the absolute value and sign terms are modeled as follows.

$$|r_t - c| = \psi_t \eta_t \quad (5.14)$$

where  $\eta_t$  is Weibull distribution and

$$\ln \psi_t = w_v + \beta_v \ln \psi_{t-1} + r_v \ln |r_{t-1} - c| + \rho_v I(r_{t-1} > c) + x'_{t-1} \delta_v \quad (5.15)$$

Note the above specification is the same as the autoregressive conditional duration (ACD) model. The  $x_t$ 's are macroeconomic predictors that may have an effect on

volatility dynamics. The specification of the indicator variable is done as follows: The indicator variable  $I(r_t > c)$  is assumed to follow Bernoulli( $p_t$ ), where  $p_t = \text{logit}(\theta_t)$ , where

$$\theta_t = w_d + \phi_d I(r_{t-1} > c) + y'_{t-1} \delta_d \quad (5.16)$$

where  $y_t$ 's can, besides the macro economic variables, include the realized variance,  $\text{RV}_t = \sum_{s=1}^m r_{t,s}^2$ , bipower variation,  $\text{BPV}_t = \frac{\pi}{2} \cdot \frac{m}{m-1} \cdot \sum_{s=1}^{m-1} |r_{t,s}| |r_{t,s+1}|$  and the realized third and fourth moments,  $\text{RS}_t = \sum_{s=1}^m r_{t,s}^3$  and  $\text{RK}_t = \sum_{s=1}^m r_{t,s}^4$ .

With the marginals specified as above with the corresponding, c.d.f.'s  $F(\cdot)$ , the joint density is given as

$$F_{r_t}(u, v) = c[F_{|r_t=c}^{(u)}, F_{I(r_t>0)}^{(v)}] \quad (5.17)$$

where  $c(w_1, w_2)$  is a Copula, a bivariate c.d.f.. For a general discussion on Copulas refer to Lai and Xing (2008, Section 12.3.3) [183].

In the trading context, the value of a strategy to buy or sell depends on how good the short term prediction of  $r_t$  is. Note that the conditional mean can be expressed as,

$$E_{t-1}(r_t) = c - E_{t-1}(|r_t - c|) + 2E_{t-1}(|r_t - c| I(r_t - c))$$

Hence,

$$\hat{r}_t = c - \hat{\psi}_t + 2\hat{\epsilon}_t \quad (5.18)$$

where  $\epsilon_t$  is the expected value of the cross product term in  $E_{t-1}(r_t)$ . If the cross-product terms are conditionally independent,  $\epsilon_t = \psi_t p_t$  and thus

$$\hat{r}_t = c + (2\hat{p}_t - 1)\hat{\psi}_t \quad (5.19)$$

Thus the prediction of  $r_t$  includes both the directional and volatility predictions.

A possible strategy is to invest in stocks if the predicted excess return is positive or invest in bonds if it is negative. In term of established measures of performance, the decomposition method is shown to yield somewhat superior results.

**Other Non-Linear Models:** There has been some interest in using Neural Networks for developing technical rules that have better predictive power than the baseline models such as random walk or the following AR( $p$ ) model:

$$r_t = \alpha + \sum_{i=1}^p \phi_i r_{t-i} + a_t \quad (5.20)$$

where  $a_t \sim N(0, \sigma_a^2)$ . This model is augmented with term used in oscillator rule that was discussed earlier. Define  $s_t^{m,n} = P_t^m - P_t^n$ , the oscillator signal resulting from the short and long term moving averages of the prices and the model,

$$r_t = \alpha + \sum_{i=1}^p \phi_i r_{t-i} + \sum_{i=1}^p \eta_i s_{t-i}^{m,n} + a_t \quad (5.21)$$

is an extended linear model and is shown to increase the predictive power. The feed forward neural network model with ‘ $d$ ’ hidden units is written as

$$r_t = \alpha_0 + \sum_{i=1}^p \phi_{ij} r_{t-i} + \sum_{j=1}^d \eta_j G \left[ \alpha_j + \sum_{i=1}^p r_{ij} s_{t-i}^{m,n} \right] + a_t \quad (5.22)$$

where the activation function  $G(x) = \frac{1}{1+e^{-\alpha x}}$  is of the logistic functional form. Gencay(1998) [130] examining the daily Dow Jones Industrial Average Index from 1877 to 1998 finds the evidence of nonlinear predictability in returns by using the oscillator signals.

## 5.5 Fair Value Models

The trading strategies are typically evaluated by comparing their performance against some benchmarks. Traders often face challenges to find trading opportunities that will outperform these benchmarks. In the model considered in Section 3.4, the quantity, ‘ $c$ ’ is taken to be transaction cost or simply it can be set to risk-free rate. The fair value models are based on the assumption that the fair price of a stock is influenced by some common factors. These factors are general market factors; examples of the linear factor models of returns are CAPM, Fama-French three and five factor models and Carhart factor models. As consistent with algorithmic trading principle, large deviations that are not explained by the dynamics of the established factors are taken to be abnormal and can be exploited as trading opportunities.

For the CAPM model, where  $r_t$  is the return on the asset,  $r_{mt}$  is the market return, and  $r_f$  is the risk-free rate,

$$r_t - r_f = \alpha + \beta (r_{mt} - r_f) + e_t \quad (5.23)$$

it is expected that  $\alpha = 0$  if the model holds. Any deviation from the null hypothesis  $H_0 : \alpha = 0$  indicates the adjusted stock performance after adjusting for the risk-return relationship. In addition it is expected in a high-frequency setting,  $e_t$ ’s are likely to exhibit some serial dependence:

$$e_t = a_t - \theta a_{t-1} \quad (5.24)$$

Estimating the model parameters can be accomplished by the methodology given in Chapter 2; the procedure is generalized least squares and is iterative. The model in (5.23) can be extended to include other factors to provide a general model:

$$r_t = \alpha + \beta' f_t + e_t \quad (5.25)$$

where  $f_t$  is an ‘ $n$ ’-dimensional vector of factors that may include in addition to the market index,  $SMB_t$ , the return on diversified portfolio of small stocks minus the

return on a diversified portfolio of big stocks,  $HML_t$  is the difference between the returns on diversified portfolios of high and low Book to Market (B/M) stocks,  $RMW_t$  is the difference between the returns on diversified portfolios of stocks with robust and weak profitability and  $CMA_t$  in the difference between the returns on diversified portfolios of the stocks of low (conservative) and high (aggressive) investment firms (see Fama and French (2015) [108]). Carhart (1997) [52] in studying the persistence of mutual fund performance uses one-year momentum in stock returns as an additional factor to Fama and French's original three factor model.

A general comment is in order: the so called factors mostly meant for explaining cross-sectional variation are slow-moving and hence the only factor that is likely to change in the high frequency context is the market factor. Also the error term may contain mostly the information about the market friction and hence the model (5.23) may itself not be reliable.

## 5.6 Back-Testing and Data Snooping: In-Sample and Out-of-Sample Performance Evaluation

With the publication of an influential paper by Brock, Lakonishok and LeBaron (1992) [46] on technical rules, the main stream finance academics have started paying attention to studying the properties of these rules. The rules considered are moving average and trading range break and applying these rules on the Dow Jones Index from 1897 to 1986, they show that technical strategies perform very well. Results they obtain are not consistent with the usual benchmark models: the random walk, AR(1), GARCH-M and Exponential GARCH. Buy signals generally tend to fare better. The p-values for the rules are estimated through Bootstrap method. But Ready(2002) [227] argues that the apparent success after transaction costs of the Brock et al. (1992) [46] moving average rules is a spurious result of data snooping. It's shown that the rules did poorly post 1986 and the average returns for 1987–2000 were higher on days when strategies suggest to be out of the market.

Data snooping occurs if the same data is used for both model selection and model validation. The significant results obtained can be due to pure chance. As Sullivan, Timmermann and White (1999) [247] suggest the data snooping may be due to survivorship bias as only successful trading rules tend to be replicated. Thus it is important to consider the almost entire universe of trading rules to study the snooping bias.

Sullivan et al (1999) [247] suggest using a bootstrap procedure studied in White (1999) [260] to check on data snooping bias. The bias comes mainly from the application of multiple trading rules on the same set of data and thus by pure chance can yield significant results. If there are ' $I$ ' trading rules under consideration and they are

validated over next ‘ $n$ ’ periods, define

$$\bar{f} = \frac{1}{n} \sum_{j=1}^n \hat{f}_{t+j} \quad (5.26)$$

where  $\hat{f}_{t+j}$  is the observed performance measure for period ‘ $t + j$ ’, based on the data up to ‘ $t$ ’. For the  $k$ th trading rule, let

$$f_{k,t+j} = \ln(1 + r_{t+j}s_k) - \ln(1 + r_{t+j}s_0) \quad (5.27)$$

where  $r_{t+j}$  is the return,  $s_k$  and  $s_0$  are signal functions for the  $k^{th}$  trading rule and the baseline or benchmark rule, that can take three positions: long, short and neutral, taking values 1, −1, and zero, respectively. The signal functions are of course based on data and information available as of ‘ $t$ ’. The proposed test assumes the null hypothesis that the best technical rule does no better than the benchmark rule:

$$H_0 : \max_{k=1,\dots,l} E(f_k) \leq 0 \quad (5.28)$$

Rejection of  $H_0$  would establish that the best technical rule is superior to the benchmark rule. Hypothesis similar to (5.28) can be formulated for the Sharpe ratio as well which will be a risk-adjusted measure. The  $p$ -values for  $H_0$  are obtained through Bootstrap resampling procedures, to construct the following statistics: If there are ‘ $B$ ’ bootstrapped values of  $\bar{f}_k$  denoted by  $\bar{f}_{k,i}^*$ ,  $i = 1, 2, \dots, B$  construct the following statistics:

$$\bar{v}_l = \max_{k=1,\dots,l} \{ \sqrt{n} \bar{f}_k \} \quad (5.29)$$

$$\bar{v}_{l,i} = \max_{k=1,\dots,l} \{ \sqrt{n} (\bar{f}_{k,i}^* - \bar{f}_k) \} \quad (5.30)$$

The values of  $\bar{v}_l$  are compared to the quantities of  $\bar{v}_{l,i}^*$  to obtain the  $p$ -values.

The number of trading rules,  $l = 7846$ , considered by Sullivan et al (1999) [247] is fairly exhaustive. The benchmark rule is being ‘always out of the market.’ The difference between the two  $p$ -values, i.e. between the Bootstrap resampling method and the nominal method is fairly substantial for the out-of-sample periods for the universe of trading rules considered. However, there are certain trading rules that are found to outperform the benchmarks both in terms of the average return and the Sharpe ratio.

Two main drawbacks of the above test are: one the test is conservative and may lose power when poor performing models are included in the universe of the tests; second the test does not identify other significant models other than the best performing model. Hsu, Hsu and Kuan (2010) [160] propose a stepwise modification of the ‘superior predictive ability (SPA)’ test suggested in Hansen (2005) [143] and the modification is suggested in Romano and Wolf (2005) [231]. The modified method works as follows; the test statistic is

$$\text{SPA}_l = \max(\max_{k=1,\dots,l} (\sqrt{n} \bar{f}_k, 0)) \quad (5.31)$$

By resampling, the critical values of  $\text{SPA}_l$  are determined. Now

- a. Rearrange  $\bar{f}_k$  in a descending order.
- b. Reject the top model ‘ $k$ ’ if  $\sqrt{n\bar{f}_k}$  is greater than the critical value; if no model is rejected stop; else go to the next step.
- c. Remove  $\bar{f}_k$  of the rejected models; regenerate the critical values. Now do the test and if no model is rejected, end the procedure; else go to the next step.
- d. Repeat (c.) until no model can be rejected.

The hypotheses stated in (5.28) are joint hypotheses and hence the tests are judged by Family Wise Error Rate (FWE) which is the probability of rejecting at least one correct null hypothesis.

Hsu et al. (2010) [160] consider two types of technical rules, filter and moving-average based, for several market indices and the ETFs. There are multiple rules that result in significant increase in mean return and in terms of Sharpe Ratio.

Bajgrowicz and Scaillet (2012) [17] use the concept of false discovery rate (FDR) to reach a negative conclusion as in Ready (2002) [227]. The FDR is studied in detail in the context of mutual fund performance by Barra, Scaillet and Wermers (2010) [23]. The methodology that we discussed last indicates if a rule outperforms a benchmark after accounting for data snooping and it does not provide any information on the other strategies. But in reality investors gather signals from multiple rules and these rules are not independent. The FDR approach by allowing for a certain number of false discoveries, accounts for the dependency and also provides an array of high performing rules. Bajgrowicz and Scaillet (2012) [17] address also the issue of persistence of these rules associating a realistic transaction cost with each trade signal. Their main finding based on the analysis of Dow Jones Industrial Average daily data (1897–1996) is that ‘we are not able to select rules whose performance is persistent and not cancelled by transaction costs.’ They conclude: “...investors should be wary of the common technical indicators present in any investment website...”, but acknowledge, “...our results say little about the existence of profitable trading strategies in other markets, using different frequencies or more sophisticated rules.”

## 5.7 Pairs Trading

Pairs trading is a market neutral trading strategy that enables traders to make profits from theoretically any market conditions. It was pioneered by a group of quants at Morgan Stanley led by Gerry Bamberger and Nunzio Tartaglia. Statistical Arbitrage (StatArb) pairs trading, first identifies pairs of similar securities whose prices tend to move together and thus uncovers arbitrage opportunities if there is a divergence in price. If their prices differ markedly, one may be overvalued, the other undervalued or both situations may occur together. Taking a short position on the relatively overvalued security and a long position on the undervalued security will result in profit

when the mispricing corrects itself in the future. The magnitude of the spread between the two securities indicates the extent of mispricing and the range of potential profit. This type of trading is called noise trading as it is not correlated to any changes in the fundamental values of the assets. A pairs trading strategy which assumes the fundamental values of the two stocks are similar can provide a measure of the magnitude of noise traders risk.

To illustrate, we consider the so called “Siamese twins” stocks: Royal Dutch/Shell and Unilever NV/PLC. Prior to 2005, Royal Dutch and Shell were jointly held by two holding companies. Royal Dutch which was domiciled in the Netherlands had an interest of 60% of the group and Shell based in the United Kingdom had a 40% interest. When the unification occurred it was agreed to divide the aggregate dividends, 60:40. If the law of one price holds, the theoretical parity ratio expressed in a common currency is,

$$\text{Theoretical Parity} = \frac{0.6(\text{outstanding Shell Shares})}{0.44(\text{outstanding Royal Dutch Shares})}.$$

Scruggs (2007) [238] defines mispricing as simply being the deviation of the actual price ratio from parity. Figure 5.1 plots the deviation from the theoretical parity which is taken to be 1.6667. Scruggs (2007) [238] also discusses another pair. Unilever NV and Unilever PLC are the twin parent companies of the Unilever group and to operate as a single company, they are governed by an equalization agreement, “one NV ordinary share equals to 1.67 ordinary shares of PLC”. Thus the theoretical parity is fixed at 6.67:1. Figure 5.2 plots the deviation from this parity. The deviation could be as high as 27% in this case, which happens in February of 1999. Both pairs clearly indicate the opportunities to exploit the divergence in the stock prices. Froot and Dabora (1999) [123] also studied the two pairs of stocks but their focus was on relating the changes in stock prices to the location of trade; the movements are more aligned with the movements of the markets where they are most intensively traded.

The first step in developing a pairs trading strategy is selecting two or more stocks that are closer to each other. The two example discussed here are natural-twins as their parent companies are the same. Generally mature companies with long trading histories and that have similar exposure to macro factors are considered for forming pairs; usually they tend to be from the same industry. We will discuss in detail how pairs can be formed later. First we present some formal procedures for implementing the pairs trading strategy.

### 5.7.1 Distance-Based Algorithms

Gatev, Goetzmann and Rouwenhorst (2006) [128] suggest using the distance based algorithm on normalized prices to identify the periods when the prices diverge. The normalization factor is a cumulative total returns index of the stock over the pairs formation period. Pairs are formed by computing the Euclidean distance between any two stocks and by selecting the pairs that are closer. A simple pairs trading algorithm is stated below, as illustration:

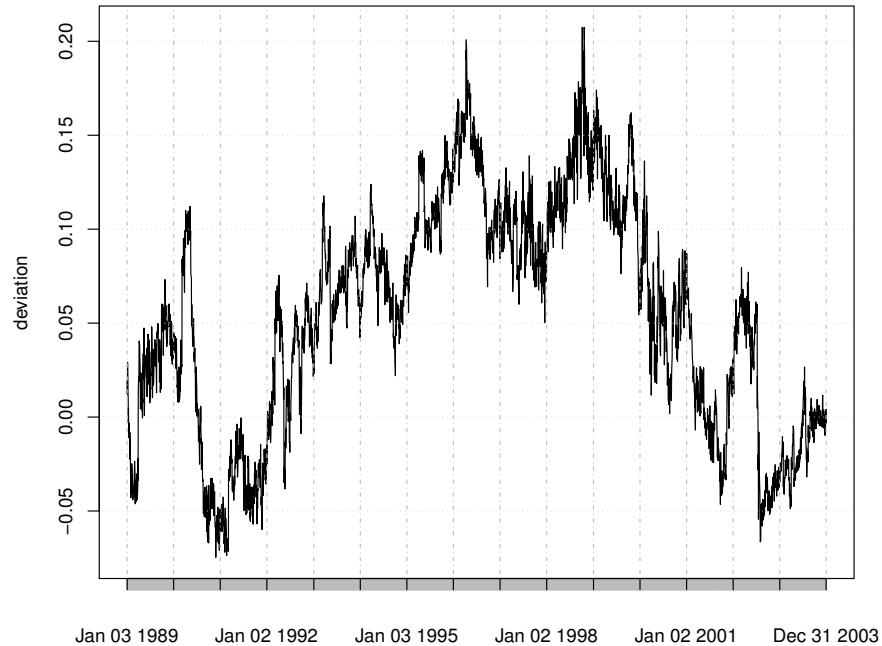
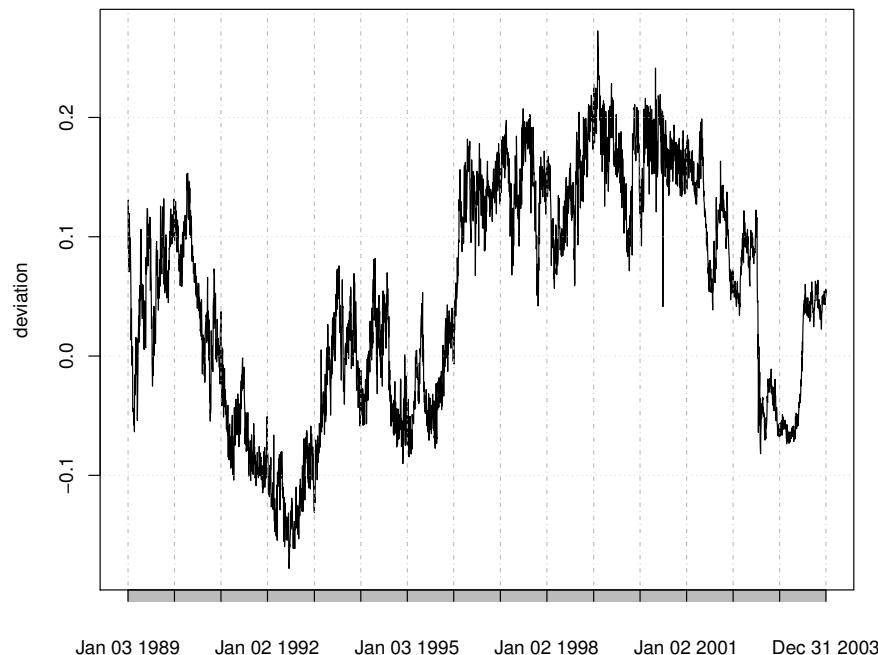
**Figure1: Royal Dutch/Shell**

Figure 5.1: Royal Dutch/Shell.

**Pairs Trading Algorithm:**

- Normalize the prices of every stock to equal unity, at the end of each calendar month,  $m$ ; use the preceding 12 months as the estimation period; let  $T_m$  be the number of trading days in the 12-month estimation associated with month  $m$ .
- For each  $t = 1, \dots, T_m$ ,  $p_t^i = \prod_{T=1}^t [1 \times (1 + r_T^i)]$ , where  $T$  is the index for all the trading days till  $t$ ;
- $\text{PD}_{i,j,m} = \sum_{i=1}^{T_m} \frac{p_t^i - p_t^j}{T_m}$   
 $\text{St.Dev PD}_{i,j,m} = \sqrt{\frac{1}{T_m - 1} \sum_{i=1}^{T_m} [(p_t^i - p_t^j)^2 - \text{PD}_{i,j,m}]^2}$
- If price difference (PD) diverges by more than 2 St.Dev PD, buy the cheaper stock and sell the expensive one; if pairs converge at later time, we unwind; if the pairs diverge but does not converge in 6 months, close the position.

**Figure2: Unilever NV/PLC****Figure 5.2: Royal Dutch/Shell.**

Gatev et al (2006) [128] study the risk and return characteristics, using daily data from 1962 to 2002, of pairs trading for various sectors. The general finding is that pairs trading is profitable in every broad sector category. In relating the excess returns to the risk factors, market, SMB, HML, momentum and reversal, it is found that the exposures are not large enough to explain the average returns of pairs trading. Thus they conclude, "...pairs trading is fundamentally different from simple contrarian strategies based on reversion."

### 5.7.2 Co-Integration

There is some evidence that prices are co-integrated for the U.S. Stock market (see Bossaerts(1988) [34]). If the price vector,  $p_t$  is co-integrated with co-integrating rank,  $r$  which imply that ' $r$ ' linear combinations of  $p_t$  are weakly dependent; thus these assets can be taken as somewhat redundant and any deviation of their price from the linear combination of the other assets prices is temporary and can be expected to revert. Writing the pairs trading concept in a model form has several advantages. It

allows us to extend pairs to triplets etc and also can accommodate incorporating other factors such as other relevant indices.

A model based co-integration approach to pairs trading is elegantly presented in Tsay (2010, Section 8-8) [252]. It is based on the assumption that prices follow a random walk and on the additional expectation that if two stocks have similar risk factors, they should have similar returns; writing  $p_{it} = \ln(p_{it})$  and with  $p_{it}$  following a random walk,  $p_{it} = p_{it-1} + r_{it}$  where  $r_{it}$  is the return. Thus price series are said to be cointegrated, if there exists a linear combination,  $w_t = p_{1t} - \gamma p_{2t}$  which is stationary and therefore mean reverting.

With the error correlation form of the bivariate series (follows from equation (90) in Ch 2),

$$\begin{bmatrix} p_{1t} - p_{1t-1} \\ p_{2t} - p_{2t-1} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} [w_{t-1} - \mu_w] + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \quad (5.32)$$

where  $\mu_w = E(w_t)$  is the mean of  $w_t$ , which is the spread between the two log prices. The quantity  $w_{t-1} - \mu_w$  denotes the deviation from equilibrium,  $\alpha$ 's show the effect of this past deviation on the current returns and should have opposite signs. Going long with one share of stock 1 and shorting  $\gamma$  shares of stock 2, will result in the return of the portfolio,  $\gamma_{p,t+i} = w_{t+i} - w_t$ , at time 't'. The increment does not depend on  $\mu_w$ .

Based on the above discussion the following strategy can be followed. Let  $\eta$  be the trading cost and  $\Delta$  be the target deviation of  $w_t$  from  $\mu_w$  and assume  $\eta < 2\Delta$ .

- Buy a share of stock 1 and  $\gamma$  shares of stock 2 at time t if  $w_t = \mu_w - \Delta$
- Unwind at time ' $t + i$ ' ( $i > 0$ ) if  $w_{t+i} = \mu_w + \Delta$

Thus the return of the portfolio would be  $2\Delta$  which is greater than the trading cost. For other modified versions of the strategy refer to Tsay (2010, Section 8.8) [252].

From a practitioner standpoint, for testing if there is a co-integrating relationship between pairs, it is easier to use the two-step test proposed by Engle and Granger (1987) [101]

- a) Run a regression model:  $p_{1t} = \alpha + \beta p_{at} + u_t$
- b) Test if the errors,  $u_t$  are stationary, via the model  $u_t = \gamma + \phi u_{t-1} + a_t$ , with the null hypothesis  $H_0 : \phi = 1$  vs the alternative hypothesis  $H_a : \phi < 1$ .

If  $u_t$  is stationary, it would imply the two prices  $p_{1t}$  and  $p_{2t}$  are co-integrated.

To illustrate pairs trading in real time, we consider the daily exchange rates between Indian rupee and U.S. dollar, GBP and Euro from December 4, 2006 to November 5, 2010. We normalize the first day's observation as unity and also note that the initial ratio is 1:1.98:1.33 in rupees for U.S. dollar, GBP and Euro respectively. From their plots in Figure 5.3, it is obvious that all three series move together until the middle of 2007, then the Euro takes off and then the U.S. dollar which is closely aligned with GBP takes off to join the Euro. For testing co-integration, one has to look at the three regimes individually. Thus tracking the co-integrating relationships over time, traders can generate profits by properly exchanging currencies.

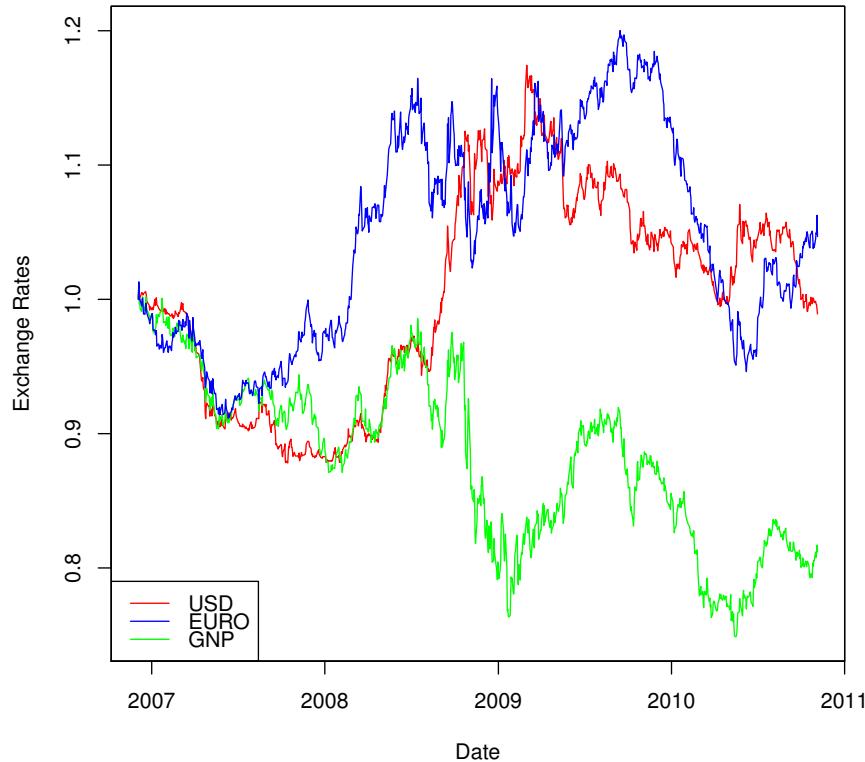
**Figure3: Rupee data**

Figure 5.3: Rupee data.

### 5.7.3 Some General Comments

**Finding Pairs:** An alternative approach is to use fundamentals of the firms to select two stocks that have similar risk factor exposures. Based on Asset Pricing Theory (APT) model,

$$r_{it} - r_{ft} = \alpha_i + \beta'_i f_t + a_{it} \quad (5.33)$$

where  $r_{ft}$  is the risk-free rate and  $f_t$  are the factor scores. Now collect the coefficient vector  $\hat{\theta}'_i = (\hat{\alpha}_i, \hat{\beta}_i)'$  for  $i = 1, 2, \dots, N$  assets under consideration and form a standardized distance measure,

$$d_{ij}^2 = \hat{\theta}'_i \text{Cov}^{-1}(\hat{\theta}_i) \hat{\theta}_i \quad (5.34)$$

and select pairs with the smaller distance. This hopefully will result in similar assets after adjusting for the risk factors. Generally the factors tend to vary much more

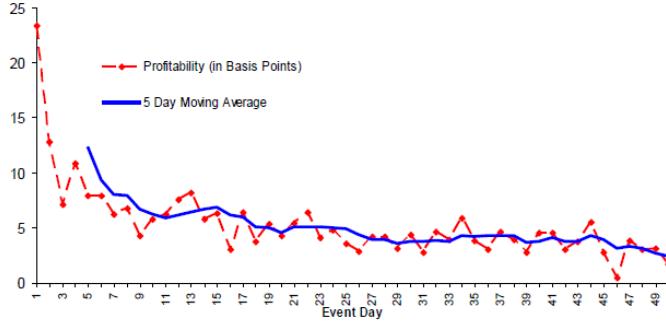


Figure 5.4: Pairs Trading Profitability in Event Time.

slowly than the excess returns,  $r_{it} - r_{ft}$ , the distance based on  $\hat{\theta}_i$  tends to move slower than short-term price movements. Recent studies confirm that the simple Euclidean distance-based pairing may be inadequate. Do and Faff (2010) [85] suggest adding two additional metrics: industry homogeneity and the frequency of the historical reversal in the price spread; the measure in (5.24) can accommodate that easily.

**High-Frequency Equity Pairs Trading:** While there is evidence of the profitability of pairs trading strategies at lower frequencies, not much is known in academia (but quite known in practice) about high-frequency intra-day pairs trading. Bowen, Hutchinson and O'Sullivan (2010) [38] consider data from the London Stock Exchange for the most liquid stocks, the constituents of the index FTSE100. The duration is divided into a formation period, that is used to identify pairs of stocks that move together, and a trading period when the actual transactions occur. The distance based criteria using normalized prices is used to identify the top pairs during the formation period. As any other strategy, pairs trading profits are affected by transaction costs, and this becomes more acute as trading frequency increases. Results appear to be also quite sensitive to the speed of execution, as excess returns disappear quickly if execute delay is greater than one period. The majority of profits seem to occur in the first and last hour of trading, when relatively larger volatility and liquidity are occurring.

Although pairs trading has been used extensively by traders, it also has its limits. In a recent study, Engelberg, Gao, Jagannathan (2009) [96] demonstrate that the opportunities for medium frequency pairs trading are not many and profits opportunities might vanish within a few days (see Figure 5.4). Scruggs (2007) [238] attributes the demise of Long-Term Capital Management to the hedge fund overconfidence in the strategy and their undue bets sizes compared to other market participants.

### 5.7.4 Practical Considerations

After identifying a potential pairs strategy via co-integration testing for instance, traders need to pay particular attention to the expected profitability of the strategy once real-life trading frictions have been taken into consideration. In practice, a significant portion of potential pairs trading strategies experience negative P&L in production due to large execution slippage, whether originating from aggressive hedging at trade initiation, or position closing out.

When back testing a pairs strategy, one of the immediate cost that might render the strategy unprofitable is the spread cost. In other words, how much will it cost to extract liquidity when the signal triggers, in particular if the signal is too short-lived to allow for sourcing liquidity passively via limit order book posting. For intraday pairs trading it is also valuable to analyze if there is any asymmetry of cost between the initiation trade and the unwind trade. Often time, it is not uncommon to witness significantly higher exit costs once the market has normalized following a short-lived entry trade opportunity.

The individual instruments microstructure also plays an important role in the strategy profitability. For instance long queue names (instruments for which the top of book liquidity is large in comparison to daily volume) are harder to trade because the time it takes for an order to reach the top of the queue and trade passively is generally larger than the time that is acceptable for the strategy to mitigate its legging risk<sup>1</sup>. As a result, most of the executions on the leg that displays a long queue will end up having to pay the full spread to be executed.

As an example, when computing the spread ratio between two instruments using the market midpoint, one might be led to believe that the ratio is currently above  $2\sigma$  of its distribution, warranting to initiate a trade. However, when considering the size of the spread and amount of the quoted volume on the passive side at the top of the book, it is not impossible that the realized ratio when crossing the spread is only a  $0.5\sigma$  event, while getting a passive execution when quoting at the end of a long queue would represent a  $5\sigma$  event that one shouldn't expect to realistically happen. These anecdotal examples give an idea of what practitioners should pay particular attention to when designing their strategies. Below, we mention a non-exhaustive list of implementation challenges also worth considering:

- **Difference in liquidity between the two legs:** Usually the selected pair has a slower moving leg and a faster moving one. This means that, in practice, one of the instruments might be ticking faster, thereby having a greater influence on the moves of the measured ratio. Additionally, faster ticking instruments also tend to have narrower spreads, meaning that in absence of control, the instrument the most likely to trigger a trade is also the one that is the cheapest to transact, forcing the strategy to then rapidly hedge out its exposure by paying the larger spread on the slowest ticking leg. One way of handling this situation is, for instance, to only

---

<sup>1</sup>legging risk can be defined as the risk of not being able to fill timely a particular leg of a pair strategy at the required price

let the strategy quote on the slow moving leg as long as it is “in ratio”, and then only trigger the aggressive spread-crossing hedge trade if and when the passive leg receives a fill.

- **Rounding/tick size granularity:** While in theory it is possible to come up with a signal threshold that is as granular as desired, in practice the step increment of the signal might be constrained by the microstructure of the instruments traded. Most exchanges globally impose a minimum tick size increment that can vary based on the liquidity of the instrument. Here too, all things being equal, a practical approach would be to quote on the instrument which constrained tick size contributes the most to the ratio moves, and only trigger the aggressive spread-crossing hedge trade if and when the most constrained tick size leg receives a fill.
- **Signal stability:** As a result of both the difference in liquidity between long and short legs, and the potential frictions induced by the instruments microstructure described above, it is not uncommon for the signal to jump around the trigger ratio frequently. Trading in and out on each signal crossing of the pre-determined ratio would incur large transaction costs that would render many strategies unprofitable. To address this issue, a common approach is to define two different thresholds sufficiently apart for entering and exiting the trade so that signal noise does not trigger excess in and out trading. It’s also possible to enforce a minimum holding duration to prevent the stop loss signal to trigger just due to local noise around the ratio threshold.
- **Time of day:** In the first 15 to 30 minutes after a market opens, many instruments have significantly wider bid-ask spreads while price discovery sets in after a night of information accumulation and no possible trading. If this significant difference in transaction costs is not properly accounted for, the trades initiated during that period can have an impact on the overall profitability of the strategy.
- **Market asynchronicity:** While pairs are usually constructed between instruments trading concomitantly, it is also possible to construct pairs strategies with instruments from different markets. For instance, a pair between a US and a Japanese stock. This raises an additional challenge, which is: how to assess the theoretical value on an instrument that is not currently traded? Practitioners commonly address this issue by first modeling the relationship between the instruments daily time series, and then inferring its stability at more granular frequencies. The common risk can be hedged with a market wide instrument such as an index future, and be unwound overnight when the second leg is entered, thereby limiting unwanted exposure to market risk.
- **Traditional Index Arb / ETF Arb:** When trading a pairs strategy between a composite instrument (Index, ETF, ...) and its constituents, the fast and efficient nature of electronic markets might render trading opportunities too small to be exploited profitably across all instruments. In particular if the number of constituents is quite large, obtaining passive fills on most of them within a reasonable

timeframe is not a realistic assumption. As a result, traders can implement partial replication strategies whereby they only trade an optimized tracking basket of the instruments that allow for the best replication under some liquidity and cost assumptions (spread size, top of book size, ...).

- **Local market idiosyncrasies:** For instance, in Japan, sector classification (which is often used as a starting point for pairs selection or as a way to avoid picking names that would have spurious autocorrelation) does not match international sector classification one-for-one. Local market participants rely on the 33 TOPIX sectors that have a closer fit to the actual structure of the Japanese economy than the commonly used 11 GICS (r) sectors.
- **Coupling market making and pairs trading strategies:** One possible approach to long queue and wide spread instruments pairs trading is to use a market making strategy to first gain queue priority in the longest queue name and achieve a certain child order density in the order book at regular volume-time intervals. For example, if only one of the instruments has a long queue, as long as the pairs strategy ratio is below the trigger threshold, the market making strategy can earn the spread on the long queue instrument while staying out of the market on the other instrument. When the ratio reaches the threshold, one can cancel the child orders created on the opposite side of the book by the market making strategy and only keep the ones on the side relevant to the pairs strategy. These orders, placed over time while market making, should already have accrued some queue priority, which will increase the probability of receiving a passive fill once the ratio is favorable. Then, the pairs strategy is able to trade aggressively the liquid leg as soon as the passive fill is received. Based on the liquidity and microstructure of the instruments in the pairs strategy, different equivalent strategies can be combined to achieve higher profitability.

---

## 5.8 Cross-Sectional Momentum Strategies

The trading strategies discussed thus far depend exclusively on the existence of time-series patterns in returns or on the relationships between returns and the risk factors. Time series strategies generally exploit the fact that stock prices may not follow random walks, at least in the short run. Even if prices follow random walks, it has been shown that strategies based on past performance contain a cross-sectional component that can be favorably exploited. For instance, momentum strategies demonstrate that the repeated purchase of past winners from the proceeds of the sale of losers can result in profits. This is equivalent to buying securities that have high average returns at the expense of securities that yield low returns on average. Therefore if there is some cross-sectional variance in the mean returns of the securities in the universe, a momentum strategy is profitable. On the other hand, a contrarian strategy will not be

profitable even when the prices follow random walks. For the demonstration of the rules of the momentum strategy in the low frequency setting, refer to Jegadeesh and Titman (1993) [168], 2002 [170] and Conrad and Kaul (1998) [64].

To better understand the sources of profit from a momentum strategy Lo and MacKinlay (1990) [198] decompose the profit function into three components. We will follow the set-up as presented in Lewellen (2002) [191]. Let  $r_{i,t-1}^k$  be the asset  $i$ 's, ' $k$ ' month return ending in ' $t - 1$ ' and let  $r_{m,t-1}^k$  be the return from the equal-weighted index's ' $k$ ' month return ending in ' $t - 1$ '. The allocation scheme over ' $N$ ' assets under the momentum strategy is,

$$w_{i,t}^k = \frac{1}{N}(r_{i,t-1}^k - r_{m,t-1}^k) \quad (5.35)$$

For ease of interpretation,  $w_{i,t}^k = w_{i,t}$  and focus on only one previous month's return for allocation; let  $r_t$  be a  $N \times 1$  vector of returns, with the mean return  $\mu = E(r_t)$  and  $\Omega = E[(r_{t-1} - \mu)(r_t - \mu)']$  the autocovariance matrix of returns. Let  $w_t$  be  $N \times 1$  vector of weights,  $w_{it}$ ; the portfolio return for month ' $t$ ' is  $\pi_t = w_t' r_t$ . It is easy to show

$$\begin{aligned} E(\pi_t) &= \frac{1}{N} E \left[ \sum_i r_{i,t-1} \cdot r_{it} \right] - \frac{1}{N} E \left[ r_{m,t-1} \cdot \sum_i r_{it} \right] \\ &= \frac{1}{N} \sum_i (\rho_i + \mu_i^2) - (\rho_m + \mu_m^2) \end{aligned} \quad (5.36)$$

where  $\rho_i$  and  $\rho_m$  are lag 1 autocovariance of asset ' $i$ ' and the equal weighted index. Note that the average autocovariance is  $\text{tr}(\Omega)/N$  and the autocovariance for the market as a whole is  $(1'\Omega 1)/N^2$ . The  $\text{tr}(\Omega)$  denotes the sum of the diagonal elements of  $N \times N$  matrix  $\Omega$ ; recall that the diagonal elements are the autocovariance,  $\rho_i$ , of individual stocks. Hence (??) can be written as,

$$E(\pi_t) = \left( \frac{N-1}{N^2} \right) \cdot \text{tr}(\Omega) + \frac{1}{N^2} [\text{tr}(\Omega) - 1'\Omega 1] + \sigma_\mu^2 \quad (5.37)$$

The three terms in (5.37) precisely identify three sources why positive profits occur (see Lo and Mackinlay (1990) [198]):

- Stock returns are positively autocorrelated, so each stock's past return predicts high future return,
- Stock returns are negatively correlated with the lagged returns on other stocks, so their poor performance predicts high returns for others
- Some stocks have simply high unconditional mean returns relative to other stocks.

The finance literature is replete with empirical verifications of the momentum strategy. However, the analytical decomposition in (??) is not informative about the

economic sources of momentum. Why the momentum strategy works has been attributed to various theories. The literature argues that firm-specific returns drive momentum. Firm-specific returns can be persistent because investors under-react to firm-specific news or stock returns covary strongly due to a common influence of a macro factor. Lewellan (2002) [191] shows that the Size and B/M portfolios also exhibit strong momentum; as these portfolios are well-diversified, their returns reflect systematic risk and therefore macroeconomic factors must be responsible for the momentum. Hvidkjaer (2006) [164] suggests that momentum could be partly driven by the behavior of small traders who under-react initially, followed by delayed reaction to buying and selling pressures. The trading pressure tend to translate into price pressures that affect the trading behavior of the small traders.

Although cross-sectional momentum is widely documented, among the three components in (27)(??), time series momentum as represented by the first term tends to dominate for every liquid instrument in a sample of 58 instruments spread over a variety of asset classes, equity index, currency, commodity and bond futures. Moskowitz, Ooi and Pedersen (2012) [214] show that the autocovariance term,  $\text{tr}(\Omega)/N$  is the dominant factor, particularly in foreign exchange and equity markets. To make meaningful comparison across asset classes, the excess returns are scaled by their ex-ante volatilities and the following models are estimated for predicting price continuation and reversal:

$$\frac{r_t^s}{\sigma_{t-1}^s} = \alpha + \beta_h \frac{r_{t-h}^s}{\sigma_{t-h-1}^s} + \varepsilon_t^s \quad (5.38)$$

and

$$\frac{r_t^s}{\sigma_{t-1}^s} = \alpha + \beta_h \text{sgn}(r_{t-h}^s) + \varepsilon_t^s \quad (5.39)$$

where  $h = 1, 2, \dots, 60$  months and the ex-ante annualized variance  $\sigma_t^2$  is calculated as follows:

$$\sigma_t^2 = 261 \sum_{i=0}^{\infty} (1 - \delta) \delta^i (r_{t-i} - \bar{r}_t)^2 \quad (5.40)$$

Here the weights  $\delta^i(1 - \delta)$  add up to one,  $\bar{r}_t$  is the exponentially weighted average return computed in a similar fashion. The exponential smoothing parameter is chosen to be  $\sum(1 - \delta)\delta^i = \frac{\delta}{1 - \delta} = 60$  days. Using  $\sigma_{t-1}$  in (??) and (??) avoids the look-ahead bias<sup>2</sup>.

We can also investigate the profitability of strategies based on time series momentum with varying look-back period ( $k$ ) and holding period ( $h$ ). If the excess return over the past  $k$  months is positive, go long or otherwise go short. The position size is taken to be,  $1/\sigma_{t-1}^s$ , inversely proportional to the ex-ante volatility of the instruments, ' $s$ '. With this choice it is argued that it will be easier to aggregate the outcome of the strategies across instruments with different volatilities and it will also lead to

---

<sup>2</sup>WE PROBABLY NEED A FOOTNOTE HERE TO ENSURE THE READER UNDERSTANDS

somewhat stable time series. For each trading strategy  $(k, h)$ , a single time series of returns,  $r_t^{\text{TSMOM}(k,h)}$ , which is the average return across all portfolios active at time, ‘ $t$ ’ is computed. To evaluate the overall value of the time series momentum strategies, the returns are adjusted for possible risk factors:

$$\begin{aligned} r_t^{\text{TSMOM}(k,h)} &= \alpha + \beta_1 \text{MKT}_t + \beta_2 \text{BOND}_t + \beta_3 \text{GSCI}_t + \beta_4 \text{SMB}_t \\ &\quad = +\beta_5 \text{HML}_t + \beta_6 \text{UMD}_t + a_t \end{aligned} \quad (5.41)$$

where  $\text{MKT}_t$  is the excess return on the MSCI World index,  $\text{BOND}_t$  is Barclay’s Aggregate Bond Index,  $\text{GSCI}_t$  is S&P GSCI commodity index and the others are Fama-French factors for the size, value and cross-sectional momentum premiums.

Using monthly data spanning January 1985 to December 2009, Moskowitz et al (2012) [214] find that the time series momentum strategies provide additional returns over and above a passive buy and hold strategy. They also show that the model (??) results in a large and significant alpha. The analysis also illustrates that speculators are likely to benefit from following trends at the expense of hedgers. The time series momentum appears to last about a year and then begins to reverse.

The momentum strategies remain profitable even after adjusting for price impact induced by trading. Although the trading decisions are done at a macro level, the actual trading takes place at a micro level. The price impact (see Chapter 8) is the impact of the current trade on the subsequent trades. Korajczyk and Sadka (2004) [180] demonstrate that certain  $(k, h)$  strategies are profitable if the proportional costs equal the effective and quoted spreads.

The role of trading volume in explaining the results of momentum strategies is studied in Lee and Swaminathan (2000) [188]. They find that, in equilibrium, both price and volume are simultaneously determined. In addition to the decile classification based on the past returns, the universe of stocks was further divided into three volume size categories, yielding thirty groups. The study provides evidence that low volume stocks tend to be under-valued by the market. Thus investor expectations affect not only a stock return but also its trading volume. High volume firms generally earn lower future returns. The general notion that volume fuels momentum is not supported but is true for losers and “it helps information “diffusion” only for winners.”

It is worth mentioning that while most documented applications of the momentum strategy are in equity markets, there is a growing academic interest in their use in foreign exchange (FX) markets. The FX markets were traditionally cited as markets where technical rules such as trend following have been successful. But studies such as Menkhoff, Sarno, Schemeling and Schrimpf (2012) [206] have explored the cross-sectional momentum and have demonstrated their profitability (up to 10%) in FX markets. The momentum portfolios in FX markets are observed to be skewed toward minor currencies that have somewhat higher transaction costs. The base currency returns,  $r_t = \ln(S_t) - \ln(S_{t-1})$  where  $S_t$  is the spot exchange rate, and the interest-

adjusted return, where  $F_{t-1}$  is the future price at  $t-1$ ,

$$\begin{aligned} r_t^* &= \ln(S_t) - \ln(F_{t-1}) \\ &= \ln(S_t) - \ln(S_{t-1}) + \ln(S_{t-1}) - \ln(F_{t-1}) \\ &\sim r_t + \frac{(r_f - r_d)}{12} \end{aligned} \quad (5.42)$$

which is the sum of pure currency appreciations and the return due to interest rate appreciation. Here  $r_f$  is the foreign interest rate and  $r_d$  is the domestic interest rate; See Okunev and White (2003) [217] for the details of this decomposition. It is found that the profitability is clearly originating from spot rate changes themselves and not necessarily driven by the interest rate differential.

In the equity markets, it is argued that momentum strategies are successful because of slow information processing and investor overreaction. However FX markets are dominated by professionals and thus any information should be impounded in the FX rate almost instantaneously. But if the portfolios are mostly made up with minor currencies, they are likely to exhibit higher volatility, higher country risk and higher interest-rate changes, thus imposing effective limits to arbitrage. It is further observed that the cross-section based currency momentum is not related to benchmark technical trading strategies and even after adjusting for them, the cross-sectional momentum profits are found to be higher.

## 5.9 Multiple Indicators and Boosting Methods

x In developing models that capture the main features of the data, it must be noted that there are no unique models. Thus they yield different forecasts. It is well-known in the forecasting literature that a combined forecast generally does perform better than the individual forecasts. If there are ' $T$ ' models and thus ' $T$ ' forecasts,  $\hat{f}_1, \dots, \hat{f}_T$ , how do we combine them to get a single forecast? Because these are all based on the same data, the estimates are likely to be correlated, resulting in a covariance matrix,  $\hat{\Sigma}_f$ . Three estimates that are proposed in the literature are: Due to the uncertainty

Table 5.1: Three proposed estimates

<b>Weighted Estimator 1:</b>	$\hat{f}_{w_1} = \hat{f}' \hat{\Sigma}_f^{-1} \hat{f}$	(??)
<b>Weighted Estimator 2:</b>	$\hat{f}_{w_2} = \hat{f}' (\text{diag} \hat{\Sigma}_f)^{-1} \hat{f}$	(??)
<b>Simple Average Estimator:</b>	$\sum_{i=1}^T \frac{\hat{f}_i}{T}$	(??)

in the estimation of  $\Sigma_f$  in regime changes that results in large estimation error, the simple average estimator is advocated. In the context of modeling the asset prices the elements of  $\Sigma_f$  represent how well the methods do in forecasting. In order to use the

above weighted estimators we need to keep track of the performance of the individual forecasting methods and in principle, the methods that yield inferior forecasts would get less attention. But a question about the span of their performance, recent versus past, remains open.

### Ensemble Learning

As mentioned, the idea of combining multiple forecasts (classifiers) into a single one is a longstanding goal of many decision-makers. This is especially important when methods that are successful at different regimes are complementary. There are two types of ensemble methods, bagging and boosting that we discuss here. Suppose the goal is to determine if we want to enter the market or not based on the information available at that time. It is somewhat easier to come up with simple rules but is harder to find a rule that has high accuracy. Under boosting, we study how to combine the simple rules uniformly into a ‘stronger’ rule. It is suggested to focus on examples that are hardest to forecast by simple rules and take a weighted majority of these rules to come up with a stronger rule.

To motivate, we will assume that we want to determine whether the price of an asset will go up or down. Under the random walk model, the error in guessing randomly is 0.5. But any rule should result in an error less than 0.5. We begin with a training set,  $S = \{(x_i, y_i), i = 1, 2, \dots, n\}$  where  $x_i$  is the information and  $y_i$  is a binary variable that indicates whether to enter the market or not. Assume that there are ‘ $T$ ’ rules,  $h_t(x)$  are the classifiers and  $D_t$  is the distribution over the training set with the error,

$$\text{err}_D(h) = P_{(x,y) \sim D}(h(x) \neq y) \quad (5.43)$$

The rule  $h_t$  is supposed to have error  $\text{err}_{D_t}(h_t)$  with respect to ‘ $D_t$ ’.

The boosting algorithm works as follows:

---

<b>Initialize:</b>	$D_1(i) = \frac{1}{n}$ for all $i$
<b>Iterate:</b>	$D_{t+1}(i) = \frac{D_t(i)}{z_t} \cdot \exp[-\alpha_t y_i h_t(x_i)]$ when $\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \text{err}_{D_t}(h_t)}{\text{err}_{D_t}(h_t)} \right)$ , $z_t$ = normalizing constant
<b>Stopping Time:</b>	Stop when the error in (1) on a validation data set does not improve
<b>Final Rule:</b>	$\text{Sign}(\sum_{t=1}^T \alpha_t h_t(x))$

---

It is possible that overfitting can happen with boosting; its performance clearly depends upon how well the two decisions (enter or not enter) are clearly separated based on the past data. The use of kernels can improve local separability. It is also important to optimally extract features,  $x$ , that can be discriminating between the two options for  $y$ . The main advantage of ensembles of different rules is that it is not likely that all rules will make the same error. The algorithm described in (33) and (34) is known as, “adaptive boosting algorithm (AdBoost)” is by Freund and Schapire (1995,

1996,1997) [117] [118] [119] and it automatically “adapts” to the data at hand. If  $t^{th}$  classifier  $h_t(x)$  is parameterized and the overall predictor takes the form,

$$h(x) = \text{Sign} \left( \sum_{t=1}^T \alpha_t \text{Sign}(w^{(t)}x + b^{(t)}) \right) \quad (5.44)$$

is a two-layer neural network with ‘ $T$ ’ hidden units. One can see the identical underlying nature of the structure between neural network and boosting algorithm.

The idea of bagging comes from training the rules on multiple data sets that supposedly have similar structure. The multiple data sets are obtained by bootstrap resampling. The bootstrap data sets are not likely to be too similar. For the data on asset prices, which are represented by an AR(1) model with the coefficient close to one, the asymptotic distribution of the sample estimate of the coefficient is non-normal. So the bootstrap samples are also useful to construct the appropriate sampling variance. The bootstrap samples are created by resampling the innovations or empirical residuals. From the single data set,  $S$  we create ‘ $M$ ’ bootstrapped samples,  $S_1, \dots, S_M$  with each set containing ‘ $n$ ’ samples. Then apply the boosting method on all these samples and average them. As noted in the literature, bagging tends to reduce variance and thus it provides an alternative to regularization, see Breiman (1996) [43].

The gradient boosting method introduced by Breiman (1999) [44] and generalized by Friedman et al. (2000) [121] and Friedman (2001) [120] further improves on the ADA boost. The method starts by fitting an additive model (ensemble),  $\sum_t \alpha_t h_t(x)$  in a forward stage-wise manner. In each stage, a weak learner is introduced to compensate for the shortcomings of existing weak learners. In gradient boosting, gradients are used to identify the shortcomings and in the regression context, these are residuals. Now the training set,  $S^* = \{(x_i, y_i - h(x_i)), i = 1, \dots, n\}$  is treated like,  $S$ , with residuals playing the role of the data. The optimization occurs by moving in the opposite direction of the gradient of the function.

## 5.10 Evaluation of Strategies: Various Measures

It is important to develop measures to evaluate the trading strategies. The standard measures stem from the capital asset pricing model:

$$r_t - r_f = \alpha + \beta(r_{mt} - r_f) + \epsilon_t \quad (5.45)$$

where  $r_t$  is the return on the asset and  $r_f$  is the return from a risk-free asset; the slope coefficient,  $\beta$  measures the riskiness of the asset and ‘ $\alpha$ ’ denotes the excess return. Thus a strategy can be simply evaluated by:

**Excess return:**  $\alpha$  with positive values of ‘ $\alpha$ ’ indicating better performance.

More popular measures are:

**Sharpe Ratio:**  $\frac{E(r_t) - r_f}{\sigma}$ : Standardized by the volatility of the asset.

**Treynor Ratio:**  $\frac{E(r_t) - r_f}{|\beta|}$ : Standardized by market risk.

with appropriate sample estimates substituted here. A modified version of excess return is known as,

**Information Ratio:**  $\frac{\alpha}{\sigma(\varepsilon_t)}$ : Standardized  $\alpha$

While the above ratios take a long-term view of the investments, two others that are used in industry focus on the execution aspects of the strategies:

**Sortino Ratio:**  $\frac{E(r_t) - r_{\text{target}}}{DR(r_t)}$ ,

where  $r_{\text{target}}$  is the target rate chosen by the investor and  $DR(r_t)$  denotes the downside volatility or the draw down. Another measure is,

**Calmar Ratio:**  $\frac{\text{Compound Annualized Rate of Return}}{|\text{Maximum Drawdown}|}$

The attention to drawdown is important because there is no guarantee that a strategy chosen through simple back-testing will work in the future. The last two ratios combine both the idea of maximizing the return and the idea of minimizing the loss. The drawdown numbers are essentially determined by the traders based on their own experience and their risk tolerance. There are some studies that focus on the statistical properties of their criteria (see Bertrand and Protopopescu (2010) [24] on the information ratio and A. Lo (2002) [197] on the Sharpe ratio). Another measure is to simply compare the strategy's returns to the returns from buy and hold.

The standard reporting of backtesting results include, number of winning and losing trades, largest winning and losing trades, net slippage, total commissions along with annualized versions of the above ratios.

## 5.11 Extraneous Signals: Trading Volume, Volatility, etc.

The trading algorithms that are presented in this chapter thus far focus on the price and return behavior of an equity. In pairs trading, we considered the behavior of price-return of a related equity to decide when to buy one and sell the other. But there is information as noted in Section 2.7 from the volume of trading. The model discussed there relates the volume to volatility. If there is no variation in the information flow about an equity, there should not be any correlation between volume traded and the return volatility. If the correlation exists we could indirectly infer that there is

information and we may want to exploit that for trading. There is a vast literature in finance studying the relationship between trading volume and serial correlation in returns. We briefly review some select studies and provide an illustrative example with some possible algorithms.

Blume, Easley and O'Hara (1994) [31] investigate the informational role of volume. If the impounding of information into stock price is not immediate, volume may provide information about the evolving process of security returns. The model discussed in Section 2.7 by Tauchen and Pitts (1983) [249] and the associated empirical tests clearly document strong relationship between volume and the absolute return. ‘But why such a pattern exists or even how volume evolves in markets is not clear.’ Recall that the model assumes that the information content affects both the return and the volume specifications; yet the correlation between the return and volume can be shown to zero. If the information content can be studied through the sequence of security prices and the associated volume, it may provide insights into the inefficiency in the market and how statistical arbitrage rules can exploit it especially if there is a lead-lag relationship. The quality of trader’s information can be captured best by combining price change with volume change. It is shown that the relationship between volume and price is in the form of V-shape thus indicating a non-linear relationship. Both ‘bad’ and ‘good’ information about the stock will result in higher volume of trade.

Campbell, Grossman and Wang (1993) [50] also explore the relationship between volume and returns; volume information is used to distinguish between price movements that occur due to publicly available information and those that reflect changes in expected returns. It is predicted that price changes with high volume tend to be reversed and this may not happen on days with low volume. Llorente, Michaely, Saar and Wang (2002) [195] extend this work by postulating that the market consists of liquidity and speculative traders. In periods of high volume but speculative trade, return autocorrelation tends to be positive and if it is liquidity trading return autocorrelation tends to be negative. In the former case, returns are less likely to become reverse. Defining the turnover ( $V$ ) as the ratio of number of shares traded to the total number of outstanding shares, the following regression model is fit for each stock:

$$r_t = \beta_0 + \beta_1 r_{t-1} + \beta_2 V_{t-1} \cdot r_{t-1} + a_t \quad (5.46)$$

For stocks associated with more speculative trading (larger bid-ask spreads),  $\beta_2$  is expected to be positive. The model can be extended to add a market proxy ( $r_{mi}$ ). Note the first order autocorrelation that we expect to hold different signs under different volume-return scenario are captured by ‘ $\beta_1$ ’ coefficient. The model’s predictions are tested for NYSE and AMEX stocks and the results are supportive of the predictions.

Gervais, Kaniel and Mingelgrin (2001) [132] show that stocks with unusually large trading activity over periods tends to experience large returns over the subsequent periods. It is argued that higher trading activity creates higher visibility and in turn creates subsequent demand for that stock. This approach where contrarian or momentum is characterized by trend in volume is also studied by Conrad, Hameed and Niden (1994) [65] and Lee and Swaminathan (2000) [188]. Assume that the change

in trading activity is measured by

$$u_{it} = (V_{it} - V_{it-1})/V_{it-1} \quad (5.47)$$

and a positive  $u_{it}$  implies higher transaction security and a negative  $u_{it}$  is taken to imply low transaction security. Form four portfolios with stocks that exhibit positive versus negative returns in conjunction with positive versus negative transaction activity. The weight for each of the form portfolios are given as:

$$w_{i,t}^* = \frac{r_{it-1}(1 + u_{it-1})}{\sum_{i=1}^{N_p} r_{it-1}(1 + u_{it-1})} \quad (5.48)$$

where  $N_p$  is the number of securities in each combination ( $p = 1, \dots, 4$ ). The empirical results confirm that autocorrelation of low and high transaction securities are different; high transaction stocks experience price reversals and low transaction stocks experience positive autocorrelations in returns. These relationships are noted to be stronger for smaller firms.

The discussion thus far refers to the use of low frequency data. Most of the studies deal with daily or weekly data that are aggregated. Simply using the end of the period closing price and the volume for the period is not appropriate as transactions occur at specified prices. The availability of high-frequency data has enabled to study the evolving influence of volume after controlling for order book imbalances. Because the parent order gets sliced into small child orders, both the number of trades and size of trades need to be considered in the place of volume. As argued in Kyle (1985) [182], the price volatility is induced by net order flow. Market makers tend to infer information from the net order flow and revise their prices. Using a random sample of approximately three hundred stocks in NYSE, Chan and Fong (2000) [55] consider the model in two steps for daily data:

$$\begin{aligned} r_{it} &= \sum_{k=1}^4 d_{ik} D_{kt} + \sum_{j=1}^{12} \beta_{ij} r_{it-j} + \epsilon_{it} \\ \epsilon_{it} &= \phi_{i0} + \phi_{im} M_t + \sum_{j=1}^{12} \phi_{ij} |\epsilon_{it-j}| + \gamma_i (V_i \text{ or } T_i) + \eta_{it} \end{aligned} \quad (5.49)$$

Here  $D$  is the day-of-the week indicator,  $M_t$  is an indicator for Monday,  $V$  and  $T$  are share volume and number of trades. The second equation that measures the volume-volatility relationship indicates that the number of trades ( $T$ ) has a better explanatory power than share volume ( $V$ ). The relationship between order book imbalance and volatility is captured by modifying the first equation in (5.49) as

$$r_{it} = \sum_{k=1}^5 \alpha_{ik} D_{kt} + \sum_{j=1}^{12} \beta_{ij} r_{it-j} + \sum_{h=1}^5 \lambda_{h,i} N T_{h,it} + \varepsilon_{it} \quad (5.50)$$

where  $N T_{h,it}$  is the net number of trades in size category ' $h$ '. The net number is

calculated by classifying the trade as buyer-initiated or seller-initiated using Lee and Ready (1992) [189] rule. It is shown that the coefficient ‘ $\lambda$ ’ is positive and significant for most stocks. But the second-stage regression results in weaker ‘ $\gamma$ ’ values.

Chordia, Roll and Subrahmanyam (2002) [62] extend this and provide a more comprehensive empirical analysis; resulting stylized facts are listed below:

- Order imbalances are strongly related to market returns; imbalances are high after market declines and low after market advances.
- Liquidity is not predictable from past imbalances but is predictable from market returns.
- Order imbalances are strongly related to volatility measures after controlling for volume and liquidity.

These stylized facts would be useful for spotting anomalies and hence for devising appropriate trading strategies.

### 5.11.1 Filter Rules based on return and volume

The rules that were discussed in Section 3.2 can all be modified to include volume as a weighing factor for the returns. That would involve tracking both the price and volume changes over time; it gets particularly cumbersome to track the volume that is appropriately normalized for the time of the day. Therefore the rules focus on defining and tracking the changes in the two quantities individually and combine them with necessary thresholds. Cooper (1999) [69] suggests four strategies based on weekly returns and another four strategies that incorporate the return (thus price) and the volume information. First the return rules are for weeks ‘ $t - 1$ ’ and ‘ $t - 2$ ’:

First-Order Price Filters:

$$\text{Return states} = \begin{cases} \text{For } k = 0, 1, \dots, 4 : & \begin{cases} \text{loser}_{k^*A} & \text{if } -k^*A > r_{i,t-1} \geq -(k+1)^*A \\ \text{winner}_{k^*A} & \text{if } k^*A \leq r_{i,t-1} < (k+1)^*A \end{cases} \\ \text{For } k = 5 : & \begin{cases} \text{loser}_{k^*A} & \text{if } r_{i,t-1} < -k^*A \\ \text{winner}_{k^*A} & \text{if } r_{i,t-1} \geq k^*A \end{cases} \end{cases}$$

Second-Order Price Filters:

$$\text{Return states} = \begin{cases} \text{For } k = 0, 1, \dots, 4 : & \begin{cases} \text{loser}_{k^*A} & \text{if } -k^*A > r_{i,t-1} \geq -(k+1)^*A \\ \text{and} & -k^*A > r_{i,t-2} \geq -(k+1)^*A \\ \text{winner}_{k^*A} & \text{if } k^*A \leq r_{i,t-1} < (k+1)^*A \\ \text{and} & k^*A \leq r_{i,t-2} < (k+1)^*A \end{cases} \\ \text{For } k = 5 : & \begin{cases} \text{loser}_{k^*A} & \text{if } r_{i,t-1} < -k^*A \\ \text{and} & r_{i,t-2} < -k^*A \\ \text{winner}_{k^*A} & \text{if } r_{i,t-1} \geq k^*A \\ \text{and} & r_{i,t-2} \geq k^*A \end{cases} \end{cases} \quad (5.51)$$

Here  $r_{i,t}$  is the non-market-adjusted return for security ‘ $i$ ’ in week ‘ $t$ ’,  $k$  is the filter counter,  $k = 0, 1, 2, \dots, 5$  and  $A$  is the lagged return grid width, set equal to 2%.

To relate if return reversals are related to trading volume, define the growth volume as the percentage changes in volume adjusted for the outstanding shares:

$$\% \Delta v_{i,t} = \left[ \frac{V_{i,t}}{S_{i,t}} - \frac{V_{i,t-1}}{S_{i,t-1}} \right] / \left[ \frac{V_{i,t-1}}{S_{i,t-1}} \right] \quad (5.52)$$

Here  $S_{i,t}$  is the number of outstanding shares. Next, we define filters based on this:

Volume Filters:

$$\text{Growth in volume states} = \begin{cases} \text{For } k = 0, 1, \dots, 4 : & \begin{cases} \text{low}_{k^*B} \text{ if } -k^*B > \% \Delta v_{i,t-1} \\ \text{high}_{k^*C} \text{ if } k^*C \leq \% \Delta v_{i,t-1} \end{cases} \\ \text{For } k = 5 : & \begin{cases} \text{low}_{k^*B} \text{ if } \% \Delta v_{i,t-1} < -k^*B \\ \text{high}_{k^*C} \text{ if } \% \Delta v_{i,t-1} \geq k^*C \end{cases} \end{cases} \quad (5.53)$$

Here  $B$  is the grid width for low growth in volume and is set to fifteen percent and  $C$  is the grid width for high growth in volume and is set to fifty percent. The asymmetry in the limits is due to skewness in the volume distribution. The value of ‘ $k^*$ ’ can be calibrated through backtesting.

Using the data for the large cap stocks, Cooper (1999) [69] examines the profitability of the various filters based on the portfolios that meet the conditions of the filters. Some general conclusions are listed below:

- Winner price-high volume strategy performs the best.
- Loser portfolios that condition on two consecutive week losses tend to do better than the strategies based on one-week losses.
- Loser price, low volume strategy results in higher profit than the strategy based on loser-price only.

### 5.11.2 An illustrative example

The methodology discussed in the last section on price-based and volume-based filters on low frequency data can be extended to high frequency data. The key issue with the volume data is to come up with proper standardization to identify abnormal level of activity in terms of volume traded. The intra-day seasonality and the overall increase or decrease in inter-day volumes need to be tracked carefully to fully exploit the deviations. We highlight the practical issues involved using 30-min price-based data for Treasury yields from June 8, 2006 to August 29, 2013. This can be taken as medium-frequency data and the price bars contrast of high, low, open and close prices along with the volume traded during the interval. Because of the time aggregation, the traditional time series methods can be used to study the trend in the price and volume series. Daily data is just a snapshot at 4 P.M. and the total volume for the day is the sum of all 30-min volumes since the previous day.

Let  $P_{t,m}$  be the price on a  $t$  in  $m$ th time unit interval. Here we take  $m = 30$  min. The returns,  $r_{t,m} = p_{t,m} - p_{t,m-1}$ , where  $p_{t,m} = \log(P_{t,m})$ . Similarly, let  $V_{t,m}$  be the volume in the interval and let  $v_{t,m} = \log(V_{t,m})$ . We define volatility within the  $m$ th time unit based on the price bars data as,

$$\hat{\sigma}_{t,m}^2 = 0.5[\ln(H_{t,m}) - \ln(L_{t,m})]^2 - 0.386[\ln(C_{t,m} - \ln(O_{t,m})]^2 \quad (5.54)$$

similar to a measure defined for the daily data.

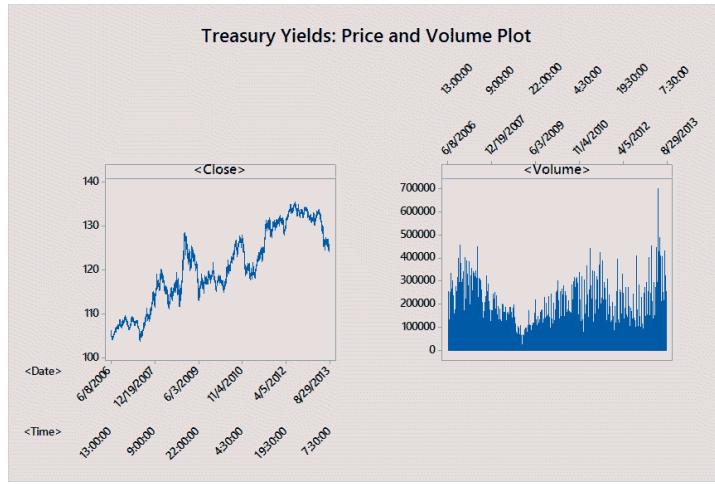


Figure 5.5: Treasury Yields: Price and Volume Plot

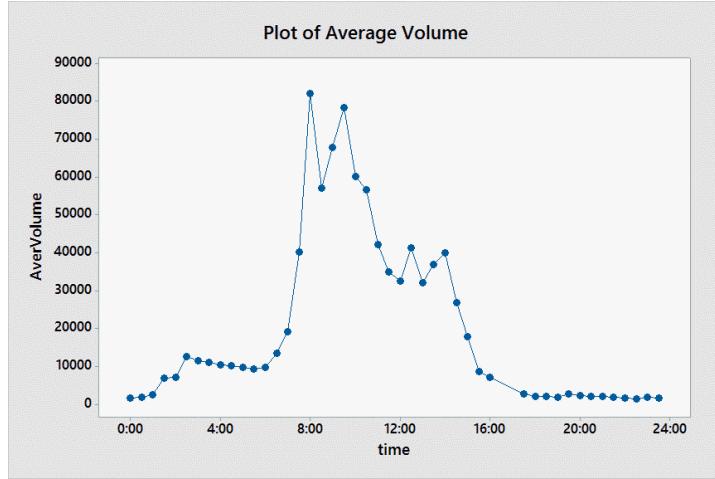


Figure 5.6: Plot of average volume.

The plot of price and volume data is given in Figure 5.5. The price data generally

shows an upward trend, but it is difficult to see the pattern in volume, especially during the course of the day. A simple average volume plot clearly indicates most of the trading occurs between 8 A.M. and 4 P.M. (Figure 5.6). The autocorrelation function of the log volume ( $v_{tm}$ ) clearly demonstrates the intra-day variations in the data. There is a great deal of interest among practitioners to model the volume accurately as a commonly used trading strategy, VWAP assumes that volume is known apriori. Two other benefits cited in Satish, Saxena and Palmer (2014) [235] are if a trader receives a large order ten minutes before the close with an instruction to complete the order, having a reliable estimate of expected volume can help evaluate the feasibility of execution. The other benefit that is mentioned is how improved volume forecast can aid alpha capture. Brownlees, Cipollini and Gallo (2011) [48] propose a dynamic model,

$$V_{tm} = \eta_t \cdot \phi_m \cdot \mu_{t,m} \cdot \epsilon_{t,m} \quad (5.55)$$

where  $\eta_t$  is a daily component,  $\phi_m$  is the periodic component and  $\mu_{t,m}$ , an intra-daily dynamic, non-periodic component. The periodic component  $\phi_m$  modeled as a Fourier (sine/cosine) function; the daily component is modeled as a function of past values of a daily volumes and the same is true with  $\mu_{t,m}$ ; they are modeled as a function of  $\mu_{t,m-1}$  and  $V_{t,m-1}$ . From the structure of the model (5.55), it is obvious that the multiplication components can be modeled additively through log transformations.

We suggest somewhat similar to (5.55) but more directly related to models discussed in Chapter 2. Before we state the results recall the main findings from past empirical studies:

- Volatility, measured by returns squared is positively correlated with volume.
- Large price movements are followed by high volume.
- Conditioning on lagged volume attenuates the leverage effect.
- After conditioning on lagged volume, there is positive risk-return relationship

Our approach to modeling is stated as follows:

- Model the volume at different frequency levels.
- Use the low frequency data for developing macro strategies.
- Use the high frequency data for making micro decisions.

Given the strong intra-day periodicity that is evident from Figure 5.6, we can simply create indicator variables for the 30-min periods and regress the volume on the indicators to capture the time of the day variation. This explains almost sixty-five percent of the variation in log volume. The residuals capture non-seasonal patterns and the partial autocorrelation function is given in Figure 5.8. This clearly indicates that the residuals can be modeled as higher order autoregressive process. As these components are additive, the forecasts from indicator regression and this autoregression can be summed up to arrive at a reasonable volume forecast. The plot of logged daily volume is given in Figure 5.9 and it is clearly stationary. The partial autocorrelation

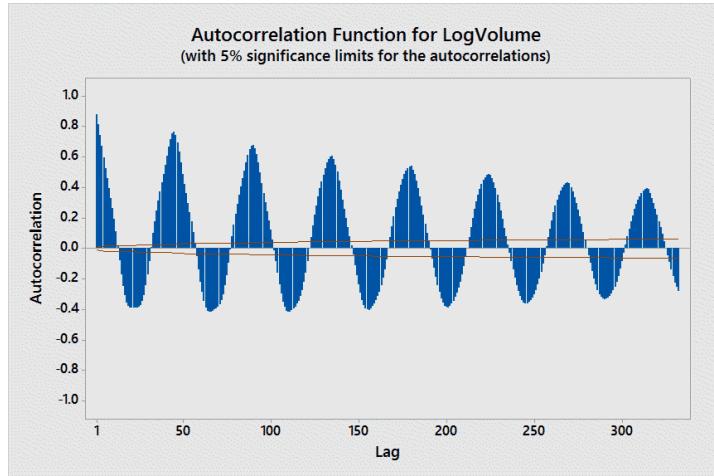


Figure 5.7: Autocorrelation function for log volume.

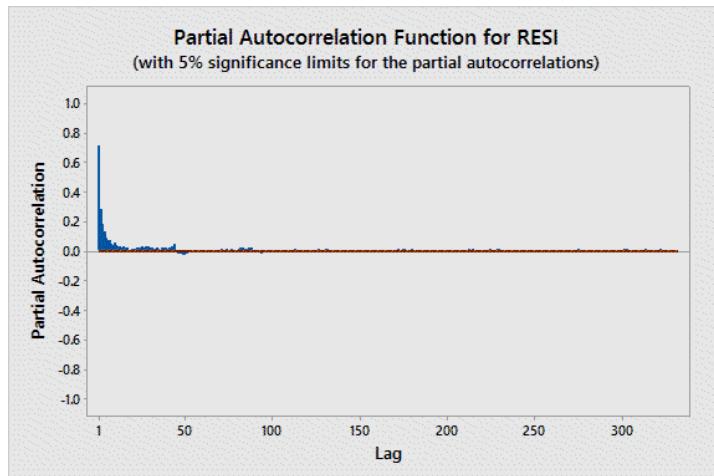


Figure 5.8: Partial autocorrelation function for RESI (with 5% significance limits for the partial autocorrelations)

function in Figure 5.10 has somewhat a simpler structure for the daily aggregated data. How do we use these results?

- Forecast next day volume using the macro model.
- Forecast the 30-min volume using the micro model.
- If the actual volume is not in line with the forecast, search if any new information has arrived!

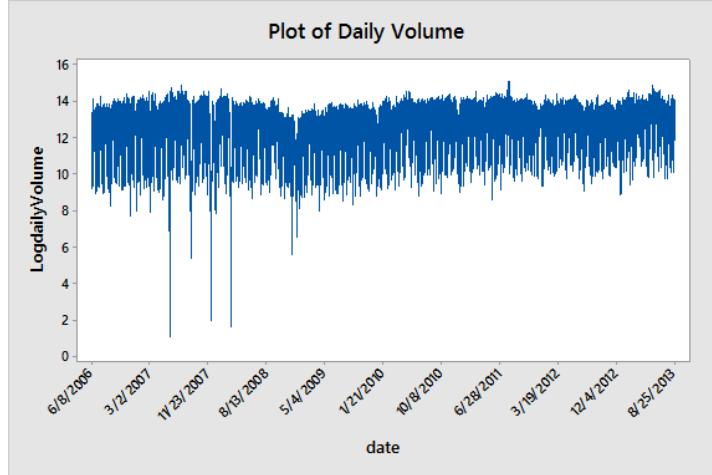


Figure 5.9: Plot of daily volume.

These methods obviously require the use of time series programs and as an alternative, we suggest using exponential smoothing approach. The details are as follows:

- Volume Forecast 1 (Direct): Use daily data to forecast  $\hat{v}_{t+1}$ , next day's volume; use exponential smoothing constant in the range of 0.25–0.35.
- Volume Forecast 2 (Aggregated): We will use a two-step procedure here:
  - (a) Compute the average volume  $\bar{v}$  over certain number of days for each ' $m$ '. Usually industry uses '22' days.
  - (b) Compute the residuals:  $e_{t-m} = v_{t,m} - \bar{v}$ ; use exponential smoothing on  $e_{t-m}$  with a smoothing constant in the range of 0.25–0.35; estimate  $\hat{e}_{t+1,m}$ .
  - (c) Forecast for next period ' $m\hat{v}_{t+1,m} = \bar{v} + \hat{e}_{t+1,m}$
  - (d) Aggregate  $\hat{v}_{t+1,m}$  over  $m$  to get  $\hat{v}_{t+1}$ .

Now we have two forecasts of daily volume  $\hat{v}_{t+1}$  and  $\hat{v}_{t+1}$ .

We want to relate the volume to volatility using the measure in (5.54). Figure 5.11, based on the plot of (5.22) indicates that there are clustering effects. From the partial autocorrelation function given in Figure 5.12, there is some memory in the variance and can be modeled via GARCH. Here our goal is to show how volume and volatility are likely to be related. The simple cross-correlation function (Figure 5.13) shows how they are closely related; here the intra-day periodicity in the residuals in volume and volatility is shown. Again in the spirit of intuitive approach to model the volatility, we suggest obtaining volatility forecasts as:

$$\begin{cases} \hat{\sigma}_{t+1}^2 = \text{weighted average of } \sigma_t^2, \sigma_{t-1}^2, \dots, \sigma_{t-22}^2 \\ \hat{\sigma}_{t+1,m}^2 = \text{weighted average of } \sigma_{t,m}^2, \sigma_{t-1,m}^2, \dots, \sigma_{t-22,m}^2 \end{cases}$$

The following algorithm that combines the price, volume and volatility is suggested; we assume the price, volume and volatility is suggested; we assume that price-based rates given in Section 3.2 are followed.

- Enter if (a) price-based rules favours and if (b)  $\sigma_{t+1,m}^2 < \hat{\sigma}_{t+1,m}^2$ , realized volatility < predicted volatility and if (c)  $\sum_{i=1}^m v_{t+1} < \sum_{i=1}^m \hat{v}_{t+1}$ , cumulative actual volume < cumulative predicted volume.
- Exit if either (a) or (b) or (c) is not true.

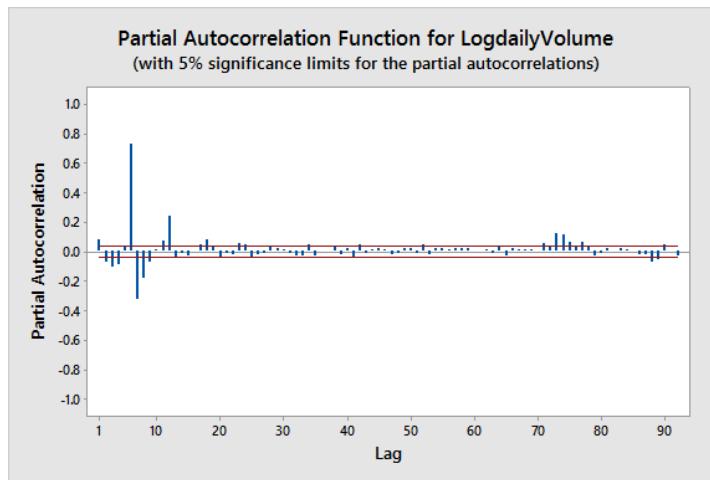


Figure 5.10: Partial autocorrelation function for log daily volume (with 5% significance limits for the partial autocorrelation).

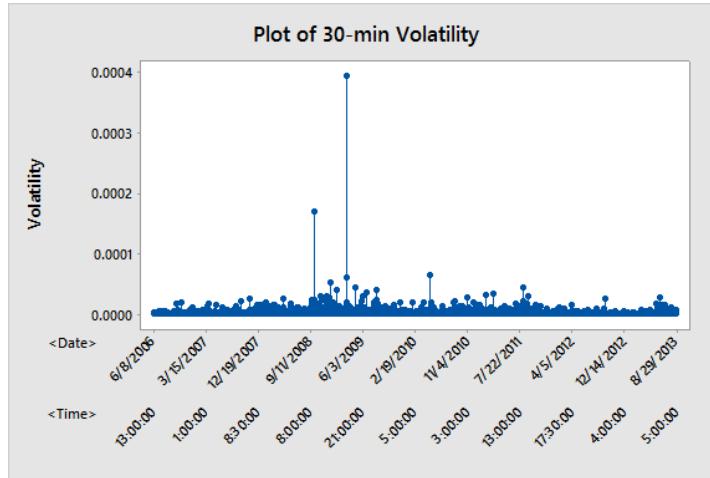


Figure 5.11: Plot of 30-min volatility.

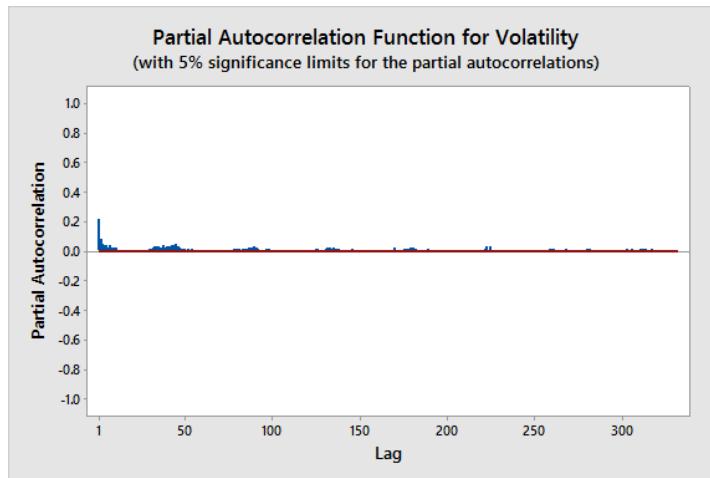


Figure 5.12: Partial autocorrelation function for volatility (with 5% significance limits for the partial autocorrelations).

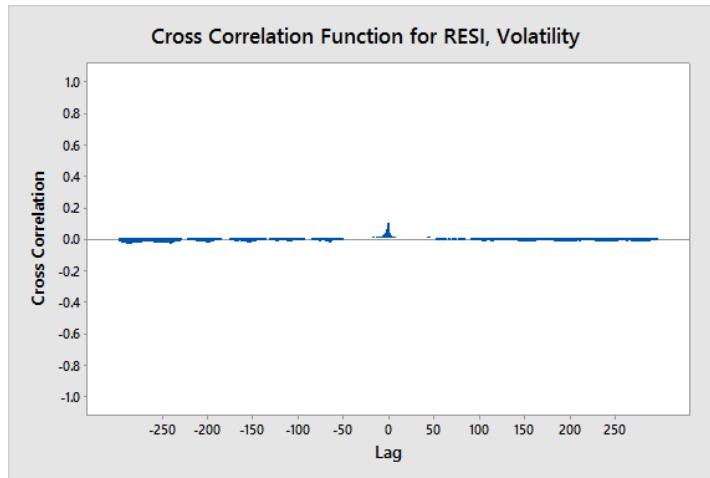


Figure 5.13: Cross correlation function for RESI, volatility.

---

## 5.12 Market Segmentation and Inconsistent Pricing across Markets

---

### 5.13 Supplements and Problems

- One of the most intriguing asset pricing anomalies in finance is so-called “price momentum effect” that was covered in this chapter. This exercise using CRSP data is meant to guide you through creating a momentum strategy. The goal is to replicate the price momentum strategy described in Jegadeesh and Titman (2001) [169] using more recent data extending to December 2014. [This problem is based on the notes from Ravi Jagannathan.]

#### A. Source of Data

The monthly stock returns from January 1965 to December 2014 can be obtained from the web. You should construct the decile ranking by yourself. For future reference, the decile ranking file is available from Ken French’s web site. The NYSE market capitalization decile contains all common shares regardless of the end of the month price.<sup>3</sup>

---

<sup>3</sup>The Fama-French factors can be downloaded from Ken French’s website. Note that you should only use the data from “U.S. Research Returns Data” section.

You should use each stock's permanent identification code (PERMNO) as a unique identifier for that stock.

Note that not all stocks have valid monthly returns for various reasons. One reason is that the stock gets delisted somewhere during the month. For the construction of your momentum portfolios, you can simply ignore the delisting returns in your portfolio construction and return accumulations for the sake of simplicity. More specifically, for this exercise, if the end of the month return of any stock is missing, you may simply assume the return is zero in your compounding. You should keep track of such cases (identify the reason for delisting/missing observation) and prepare a table summarizing how many such cases occurred.

## B. Stocks Selection Criterion

Consider all the common stocks traded on AMEX, NYSE and NASDAQ but subject to the following filtering rules:

- Exclude all stocks priced lower than \$5 at the end of month ( $t$ ).
- Exclude all stocks belonging to the smallest NYSE market decile at the end of month ( $t$ ). You should first find the cutoff for the NYSE smallest market decile, then delete all the stocks (including stocks on AMEX, NYSE, and NASDAQ) with a market cap that's below the cutoff.
- Another thing to take in consideration is that when you form your portfolios, you should only apply the filters to the stocks when you form the portfolio (i.e. at the end of month  $t$ ). The future returns (stock returns at month  $t + 2$  to  $t + 7$ ) should not be filtered. Otherwise it creates a look-ahead bias. For example, for portfolio returns at month  $t + 2$  to  $t + 7$ , you drop stocks with a price <5 at any month between  $t + 2$  to  $t + 7$ . First, when forming the portfolio at time  $t$ , you use information you don't have at month  $t$ . Second, you eliminate stocks with likely poor performance (low price) in the future. The portfolio returns (especially the loser portfolio returns) are likely to be biased upward from this action.

Again, you should keep track of such cases and prepare a table summarizing how many such cases occurred..

## C. Portfolio Formation

At the end of each month ( $t$ ), sort the stocks into 10 decile portfolios based on the six month cumulative returns between month  $(t - 5)$  to  $(t)$ . The first decile contains the stocks with the lowest past six-month returns, and the tenth decile contains the stocks with the highest past six-month returns.

- Start by accumulating the return of each decile portfolio between  $(t + 2)$  to  $(t + 7)$ , i.e. skipping the first month following the portfolio formation.

- At the beginning of each month, construct overlapping decile portfolios as equally weighted averages of the past six month months' decile portfolios. That is, at each month  $\tau$ , the  $i$ th overlapping decile portfolio consists of an equal weighted average of the  $i$ th decile portfolios of the previous six month  $(\tau - 5)$  to  $(\tau)$ .<sup>4</sup>
- For each month and of the 10 overlapping decile portfolios, record the portfolio returns between January 1995 and December 2014.

#### D. Questions of Interests

After constructing the relevant portfolios, answer the following questions. When you answer these questions, note any additional assumptions/choices you made.

#### E. Momentum strategy profits.

- Report the monthly average returns of your overlapping decile portfolio, as well as the winner ( $P_1$ , highest past six month return portfolios) minus losers ( $P_{10}$ , lowest past six month return portfolios).
- Report the monthly average returns of your overlapping decile portfolio for January and non-January months respectively, as well as the winner ( $P_1$ , highest past six month return portfolios) minus losers ( $P_{10}$ , lowest past six month return portfolios).
- Using CAPM and Fama-French version of multifactor asset pricing model to test the spreads between winner and loser portfolio ( $P_{10} - P_1$ ). The Fama-French factors can be obtained from Kent French's web site.
- Tabulate these results in a table similar to table I in Jegadeesh and Titman (2001).
- Compared your results to the results in Jegadeesh and Titman (2001) *and comment on your findings*.

#### F. Momentum strategy and market capitalization.

- Repeat the above exercise but split the sample into small-cap and big-cap portfolio, where the small-cap portfolio contains the stocks in the lower five NYSE market capitalization deciles.

---

<sup>4</sup>These 10 overlapping decile portfolios are the testing assets in our asset pricing test exercises.

- (b) Tabulate these results in a table similar to table *I* in Jegadeesh and Titman (2001).
  - (c) Compare the simple momentum strategy's profits, and momentum strategy cut by the market capitalization profits to value-growth strategy's return. The value-growth strategy return can be proxied by the HML factor obtained from Ken French's website.
  - (d) *Comment on your findings.* In particular, briefly suggest at least three reasons why momentum profits may be related to market capitalization, and how would you go ahead and test it further.
2. One of the commonly used trading strategies in finance the so-called “pairs trading” is covered in this chapter. This is an exercise to replicate the strategy. Obtain the daily data from January 2000 to June 2014 on stock prices, returns, trading volume and shares outstanding on the following companies only:

Oil stocks: HAL, SLB, EQT, OXY, CVX

Tech stocks: AAPL, CSCO, GOOG, INTL, MSFT

Bank stocks: BAC, JPM, WFC, C, STI

Retail stocks: ANF, GES, SKS, SHLD, URBN

Begin forming pairs in January 2000 based on the estimation period, January 2000 to December 2000. The last estimation period is January 2014 to December 2014. Eligible pairs corresponding to the last estimation remain eligible till December 2014. Pairs that opened in December 2013 and did not converge with the stipulated maximum of six months would have been closed in June 2014.

As you recall, the pairs trading strategy works as follows:

- (a) Within each sector, match pairs based on normalized price differences over the past year.
- (b) When a tracked pair diverges by more than two standard deviations of their normalized price difference, buy the “cheap” stock and sell the “expensive” one after a one-day wait period (to mitigate the effects of any market irregularities).
- (c) If prices for a traded pair re-converge, exit the paired positions.
- (d) If prices for a traded pair do not re-converge within six months, close the positions. An alternative strategy closes all traded pairs in ten trading days.

Questions of interest:

- a. Quantify the profitability of the pairs trading strategy and on average, how long it takes for pairs to re-converge?
- b. Can you provide some insight into why the strategy works based on trading volume, shares outstanding and the market trend?

Note: As minimum, you are supposed to produce figures similar to figures 1-5 of the paper as well as Table 1

Another Question of interest

- c. Is there a pair across sectors that can outperform the pairs formed within each sector?

3. The file Exchange rates.csv contains exchange rates between US\$ and 25 major currencies; the daily data spans from Jan 3, 2000 to April 10, 2012. Consider the data from Jan 3, 2002 for the following currencies: EUR, GBP, DKK, JPY and CNY. Compute  $r_t = \ln p_t - \ln p_{t-1}$ , returns for each series.

**A(a)** Estimate the mean vector and the covariance matrix of the returns.

- (b) Construct a portfolio with Markowitz's minimum variance objective with expected returns at least R, the median of the elements in the mean vector. The weights must be positive and should sum to 1.
- (c) Repeat (a) and (b) for each year and comment on the variation in the composition of the portfolios.

**B. Evaluation of Trading Rule:** Consider the data starting June 11, 2003.

- (a) Compute the average return for each day. We will treat this a 'market' return.
- (b) Consider the two currencies: EURO and JPY and assume that your investment is restricted to only these two currencies. We will follow the simple algorithm: if

$$\frac{r_i - \bar{r}}{sd(r_i)} = \begin{cases} > 2 & sell \\ < -2 & buy \\ o.w. & hold \end{cases}$$

The average and standard deviation of the returns are based on last 30-day observations. Evaluate this algorithm by comparing it with simple holding.

- (c) What is the value of the 'market' return? How could this information be used?

**C. Trading Rule with other factors:** In addition to exchange rates from June 11, 2003, consider the data on commodity prices: Gold, Oil, Copper and a commodity index. Among the following currencies, CAD, BRL, ZAR, AUD, NZD, CAF and

RUB, which are most influenced by the commodity prices. Discuss how you would incorporate this into a trading strategy.

4. This homework consists of an exercise on ‘pairs trading’ using CRSP data which will guide you through creating that strategy. The goal is to replicate the strategy as described in Engelberg, Gao and Jagannathan (2009) [96]. Focus only on sections 1,2 and 4 of the paper. Obtain the daily data from January 1992 to June 2010 on stock prices, returns, trading volume and shares outstanding on the following companies only:

DELL, HP, ADBE, MSFT, AAPL  
JPM, WFC, BAC, C, BRK.A  
JNJ, WMT, COST, KO, PEP  
GM, TM, F  
XOM, COP

“Begin forming pairs in January 1993 based on the estimation period, January 1992 to December 1992. The last estimation period is January 2009 to December 2009. Eligible pairs corresponding to the last estimation remain eligible till December 2009. Pairs that opened in December 2009 and did not converge with the stipulated maximum of six months would have been closed in June 2010.”

As you recall, the pairs trading strategy works as follows:

- (a) Match pairs based on normalized price differences over the past year.
- (b) When a tracked pair diverges by more than two standard deviations of their normalized price difference, buy the “cheap” stock and sell the “expensive” one after a one-day wait period (to mitigate the effects of any market irregularities)
- (c) If prices for a traded pair re-converge, exit the paired positions.
- (d) If prices for a traded pair do not reconverge within six months, close the positions. An alternative strategy closes all traded pairs in ten trading days.

Question of interest:

1. Quantify the profitability of the pairs trading strategy and on average, how long it takes for two pairs to reconverge?
2. Can you provide some insights into why the strategy works based on trading volume, shares outstanding and the market trend?

Note: as a minimum, you are supposed to produce figures similar to figure 1-5 of the paper as well as table 1.

#### 5. Technical Rules

Consider the data on exchange rates between the Indian Rupee and the US Dollar,

British Pound and Euro. In order to compare the performance of a strategy, appropriately normalize the entry point exchange rate: invest equal amounts of capital (counted in INR, which we will call the “base currency”) in each trade. Total portfolio return on each day should be computed as a weighted sum of returns from each active trade (weighted by the current value of each trade). Evaluate the signals discussed in (a) and (b) below using the Sharpe ratios and PNL with and without including transaction costs of 15 basis points. Assume a risk free rate of 3%. (1 basis point =  $10^{-4}$ ; so each dollar traded costs 0.0015 in transaction costs)

(a) *Moving average trading rules:* Trading signals are generated from the relative levels of the price series and a moving average of past prices.

- Rule 1: Let  $\bar{p}_t^{(m)}$  as defined in (4); use the strategy to sell if  $p_t > \bar{p}_t^{(m)}$  and to buy if  $p_t < \bar{p}_t^{(m)}$ . Evaluate this strategy. What is the optimal “ $m$ ”?
- Rule 2 (Bollinger rule): Use the Bollinger rule as outlined in the text; what is the optimal “ $m$ ”?
- Rule 3 (Resistance-Support): Buy if  $p_t > \max_{0 \leq i < m} p_{t-i}$  and sell if  $p_t < \min_{0 \leq i < m} p_{t-i}$ . Evaluate this rule. What is the optimal “ $m$ ”?
- Rule 4 (Momentum): Compute a short-term moving average over the last  $m$  prices,  $\bar{p}_t^{(m)}$  and a long-term average over the last  $n$  prices,  $\bar{p}_t^{(n)}$  with  $m < n$ . If  $\bar{p}_t^{(m)}$  crosses  $\bar{p}_t^{(n)}$  from below, sell and if it crosses  $\bar{p}_t^{(n)}$  from above, buy. Evaluate this strategy. What are the optimal values of  $m$  and  $n$  ?

(b) *RSI Oscillator Rule:* Use  $RSI_t$  as defined in (5).

If  $RSI_t < 30$ , buy and if  $RSI_t > 70$ , sell. Evaluate this trading strategy. What is the optimal “ $m$ ”?

(c) *Pairs Trading: Distance Method*

We want to develop a pairs trading strategy and check if it does any better than the strategies in Problem 1. At the last trading day of the month, look back 3 months and identify if any pairs are worth trading. Use the distance measures, thresholds, and portfolio return accounting as given in Engleberg, et al. (2009, p2-4). Summarize the results using the same format as in Problem 1.

(d) *Pairs Trading: Cointegration Method*

In the cointegration method, we can consider the possibility of trading more than 2 assets. Check using a moving window of 3 months which series are cointegrated. Develop appropriate trading strategies to trade portfolios formed from the cointegrating vectors and compare with the pairs trading strategies from Problem 2. (You may use and extend the ideas from Problem 2 to concretely define the trading indicators and thresholds for entry and exit.) Summarize the performance results as in Problem 1.

(e) *Pairs Trading: APT method*

Consider the data with the S&P500 market index returns and assume that the

risk free rate is 3%. Regress each currency's excess return (over the risk free rate) on the excess market return over a moving window of 3 months. Find pairs based on the similarity of the regression coefficients.

- i. Compare the pairs that result from the APT method with those obtained in Problem 2.
- ii. After finding the pairs, use the same rules as in Problem 2 to decide when to enter and exit specific trades.

Summarize the performance results as in Problem 1.

(f) *Summing up*

Provide a summary table with the main findings and drawbacks of all the methods that are explored in this assignment. In particular, comment on the performance of single-asset strategies versus multi-asset strategies. Also comment on whether the market factor adds any value.

6. This exercise using data for sixteen stocks will guide you through creating a momentum strategy. The goal is to replicate the price momentum strategy described in Jegadeesh and Titman (2001) using this data, albeit in a small scale. [Solution available on Web.]

### 6.1 Data

The daily returns from January 11, 2000 to November 20, 2005 for sixteen stocks are in the file `d_logret_16stocks`; to this add two more columns: SP500 returns and risk free rate.

### 6.2 Portfolio Formation

On the last day of each month, sort the stocks into four quartile portfolios based on the past six month cumulative returns. The first quartile contains the stocks with the lowest past six-month returns, and the fourth quartile contains the stocks with the highest past six-month returns.

- (a) Start by accumulating the return of each quartile portfolio for next six months
- (b) For each month and of the four quartile portfolios, record the average portfolio returns between January 2002 and November 2005.

### 6.3 Questions of Interests

After constructing the relevant portfolios, answer the following questions. When you answer these questions, note any additional assumptions/choices you made.

#### 6.3.1 Momentum strategy profits

1. Report the monthly average returns of your quartile portfolio, as well as the winner (Q4, highest past six month return portfolios) minus losers (Q1, lowest past six month return portfolios).
2. Report the monthly average returns of quartile portfolio for January and non-January months respectively, a well as the winner (Q4, highest past

- six month return portfolios) minus losers (Q1, lowest past six month return portfolios).
3. Using CAPM test the spreads between winner and loser portfolio (Q4-Q1).
  4. Tabulate these results in a table similar to table I in Jegadeesh and Titman (2001).
  5. Compare your results to the results in Jegadeesh and Titman (2001) and comment on your findings.
7. Consider the 30-min price bar data for Treasury Yields from June 8, 2006 to August 29, 2013. Daily data for Treasury Yields is just a snapshot at 4 pm and the total volume for the day is the sum of 30-min volumes since the previous day. [Solution available on Web.]
- (a) Develop ARMA time series models for both daily and the 30-min volumes. We can use daily data for developing macro strategies and 30-min data for micro strategies.
  - (b) Compare the exponential smoothing approach suggested in the text with ARMA modeled in (a).
  - (c) Compute volatility measure,  $\sigma_t^2 = 0.5[\ln \frac{H_t}{L_t}]^2 - 0.386[\ln \frac{C_t}{O_t}]^2$  for both levels and compare them
  - (d) Compute volatility forecasts,  $\hat{\sigma}_{t+1}^2 = \text{weighted average of } \sigma_t^2, \sigma_{t-1}^2, \dots, \sigma_{t-22}^2$ , for daily data and  $\hat{\sigma}_{t+1,m}^2 = \text{weighted average of } \sigma_{t,m}^2, \dots, \sigma_{t-22,m}^2$  for  $m^{th}$  30-min interval. [ In (a), you should notice strong seasonality in the 30-min data]
  - (e) Develop trading strategies that incorporate
    - i. price information only
    - ii. price and volume information and
    - iii. price, volume and volatility information
  - (f) Demonstrate that volume and volatility add value to the enter/exist strategies.



# 6

---

## *News Analytics: From Market Attention and Sentiment to Trading*

---

### CONTENTS

6.1	Introduction to News Analytics: Behavioral Finance and Investor Cognitive Biases .....	157
6.2	Market Micro Structure and Liquidity .....	162
6.3	Automated News Analysis and Market Sentiment .....	162
6.4	News Analytics and its Applications to Trading .....	164
6.5	Discussion / Future of Social Media News .....	173

---

### **6.1 Introduction to News Analytics: Behavioral Finance and Investor Cognitive Biases**

The determinants of large sporadic movements in stock prices that are not justified by fundamentals have been speculated and studied by many economists, starting with Keynes (1937) [179]. The early empirical studies on this (see Cutler, Poterba and Summers (1989) [73]) linking stock news to stock prices did not find any significant relationships between large returns and the macroeconomic events. The rational financial theory that assumes that the market price of a stock is equal to the present value of future expected cash flows is generally unable to explain these patterns. So alternative explanations are sought from the behavioral finance perspective. It is argued that investors make decisions based on their sentiment that may not be related to data at hand about the stock performance and generally competing with these sentimental investors can be costly. Several market crashes can attest to these premises.

There are a number of studies that find evidence to support the hypothesis that the content of news media can predict stock market activities. The communication channels have evolved from slow print media where the information is screened and edited to almost instantaneous social media, where participation is high but the information is of varying quality. The web which is in-between can be slower than social media but a certain degree of quality can be normally expected. Antweiler and Frank (2004) [15] find evidence for the relationship between internet message activities that can be characterized into “buy”, “sell” or “hold” recommendations and the trading volume and the volatility.

Tetlock (2007) [250] provides a nice summary of theoretical models of the effect of investor sentiment on stock prices. They assume that there are two types of traders, noise and rational traders, who hold random beliefs about the future stock performance and hold informed beliefs based on the past performance and the future potential respectively. The difference between the two beliefs when measured right is taken to reflect the investor sentiment. The changes in the stock behavior can be due to a number of other factors as well, such as risk aversion. The so-called media pessimism can be a proxy for such stock-related or in general market-related factors as well but its timing is used to distinguish sentiment related influence on the stock behavior. Lower sentiment generally leads to downward pressure on prices resulting in low returns at short horizons but will be reversed in the long run. If it is due to true information about the stock, the trend in returns will persist. The investor sentiment is also likely to influence the trading volume with irrational traders trading heavily or “rational” traders with shorter horizon.

Baker and Wurgler (2006) [18] investigate how investors’ beginning-of-period sentiment affects the cross-section of future stock returns. A number of proxies for sentiment are considered. These are closed-end fund discount measured by the difference between the net asset value of closed-end stock fund shares and their market prices, NYSE share turnover, the number and average returns of the IPOs, the equity share in the new issues and the dividend premium which is measured by the log difference of the average market-to-book ratios of payers and non-payers. These six proxies for investment are used to form a sentiment index via principal component method. The first component is then shown to be related to future stock returns after adjusting for the stock and firm characteristics such as firm size, age, profitability, book to market equity, etc.. Baker and Wurgler (2007) [19] consider some additional proxies as well and these are able to better discriminate the two factors that could lead to mispricing, a change in sentiment of the irrational traders and a limit to arbitrage from the rational investors. Market turnover, defined as the ratio of trading volume to the number of listed shares, Market Volatility Index (VIX), equity issues as a ratio of total new issues that includes debt issues as well and a measure of insider trading pattern are all considered to be other proxies for the sentiment. The constructed index is orthogonalized to a set of state variables: real growth in durables, nondurables and services consumption; growth in employment and the National Bureau of Economic Research recession indicator. The time-varying index constructed from these is shown to match up well with the anecdotal accounts of the timing of bubbles and crashes.

It must be noted that the sentiment index constructed by Baker and Wurgler (2006, 2007) [18] [19] is based on monthly data as the underlying variables are somewhat slow-moving. At low frequency, the index does provide a predictive direction for the stock performance. The index is updated and posted in the authors’ websites; it is given in Figure 6.1. Yu and Yuan (2011) [264] and Stambaugh, Yu and Yuan (2012) [244] show how the index can shed light on well-documented anomalies. These anomalies include unexpected deviations in performance during the time of financial distress; the firms with high probability of failure have lower, not higher, subsequent returns which is not expected by the standard asset pricing models. An-

other one is the momentum effect which refers to the observation that high past returns forecast high future returns. Also it is noted that higher past investment predicts abnormally lower future returns.

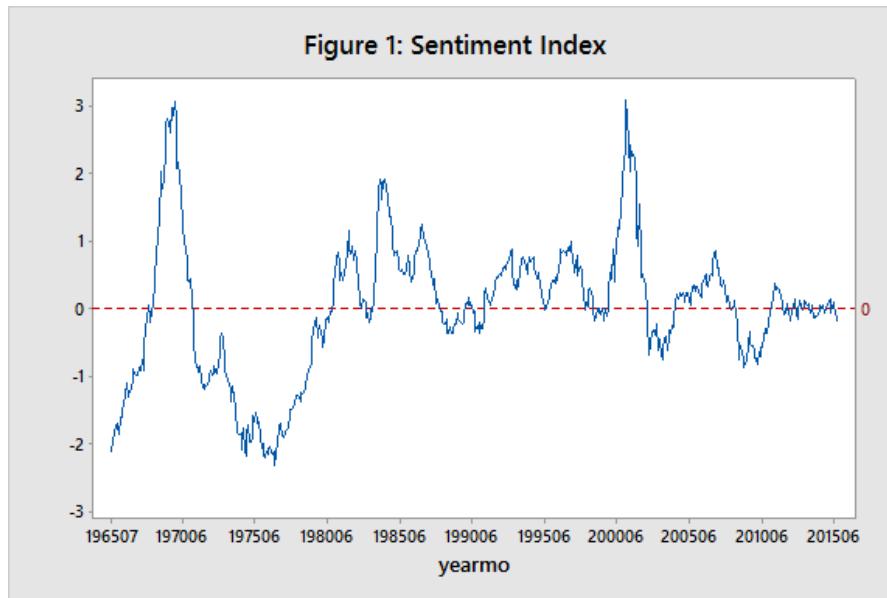


Figure 6.1: Sentiment Index.

While the extensive discussion of these results at the stock level is not possible here, we want to present the descriptive statistics of key performance indicators and other associated characteristics at the market level. The goal of this exercise is to show how conditional on the sentiment scores the averages over different key characteristics differ. The data on Fama-French factor model, see Fama and French (2015) [109], is taken from French's website. It consists of (details are directly from the site):

- SMB (Small Minus Big) is the average return on the nine small stock portfolios minus the average return on the nine big stock portfolios.
- HML (High Minus Low) is the average return on the two value portfolios minus the average return on the growth portfolios.
- RMW (Robust Minus Weak) is the average return on the two robust operating profitability portfolios minus the average return on the two weak operating profitability portfolios.
- CMA (Conservative Minus Aggressive) is the average return on the two conservative investment portfolios minus the average return on the two aggressive investment portfolios.

- $R_m - R_f$ , the excess return on the market, value-weight return of all CRSP firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of  $t$ , and good return data for  $t$  minus the one-month Treasury bill rate (from Ibbotson Associates).

The daily data on the above factors for the same duration as the duration of the sentiment data given in Figure 6.1 was collected and aggregated to match the monthly level sentiment data. For the excess return,  $R_m - R_f$  we compute the monthly averages and as in Yu and Yuan (2011) [264], we compute the realized variance of the excess return as,

$$\hat{\sigma}_t^2 = \frac{22}{N_t} \sum_{t=1}^{N_t} (R_{mt} - R_{ft})^2 \quad (6.1)$$

where  $N_t$  is the number of trading days in a month. The aggregated factors are plotted in Figure 6.2.

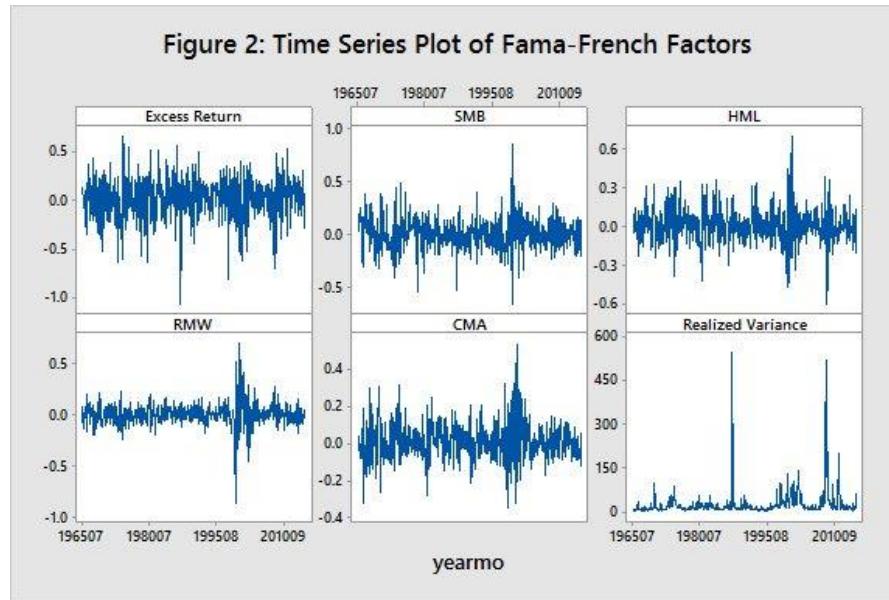


Figure 6.2: Time Series Plot of Fama-French Factors.

To classify the sentiment into low and high Baker and Wurgler (2006) [18] simply use the negative and positive scores. In order to allow for the natural variation in the sentiment scores, we classify the score below the first quartile ( $-0.5$ ) as low sentiment and the score above the third quartile ( $+0.5$ ) as high sentiment and in between as normal sentiment. Table 4.1 provides the summary. The contrast between the low and high sentiment periods is quite evident.

Some observations follow: While the pattern in excess return is consistent with what is observed in Yu and Yuan (2011) [264], the realized variance in high sentiment

Table 6.1: Fama &amp; French Factors Across Sentiment Levels

	Low	Normal	High	All
N	152	283	168	603
Sentiment	-1.297	0.031	1.120	0.000
Excess Return	0.0127	0.044	-0.004	0.023
SMB	0.031	0.009	-0.010	0.009
HML	0.014	0	0.052	0.018
RMW	-0.001	0.007	0.035	0.013
CMA	0.011	0.003	0.040	0.016
Realized Variance	23.89	24.58	16.83	22.25

period is lower than the other periods, but this is consistent with the mean-variance relationship as expected by the asset pricing models. The SMB which measures the difference in returns between small stocks and large stocks is negative during high sentiment periods indicating the portfolio of small stocks does not do well in these periods. The HML, the difference in returns between high and low B/M stocks tends to be positive and higher during high sentiment periods. The RMW which measures the performance difference between robust and weak profitability portfolios is also higher during high sentiment periods. Finally, CMA, which measures the difference in performance between conservative and aggressive firms continue to do well in high sentiment periods.

Da, Engleberg, and Gao (2011) [74] review studies on investor attention and its impact on stock performance. The proxies such as extreme returns, trading volume, news and headlines, advertising expense etc. were used as indirect measures of attention. Because some of these proxies can be driven by factors unrelated to investor attention, it must be noted that they, at best, are proxies. Instead it is suggested that search frequency of key words related to a stock in a search engine is a more direct measure of the attention as the users who search directly express their intent and thus their interest.

Google publishes the search volume index (SVI) for the key words and except for rarely-searched tail terms, the index is fairly reliable due to the significant market share of Google in terms of search volume compared to other search engines. Da et al. (2011) [74] find SVI is a leading indicator of extreme returns and news and the abnormal SVI matches with the known anomalies. Generally we must emphasize that what are meant by anomalies are any empirical deviations from what are postulated by economic theory. A positive abnormal SVI generally predicts higher stock prices in the short term and price reversals in the long term.

Da, Engleberg, and Gao (2015) [75] extend this idea to study the market-level sentiment and its correlation to aggregate market returns. The market-level index is based on the top thirty search terms that relate to general public's concern, such as unemployment, recession, bankruptcy, etc. The index is shown to predict short-term return reversals, increase in volatility and flows from equity funds to bond funds. The change in the index is demonstrated to be correlated with certain sentiment indicators, of Baker and Wurgler (2006) [18].

Stambaugh et al. (2012) [244] consider several asset-pricing anomalies that may be attributed to financial distress, momentum, asset growth etc. For each anomaly, the strategy that goes long in the highest-performing decile and goes short in the lowest-performing decile is examined. However as Huang et al. (2015) [161] observe, “... whether investor sentiment can predict the aggregate stock market at usual monthly frequency is still an open question.” It must be noted that the market is efficient for the most part and impounding any information occurs rather immediately at the higher frequency level. Thus there is a great deal of interest to mine news, attention and sentiment as they occur and the information from social media has come to play an effective role. We present some evidence to that effect in the next section.

---

## 6.2 Market Micro Structure and Liquidity

---

### 6.3 Automated News Analysis and Market Sentiment

Trading now mostly occurs at a relatively high frequency level and hence any news related to a stock gets processed almost instantaneously. The literature reviewed earlier refers to low frequency sentiment data and hence its influence on returns is not easy to notice. New technologies that can handle automatic news collection, extraction and categorization of relevant information are fast-emerging. Quantitative models that incorporate this information in investment decisions are being actively researched and automation of these activities can greatly help traders shortening their reaction time to emerging news feeds.

There has been a surge of academic interest in this topic (see Mitra and Mitra (2011) [211] and Petersen (2016) [224]). Several technology companies (iSentium, MarketPsych, RavenPack, etc.) have sprung up in the last few years offering the service of aggregating the web information. Financial news can be classified into two main groups. Some financial news related to a stock are released on a regular basis, such as earnings statements and are well structured. The expectations on the earnings are formed based on the stock’s other performance indicators and when the actual earnings differ from the trader’s expectations, the trading decisions are normally adjusted for the a priori speculation. Then there are other news related to the product or personnel of the company that may come unscheduled and may affect the trading decisions. Although these are unstructured due to limited sources, they can be easily processed. Usually the early exploiters of the information can expect to gain some advantage over others. These are generally termed as news-based event strategies and have been fairly well-studied in the finance literature.

Unstructured news streams come at irregular intervals and are usually textual and qualitative. In order to correlate this information with the stock performance, they need to be appropriately quantified and because this information originates from numerous sources such as in social media, it needs to be properly aggregated. Without

any aggregation it may be difficult to extract the signal from the noise. Some common steps involved in making the news useable for trading are:

- Identify the unique and relevant news on a timely basis.
- Quantify the textual information into informative scores, whose variation would cover negative to positive sentiment/connotation.
- Aggregate the scores from different sources.
- Develop a baseline model adjusting for seasonality of the sentiment scores.
- Relate the abnormal deviation of the sentiment scores to abnormal stock price movements.

As many of these concepts and related issues are discussed in the references cited in this chapter, we will be brief in our discussion. Das and Chen (2007) [79] show how it is possible to capture the net sentiment from positive and negative views on message boards using statistical language techniques. The emotive sentiment is elicited using different classification algorithms and the results are pooled via majority voting. Opinions are extracted from message board postings and they are classified into three types: bullish, bearish and neutral. The classification algorithms include parts of speech tagging and support vector machines. The algorithms are initially trained using humanly preclassified messages and are designed to learn and apply these rules out-of-sample. Relating the sentiment to the performance of stocks in the Morgan-Stanley High-Tech Index it is observed that there is no strong relationship to individual stock prices but there is a statistical relation to the aggregate index performance. There is also a strong relationship between message volume and volatility consistent with Antweiler and Frank (2004) [15].

We want to briefly provide some details on a few providers of web analytics; interested readers should consult these providers' websites for further details.

- Ravenpack (NewsScore): The key information on entities (company, organization, currency, commodities and place) from major news sources are gathered and processed for their relevance and are assigned event sentiment score. The scores are aggregated on a rolling window basis. In addition, the aggregate count as event volume, the event novelty score, event novelty elapsed time etc. are provided. For equity markets, composite sentiment score and the impact projections of the news are also available. The main news providers are Dow Jones Newswires, The Wall Street Journal (all editions) and Barron's.
- Thomson Reuters (News Analytics): Data fields include relevant score of the news item to the asset, number of sentiment words or tokens, sentiment classification, novelty of the content, feed volume etc. The sources are Reuters, PR Newswire etc.
- Thomson Reuters (MarketPsych Indices): News and social media information in real time are delivered as data series; there are categorized into three types of indicators: Emotional indicators, Macroeconomic metrics and Buzz metrics on the

asset level. The data is updated on a minute basis. The social media data includes blogs, internet-forums and finance-specific tweets.

- iSentium (iSense): Extracts sentiment signals using natural language processing architecture on Twitter data; the data feed includes retweets, if the author has a finance-related bio, number of followers etc. The impact is measured by the number of retweets for each author's tweets.

In the next section, we will summarize studies that use the information provided by these agencies for developing trading strategies. To provide the theoretical underpinning of the effect of investor sentiment, two assumptions are made (see Delong, Shleifer, Summens and Walman (1990) [80]): market consists of two types of traders, noise and rational and they both are risk averse. Tetlock (2007) [250] defines sentiment as the level of noise traders' random beliefs to rational traders' updated beliefs; it is argued that media (collective) pessimism is a proxy for either the sentiment or risk aversion and that it is expected that high media pessimism will predict low returns at short horizons and at longer horizons, a reversion to fundamentals will occur.

## 6.4 News Analytics and its Applications to Trading

We have reviewed studies that have clearly documented the influence of news on stock market behavior. In this section, we will outline first the key methodologies used in relating the data on the inflow of news to key characteristics of stock trading and the stock performance indicators. A simple framework that we will follow is from Gloss-Klussmann and Hautsch (2011) [139] and is depicted in Figure 6.3:

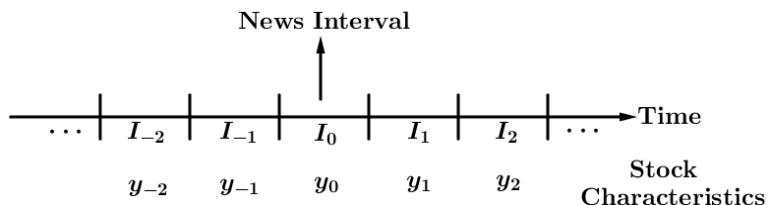


Figure 6.3: News Interval.

The news arrival can generally be modeled as a point process but to remove any noisy friction in the data, both in the newsfront and in the stock price movement, it may be advisable to aggregate the information in short intervals. The stock characteristics can include the change in the price as well as volume, volatility, etc. during the interval. In addition to time aggregation, we also need to consider aggregation over different variables or dimensions of sentiment as in Baker and Wurgler (2006) [18]. We describe these in two contexts below.

### **Aggregation Over Dimensions:**

In the use of monthly data from Google, Da et al. (2015) [75] aggregate the daily change in a related search term,  $i$ , over the top thirty terms to come up with an index called, FEARS (financial and economic attitudes revealed by search) as

$$\text{FEARS}_t = \sum_{i=1}^{30} R^i(\Delta \text{ASVI}_{it}) \quad (6.2)$$

where  $\Delta \text{ASVI}_{it}$  is an adjusted (winsorized, deseasonalized, and standardized) daily change in search volume,

$$\Delta \text{ASVI}_{it} = \ln(\text{SVI}_{it}) - \ln(\text{SVI}_{it-1}) \quad (6.3)$$

Here  $R(\cdot)$  denotes the ranking function. Generally search volume data exhibits non-stationarity as well as strong weekday seasonality. It has, for instance, been found that web searches tend to be higher on Monday through Wednesday than on other days. The above adjustments to search volume take care of both concerns. The Equation (6.2) is essentially a differencing operator that addresses the stationarity and a similar differencing for the weekday effect can be easily incorporated. These ideas were discussed in Chapter 2.

The sentiment index in Baker and Wurgler (2006) [18] is a composite index based on the five standardized sentiment proxies mentioned earlier. Because some variables take longer to express the same sentiment, the proxies are lagged. The series are smoothed over the prior twelve months and the index uses the  $t - 12$  value. The new series do not use turnover, due to fragmentation, NYSE is no longer the dominant venue.

In the construction of this index, it appears the principal component of the five variables and their lags are used. Thus the first stage index has ten loadings; using the correlation between the index and current and lag values of the variables, the time index that is most appropriate for each of the variables is selected. In the original construction, the variables RIPO and PD-ND are lagged but the rest are used as of current time. The sentiment index is also further refined to adjust for a common business cycle components measured through growth in the industrial production index, growth in consumer durables, nondurables and services and an indicator variable for NBER recessions. The regression residuals are argued as better, cleaner measures of sentiment. These are then used in the principal component analysis. Thus mathematically, if  $x_t = (x_{1t}, x_{2t}, \dots, x_{5t})'$ , the first principal component is obtained by

$$\max_l l' \sum_{xx} l \quad ; \quad l'l = 1 \quad (6.4)$$

Because the variables are standardized, it should be noted that the covariance matrix,  $\sum_{xx}$  is also the correlation matrix. The first linear combination is associated with the eigenvector that corresponds to the largest eigenvalue of the matrix,  $\sum_{xx}$ , gives the sentiment score,

$$s_t = \frac{l' x_t}{\sqrt{\lambda}} \quad (6.5)$$

in Figure 6.3. Here  $\lambda$  is the largest eigenvalue of  $\sum_{xx}$  so that the sentiment score,  $s_t$  is taken to have unit variance.

### Relating Sentiment to Stock Performance:

The general framework relating the sentiment scores ( $s_t$ ) to performance indicators ( $y_t$ ) is simply through lagged regression. For example, the returns  $R_t(y_t)$  are related to ( $s_t$ ) and other control variables and can take the form:

$$R_{t+k} = \beta_0 + \beta_1 s_t + \gamma' z_t + u_{t+k} \quad (6.6)$$

where  $z_t$  is a set of control variables that may be related to the stock, such as Fama and French factors. The left side can also be (annualized) realized volatility, calculated by sampling the price over ‘ $N$ ’ times during a day.

$$rv_t = 250 \sum_{d=1}^N r_{t,d}^2 \quad (6.7)$$

Because of the persistence of the volatility, some adjustments via differencing should be considered when relating sentiment scores to volatility. Finally, it is possible that the effect of positive sentiment on stock returns could be different from that of negative sentiment. Fitting models similar to (6.6) for both scenarios separately would be a worthwhile exercise.

### Aligned Investor Sentiment:

In the above set-up, the construction of the sentiment scores (Equation 6.5) or FEARS (Equation ??) is done first and then it is related to—for instance—the returns via (6.6) subsequently. Huang, Jiang, Tu and Zhou (2015) [161] combine the two steps, together arguing the linear combination in (6.5) is constructed better in referencing to the stock characteristics and not on its own. They propose the partial least squares (PLS) method advanced by Wold (1984) [261].

**Partial Least Squares:** Partial least squares by Wold (1984) [261] is an iterative computational algorithm for linear regression modeling, especially for situations where the number  $n$  of predictor variables ( $X_t$ ) is large. The  $m \times 1$  multiple response,  $X_t$  ( $m > 1$ ) version of partial least squares begins with a ‘canonical covariance’ analysis, a sequential procedure to construct predictor variables as linear combinations of the original  $n$ -dimensional set of predictor variables  $X_t$ . At the first stage, the initial predictor variable  $X_1^* = b_1' X_t$  is determined as the first ‘canonical covariance’ variate of the  $X_t$ , the linear combination of  $X_t$  that maximizes  $\{(T - 1) \times \text{sample covariance of } b' X_t \text{ and } l' Y_t\} \equiv b' \mathbf{X} \mathbf{Y}' l$  over unit vectors  $b$  (and over all associated unit vectors  $l$ ). For given vectors  $b$ , we know that  $l = \mathbf{Y} \mathbf{X}' b / (b' \mathbf{X} \mathbf{Y}' \mathbf{X}' b)^{1/2}$  is the maximizing choice of unit vector for  $l$ , so that we need then to maximize  $b' \mathbf{X} \mathbf{Y}' \mathbf{X}' b$  with respect to unit vectors  $b$ . Thus, the first

linear combination vector  $b_1$  is the (normalized) eigenvector of  $\mathbf{XY}'\mathbf{YX}'$  corresponding to its largest eigenvalue. (In the univariate case,  $m = 1$  and  $\mathbf{Y}$  is  $1 \times T$ , and so this reduces to  $b_1 = \mathbf{XY}' / (\mathbf{YX}'\mathbf{YX}')^{1/2}$  with associated “eigenvalue” equal to  $\mathbf{YX}'\mathbf{YX}'$ .)

At any subsequent stage  $j$  of the procedure, there are  $j$  predictor variables  $x_{it}^* = b_i' X_t$ ,  $i = 1, \dots, j$ , determined and the next predictor variable  $x_{j+1}^* = b_{j+1}' X_t$ ; that is, the next unit vector  $b_{j+1}$  needs to be found. It is chosen as the vector  $b$  to maximize the sample covariance of  $b' X_t$  and  $l' Y_t$  (with unit vector  $l$  also chosen to maximize the quantity); that is, we maximize  $b' \mathbf{XY}'\mathbf{YX}' b$ , subject to  $b_{j+1}$  being a unit vector and  $x_{j+1,t}^* = b_{j+1}$ ;  $X_t$  being orthogonal to the previous set of  $j$  predictor variables  $x_{it}^*$ ,  $i = 1, \dots, j$  (i.e.  $b_{j+1}' \mathbf{XX}' b_i = 0$ , for  $i = 1, \dots, j$ ). The number  $w$  of predictors chosen until the sequential process is stopped is a regularization parameter of the procedure; its value is determined through cross-validation methods in terms of prediction mean square error. The final estimated regression equations are constructed by ordinary LS calculation of the response vectors  $Y_k$  on the first  $w$  ‘canonical covariance’ component predictor variables,  $X_k^* = (x_{1k}^*, \dots, x_{wk}^*)'$   $\equiv \hat{B} X_k$ , where  $\hat{B} = [b_1, \dots, b_w]'$ ; that is, the estimated regression equations are  $\hat{Y}_k = \hat{A} X_k^* \equiv \hat{A} \hat{B} X_k$  with

$$\hat{A} = \mathbf{YX}'(\mathbf{X}^*\mathbf{X}^*)^{-1} \equiv \mathbf{YX}'\hat{B}'(\hat{B}\mathbf{XX}'\hat{B}')^{-1}$$

and  $\mathbf{X}^* = \hat{B} \mathbf{X}$  is  $w \times T$ . In practice, the predictor variables  $X_k$  and response variables  $Y_k$  will typically already be expressed in terms of deviations from sample mean vectors.

Discussion of the sequential computational algorithms for the construction of the ‘canonical covariance’ components  $x_{jk}^* = b_j' X_k$  required in the partial least squares regression procedure, and of various formulations and interpretations of the partial least squares method are given by Wold (1984) [261], Helland (1988,1990) [150] [151], and Garthwaite (1994) [127]. In particular, for univariate partial least squares, Garthwaite (1994) [127] gives the interpretation of the linear combinations of predictor variables,  $x_{jk}^* = b_j' X_k$ , as weighted averages of predictors, where each individual predictor holds the residual information in an explanatory variable that is not contained in earlier linear combination components ( $x_{ik}^*$ ,  $i < j$ ), and the quantity to be predicted is the  $1 \times T$  vector of residuals obtained from regressing  $y_k$  on the earlier components. Also, for the univariate case, Helland (1988,1990) [150] [151] showed that the linear combination vectors  $\hat{B} = [b_1, \dots, b_w]'$  are obtainable from nonsingular ( $w \times w$ ) linear transformation of  $[\mathbf{XY}', (\mathbf{XX}')\mathbf{XY}', \dots, (\mathbf{XX}')^{w-1}\mathbf{XY}']'$ .

Huang et al. (2015) [161] demonstrate that the new combination of predictors based on PLS method has much greater predictive power for the aggregate stock market and its predictability is both statistically and economically significant.

**Social Media/Twitter Data:** Over the past years, significant development has been made in sentiment tracking technologies that can extract indicators of public mood from social media such as blogs and Twitter feeds. Although each Tweet is of limited number of characters, the aggregate of Tweets may provide an accurate representation of public sentiment. This has led to the development of real-time sentiment tracking

tools (Opinion Finder, GPOMS (Google Profile of Mood Status), etc.) to track the public mood and relating them to economic indicators. Crowd-sourced data generated by social network sites such as Twitter/StockTwits is being increasingly used by quantitative researchers to generate trading signals. The wisdom of the crowd surprisingly appears to outperform the wisdom of the experts in a variety of instances. For a recent application relating sentiment data to earnings announcement, see Liew, Guo and Zhang (2017) [192]. We now illustrate a simple trading algorithm using the iSentium data.

iSentium expertise lies in providing market sentiment indicators. Market-related texts from Twitter are processed through Natural Language Processing Techniques and the contents are assigned sentiment scores in the range of  $-30$  to  $30$  with positive score indicating optimistic view and the negative score, a pessimistic view. For a specific keyword the data contains the following information:

- Time when the Tweet is recorded
- Sentiment Score
- Total number of Tweets since Jan 1, 2012
- Whether it is a retweet
- Whether the author has a finance-related bio
- Number of followers of the author
- Number of Tweets posted by the author
- Average number of retweets for each of the author's Tweets, measuring the author's impact.

In addition to per-tweet data, the company also provides binned data, binned over various time durations. Surprisingly tweet volume is high for popular ETFs and large cap technology stocks. To illustrate, we consider the per-tweet data from Jan 1, 2012 to September 10, 2014 for the **ESMini** (ticker symbol: **QQQ**). The typical data is listed in Table 6.2:

It is quite apparent from the short list of entries in Table 6.2, the same author of a tweet can have varying sentiments in a short duration of time. It is also likely that a few authors can dominate the number of tweets in a short duration ( $\approx 90$  minutes in the sample data in Table 6.2). Consequently in the application we describe here, the daily summaries of sentiment scores are calculated and matched with the daily performance of ESMini. Note that twitter activity occurs every day but the markets are open only on weekdays. No attempt is made to account for the cumulative effect of weekend sentiment scores on the beginning-of-the-week ESMini performance. The ESMini data contains daily price bars (with open, high, low and close) and the volume of transactions.

Table 6.2: Typical Elements of iSentium Data

Date	Time	SS	Retw	Retwld	authid	FinBio	nfollow	ntweets	impact
1/1/2012	3:04:41	9	0	-1.00E + 00	202647426	1	222	16	0
1/1/2012	3:10:06	9	0	-1.00E + 00	23059499	1	10501	27	0
1/1/2012	3:19:46	-11	0	-1.00E + 00	202647426	1	222	25	0
1/1/2012	3:43:52	0	0	-1.00E + 00	202647426	1	222	30	0
1/1/2012	3:50:01	0	0	-1.00E + 00	23059499	1	10501	48	0
1/1/2012	4:10:01	-5	0	-1.00E + 00	23059499	1	10500	89	0
1/1/2012	4:11:02	-5	0	-1.00E + 00	207429198	0	51	2	0
1/1/2012	4:12:10	-5	1	1.53E + 17	16704290	1	1170	3	0
1/1/2012	4:26:07	-10	0	-1.00E + 00	207429198	0	51	9	0
1/1/2012	4:30:03	-10	0	-1.00E + 00	23059499	1	10500	104	0.0625

The time series plot of the sentiment related variables and the ESMini trade related variables are given in Figure 6.4. Here the variables are defined as:

- RetO=  $\ln(P_{t_0}) - \ln(P_{t-1,0})$  based on opening price
- Volume is logged
- Volatility=  $0.5(H_t - L_t)^2 - 0.386(C_t - O_t)^2$ , where  $H_t = \ln(P_{tH})$ , the high price observed during the day etc.
- SSAav is the average sentiment score for the day
- FBav is the proportion of tweets by the tweeters with a finance bio
- Retav is the average number of retweets per day

While one focus is on sentiment score, the other two sentiment variables indicate the intensity of activity by the twitter participants in the finance field.

It can be seen from Figure 6.4, that the sentiment average (SSAav) exhibits some variation over time but it is generally stationary. There is some memory in the series as the first few autocorrelations are found to be significant. All the other sentiment related series also exhibit some memory. The same can be noted for the volume and volatility series as well. One of the key tools to connect two series in terms of lead-lag relationships is the cross-correlation function defined in Chapter 2. For the sentiment data ( $s_t$ ) to be of any predictive use, it should be correlated to lead return ( $r_{t+k}$ ) and lead volatility ( $V_{t+k}$ ). Figures 6.5 and 6.6 capture the cross-correlation, but one can observe that there is some concurrent correlation between  $s_t$  and  $V_t$ . This may be due to the fact that the aggregation of data was done over the entire day, which may be too long a duration for aggregation. These relationships can be confirmed in Table 6.3 with stepwise regression model fit.

We propose two naïve strategies that can be used for trading using the sentiment data are proposed; both are in the spirit of the discussion related to sentiment data by Baker and Wurgler (2006) [18]. The first strategy develops rules based on the

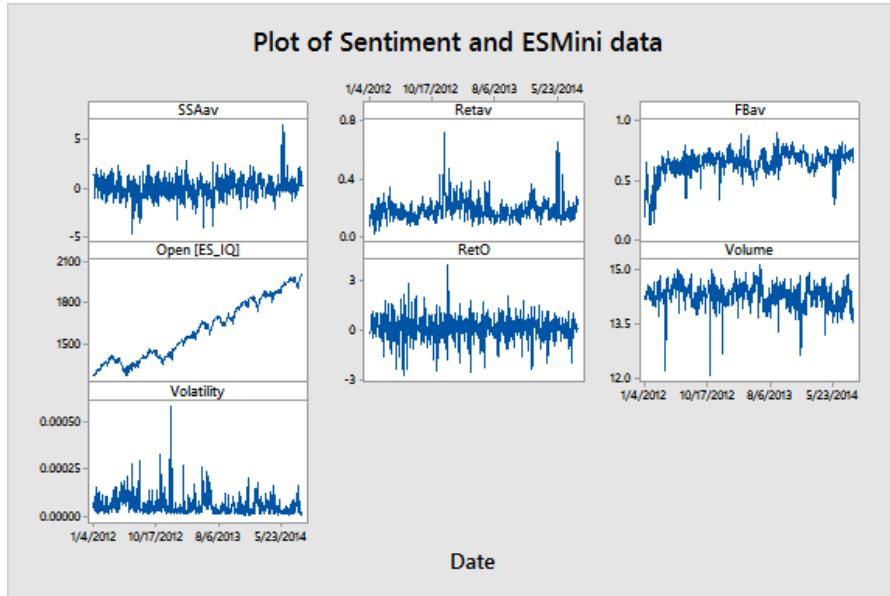


Figure 6.4: Sentiment score and ESMini.

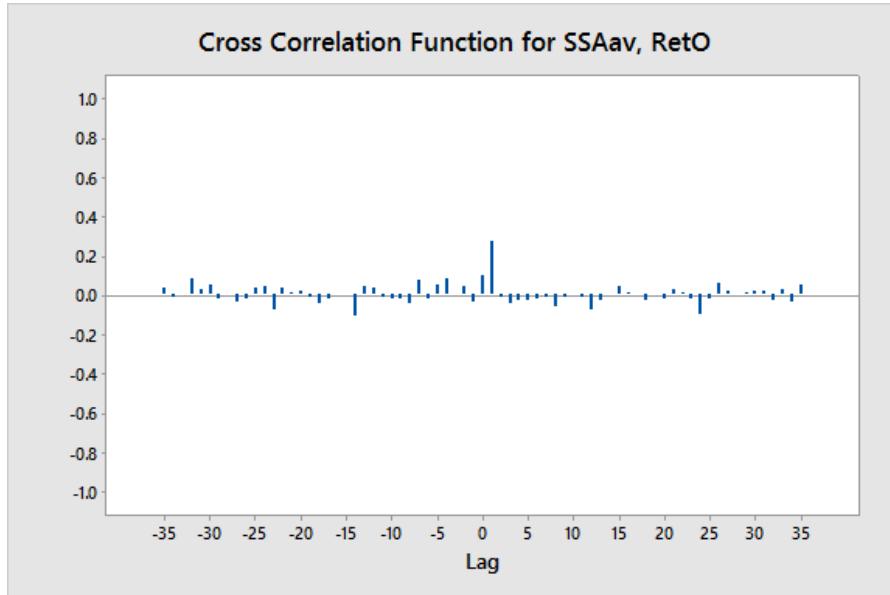


Figure 6.5: CCF(SSAve → Reto).

sentiment scores and the second is based on the more direct regression relationship between returns and the sentiment scores:

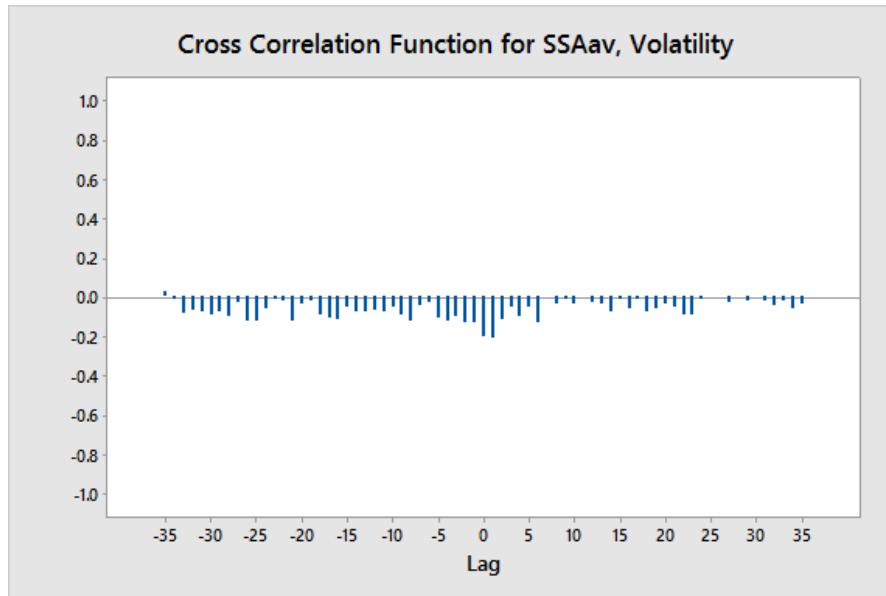


Figure 6.6: CCF(SSAve – Reto).

Table 6.3: Regression of ESMini variables on sentiment factors

Predictors	Response Variable		
	Return	Volume	Volatility
SSAav	—	-0.0482	-0.00005
SSAverlag	0.1895	-0.0291	-0.00006
AuHTAV	—	-0.000017	-0.000001
RetAV	—	—	-0.000063
NfollowAV	—	—	-0.000001
AVetW	—	—	0.000047
Constant	0.0759	14.41	0.000085
<i>R</i> <sup>2</sup>	0.0732	0.1154	0.1357

\*All indicated factors are significant at 0.05 level

1. Crossover Strategy: If the difference between 3-day EWMA and 30-day EWMA of sentiment scores is positive, go long; otherwise go short.
2. Open-to-open Strategy: Regress open-to-open returns on 10-day EWMA and its lag to predict the return; if positive, go long. Hold the position until the next day open.

It is easy to note that the moving average crossover strategy discussed in Chapter 3 is comparing the short term trend with the long term trend. The strategy exploits the deviation between the two. Both strategies can be evaluated on a moving window ba-

sis. Here we present results for the cross-over strategy only. As noted from Figure 6.4, the ESMini price has been steadily going up and thus a simple buy and hold strategy over time would be profitable. So we compare performance of the strategy with the buy and hold. The simple averages over buy and sell signals are quite revealing as seen from Table 6.4

Table 6.4: Descriptive Statistics (Averages): Cross-over strategy

	SSAve	RetAve	FBAC	Ret0	Volume	Volatility	<i>N</i>
Short	-0.5223	0.1783	0.6539	-0.0866	14.327	0.000061	333
Long	0.4398	0.1793	0.6477	0.2206	14.220	0.00044	309
All	-0.0592	0.1788	0.6509	0.06125	14.28	0.000053	642

The negative signal from sentiment score matches with the negative return and higher volatility. The equity curve in Figure 6.7 clearly demonstrates the value of the strategy.

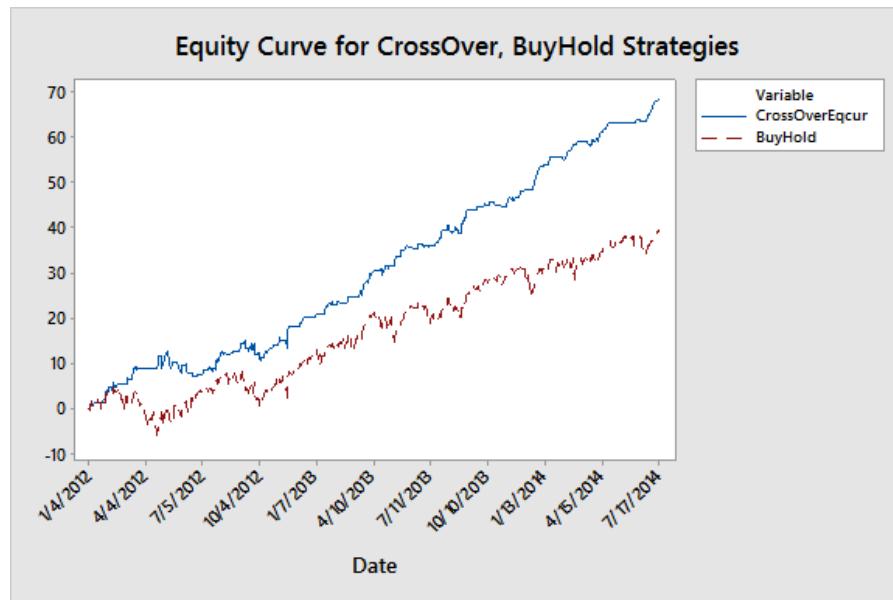


Figure 6.7: Equity Curve for ESMini.

**Bloomberg News and Social Sentiment Data:** The news agency applies statistical machine-learning techniques to process the textual information related to equities and quantifies the sentiment both at the story-level and at the equity-level. The scores indicate at the story-level if the sentiment is positive, negative, or neutral with a confidence level; these are aggregated to provide the company-level sentiment. The computation is based on both news and Twitter feeds with a rolling window of half-an-hour. The predictive strength is tested out using the performance of the equity at a

daily level as well as through changes in the stock price during the day. The approach is similar to cross-sectional momentum strategy that is outlined in Chapter 3. It can be shown that the stocks in the top decile of the day's sentiment scores tend to do better during that day.

The sentiment data also can be used to predict the nature of earnings report. The event driven strategies are still a significant part of trading strategies and it can be shown that sentiment scores prior to earnings releases can augment these strategies. The intraday strategy is based on using the arrival of information during an interval of pre-defined duration and making trading decisions based on the sentiment scores and their confidence. Generally traditional news combining with a twitter feed can generate stronger signals. A methodology used to relate sentiment scores whether from traditional news or from twitter is to first construct an index,

$$I_t = \ln \left( \frac{1 + \|B_{u_t}\|}{1 + \|B_{e_t}\|} \right) \quad (6.8)$$

that adjusts for the baseline activities, here  $B_{u_t}$  denotes the number of bullish tweets and  $B_{e_t}$  denotes the number of bearish tweets. Defining the intraday return,  $R_t = \ln(s_t^{\text{close}}) - \ln(s_t^{\text{open}})$ , where  $s_t$  refers to the stock price on a given time unit. Then forming,  $y_t = (I_t^N, I_t^{TW}, R_t)'$ , VAR model that was discussed in Chapter 2 is used to identify the lead-lag relationships; here  $I_t^N$  is the index for the news sentiment, such as Google search volume for relevant keywords and  $I_t^{TW}$  refers to the index based on Twitter data. The VAR model provides a neat framework to consider other information such as market volume, market volatility, etc. that can be easily incorporated.

Thomson Reuters MarkerPsych Indices (TRMI): The TRMI are also based on news and social media that produce three key types of indicators: Emotional indicators, Macroeconomic metrics and the so called Buzz metrics. These are updated on a minute basis and covers various economic entities. The indices cover over fifty items and they are scored from -1 to 1 or from zero to 1.

---

## 6.5 Discussion / Future of Social Media News



# 7

---

## *Active Portfolio Management and Dynamic Investment Strategies*

---

### CONTENTS

7.1	Introduction to Modern Portfolio Theory .....	175
7.1.1	Mean-Variance Portfolio Theory .....	176
7.1.2	Multifactor Models .....	179
7.1.3	Tests Related to CAPM and APT .....	179
7.1.4	An Illustrative Example .....	182
7.1.5	Summary of Models and Implications for Investing .....	184
7.2	Statistical Underpinnings .....	187
7.2.1	Portfolio Allocation Using Regularization .....	191
7.2.2	Portfolio Strategies: Some General Findings .....	192
7.3	Dynamic Portfolio Selection .....	195
7.4	Portfolio Tracking and Rebalancing .....	197
7.5	Transaction Costs, Shorting and Liquidity Constraints .....	199
7.6	Alphas, Betas, Information Ratios and Realized Returns .....	202
7.7	Supplements and Problems .....	202

---

### 7.1 Introduction to Modern Portfolio Theory

Of fundamental interest to financial economists is to examine the relationship between the risk of a financial security and its return. While it is obvious that risky assets can generally yield higher returns than risk-free assets, a quantification of the tradeoff between risk and expected return was made through the development of the Capital Asset Pricing Model (CAPM) for which the groundwork was laid by Markowitz (1959) [204]. A central feature of the CAPM is that the expected return is a linear function of the risk. The risk of an asset typically is measured by the covariability between its return and that of an appropriately defined ‘market’ portfolio. Examples of expected return-risk model relationships include the Sharpe (1964) [240] and Lintner (1965) [194] CAPM, the zero-beta CAPM of Black (1972) [29], the arbitrage pricing theory (APT) due to Ross (1976) [232] and the intertemporal asset pricing model by Merton (1973) [209]. The economy-wide models developed by Sharpe, Lintner and Black are based on the work of Markowitz which assumes that investors would hold a

mean-variance efficient portfolio. The main difference between the works of Sharpe and Lintner and the work of Black is that the former assume the existence of a risk-free lending and borrowing rate whereas the latter derived a more general version of the CAPM in the absence of a risk-free rate.

In this section, we briefly review the CAPM model and the implications for empirical research in the area of portfolio construction and testing.

### 7.1.1 Mean-Variance Portfolio Theory

A portfolio which is composed of individual assets has its risk and return characteristics based on its composition and how the individual asset characteristics correlate with each other. The optimal combination is designed to produce the best balance between risk and return. For a given level of return, it will provide the lowest risk and for an acceptable level of risk, it will provide the maximum return. The locus of the combination of risk and reward that characterizes the optimal portfolios is called the “Efficient Frontier.”

We follow the same conventions as before to denote the return as  $r_t = \ln(P_t) - \ln(P_{t-1})$ ; assume we have ‘ $m$ ’ assets in the portfolio with ‘ $w_i$ ’ denoting the share value invested in asset, ‘ $i$ ’, if  $R_t = (P_t - P_{t-1})/P_{t-1}$ , the portfolio return is  $R_{pt} = \sum_{i=1}^m w_i R_{it}$  and thus  $r_t = \ln(1 + \sum_{i=1}^m w_i R_{it}) \approx \sum_{i=1}^m w_i r_{it}$ . If  $r_t$  denotes the vector of ‘ $m$ ’ asset returns and with weights stacked up as a vector,  $w$ , then  $r_{pt} = w' r_t$  resulting in  $\mu_p = E(r_{pt}) = w' \mu$  and  $\sigma_p^2 = w' \Sigma w$ , where  $\Sigma$  is the  $m \times m$  variance-covariance matrix of  $r_t$ . If the returns are uncorrelated or negatively correlated, then observe that

$$\sigma_p^2 = w' \Sigma w \leq \sum_{i=1}^m w_i \text{Var}(r_{it}) \leq \frac{v}{m} \quad (7.1)$$

where ‘ $v$ ’ is the maximum of  $\text{Var}(r_{it})$ , clearly indicating that diversification tends to reduce risk. The power of the diversification can be seen clearly if we assume the covariance matrix,  $\Sigma$ , has all variances equal and if all off-diagonal covariance elements are the same. Then,  $\sigma_p^2 = \frac{1}{n} \cdot \sigma^2 + \frac{n-1}{n} \cdot \rho \cdot \sigma^2$ . Observe that if  $\rho = 0$ ,  $\sigma_p^2 \rightarrow 0$  as  $n \rightarrow \infty$  and if  $\rho = 1$ ,  $\sigma_p^2 = \sigma^2$  that results in no benefit. When  $\rho < 0$  as shown in (7.1), the portfolio variance is less due to diversification. But it should be noted that we cannot completely eliminate portfolio risk when the correlations among the assets are positive. Observe that  $\sigma_p^2 = \frac{1}{m^2} \sum_{i=1}^m \sigma_i^2 + \frac{1}{m^2} \sum_{i \neq j} \sigma_{ij} \leq \frac{\sigma_{\max}^2}{m} + \frac{m-1}{m} \cdot A \rightarrow A$  as  $m \rightarrow \infty$ .

**Minimum Variance Portfolio:** The efficient portfolio is obtained through the constrained optimization:

$$\min_w w' \Sigma w \ni w' \mu = \mu^*, \quad w' 1 = 1 \quad (7.2)$$

Here  $1$  is the  $m \times 1$  unit vector. The solution to this can be obtained via Lagrangian function and can be found in Campbell, Lo and MacKinley (1996, Section 5.2) [51].

With  $A = 1'\Sigma^{-1}\mu$  (weighted mean),  $B = \mu'\Sigma^{-1}\mu$  ( $F$ -ratios),  $C = 1'\Sigma^{-1}1$  and  $D = BC - A^2$ , the ‘weight’ vector,  $w$ , is given as,

$$w_{\text{eff}} = \{B\Sigma^{-1}1 - A\Sigma^{-1}\mu + \mu^*(C\Sigma^{-1}\mu - A\Sigma^{-1}1)\}/D \quad (7.3)$$

and

$$\sigma_{\text{eff}}^2 = (B - 2\mu^*A + \mu^{*2}C)/D$$

The graph of the efficient frontier ( $\mu^*$ ,  $\sigma_{\text{eff}}$ ) in the mean-standard deviation space is the right side of a hyperbola (see Figure 7.1).

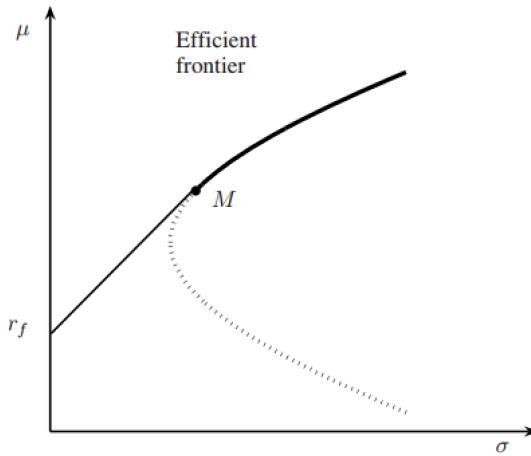


Figure 7.1: Efficient frontier with no short selling.

It must be noted that in the formulation (7.2), the weights can be negative implying that short selling is allowed. With no short selling,  $w_i \geq 0$  for all ‘ $i$ ’, there is no explicit solution but can be numerically solved via quadratic programming.

Suppose an investor is interested in a portfolio with maximum Sharpe ratio = average return/standard deviation, which represents the return per unit risk. Graphically, the portfolio is the point where a line through the origin is tangent to the efficient frontier, because this point represents the highest possible Sharpe ratio. So it is called the tangency portfolio. The inverse of the slope is obtained by setting  $\frac{\delta\sigma_{\text{eff}}^2}{\delta\mu^*} = 0$  which yields  $\mu^* = A/C$  and  $\sigma_{\text{MV}}^2 = \frac{1}{C}$  and the efficient combination is simply,  $w_{\text{MV}} = \Sigma^{-1}\mu/C$ . Thus we have described two portfolios that an investor can prefer. If minimum amount of risk is desired, the minimum variance portfolio is taken and if the objective is to maximize the Sharp ratio, the tangency portfolio is taken. These can be formulated from the point of view of unified theory of utility maximization.

Assume that the utility function is given as,

$$u = w'\mu - \frac{\lambda}{2} w'\Sigma w \quad (7.4)$$

where ‘ $\lambda$ ’ is called the parameter of absolute risk-aversion. The greater the ‘ $\lambda$ ’, the more risk averse the investor is. Usually  $2 < \lambda < 4$ . Thus assuming all investors are risk-averse, the objective function is,

$$\max_w u \ni w'1 = 1 \quad (7.5)$$

It can be easily shown that the resulting weight vector is,

$$w_{\text{eff}} = \frac{1}{\lambda} (\Sigma^{-1}\mu + \Sigma^{-1}1 \cdot \gamma^*) \quad (7.6)$$

where  $\gamma^* = (\lambda - A)/C$ . The results can be obtained again using Lagrangian multipliers and the matrix differentiation results  $\left(\frac{\delta \Sigma w}{\delta w} = \Sigma; \frac{\delta(w'\Sigma w)}{\delta w} = 2 \cdot w'\Sigma\right)$ . From (7.6) observe that it is a combination of minimum variance portfolio and the tangency portfolio. At the optimum point, note

$$\mu_{\text{opt}} = w'\mu = \frac{D}{\lambda C} + \mu_{\text{MV}} \quad \text{and} \quad (7.7)$$

$$\sigma_{\text{opt}}^2 = w'\Sigma w = \sigma_{\text{MV}}^2 + \frac{1}{\lambda^2} \cdot \frac{D}{C}$$

If an investor is fully risk-averse,  $\lambda \rightarrow \infty$  and the optimal portfolio is the minimum variance portfolio and when  $\lambda = A$ , the optimum portfolio is the tangency portfolio. Thus both are special cases of Markowitz strategy.

For realistic situations related to trading we need to consider borrowing and lending at the risk-free rate, ‘ $r_f$ ’. We assume that in addition to ‘ $m$ ’ risky assets, where investment is made, there is a possibility of putting the funds in the risk-free assets such as treasury bonds or CD’s etc. The optimization problem can be stated as,

$$\min_w w'\Sigma w \ni w'\mu + (1 - w'1)r_f = \mu^* \quad (7.8)$$

This results in efficient weights:

$$w_{\text{eff}} = \frac{\mu^* - r_f}{(\mu - r_f 1)' \Sigma^{-1} (\mu - r_f 1)} \cdot \Sigma^{-1} (\mu - r_f 1) \quad (7.9)$$

If  $w_i$ ’s are set  $\geq 0$ , there is no explicit analytic solution and the Figure 7.1 depicts the optimal allocation. When  $w_i$ ’s are unrestricted, the efficient frontier based on all risky assets is given as,

$$\mu = r_f + \frac{\mu_m - r_f}{\sigma_m} \cdot \sigma \quad (7.10)$$

which is known as capital market line or the security market line; here  $\mu_m$  can be the return from S&P500 index. The form of (7.10) lends itself for empirical testing via the regression model,

$$r_{it} - r_f = \beta_i(r_{mt} - r_f) + \epsilon_{it} \quad (7.11)$$

where  $\beta_i = \text{Cov}(r_{it}, r_{mt})/\sigma_m^2$ ; note the volatility in ' $r_i$ ' can be decomposed as

$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_e^2. \quad (7.12)$$

If  $\beta_i > 1$ , the asset is taken to be aggressive. The commonly used performance indices follow from the population version of the model (7.11) which is  $\mu - r_f = \alpha + \beta(\mu_m - r_f)$ . Sharpe index is given as  $(\mu - r_f)/\sigma$ , Treynor's index is  $(\mu - r_f)/\beta$  and Jensen's index is simply the ' $\alpha$ '. If  $\alpha > 0$ , it is taken that the portfolio performs better than the market after adjusting for the risk.

The CAPM model is empirically tested via (7.11) with added intercept terms:

$$r_{it} - r_f = \alpha_i + \beta_i(r_{mt} - r_f) + \epsilon_{it} \quad (7.13)$$

If CAPM holds then  $H_0 : \alpha_i = 0, i = 1, 2, \dots, m$  should be supported. But empirical evidence does not seem to support  $H_0$ ; various explanations are possible. It is argued that the proxies such as S&P500 index are not theoretically true market portfolio which is supposedly based on all assets. It has been shown that other asset related information provide better explanatory power for the cross-sectional variation of the asset returns. This has led to multi-factor models that is discussed next.

### 7.1.2 Multifactor Models

Ross (1976) [232] introduced the Arbitrage Pricing Theory (APT) which is more general than CAPM; the cross-sectional variation in the asset returns, it is postulated may be due to multiple risk factors. There is no need to identify a market portfolio nor the risk-free return. The model can be written as:

$$r_{it} = \alpha_i + \beta'_i f_t + \epsilon_t \quad (7.14)$$

and given  $f_t$  is a  $n \times 1$  vector of factors,  $E(\epsilon_{it}) = 0$ ,  $E(\epsilon_{it}\epsilon_{jt}) = \sigma_{ij}, i, j = 1, 2, \dots, m$  and  $t = 1, 2, \dots, T$ . In the absence of arbitrage, Ross (1976) [232] shows that the following relationship should hold:

$$\mu_i \sim r_f + \beta'_i \lambda_k \quad (7.15)$$

where  $\lambda_n$  is a  $n \times 1$  vector of factor risk premia.

Generally, there are two approaches to empirically model (7.14) and test if (7.15) holds. In the first approach, it is assumed that the factors ( $f_t$ ) are unknown and are estimated from the  $(m \times m)$  covariance matrix of the returns via principal components. The number of factors,  $n$ , is generally determined via subjective considerations. In the second approach, it is taken that the factors are determined by macroeconomic variables that reflect systematic risk and by asset characteristics. Chen, Roll and Ross (1986) [59] consider expected inflation, spread between high and low grade bond yields, spread between long and short interest rates for U.S. government bonds etc. as known factors. The asset characteristics that are noted to be useful are: ratio of book to market value, price-earnings ratio, momentum, Farma-French factors. These factor models with known or unknown factors tend to fare better than CAPM. As we will demonstrate, they are also useful for the construction of the portfolios as these factors are likely to represent different risk dimensions.

### 7.1.3 Tests Related to CAPM and APT

To test if any particular portfolio is ex ante mean-variance efficient, Gibbons, Ross and Shanker (1989) [134] provide a multivariate test statistic and study its small sample properties under both null and alternative hypotheses. This follows from the multivariate regression model and the associated test procedures discussed in the last section. Recall the null hypothesis is:

$$H_0 : \alpha_i = 0 \quad \forall i = 1, 2, \dots, m \quad (7.16)$$

The test statistic is a multiple of Hotelling's  $T^2$  stated as,

$$F = \frac{T - m - 1}{m} \cdot \frac{\hat{\alpha}' \hat{\Sigma}_{\epsilon\epsilon}^{-1} \hat{\alpha}}{1 + \hat{\theta}_p^2} \sim F(m, T - m - 1) \quad (7.17)$$

where  $\hat{\theta}_p = \bar{r}_p / s_p$ , the ratio of sample average and standard deviation of  $r_{pt}$ . The noncentrality parameter depends on  $\alpha' \hat{\Sigma}_{\epsilon\epsilon}^{-1} \alpha$  which is zero under  $H_0$ . The statistical power of the  $F$ -test thus can be studied. The test statistic in (7.17) can be easily derived from the results on multivariate regression by setting  $F_1 = I_m$  and  $G_2 = \gamma^* = (1, 0)'$ .

The constrained ML estimator of  $C$  under (7.16) is  $\hat{C} = \tilde{C} \left[ I_2 - \frac{\gamma^* \gamma^{*'} (x \gamma')^{-1}}{\gamma^{*'} (x \gamma')^{-1} \gamma^*} \right]$ . Thus we can compute the residual covariance matrices both under the null and alternative hypotheses to arrive at the LR statistic.

A practical question that needs to be resolved is on the choice of 'm' and 'T'. With over 5000 stocks listed in NYSE and with daily data available since 1926, the choice is restricted only by the condition  $T > m$  for the variance-covariance matrix in the estimation of the model (7.13) to be non-singular. Gibbons et al. (1989) [134] suggest 'm' should be roughly a third to one half of 'T'. Then it is also important to decide which 'm' assets are to be chosen. One suggested approach is to use beta-sorted assets but it may not necessarily give better power to the  $F$ -test.

Note that  $H_0 : \alpha = 0$ , where  $\alpha' = (\alpha_1, \dots, \alpha_m)$  is violated if and only if some portfolio of the assets considered in the regression model (7.13) have non-zero intercept. If we write  $\hat{\alpha}$  as the  $m \times 1$  vector of estimates of the intercepts, then for a linear combination of  $a' r_t$ , where  $r_t$  is the  $m \times 1$  vector of asset returns, [observe that if we let  $Y_t = r_t - r_f 1$  and  $X_t = [1, r_{mt} - r_f]' = [1, x_t]', a' Y_t = a' \alpha + (a' \beta) x_t + \epsilon_t$ , and thus]  $\text{Var}(a' \hat{\alpha}) = (1 + \hat{\theta}_p^2) \frac{a' \hat{\Sigma} a}{T}$  and hence

$$t_a^2 = \frac{T(a' \hat{\alpha})^2}{(1 + \hat{\theta}_p^2)(a' \hat{\Sigma} a)} \quad (7.18)$$

Maximizing  $t_a^2$  is therefore equivalent to minimizing  $a' \hat{\Sigma} a$  subject to  $a' \hat{\alpha} = C$  as in (7.1), this is equivalent to the portfolio minimization problem. The solution is

$$a = \frac{C}{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha}} \cdot \hat{\Sigma}^{-1} \hat{\alpha} \quad (7.19)$$

and thus  $t_a^2 = T \cdot \frac{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha}}{1 + \hat{\theta}_p^2} = \frac{T}{T-m-1} \cdot m \cdot F$ . Thus ‘ $a$ ’ can provide information about creating a better model. The portfolio based on ‘ $a$ ’ can provide information about creating a better model. The portfolio based on ‘ $a$ ’ is termed as an active portfolio and Gibbons et al. show that when this is combined with the ‘market’ portfolio series in ex post efficient portfolio. To make the comparison between the three portfolios, average returns for all three are set equal which results in the choice of  $C = \bar{r}_m \cdot \frac{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha}}{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\gamma}}$  and thus the active portfolio results in

$$a = \bar{r}_m \cdot \frac{\hat{\Sigma}^{-1} \hat{\alpha}}{\hat{\alpha}' \hat{\Sigma}^{-1} \hat{\gamma}} \quad (7.20)$$

which can be compared with  $w_{\text{eff}}$  in (7.3). It must be noted that the weights obtained in (7.3) are not in reference to any other portfolio and not adjusted for any comparison with other portfolios. Thus one could expect that the two portfolios would be the same only when the assets selected are not correlated to market index or to any benchmark portfolio.

The test for APT model for known factors in (7.14) also follows easily from the multivariate regression results discussed in the last section. The model can be restated as

$$r_t = \alpha + B f_t + \epsilon_t \quad (7.21)$$

If  $r = [r_1, \dots, r_T]$  and  $f = [f_1, \dots, f_T]$  are  $m \times T$  and  $n \times T$  data matrices, then

$$\hat{\alpha} = (r M_f 1_T)(1'_T M_f 1_T) \quad (7.22)$$

where  $M_f = I_T - f'(ff')^{-1}f$  is a  $T \times T$  idempotent matrix ( $M_f * M_f = M_f$ ). The null hypothesis,  $H_0 : \alpha = 0$  is tested through Jensen’s test:

$$F = \left( \frac{T-m-n}{m} \right) (1 + \bar{f}' \hat{\Omega}^{-1} \bar{f})^{-1} \hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha} \sim F(m, T-m-n) \quad (7.23)$$

where  $\bar{f} = \frac{1}{T} \sum_{t=1}^T f_t$ ,  $\hat{\Omega} = \frac{1}{T} \sum_{t=1}^T (f_t - \bar{f})(f_t - \bar{f})'$  and  $\hat{\Sigma}$  is the residual covariance matrix. Recall that the known factors could be macroeconomic variables or the financial variables such as Fama and French factors discussed in Section 4.1.

The key tool for estimation of unknown factors is Principal Component Analysis (PCA). The main difference between Factor Analysis (FA) and PCA is that the former is meant to capture the covariances among the returns,  $r_t$  whereas PCA’s focus is on capturing the variances or volatilities of  $r_t$ . Assume that  $\text{Var}(r_t) = \Sigma_{rr}$ ,  $m \times m$  covariance matrix. Now transform the returns by  $P r_t$  where  $P$  is orthogonal,  $P' P = I_m$  and  $P \Sigma_{rr} P' = \Lambda$ , where  $\Lambda$  is the diagonal matrix. Note the transformed returns,  $Z_t = P r_t$  has  $\text{Var}(Z_t) = \Lambda$  and thus uncorrelated. The goal is to recover much of the total variance  $\sum_{i=1}^m \sigma_i^2$  with a few linear combinations of  $r_t$ .

**Result 7.1.** The solution to  $\max_w w' \Sigma w \ni w' w = 1$  is obtained from solving  $\Sigma w = \lambda w$ . Thus ‘ $w$ ’ is the eigenvector that corresponds to the largest eigenvalue of  $\Sigma$  obtained from  $|\Sigma - \lambda I| = 0$  where  $|\cdot|$  denotes the determinant of the matrix; note  $\text{tr}(\Sigma) = \sum_{i=1}^m \sigma_i^2 = \sum_{i=1}^m \lambda_i$  and the values of ‘ $\lambda_i$ ’ are disproportional such that we can approximate  $\text{tr}(\Sigma) \sim \sum_{i=1}^r \lambda_i$  where ‘ $r$ ’ is much less than ‘ $m$ ’.

In the factor model set-up given in (7.20), it is assumed that the factors  $f_t$  are random with  $E(f_t) = 0$ ,  $\text{Var}(f_t) = \Omega$  and  $E(f_t' \epsilon_t) = 0$ . This results in,

$$\begin{aligned} \text{Cov}(r_t) &= B\Omega B' + V = \Sigma_{rr} \\ \text{and} \quad \text{Cov}(r_t, f_t) &= B\Omega \end{aligned} \tag{7.24}$$

where  $\Omega$  and  $V$  are taken to be diagonal; note because  $r_t \sim N(\alpha, \Sigma_{rr})$  the generalized least-squares estimates are given as:

$$\begin{aligned} \hat{f}_t &= (\hat{B}\hat{V}^{-1}\hat{B})^{-1}\hat{B}'\hat{V}^{-1}r_t \\ \text{and} \quad \hat{\alpha} &= \bar{r} - \hat{B}\bar{f} \end{aligned} \tag{7.25}$$

From (7.25), it follows that  $\text{Var}(\hat{f}_t) = (B'\bar{V}^{-1}B)^{-1}/T$  and  $\text{Var}(\hat{\alpha}) = \frac{1}{T}(V - B(B'V^{-1}B)^{-1}B')$ . To decide on the number of factors, the following test statistics is suggested

$$\begin{aligned} \chi^2 &= -\left((T-1) - \frac{2r+5}{6} - \frac{1}{3}r\right)(\log|\hat{\Sigma}_{rr}| - \log|\hat{B}\hat{B}^{-1} + \hat{V}|) \\ &\sim \chi^2_{\{(m-r)^2-m-r]/2\}} \end{aligned} \tag{7.26}$$

The number of adequate factors can also be verified through the significant ‘λ’s in Result 7.1. We will discuss an application in the next section that will clarify many of these concepts.

### 7.1.4 An Illustrative Example

We consider monthly data from July 1926 to October 2012 on returns from six portfolios that are formed based on the size and book-to-market (BM) ratio. This data is taken from French’s data library. The titles follow the naming convention, for example, SML contains small size plus low BM to BGH contains big size with high BM ratio. The data also contains  $r_f$ ,  $r_{mt} - r_f$  and the two factors SMB (small-minus-big size factors) and HML (high-minus-low BM factor).

The histogram of the portfolio returns is given in Figure 7.3. The descriptive are stated in Table 7.1. The returns are generally positively skewed with more sharply peaked than the normal distribution. The Jarque-Bera test (Equation 2.44) rejects the null hypothesis that returns are normally distributed. The efficient frontier with the risk-free rate using the weights in (7.9) along with the portfolios is plotted in Figure 7.3, by setting  $\mu^* = \hat{\mu}_m$ , the average of the market portfolio. The weights for the assets are  $(-1.06, 1.08, 0.40, 0.58, -0.36, -0.32)$  and the share of the risk-free is 0.68. The scatter plot of the actual mean value versus standard deviation of the six portfolios yields the slope of 0.1471 which is larger empirical Sharpe ratio  $\left(\frac{\bar{r}_m - r_f}{\hat{\sigma}_m}\right)$ , 0.1147. This clearly indicates that the market portfolio is not mean-variance efficient.

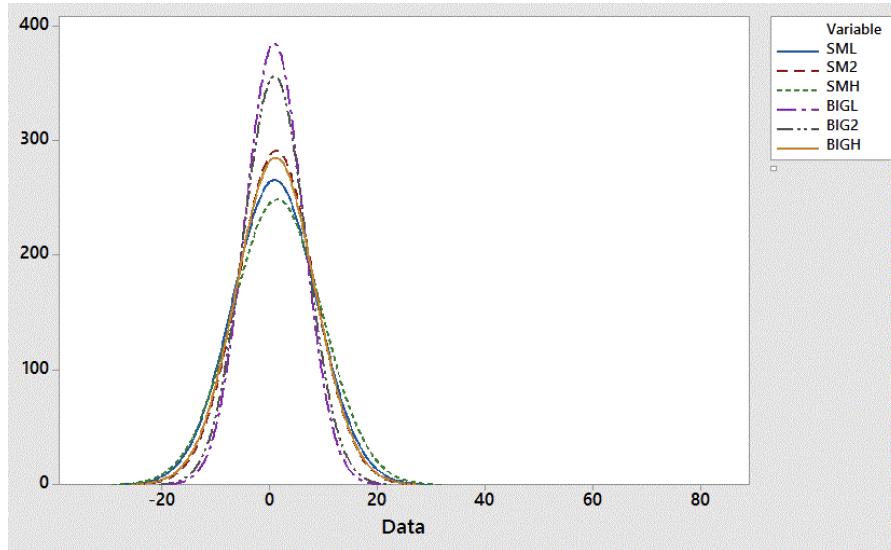


Figure 7.2: Histogram of Monthly Returns.

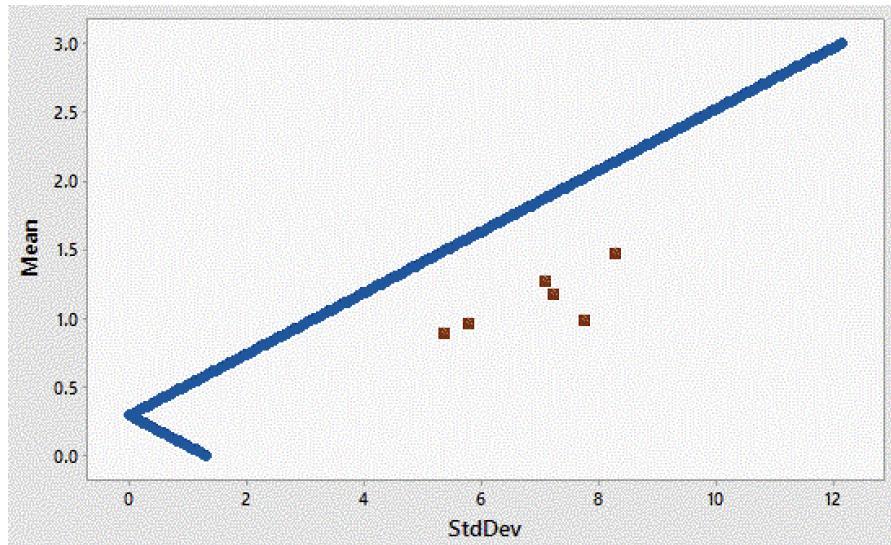


Figure 7.3: Efficient Frontier-Empirical.

Note that while three portfolios (BigL, Big2, BigH) lie on the line, two others (SM2 and SMH) lie above the line. This generally implies that it is possible to outperform the empirical Markowitz portfolio.

The three known factor model results are presented in Table 7.2. There are several

Table 7.1: Descriptive statistics ( $T = 1036$ )

Variable	Mean	StDev	Median	Skewness	Kurtosis
SML	0.983	7.771	1.100	0.98	10.23
SM2	1.271	7.092	1.530	1.27	14.51
SMH	1.472	8.303	1.660	2.13	21.64
BIGL	0.892	5.365	1.195	-0.13	5.16
BIG2	0.960	5.799	1.225	1.25	16.82
BIGH	1.173	7.243	1.395	1.53	18.05

Table 7.2: Fama French model (estimates)

Portfolio	Const	Market	SMB	HML	$R^2$
SML	-0.168*	1.09*	1.05*	-0.167*	0.974
SM2	0.0625	0.977*	0.818*	0.702*	0.978
SMH	0.0202	1.03*	0.933*	0.783*	0.991
BIGL	0.077*	1.02*	-0.0965*	-0.231*	0.978
BIG2	-0.0507	0.966*	-0.122*	0.729*	0.954
BIGH	-0.112*	1.079*	0.0205	0.819*	0.968

conclusions that can be drawn from this table. With the three factors, the variations in the cross-sectional returns are well-captured by these factors. The coefficients of the market factors are closer to unity indicating that the six portfolios are closely aligned with the market index. But the estimates of  $\alpha$  do indicate some mixed results. The value of the test statistics in (7.23) is 5.54 which is larger than  $\chi^2_6(0.05) = 2.1$  and hence indicating that there should be other factors.

To identify the number of factors required to capture the total variance, we perform PCA of the covariance matrix. The Scree plot is given in Figure 7.5. The first component is responsible for 91% of the total variance and if three components are included we can capture 98% of the overall variance. The first factor trading is almost proportional to unit vector indicating that it can be interpreted as a market factor. It is easy to show with three factor model, the difference between  $\hat{\Sigma}$  and  $\hat{B}\hat{B}' + V$  is very small. This sort of parameter reduction will become useful when we deal with large number of assets, a topic that is taken up in the next section.

### 7.1.5 Summary of Models and Implications for Investing

As the focus of this book is on trading and investing we want to summarize some main takeaways from the widely used theoretical models. Some implications of these models are listed below:

**One-fund Theorem:** There is a single investment of risky assets such that any efficient portfolio can be constructed as a combination of this fund and the risk-free asset. The optimal one-fund is the market fund which implies that the investor should purchase every stock which is not practical; this has given rise to the so-called index

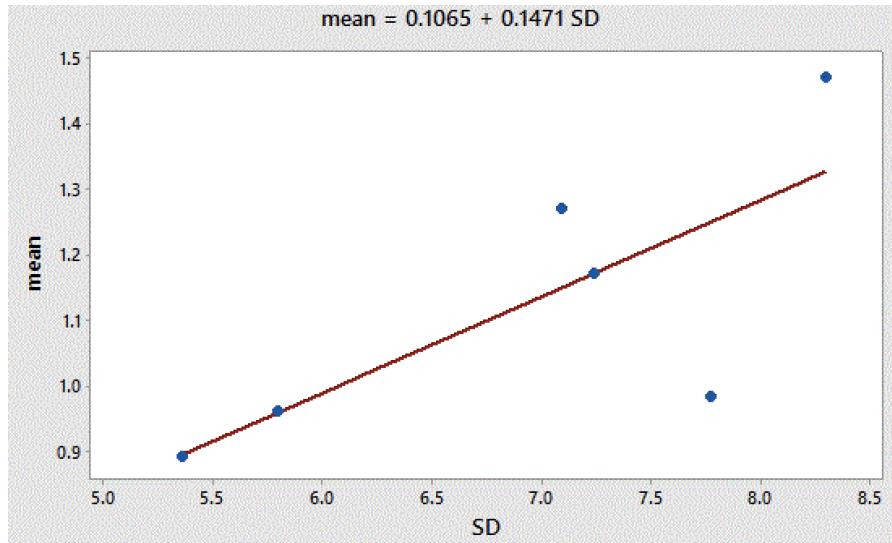


Figure 7.4: Capital Market Line.

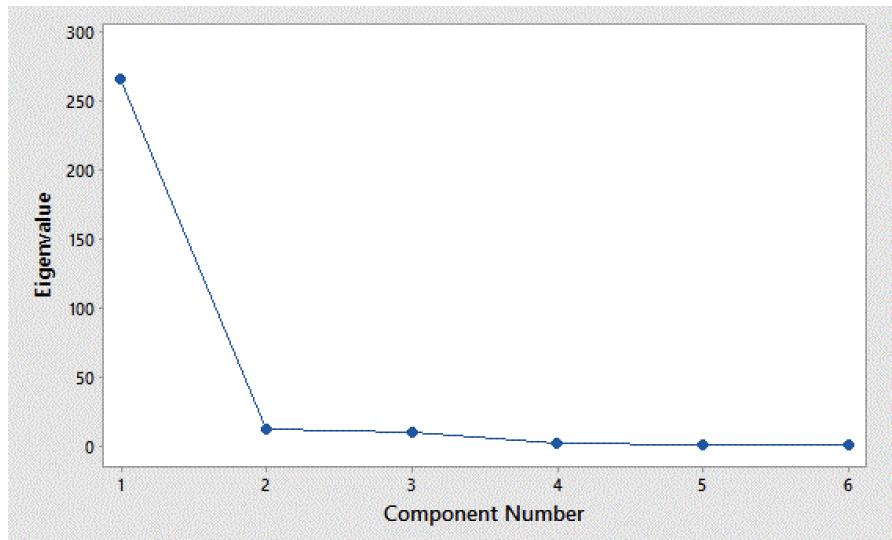


Figure 7.5: Scree Plot.

funds.

**Two-fund Theorem:** Two efficient portfolios (funds) can be established so that any efficient can be duplicated; all investor seeking efficient portfolios need only to invest

in combinations of these two funds.

**No short-selling:** When the weights,  $w_i$ , are restricted to be positive, typically many of them tend to be zeros, by contrast with short-selling, most of the weights tend to be non-zeros. The resulting concentrated portfolios are more likely to have higher turnover rates. This has implications for the increase in trading costs.

**CAPM:** Recall the model is  $r_i - r_f = \beta_i(r_m - r_f) + \epsilon_i$  with  $\sigma_i^2 = \beta_i^2\sigma_m^2 + \sigma_\epsilon^2$ ;  $\beta_i = \frac{\sigma_{im}}{\sigma_m^2}$  is the normalized version of covariance. If  $\beta_i = 0$ ,  $r_i \sim r_f$  even if  $\sigma_i^2$  is large. There is no risk premium; the risk that is uncorrelated with the market can be diversified away. if  $\beta_i < 0$ ,  $r_i < r_f$  even if  $\sigma_i^2$  is large; such an asset reduces the overall portfolio risk when it is combined with the market portfolio; they provide a form of insurance. Portfolio ‘Beta’ is  $\beta_p = \sum_{i=1}^m w_i \beta_i$ . The CAPM can be used as a pricing formula. Note if the purchase price of an asset is ‘ $P$ ’ (known) which may be sold at ‘ $Q$ ’ (unknown), from CAPM,  $r = \frac{Q-P}{P} = r_f + \beta(r_m - r_f)$  which results in the pricing formula

$$P = \frac{Q}{1 + r_f + \beta(r_m - r_f)} \quad (7.27)$$

This concept to evaluate a single asset can be extended to jointly evaluating multiple assets.

**Investment Pattern:** The minimum-variance portfolio method is likely to invest into low residual risk and low beta stocks. The CAPM model, (7.11) written in the vector form leads to

$$r_t - r_f 1 = \alpha + \beta(r_m - r_f) + \epsilon_t \quad (7.28)$$

and with the assumption that the error-covariance matrix is diagonal, the covariance matrix of the excess returns is

$$\Sigma_{rr} = \beta \beta' \sigma_m^2 + D \quad (7.29)$$

where  $D$  is diagonal with the element,  $\sigma_{ei}^2$ . The inverse of the covariance matrix has a simpler structure:

$$\Sigma_{rr}^{-1} = D^{-1} - \frac{\sigma_m^2}{1 + a} \cdot bb' \quad (7.30)$$

where  $a = \sigma_m^2 \cdot \sum_{i=1}^m b_i \beta_i$  and  $b_i = \beta_i / \sigma_i^2$ . We have shown earlier that the portfolio weights under minimum-variance is

$$w = \frac{\Sigma_{rr}^{-1} 1}{1' \Sigma_{rr}^{-1} 1} \quad (7.31)$$

and

$$\sigma_{MV}^2 = \frac{1}{1' \Sigma_{rr}^{-1} 1}$$

Substituting (7.30) in (7.31), notice that

$$w = \sigma_{\text{MV}}^2 \left( D^{-1} \mathbf{1} - \frac{\sigma_m^2}{1+a} \cdot b b' \mathbf{1} \right) \quad (7.32)$$

with a typical element,  $w_j = \frac{\sigma_{\text{MV}}^2}{\sigma_j^2} \left( 1 - \beta_j \left( \frac{\sigma_m^2}{1+a} \cdot \sum_{i=1}^m b_i \right) \right)$ . It is easy to show that the second term in parenthesis above,  $\frac{\sigma_m^2}{1+a} \cdot \sum_{i=1}^m b_i \sim 1$  and so

$$w_j = \frac{\sigma_{\text{MV}}^2}{\sigma_j^2} (1 - \beta_j) \quad (7.33)$$

Thus when  $\sigma_j^2$  is small and low  $\beta_j$ ,  $w_j$  will be large.

## 7.2 Statistical Underpinnings

The implementation of the portfolio theory in practice when the means and covariances of returns are unknown has raised some challenging issues. It is a theory that applies to a single period and the choice of portfolio weights that would determine the portfolio returns for a future period. Because of uncertainty involved in the future returns the use of the estimates of the means and covariances of the returns based on the past data must be carefully evaluated. The sample average and the sample covariance matrix which are also maximum likelihood estimates under the assumptions that returns are normally distributed when plugged in, the main thesis that the combination of the assets would yield efficient portfolio does not seem to hold empirically.

The following numerical illustration taken from Bruder, Gausse, Richard and Roncalli (2013) [49] is quite telling. Consider the universe of four assets with the average return vector and the covariance matrix as given below.

$$\bar{r}' = (0.05, 0.06, 0.07, 0.08)$$

$$\hat{\Sigma} = \begin{pmatrix} 0.01 & 0.0084 & 0.0098 & 0.0105 \\ & 0.0144 & 0.01176 & 0.0126 \\ & & 0.0156 & 0.0147 \\ & & & 0.0225 \end{pmatrix} \quad (7.34)$$

If the risk tolerance parameter ‘ $\lambda$ ’ is set at 0.1, then the resulting optimal portfolio has the following weights:  $w^* = (0.2349, 0.1957, 0.1878, 0.2844)$ . If the standard deviation (volatility) of the second asset increases by 3%, the weight attached to this asset changes to  $-0.1404$  and if the return of the first asset changes from 5% to 6%, the optimal weight goes to 0.6319 from 0.2349. The sensitivity of the efficient portfolio weights even to minor perturbations is quite obvious.

This may be due to a number of factors. We have to keep in mind that the performance of the portfolio is related to future means and covariances of the returns. As seen from the plot of the six portfolio returns, Figure 7.2, the distributions may not follow normal distributions. There are portfolios which are linear combinations of several individual assets and therefore individual assets can even more easily deviate from normality. Secondly, this could be result of curse of dimensionality. When ‘ $m$ ’ the number of assets is fairly large, the number of unknown parameters,  $m(m + 1)$  is also large and to estimate them accurately it requires to have a fairly large data. An extreme example would be tracking Russell index with 3000 stocks would require estimating 4.5 million covariances. Another aspect that needs to be looked into is regarding the time dependence of the returns. Under efficient market hypothesis, returns are expected to be independent; but as seen in previous chapters that there is some dependence, an anomaly exploited by the algorithms, need to be taken into account in the portfolio construction. We address these issues in this section.

It has been shown that because of the large sampling error of the estimates, an equally weighted portfolios can perform closer to the efficient portfolio. Studies have incorporated the estimation risk via Bayesian approach and evaluated the impact of this risk on final choice of the portfolios. The exact distribution of the maximum likelihood estimators of the means and variance is well-known, but the exact distribution of the returns and the volatility of the efficient portfolio is not easy to obtain. Jobson and Korkie (1980) [172] derive the asymptotic distribution and examine its properties via extensive simulations. The convergence of the sample estimates is shown to depend upon the magnitude of  $T$  to  $1/A^2$ , where  $A = 1'\Sigma^{-1}\mu$ , the sum of weighted averages. Generally if the covariances among the assets are relatively small but the mean returns are large, the estimators tend to perform better. In any case, good estimates of ‘ $\mu$ ’ and ‘ $\Sigma$ ’ are essential for the construction of the efficient portfolio. Michaud (1989) [210] suggests bootstrap methods to repeatedly resample with replacement from the observed sample of returns  $(r_1, \dots, r_m)$  and use the resulting average and the variance as the the bootstrap replications.

**Shrinkage and Bayesian Methods:** The shrinkage methods follow the same optimization criterion but impose a penalty on the size of the coefficients. For example, we can add following additional constraints to (7.2):

$$\begin{aligned} \text{Ridge: } & \sum_{i=1}^m w_i^2 \leq w \\ \text{Lasso: } & \sum_{i=1}^m |w_i| \leq w \end{aligned} \tag{7.35}$$

Both procedures can be unified as,  $\sum_{i=1}^m |w_i|^q \leq w$  for  $1 \leq q \leq 2$ . As it has been shown that these methods yield estimates that are Bayes estimates with different priors, we will now discuss Bayesian approach.

The Bayesian approach combines the prior distribution and the likelihood to obtain posterior distribution of the unknown parameters. Usually the mode of this distribution is taken as a reasonable estimate. In a influential paper, Black and Litterman

(1992) [30] describe the need for Bayesian approach. The straight solution to (7.2) almost always results in large short positions in many assets which would increase the trading costs. If the weights are restricted to be positive the solution results in zero weights for many assets and large weights in the assets with small capitalizations. The results generally are not reasonable mainly due to poor estimate of expected future returns. In practice, investors augment their views using their prior experience and their intuition to forecast the likely returns based on other asset related data. The Bayes Theorem in a natural way combines these two aspects. The details that follow are from Lai, Xin and Chen (2011) [185] and we will comment on the practical implementation issues after presenting mathematical results.

The model begins with the following distributions:

$$r_t \sim N(\mu, \Sigma), \quad \mu | \Sigma \sim N(v, \Sigma/\kappa) \text{ and } \Sigma \sim IW_m(\psi, n_0) \quad (7.36)$$

where  $IW_m(\psi, n_0)$  is the inverted Wishart with ' $n_0$ ' degrees of freedom with  $E(\Sigma) = \psi/(n_0 - m - 1)$ . These prior are known as conjugate priors as their functional forms closely match the likelihood functional form. When the degree of freedom ' $\kappa$ ' goes to zero, the prior is non-informative. The posterior means are:

$$\begin{aligned} \hat{\mu} &= \frac{\kappa}{T + \kappa} \cdot v + \frac{T}{T + \kappa} \bar{r} \\ \hat{\Sigma} &= \frac{\psi}{T + n_0 - m - 1} + \frac{T}{T + n_0 - m - 1} \cdot \left\{ s + \frac{\kappa}{T + \kappa} (\bar{r} - v)(\bar{r} - v)' \right\} \end{aligned} \quad (7.37)$$

where  $S = \frac{1}{T} \sum_{t=1}^T (r_t - \bar{r})(r_t - \bar{r})'$ , the residual covariance matrix. If ' $\kappa$ ' and ' $n_0$ ' are taken to be large, it implies that the investor puts more emphasis on priors. To use (7.37), the investor has to specify these values.

Black and Litterman (1992) [30] propose shrinking the prior estimate of  $\mu$  to a vector ' $\pi$ ' representing excess equilibrium returns such as the one implied by CAPM which is termed as the return on the global market. Thus

$$\pi = \frac{\mu_m - r_f}{\sigma_m^2} \cdot \Sigma \cdot w_m \quad (7.38)$$

where  $w_m$  are the weights on the global market and  $\sigma_m^2$  is the variance of the rate of return on the global market. To get the posterior distribution of  $\mu | \pi$  via Bayes Theorem, as mentioned in Satchell and Snowcraft (2000) [234], the approach by Black and Litterman (1992) [30] was to place this in a tractable form that investors can operationalize. Assume that  $k \times m$  matrix  $P$  represents investor's beliefs, then  $P\mu \sim N(q, \Omega)$ , where ' $\Omega$ ' is diagonal. The parameters ' $q$ ' and ' $\Omega$ ' are known as hyperparameters. Also it is assumed that  $\pi | \mu \sim N(\mu, \tau\Sigma)$ , where  $\tau$  is a scaling factor. Thus the p.d.f. of ' $\mu$ ' is given as  $\mu | \pi \sim N(\hat{\mu}^{BL}, \hat{\Sigma}^{BL})$ , where

$$\begin{aligned} \hat{\mu}^{BL} &= (\hat{\Sigma}^{BL})^{-1}[(\tau\Sigma)^{-1}\pi + P'\Omega^{-1}q] \\ \text{and} \quad \hat{\Sigma}^{BL} &= (\tau\Sigma)^{-1} + P'\Omega^{-1}P \end{aligned} \quad (7.39)$$

For various interesting practical applications, see Satchell and Snowcraft (2000) [234]. But the requirement for subjective judgements makes it difficult to use Black-Litterman model in asset management practice.

It has been noted in the literature that estimating the expected returns using the past returns fails to account for likely changes in the level of market risk. Merton (1980) [208] based on a detailed exploratory study concludes that the investors should recognize that the expected market return in excess of risk free return should be positive and that there is heteroscedasticity in the realized returns. It must be noted that the price series is random walk and therefore with shifting means, the successive differences (essentially the returns) can exhibit different levels of variances. Recent focus has been on the improved estimation of the covariance matrix,  $\Sigma$ . As the sample covariance matrix with  $m(m + 1)/2$  elements requires substantially large ' $T$ ' for estimation, alternatives for a structural covariance matrix with fewer parameters are proposed. There are all in the same spirit of shrinkage estimators:

$$\hat{\Sigma} = \alpha s_0 + (1 - \alpha)s \quad (7.40)$$

where ' $\alpha$ ' is called a shrinkage constant and ' $s_0$ ' matrix is restricted to have only a few 'freer' elements. With the CAPM model that leads to covariance structure, (7.29), the resulting covariance matrix which is the sum of a diagonal matrix and a matrix of rank one, will have only ' $2m$ ' independent elements that need to be estimated. Ledoit and Wolf (2003) [187] describe a procedure to arrive at the estimate of  $\alpha$  in (7.40) which we briefly describe below.

Minimizing the following risk function,

$$R(\alpha) = \sum_{i=1}^m \sum_{j=1}^m E[\alpha s_{ij}^0 + (1 - \alpha)s_{ij} - \sigma_{ij}]^2 \quad (7.41)$$

we arrive at ' $\alpha$ ' as

$$\alpha^* = \frac{\sum_i \sum_j [\text{Var}(s_{ij}) - \text{Cov}(s_{ij}^0, s_{ij})]}{\sum_i \sum_j [\text{Var}(s_{ij}^0 - s_{ij}) + (\phi_{ij} - \sigma_{ij})^2]} \quad (7.42)$$

where  $\phi_{ij} = E[\alpha s_{ij}^0 + (1 - \alpha)s_{ij}]$ . It is further shown that

$$\alpha^* \sim \frac{1}{T} \cdot \frac{\pi - \rho}{\gamma} \quad (7.43)$$

where  $\pi = \sum_i \sum_j \text{Asy Var}(\sqrt{T}s_{ij})$ ,  $\rho = \sum_i \sum_j \text{Asy Cov}(\sqrt{T}s_{ij}^0, s_{ij})$  and  $\gamma = \sum_i \sum_j (\phi_{ij} - \sigma_{ij})^2$ . Note that with CAPM specification which can be taken as a single index model, the ' $\gamma$ ' quantity represents the misspecification of the single index.

Fan, Fan and Lv (2008) [110] consider high dimensional covariance matrix estimation via factor model (7.14). The multifactor model generally captures well the cross-sectional variations and thus the number of parameters that need to be estimated can be significantly reduced. It is possible that the number of factors ' $n$ ' can grow with

the number of assets, ‘ $m$ ’, and the sample period, ‘ $T$ ’. Comparing  $\hat{\Sigma}_{rr} = \hat{B}\hat{\Omega}\hat{B}' + \hat{V}$  in (7.24) with the sample covariance matrix of the returns, ‘ $S$ ’, it is shown that the former performs better if the underlying factor model is true. In calculations that involve the inverse of the covariance matrix, such as the calculation of portfolio weights, factor model based estimates tend to perform better but it does not make a difference in the actual risk calculation ( $\hat{w}'\hat{\Sigma}\hat{w}$ ), which involves the estimate of the covariance matrix. Recent studies have focused on proving bounds on the risk; the calculations are somewhat complicated as they involve estimates of the weight vector as well as the estimates of the covariance matrix. See Fan, Han, Liu and Vickers (2016) [111] and the references therein.

### 7.2.1 Portfolio Allocation Using Regularization

It is noted that the poor performance of Markowitz’s conceptual framework is due to the structure of the optimization problem. It is an ill-conditioned inverse problem as the solution requires the inverse of the covariance matrix that may involve highly correlated securities. A number of regularization procedures have been proposed in the literature to address the instability in the solution and we present a few of them here.

**The L<sub>2</sub> Constraint:** This is also known as ridge constraint that imposes the penalty on the size of the estimates as given in (7.35). The problem is defined as follows:

$$\min_w w'\Sigma w \ni w'\mu = \mu^*, w'\mu = 1 \text{ and } \sum_{i=1}^m w_i^2 \leq w_0 \quad (7.44)$$

when there are many correlated assets, the weights are poorly determined and can exhibit high variance. A large positive weight for an asset can be canceled by a large negative weight for the correlated asset. The size constraint added in (7.44) can address this issue. If the added constraint has the associated Lagrangian multiplier,  $\gamma^*$ , the solution in (7.6) is modified by replacing  $\Sigma$  by  $(\Sigma + \gamma^* I_m)$ . The solution adds a positive constant to smaller eigenvalues of possibly a singular matrix,  $\Sigma$ . Observe that the eigen decomposition of  $\Sigma = V\Lambda V'$  essentially gives us the principal components of the asset universe considered for the formation of the portfolio. The optimization can also be extended to accommodating to mimic a target portfolio with weights  $w^*$ . The set-up would be the same except for the last size constraint is replaced by

$$(w - w^*)' A (w - w^*) \leq w_0 \quad (7.45)$$

If ‘ $\lambda$ ’ and ‘ $\gamma$ ’ are the Lagrangian multipliers associated with the mean specification constraint in (7.45) and the constraint above, the solution is

$$\hat{w}(\lambda, \gamma) = (\hat{\Sigma} + \gamma A)^{-1}(\lambda\hat{\mu} + \gamma Aw_0) \quad (7.46)$$

which results from shrinking both the mean vector and the covariance matrix.

**The L<sub>1</sub> Constraint:** The method popularly called as LASSO (least absolute shrinkage and selection operator) is a shrinkage method like the ridge but there are some important differences. The problem can be stated as follows:

$$\min_w w' \Sigma w \ni w' \mu = \mu^*, w' \mu = 1 \text{ and } \sum_{i=1}^m |w_i| \leq w_0 \quad (7.47)$$

The added constraint in (7.47) leads to nonlinear solutions and no explicit solution as with ridge penalty is possible. The numerical solutions are obtained via quadratic programming.

Both  $L_1$  and  $L_2$  constraint versions yield solutions that can be justified as Bayes with different priors. The LASSO method produces sparse solution that is easy to interpret. Sparser solution in the portfolio context means lower transaction costs when the portfolios are rebalanced. A convenient way to solve these optimization problems is to formulate the original set-up in (7.2) in the regression framework; this will help to borrow regularization ideas used in the general area of multivariate regression. Observe that  $\Sigma = E(r_t r_t') - \mu \mu'$  and thus the minimization problem is equivalent to (in the sample version)

$$\hat{w} = \arg \min_w \frac{1}{T} \| \mu^* 1_T - R' w \|_2 \ni w' \hat{\mu} = \mu^*, w' 1_m = 1 \quad (7.48)$$

Here  $R$  is the  $m \times T$  matrix of the row returns and  $\hat{\mu} = \frac{1}{T} \sum_{t=1}^T r_t$ . The regularization constraint in (7.47) when added to (7.48) results in the following optimization problem:

$$\hat{w}^{(\tau)} = \arg \min_w [\| \mu^* 1_T - R' w \|_2 + \tau \| w \|_1] \ni w' \hat{\mu} = \mu^*, w' 1_m = 1 \quad (7.49)$$

As stated in Brodie, Daubechies, De Mol, Giannone and Loris (2009) [47] adding the  $l_1$ -constraint results in several useful advantages. In addition to introducing sparsity that is helpful for managing large portfolios, the procedure provides a framework for regulating the amount of shorting which can be a proxy for the transaction costs. More importantly it provides a solution that is more stable which was a main concern as small perturbations in the estimates of  $\mu$  and  $\Sigma$  can lead to very different solutions.

The algorithm that is most convenient to solve (7.49) is Least Angle Regression (LAR) developed in Efron, Hastie, Johnstone and Tibshirani (2004) [95]. It starts with a large value of ‘ $\tau$ ’ and as ‘ $\tau$ ’ decreases the optimal solution,  $\hat{w}^{(\tau)}$ , changes in a piecewise linear fashion and thus it is only essential to find critical points where the slope changes. More relevant details can be found in Brodic et al. (2009) [47]. The LASSO procedure was evaluated using Fama-French 100 portfolios sorted by size and book-to-market as the universe. The efficient portfolio was constructed on a rolling basis using the monthly data for five years and its performance is evaluated in the following year. It is concluded, see Brodie et al. (2009; Figure 2) [47], that optimal sparse portfolios that allow short portfolios tend to outperform both optimal no-short-positions portfolio and the naïve evenly weighted portfolio.

### 7.2.2 Portfolio Strategies: Some General Findings

The following wisdom advocated by Rabbi Isaac bar Aha that “one should always divide his wealth into three parts: a third in land, a third in merchandise and a third ready to hand”, the equal division among the assets has been tested out in the case of equity allocation as well. Although this diversification strategy appears to be naïve because it is based on neither any theory nor any data, has done relatively well. DeMiguel, Garlappi and Uppal (2009) [81] compare this naïve strategy with other theory-based strategies that were discussed in earlier sections in terms of its performance using various data sets. The summary of their findings is worth noting: “...of the strategies from the optimizing models, there is no single strategy that dominates the  $1/N$  strategy in terms of Sharpe ratio. In general,  $1/N$  strategy has Sharpe ratios that are higher ... relative to the constrained policies which in turn have Sharpe ratios that are higher than those for the unconstrained policies.”

The naïve portfolio has also very low turnover. In addition to keeping naïve portfolio as a benchmark, a practical implication is that the estimation of the moments of asset returns needs improvement using other information based on asset characteristics.

**Combination of Strategies:** Tu and Zhou (2011) [253] combine the rules based on Markowitz Theory with the naïve strategy to obtain new portfolio rules that can perform uniformly better. The combination can be interpreted as a shrinkage estimator, shrinking toward the naïve estimator. The linear combination

$$\hat{w}_c = \delta\hat{w} + (1 - \delta)w_N \quad (7.50)$$

where  $\hat{w}$  is the weights that come from variations of efficient portfolios and  $w_N$  has all its elements equal. With the criterion of minimizing the loss function,

$$L(w^*, \hat{w}_c) = U(w^*) - E[U(\hat{w}_c)] \quad (7.51)$$

where  $U(\cdot)$  is the expected utility, the combination coefficient  $\delta$ , is obtained. This strategy was tested out using the same data as in DeMiguel et al. (2009) [81] and generally the combination rules perform well.

**Familiarity:** It has been advocated by many economists and practitioners such as Warren Buffet that investors should focus on a few stocks that they are familiar with rather than a wide scale of diversification. Boyle, Garlappi, Uppal and Wang (2012) [40] develop a theoretical framework to answer some relevant questions related to familiarity and diversification. Writing the return as sum of components, systematic and idiosyncratic,

$$r_i = r_s + u_i \quad i = 1, 2, \dots, m \quad (7.52)$$

the utility function in (7.4) is modified to take into account the investors familiarity with some assets. This is done via the assumption that the investor is ambiguous about

the mean,  $\mu$ , but has perfect knowledge about  $\sigma$ . This information alters the criterion in (7.5) as,

$$\max_w \min_{\mu} \left( w' \mu - \frac{\lambda}{2} \cdot w' \Sigma w \right) \ni \frac{(\mu - \hat{\mu})^2}{\sigma_{\hat{\mu}}^2} \leq \alpha_i^2, \quad i = 1, 2, \dots, m \quad (7.53)$$

where  $\sigma_{\hat{\mu}}^2 = (\sigma_s^2 + \sigma_u^2)/T$ . Two key quantities that play important roles are the level of ambiguity ‘ $\alpha_i$ ’ and the inverse coefficient of variation,  $\hat{\mu}/\sigma_{\hat{\mu}}$ . Some conclusions from this study are

- When ambiguity is present, the investor tends to hold disproportionate (to Markowitz allocation) amount in the familiar assets.
- When ambiguity is present, the number of assets tends to be smaller.

Huberman (2001) [162] provides several instances of investment in the familiar options rather than choices that lead to efficient outcome. Workers generally tend to choose their employers stock as they probably think that they know about their behavior and may believe they have information that the market does not have. Thus various examples are provided that indicate that although investment in the familiar contradicts the advice to diversify. This adds a dimension to the traditional risk-return formulation that one has to recognize.

**Large Portfolio Selection with Exposure Constraints:** Several techniques have been discussed in this section to reduce the sensitivity of the optimal portfolio weights to uncertainty in the inputs, such as the mean returns and the covariance matrix of returns. Generally these techniques are not adequate to address the effect of estimation errors that accumulate due to large portfolio size. Jagannathan and Ma (2003) [165] impose no short-sales constraints as well as upper bounds on portfolio weights as commonly imposed by practitioners. The empirical evidence indicates that simply imposing upper bounds does not lead to significant improvement in the out-of-sample performance. For an estimated covariance matrix,  $S$ , the formulation of the model is as follows:

$$\min_w w' S w \ni w' 1 = 1, \quad 0 \leq w_i \leq \tilde{w} \quad \text{for } i = 1, 2, \dots, m \quad (7.54)$$

Note the variation in its formulation; no expectation on the average return of the portfolio as in Markowitz set-up is specified. Fan, Zhang and Yu (2012) [112] argue that the resulting optimal portfolio, although performs well, it is not diversified enough.

Defining the total proportions of long and short positions by  $w^+ = (\|w\|_1 + 1)/2$  and  $w^- = (\|w\|_1 - 1)/2$ , we have  $w^+ + w^- = \|w\|_1$  and  $w^+ - w^- = 1$ . The optimization problem is stated as,

$$\max_w E[U(w'r)] \ni w' 1 = 1, \|w\|_1 \leq C \text{ and } Aw = a \quad (7.55)$$

Here  $U(\cdot)$  is a utility function and the extra-constraints  $Aw = a$  can capture allocations to various sectors. But the key added constraint,  $\|w\|_1 \leq C$  can accommodate

from no short sales ( $C = 1$ ) to unrestricted sales ( $C = \infty$ ) constraints. The solution to (7.55) can be obtained via quadratic programming. Fan et al. (2012) [112] advocate data-driven choice of ‘ $C$ ’ via minimizing the risk for a sample learning period. Using select 600 stocks from the pool of stocks that constitute Russell 3000, the constructed portfolios tend to perform well for  $C \sim 2$ .

### 7.3 Dynamic Portfolio Selection

The static Markowitz framework, as discussed in the last section, is not realistic for practical implication. Dynamic approaches that have been discussed in the literature will be presented here briefly. The single period optimization problem can be formulated as,

$$\max_{w_t} E_t \left[ w_t' r_{t+1} - \frac{\gamma}{2} (w_t' r_{t+1})^2 \right] \quad (7.56)$$

which results in the optimal solution,

$$w_t = w = \frac{1}{\gamma} \cdot E [r_{t+1} r_{t+1}']^{-1} E[r_{t+1}] \quad (7.57)$$

that can be estimated using the sample counterpart. Here  $r_{t+1}$  is the excess returns  $R_{t+1} - R_t^f$ . It is assumed here that the returns are independent and identically distributed. Brandt and Santa-Clara (2006) [42] extend the above set-up to multiperiods. Simply stated the two-period mean-variance objective function is,

$$\max E_t \left[ r_{t \rightarrow t+2}^p - \frac{\gamma}{2} (r_{t \rightarrow t+2}^p)^2 \right] \quad (7.58)$$

where the excess return for the two period investment strategy,

$$r_{t \rightarrow t+2}^p = (R_t^f + w_t' r_{t+1})(R_{t+1}^f + w_{t+1}' r_{t+2}) - R_t^f R_{t+1}^f \quad (7.59)$$

It can be seen that the above expression can be written as

$$r_{t \rightarrow t+2}^p = w_t' (r_{t+1} R_{t+1}^f) + w_{t+1}' (r_{t+2} R_t^f) + (w_t' r_{t+1})(w_{t+1}' r_{t+2}) \quad (7.60)$$

with the first two terms denoting excess returns from investing in the risk free asset in the first period and in the risky asset in the second period and vice versa. The third term that captures the effect of compounding is of smaller magnitude. Ignoring that term results in the optimal portfolio weights,

$$w = \frac{1}{\gamma} E \left[ r_{t \rightarrow t+2}^p r_{t \rightarrow t+2}^{ip} \right] E[r_{t \rightarrow t+2}^p] \quad (7.61)$$

The solution reflects the the choice between holdings the risky assets in the first period only and holding them in the second period only and thus is termed as “timing

portfolios". The above result can be extended to any number of periods.

**Conditioning Information in Portfolio Construction:** It has been debated that the portfolio managers may use conditioning information they have about assets in forming the portfolios, not simply using the mean and variance of the past returns. Thus the managers' conditionally efficient portfolios may not appear efficient to the investor with no knowledge of that information. Ferson and Siegel (2001) [114] study the properties of unconditional minimum-variance portfolios when the conditioning information is present. "For extreme signals about the returns of risky assets, the efficient solution requires a conservative response." Managers reduce risk by taking a small position in the risky asset while keeping the same level of portfolio performance. In its basic formulation, assuming that the investment alternatives are a risky asset and risk-free return, with  $w(\tilde{S})$ —the proportion invested in the risky-asset, the portfolio return and the variance are:

$$\begin{aligned}\mu_p &= E[r_f + w(\tilde{S})(r_t - r_f)] \\ \text{and} \quad (7.62) \quad \sigma_p^2 &= E[(w(\tilde{S})(r_t - r_f))^2] - (\mu_p - r_f)^2\end{aligned}$$

Here  $\tilde{S}$  is the observed signal and  $w(\tilde{S})$  is a function of that signal that the manager considers at the time of allocation. The minimization criterion leads to the following weight for the risk asset:

$$w(\tilde{S}) = \frac{\mu_p - r_f}{C} \left( \frac{\mu(\tilde{S}) - r_f}{(\mu(\tilde{S}) - r_f)^2 + \sigma_e^2(\tilde{S})} \right) \quad (7.63)$$

where  $C = E \left[ \frac{(\mu(\tilde{S}) - r_f)^2}{(\mu(\tilde{S}) - r_f)^2 + \sigma_e^2(\tilde{S})} \right]$  and the minimized variance is  $\sigma_p^2 = (\mu_p - r_f)^2 \left( \frac{1}{C} - 1 \right)$ .

Here  $\sigma_e^2(\tilde{S})$  is the conditional variance given the signal. Note when this is zero, which signifies that there is no conditional information,  $w(\tilde{S}) = w = (\mu_p - r_f)/(\mu - r_f)$ , that is the result of Markowitz theorem. The above results can be easily extended to multiple risky asset scenario.

To implement this idea, we assume that the weight vector is a linear function of 'k' state variables:

$$w_t = \theta z_t \quad (7.64)$$

and the maximization criterion in (7.56) can be modified as,

$$\max_{\theta} E_t \left[ (\theta z_t)' r_{t+1} - \frac{\gamma}{2} (\theta z_t)' r_{t+1}' r_{t+1} (\theta z_t) \right] \quad (7.65)$$

Denoting  $\bar{r}_{t+1} = z_t \otimes r_{t+1}$  and  $\tilde{w} = \text{vec}(\theta)$ , we can obtain the optimal solution as,

$$\begin{aligned}\tilde{w} &= \frac{1}{\gamma} E[\tilde{r}_{t+1} \tilde{r}_{t+1}']^{-1} E(\tilde{r}_{t+1}) \\ &= \frac{1}{\gamma} E[(z_t z_t' \otimes (r_{t+1} r_{t+1}'))^{-1}] E[z_t \otimes r_{t+1}] \quad (7.66)\end{aligned}$$

Here  $\otimes$  is meant for a Kronecker product. Brandt and Santa-Clara (2006) [42] also show how both single period and multi-period problems can be formulated as the constrained VAR.

Using the established evidence from the literature that market returns can be predicted by conditioning on divided-price ratio, short-term interest rate, term spread and credit spread, to study the market returns for both stocks and bonds, the authors evaluate the conditional and unconditional policies at monthly, quarterly and annual holding periods, with the value of  $\gamma = 5$ . Some general conclusions are:

- Conditional policies are quite sensitive to state variables. Because of the predictability of returns, conditional policies allow for the investor to be more aggressive on average. Sharpe ratios are generally higher for conditional policies.
- Results are less pronounced for the larger holding periods. Conditional policies can be very different at different time frequencies.
- The above stated conclusions hold for multiperiod portfolio policies as well with monthly rebalancing.

## 7.4 Portfolio Tracking and Rebalancing

The proportions of capital allocated to various assets can change due to their performances and investors risk aversion or objectives also can change over time. Thus investors often engage in portfolio rebalancing to keep the portfolio on target risk exposure and exploit the market opportunities to enhance returns in the long run. First we formulate the problem with constraints and then present the results of some empirical studies.

The portfolio returns ( $r_t$ ) are generally compared with some index returns ( $I_t$ ); the excess return ( $ER$ ) are the tracking error ( $TE$ ) are defined as follows:

$$\begin{aligned} ER &= \frac{1}{T} \sum_{t=1}^T (r_t - I_t) \\ TE &= \frac{1}{T} \sum_{t=1}^T (r_T - I_t)^2 \end{aligned} \tag{7.67}$$

The optimization function is to minimize  $\lambda\sqrt{TE} - (1 - \lambda)ER$ , via the choice of the portfolio weights. Another optimization set-up is to balance between tracking error and transaction costs. Let  $x = w - w_0$ , where ' $w_0$ ' is the current portfolio weights and ' $w$ ' is the target weights; the problem with constraints can be stated as:

$$\min_w (w - w_0) \Omega (w - w_0) \ni w' 1 = 1 \tag{7.68}$$

subject to the following constraints:

$$\begin{aligned} w &\geq 0 \text{ long only} \\ \sum |x_i| &\leq u \text{ turn-over} \\ \sum_{i=1}^m \beta_{ik} w_i &\leq u_k \text{ maximum exposure to a factor} \\ l_i &\leq w_i \leq u_i \text{ holding} \end{aligned} \tag{7.69}$$

Note the criterion in (7.68) can also be taken as a version of tracking error.

Some practical implementation ideas are discussed in Pachanova and Fabozzi (2014) [219]. In addition to naïve allocation strategy discussed earlier in this chapter, other ad-hoc methods simply imitate the composition of some market indices such as S&P 500 that weighs stocks by their market capitalizations or Dow-Jones that weighs stocks by their prices. To imitate, select the stocks with the largest weights in the benchmark. Other methods include stratification of benchmarks into groups and select stocks from each group that ensures diversification. The groups can be industry-based or if it is based on past performance, the selected assets are called basis assets. The difference between the original allocation model and the rebalancing models can be due to change in the objectives and also change in the transaction costs and taxes. If a manager wants to move toward higher alpha, the restriction on the tracking error should be added to (7.69). Like other calculations related to portfolio management, the rebalancing also involves backward-looking estimates. To reflect the future more accurately, managers should arrive at forward-looking estimates of tracking error using conditioning arguments given in the earlier section.

**Transaction Costs:** Bid-ask spreads, brokerage fees, taxes and the price impact on trading are covered in Chapter 7. Here we present some observations related to trading costs as noted in the literature. Defining the turnover as,

$$\mathcal{T} = \sum_{i=1}^m |w_i - w_{0i}| = \sum_{i=1}^m \tau_i \tag{7.70}$$

and the costs of rebalancing is

$$C = \sum_{i=1}^m k_i \tau_i \tag{7.71}$$

Note  $w_{0i} = \frac{w_i r_i}{\sum w_i r_i}$  and  $\tau_i = \left| \frac{w_i(r_i - r_{0i})}{\sum w_i r_i} \right|$  and thus a key term here is  $r_i - r_{0i}$ , the difference between the actual and target return for asset ‘*i*’. Kourtis (2014) [181] studies the distribution of  $\tau$  and  $C$  under the naïve allocation strategy. Denoting  $r^p = \sum w_i r_i$  the portfolio return, the distribution of  $\tau_i$  depends on  $(r_i - r_{0i})/r^p$ , the ratio of excess return to the portfolio return. A general formulation that accounts for the transaction cost is

$$\min_w \delta(w - w_0)' \Omega (w - w_0) + \sum_{i=1}^m k_i E[\tau_i(r_i, w)] \tag{7.72}$$

with the controlling parameter,  $\delta$ , on the tracking error. By explicitly incorporating the trading cost in the criterion, it is shown that the total cost of transaction is reduced.

Moorman (2014) [212] considers three portfolio strategies: a levered-momentum strategy, a zero-cost-decile momentum strategy and equally-weighted naïve strategy. Because the momentum strategies adapt to changing market conditions swiftly requires frequent rebalancing and can reveal the drawbacks of any method to reduce transaction costs. First there is a no-trade region where the distance between target and the current weight is not significant. Write the rebalanced weight as,

$$w_t^* = \alpha_t w_{0t} + (1 - \alpha_t) w_t \quad (7.73)$$

where ' $\alpha_t$ ' is the rate of rebalancing; if  $\alpha_t = 1$ , then there is no rebalancing. It is easily seen that the current month's unbalanced weights are related to prior month's rebalanced weights as,

$$w_{0t} = w_{t-1}^* \left( \frac{1 - r_t}{1 + r_{pt}} \right) \quad (7.74)$$

where  $r_{pt}$  is the before-transaction cost return on the portfolio. Note

$$r_{pt} = \sum_{i=1}^{N_t} w_{it}^* r_{it} - c_{it} |w_{it}^* - w_{0t}| \quad (7.75)$$

where  $c_{it}$  is the cost of transaction. The ' $\alpha$ ' is chosen using the following objective function that results from constant relative risk aversion utility:

$$\max_{\alpha} \frac{1}{T} \sum_{t=0}^{T-1} \frac{(1 + r_{pt+1})^{1-\gamma}}{1 - \gamma} \quad (7.76)$$

where ' $\gamma$ ' is the coefficient of relative risk-aversion. The parameter ' $\alpha$ ' is chosen via grid search, outside no-trade zone. An empirical investigation of various distance and rebalancing methods reveal that transaction cost reduction is most successful for the lowered momentum and equally-weighted market portfolios but more difficult for the momentum portfolios because of higher turnover.

## 7.5 Transaction Costs, Shorting and Liquidity Constraints

It is clear, based on the discussion in this chapter that the managers tend to mimic better performing portfolios or use information that they have about future values of the assets in the construction and rebalancing of the portfolios. But at the end of investment horizon, the performance evaluation of the selected portfolio depends on the average returns after adjusting for the risk premia. Consider the  $k$ -factor models, (7.14),  $r_{it} = \alpha_i + \beta'_i f_t + \epsilon_{it}$  and the equation that relates the average return to risk

premia ( $\lambda_k$ ),  $\mu_i = r_f + \beta'_i \lambda_k$  in (7.15). The equation (7.14) can be also written more compactly as,

$$r_t = \alpha + B f_t + \epsilon_t \quad (7.77)$$

where  $\alpha$  is a  $m \times 1$  vector and  $B$  is the  $m \times k$  matrix of factor premia. Thus the portfolio return,

$$\begin{aligned} r_{pt} &= w' r_t = w' \alpha + w' B f_t + w' \epsilon_t \\ &= \alpha_p + \beta'_p f_t + \epsilon_{pt} \end{aligned} \quad (7.78)$$

Because ' $w$ ' is random, note

$$\begin{aligned} E(r_{pt}) &= E(\alpha_p) + E(\beta'_p f_t) \\ &= \alpha' E(w) + \text{Cov}(\beta_p, f_t) + E(\beta'_p) E(f_t) \end{aligned} \quad (7.79)$$

Note the first term is the result of asset choice, the second term is the result of the factor model theory and the last term reflects the risk premia.

**Choice of Assets:** The screening of candidate assets relies upon assessing the underlying factor for return behavior of a portfolio. Factors are grouped into generally as fundamental factors and macroeconomic factors, although there are other considerations such as if investing in certain assets are considered to be socially responsible, or environmentally safe, etc.. Fundamental factors are related to the asset characteristics such as analysts forecasts and the macroeconomic factors relate to activities in the economy that are likely to be closely associated with the asset behavior. These may be industry factors or sentiment factors. These factors can be standardized and an aggregate store can be used to sort out the assets. Several sorting methodologies are available in the literature. Using the factor scores, sort the stocks and go long on stocks with high scores and go short on stocks with low values. This method is easy to implement and it diversifies away risk of individual stocks. Some disadvantages are that it may be difficult to infer which variables provide unique information and portfolios may pick up different undesirable risks.

Patton and Timmermann (2010) [223] test for positive relation between ex-ante estimates of CAPM beta and returns that follow using daily data. Stock are sorted at the start of each month into deciles on the basis of their betas using prior one year daily data and value-weighted portfolios are formed. Returns for these portfolios are recorded for the following month. If the CAPM holds, one should expect the average returns should increase with the betas, which is empirically confirmed. The authors also examine if the post-ranked betas of portfolios ranked by ex-ante beta estimates are monotonic confirming the predictability of the betas. Sorting the stocks based on their betas and forming decile portfolios, the performance is evaluated in the subsequent month. Specifically from the CAPM model,  $r_{it} = \alpha + \beta_i r_{mt} + \epsilon_{it}$  for  $i = 1, 2, \dots, 10$ , the following hypothesis,  $H_0 : \beta_1 \geq \beta_2 \geq \dots \geq \beta_{10}$  is tested. The estimates range from 1.54 to 0.6 and the hypothesis is confirmed. This approach is extended to two-way sorts; firm size and book-to-market ratio or firm size and momentum. It is concluded that the size effect in average returns is absent among loser

stocks. Also momentum effects are stronger for small and medium size forms.

**Alpha and Beta Strategies:** Portfolio managers have traditionally developed strategies to maximize ‘ $\alpha$ ’ in the factor model. One of the measures, information ratio defined in Chapter 3 is taken as an index for the portfolios performance. But after the 2007–2008 financial crisis, there is a great deal of interest in smart-beta (referring to the coefficients of the factor model) strategies. This is a broad term that refers to from ad-hoc strategies to a mix of optimized simpler portfolios and investment in non-traditional asset classes. The construction is generally rules-based and is transparent. The commonality among them is in their exposure to factors that investors are comfortable to choose from. Kahn and Lemmon (2015) [176] provide a framework to understand the beta strategies along the decomposition of return in (7.79) into three returns. To summarize:

- “The smart-beta return arises from statics (i.e. long-term average) exposures to smart-beta factors.
- The pure alpha consists of three pieces:
  1. The average security selection return (beyond smart-beta factors).
  2. The average macro, industry and country returns (beyond smart-beta factors)
  3. The return due to smart-beta timing.”

The authors advocate an optimal blend of active smart-beta and index products. The optimality depends on the understanding of expected returns and risks. Other considerations include how the impact of the underlying factors will change and how the volatility and correlations among the factors are likely to change.

Frazzini and Pedersen (2014) [116] present a model with leverage and margin constraints that can vary over time and can be different for different investor and formally study the implications. These implications are indeed verified by the empirical analysis as well. The CAPM model assumes that an investor is interested in maximizing the expected return for a given leverage on her risk tolerance. But some investors may be constrained in their leverage (such as insurance companies have to maintain certain amount of reserves) and may over-invest in risky securities. The growth of ETF’s indicates that some investors cannot use leverage directly. The inclination to invest in high-beta assets require lower risk-adjusted returns than low-beta assets that require leverage. The authors set out to study how an investor with no leverage constraint can exploit this and other related questions.

Consider the two period model agents trade ‘ $m$ ’ securities that each pays ‘ $\delta_t$ ’ as dividend and ‘ $W$ ’ shares are outstanding. An investor chooses a portfolio of shares,  $w = (w_1, \dots, w_m)'$  and the rest is invested at the risk-free rate,  $r_f$ . Denoting ‘ $P_t$ ’ as the  $m \times 1$  price vector, the following utility function criterion is optimized.

$$\max w'(E_t(P_{t+1} + \delta_{t+1}) - (1 + r_f)P_t) - \frac{\gamma}{2} w' \Omega w \quad (7.80)$$

subject to the portfolio constraint

$$m_t \cdot w' P_t \leq W \quad (7.81)$$

The quantity ‘ $m_t$ ’ is a leverage constant and if  $m_t = 1$ , it implies that the investor simply cannot use any leverage. If  $m_t > 1$  would imply the investors have their own cash reserves. Solving above with  $\psi_t$  as the Lagrangian multiplier for the constraint in (7.81), we obtain:

$$\begin{aligned} \text{Equilibrium Price: } P_t &= \frac{E_t(P_{t+1} + \delta_{t+1}) - \gamma \Omega w^*}{1 + r^f + \psi_t} \\ \text{Optimal Position: } w^* &= \frac{1}{\gamma} \Omega^{-1} (E_t(P_{t+1} + \delta_{t+1}) - (1 + r^f + \psi_t) P_t) \quad (7.82) \\ \text{Aggregate Lagrangian: } \psi_t &= \sum_i \left( \frac{\gamma}{\gamma^i} \right) \psi_t^i \end{aligned}$$

Some key implications are:

- (i) An asset’s alpha in the market model is  $\alpha_t = \psi_t(1 - \beta_t)$ ; thus alpha decreases as beta increases.
- (ii) Sharpe ratio increases with beta until beta is unity and then decreases.
- (iii) If portfolios are sorted out in terms of betas, expected excess return is  $E_t(r_{t+1}) = \frac{\beta_t^H - \beta_t^L}{\beta_t^L \beta_t^H} \cdot \psi_t$ , where the superscripts denote high and low beta portfolios.

These findings are useful for studying how investment decisions can be made. Investors who are constrained in their access to cash tilt toward riskier securities with higher betas. Empirically it has been shown that portfolios with higher betas have lower alphas and lower Sharpe ratios than portfolios of low-beta assets.

## 7.6 Alphas, Betas, Information Ratios and Realized Returns

## 7.7 Supplements and Problems

1. Using the historical log returns of the following tickers: CME, GS, ICE, LM, MS, for the year 2014 (Exercise 5.1.csv), estimate their means  $\mu_i$  and covariance matrix  $\Sigma$ ; let  $R$  be the median of the  $\mu_i$ ’s.
  - (a) Test if the returns series are random via the autocorrelation function.
  - (b) Solve the Markowitz’s minimum variance objective to construct a portfolio with the above stocks, that has expected return at least  $R$ . The weight  $\omega_i$  should sum to 1. Assume short selling is possible.
  - (c) Generate a random value uniformly in the interval  $[0.95\mu_i, 1.05\mu_i]$ , for each stock  $i$ . Resolve Markowitz’s objective with these mean returns, instead of  $\mu_i$  as in (b). Compare the results in (b) and (c).

- (d) Repeat three more times and average the five portfolios found in (a), (b) and (c). Compare this portfolio with the one found in (a).
- (e) Repeat (a), (b) and (c) under no short selling and compare the results obtained with short selling.
- (f) Use the LASSO method to construct the portfolio and compare it with the result in (e).
2. Suppose that it is impractical to construct an efficient portfolio using all assets. One alternative is to find a portfolio, made up of a given set of  $n$  stocks, that tracks the efficient portfolio closely—in the sense of minimizing the variance of the difference in returns.
- Specifically, suppose that the efficient portfolio has (random) rate of return  $r_M$ . Suppose that there are  $N$  assets with (random) rates of return  $r_1, r_2, \dots, r_n$ . We wish to find the portfolio whose rate of return is  $r = \alpha_1 r_1 + \alpha_2 r_2 + \dots + \alpha_n r_n$  (with  $\sum_{i=1}^n \alpha_i = 1$ ) by minimizing  $\text{Var}(r - r_M)$ .
- Find a set of equations for the  $\alpha_i$ 's.
  - Another approach is to minimize the variance of the tracking error subject to achieving a given mean return. Find the equation for the  $\alpha_i$ 's that are tracking efficient.
  - Instead of minimizing  $\text{Var}(r - r_M)$ , obtain  $\alpha$ 's that result from minimizing mean-squares of the difference  $(r - r_M)$ .
- 3.
- The file `d_15stocks.csv` and `m_15stocks.csv` contain daily and monthly stock returns data for 15 stocks from Jan 11, 2000 to Mar 31, 2013.
  - The file `d_indexes.csv` and `m_indexes.csv` contain daily and monthly returns data for the volume-weighted and equal-weighted S&P500 market indices (VWRETD and EWRETD, respectively) from Jan 11, 2000 to Mar 31, 2013.

Using daily and monthly returns data for 15 individual stocks from `d_15stocks.csv` and `m_15stocks.csv`, and the equal-weighted and value-weighted CRSP market indexes (EWRETD and VWRETD, respectively) from `d_indexes.csv` and `m_indexes.csv`, perform the following statistical analyses using R. For the subsample analyses, split the available observations into equal-sized subsamples.

- Compute the sample mean  $\hat{\mu}$ , standard deviation  $\hat{\sigma}$ , and first-order autocorrelation coefficient  $\hat{\rho}(1)$  for daily simple returns over the entire sample period for the 15 stocks and two indexes. Split the sample into 4 equal subperiods and compute the same statistics in each subperiod – are they stable over time?
- Plot histograms of daily simple returns for VWRETD and EWRETD over the entire sample period. Plot another histogram of the normal distribution with mean and variance equal to the sample mean and variance of the returns plotted in the first histograms. Do daily simple returns look approximately normal? Which looks closer to normal: VWRETD or EWRETD?

- (c) Using daily simple returns for the sample period, construct 99% confidence intervals for  $\hat{\mu}$  for VWRETD and EWRETD, and the 15 individual stock return series. Divide the sample into 4 equal subperiods and construct 99% confidence intervals in each of the four subperiods for the 17 series – do they shift a great deal?
  - (d) Compute the skewness, kurtosis, and studentized range of daily simple returns of VWRETD, EWRETD, and the 15 individual stocks over the entire sample period, and in each of the 4 equal subperiods. Which of the skewness, kurtosis, and studentized range estimates are statistically different from the skewness, kurtosis, and studentized range of a normal random variable at the 5% level? For these 17 series, perform the same calculations using monthly data. What do you conclude about the normality of these return series, and why?
4. The file `FF_Data_ForGRStest.csv` contains historical monthly returns for one set of 6 portfolios and another set of 25 portfolios, formed based on the size and book-to-market ratio (BM). The data is obtained from French's data library. The portfolios are formed as the intersection of size (or market equity, ME) based portfolios and book equity to market equity ratio (BE/ME) based portfolios ( $2 \times 3$  forming the first set of 6 portfolios and  $5 \times 5$  forming the second set of 25 portfolios). These portfolios are discussed in their 1993 paper by Fama and French.
- In this exercise we will only work with the first set of 6 portfolios, which are contained in the columns named beginning with "PF6", with the rest of the column name following French's naming convention about the size and BM of the corresponding portfolios – SML contains small size + low BM, SM2 contains small size + medium BM, SMH contains small size + high BM, BIGL contains big size + low BM, etc.
- Finally, the last 4 columns of the data set contain the Fama-French factors themselves along with the risk-free rate: MktMinusRF contains the excess return of the market over the risk-free rate, SMB contains the small-minus-big size factor, HML contains the high-minus-low BM factor and RF contains the risk-free rate.
- (a) Using the entire sample, regress the excess returns (over the risk-free rate) of each of the 6 portfolios on the excess market return, and perform tests with a size of 5% that the intercept is 0. Report the point estimates,  $t$ -statistics, and whether or not you reject the CAPM. Perform regression diagnostics to check your specification.
  - (b) For each of the 6 portfolios, perform the same test over each of the two equi-partitioned subsamples and report the point estimates,  $t$ -statistics, and whether or not you reject the CAPM in each subperiod. Also include the same diagnostics as above.
  - (c) Repeat (a) and (b) by regressing the excess portfolio returns on all three Fama-French factors (excess market return, SMB factor and HML factor).

- (d) Jointly test that the intercepts for all 6 portfolios are 0 using the  $F$ -test statistic or Hotelling's  $T^2$  for the whole sample and for each subsample when regressing on all three Fama-French factors.
  - (e) Are the 6 portfolio excess returns (over the risk-free rate) series cointegrated? Use Johansen's test to identify the number of cointegrating relationships.
5. The file `FF_Data_ForGRStest.csv` contains returns on 25 size sorted portfolios along with the risk free rate, excess market return along with two Fama-French factors. The monthly data spans from 1926 to 2012.
- (a) Fit CAPM to the 25 portfolios. Give point estimates and 95% confidence intervals of  $\alpha$ ,  $\beta$ , the Sharpe index, and the Treynor index. (Hint: Use the delta method for the Sharpe and Treynor indices.)
  - (b) Test for each portfolio the null hypothesis  $\alpha = 0$ .
  - (c) Use the multivariate regression model to test for the 25 portfolios the null hypothesis  $\alpha = 0$ .
  - (d) Perform a factor analysis on the excess returns of the 25 portfolios. Show the factor loadings and the rotated factor loadings. Explain your choice of the number of factors.
  - (e) Consider the model

$$r_t^e = \beta_1 \mathbf{1}_{t < t_0} r_M^e + \beta_2 \mathbf{1}_{t \geq t_0} + \epsilon_t$$

in which  $r_t^e = r_t - r_f$  and  $r_M^e = r_M - r_f$  are the excess returns of the portfolio and market index. The model suggests that the  $\beta$  in the CAPM might not be constant (i.e.  $\beta_1 \neq \beta_2$ ). Taking February 2001 as the month  $t_0$ , test for each portfolio the null hypothesis that  $\beta_1 = \beta_2$ .

- (f) Estimate  $t_0$  in (e) by the least squares criterion that minimizes the residual sum of the squares over  $(\beta_1, \beta_2, t_0)$ .
  - (g) Fit the Fama-French model and repeat (a)–(c).
6. Portfolio Rebalancing: Consider the 15 stocks in file from Exercise 3: Divide the duration of the data into four time intervals.
- (a) At the end of each interval, compute the optimal portfolio weights using the risk-aversion formulation; choose  $2 \leq \lambda \leq 4$  (evaluate at  $\lambda = 2, 3$  and 4). Comment on how the portfolio weights have changed and why.
  - (b) For each interval, construct factor models; sort the stocks based on the first factor. Follow 130/30 strategy and evaluate the allocation procedure.
  - (c) Consider the technology stocks: IBM and MSFT; use the daily data after the first interval to adaptively compute the portfolio weights using risk-aversion formulation as well as VaR. Evaluate the two criteria using the actual vs expected returns; also comment on how weights change over time.

7. Consider the exchange rate data in the file `Forex1.csv` for 24 currencies; consider the returns on these currencies along with the market return and the risk free rate. The daily data spans from 3/1/2005 to 10/4/2012. We will use the data from 2005 to 2010 to construct a portfolio and the data from 2011 to 2012 to evaluate the portfolio.
  - (a) Compute the mean  $\hat{\mu}$  and sample covariance matrix,  $\hat{\Sigma}$  of the log returns.
  - (b) Compute the shrinkage estimate  $\hat{\Sigma}^*$  of the covariance matrix, under one factor model.
  - (c) Estimate the optimal weights, both with and without short selling under the regular estimate of  $\Sigma$  and as well as under the shrinkage estimate. Also consider equal weighted portfolio and the portfolio weighted by the inverse of variances. Compute the optimal weights for all combinations and discuss them.
  - (d) Estimate the optimal weights using LASSO both for short and no-short positions.
  - (e) Evaluate the different weighting schemes in terms of their performance during the validation period, 2011–2012. Summarize your findings and offer some intuitive explanations.
8. Consider the data in Problem 7.
  - (a) Perform a PCA on the 24 exchange returns; use the first two principal components on factors in a two-factor model for  $F$ , estimate  $F$ .
  - (b) Using the estimated  $\hat{F}$  in (a) as the shrinkage target, compute a new shrinkage coefficient and the new shrinkage estimate of  $\Sigma$ .
  - (c) Compare the efficient frontier corresponding to this estimate with those obtained in (1.c).
  - (d) Group the currencies in terms of their performance during 2005–2010 into four quantiles. Construct a portfolio based on the top performing quantile and another portfolio based on the bottom quantile; evaluate their performance using the data for 2011–2012.
9. The file `m_logret_10stocks.txt` contains the monthly returns of ten stocks from January 1994 to December 2006. The ten stocks include Apple Computer, Adobe Systems, Automatic Data Processing, Advanced Micro Devices, Dell, Gateway, Hewlett-Packard Company, International Business Machines corp., and Oracle Corp.. Consider portfolios that consists of ten stocks.
  - (a) Compute the sample mean  $\hat{\mu}$  and the sample covariance matrix  $\hat{\Sigma}$  of the log returns.
  - (b) Assume that the monthly target return is 0.3% and that short selling is allowed. Estimate the optimal portfolio weights by replacing  $(\mu, \Sigma)$  in Markowitz's Theory by  $(\hat{\mu}, \hat{\Sigma})$ .

- (c) Do the same as in (b) for Michaud's resamples weights (??) using  $B = 500$  bootstrap samples.
  - (d) Plot the estimated frontier (by varying  $\mu_*$  over a grid that uses  $(\hat{\mu}, \hat{\Sigma})$  to replace  $(\mu, \Sigma)$ ) in Markowitz's efficient frontier.
  - (e) Plot Michaud's resampled efficient frontier using  $B = 500$  bootstrap samples. Compare the plot in (d).
10. Consider again the data in the file `FF_Data_ForGRStest.csv`. Consider the returns on 25 size sorted portfolios along with the risk free rate and Fama-French factors. The monthly data spans from 1926 to 2012. We will use the data from 1926 to 2000 to construct a portfolio and the data from 2001 to 2012 to evaluate the portfolio.
- (a) Compute the mean  $\hat{\mu}$  and sample covariance matrix,  $\hat{\Sigma}$  of the log returns.
  - (b) Compute the shrinkage estimate  $\hat{\Sigma}^*$  of the covariance matrix, under one factor model.
  - (c) Estimate the optimal portfolio weights, both with and without short selling under the regular estimate of  $\Sigma$  and as well as under the shrinkage estimate. Also consider equal weighted portfolio and the portfolio weighted by the inverse of variances. Compute the optimal weights for all combinations and discuss them.
  - (d) Evaluate the different weighting schemes in terms of their performance during the validation period, 2001–2012. Summarize your findings and offer some intuitive explanations.
11. Consider the data in Problem 10.
- (a) Perform a PCA on the 25 portfolio returns; use the first two principal components on factors in a two-factor model for  $F$ , estimate  $F$ .
  - (b) Using the estimated  $\hat{F}$  in (a) as the shrinkage target, compute the new shrinkage estimate of  $\Sigma$ .
  - (c) Compare the efficient frontier corresponding to this estimate with those obtained in (10.c)



---

---

## **Part III**

# **Execution Algorithms**

---

---



# 8

---

## *Electronic Exchanges and Dynamics of Limit Order Books*

---

### CONTENTS

8.1	Limit Order Book (LOB): Basic Description and Characteristics .....	213
8.2	Data from LOB and Relevant Questions .....	216
8.3	Models for LOB dynamics, based on Level I and Level II data .....	219
8.3.1	Application of a Hawkes process .....	228
8.3.2	One-Dimensional Models .....	233
8.3.3	An extension .....	237
8.3.4	Two-Dimensional models .....	237
8.3.5	Limitations and findings .....	240
8.4	Models for Level II and Level III data .....	241
8.5	Estimation of Hidden Orders .....	241
8.6	Volume Imbalance and Other Indicators .....	241
8.7	Hidden Markov Models (HMM) for dynamic market microstructure: Estimation of Parameter and States .....	241
8.8	Appendix: Adaptive Particle Filters in HMM .....	241
8.9	Supplements and Problems .....	241

Stoll (2006) [246] in an instructive essay traces the evolution of trading that started under a tree in 1792 with 24 brokers. This has become later on the New York Stock Exchange (NYSE). Modern electronic technology has changed how trading is done. Now all transactions are carried out through a computer system with little human intervention (see the snapshots of NYSE over two different decades). This has also changed how stock market trading is done. The Electronic Communication Networks (ECN) automate the trading process but the fundamental functions of trading are not changed except for the instant dissemination of the demand and supply information to all investors. The identity of traders is kept anonymous. Executions are done fast and very complex orders can now be submitted to the market. The trading costs—both direct such as commissions, brokerage fees, etc. and indirect cost due to price-impact—are expected to be lower. Over the period of two decades (1980–2000), the commissions have gone down from 1.2% to 0.2%. But with higher turnover and thus with increased volume, the total commissions have increased four-fold. The reduction in bid-ask spread although may be primarily due to tick size is accelerated under electronic trading.



Figure 8.1: Treasury bind pit 1989.



Figure 8.2: The NYSE trading floor in August 2008.

It must be observed that electronic trading does not necessarily imply automated trading. But it has certainly led to more automated trading due to easy execution of the algorithms. Although there is a concern that the traders can game the prices, but

computer audit trails can be used to monitor unlawful behaviors. The debate over whether electronification of trading will lead to more consolidation or fragmentation has favored the latter. The apprehension over the possibility of inefficient price formation due to fragmentation is empirically shown not to hold. The order protection rule (SEC 2005) that dictates that an order submitted to any market should get the best of prices in all markets.

Thus large number of equity and derivative markets are now organized as electronic limit order books. The Electronic Communications Networks (ECNs) in the United States, Hong Kong and Toronto Stock Exchanges are examples of equity markets. Chicago Mercantile Exchange's (CME) Globex platform, Euronext Liffe's centralized limit order book are example of derivative markets. Open limit order books have become popular due to the greater transparency of the market when compared with dealer market settings, that rely upon the information on the dealers' best quotes. The limit order book provides its users to view the depth of the book at various price level away from the best quotes. In dealer market prices for trades beyond the quoted size must be obtained through negotiation with market maker.

There has been a great deal of interest in studying the information content of an open limit order book and if it helps investors in tracking short-term price predictions. Assuming that the market consists of informed and noise traders, and the informed traders actively submit market orders, Glosten (1994) [135] and Seppi (1997) [239] argue that the order book does not contain any information beyond the best bid and best offer. But other studies such as Harris and Panchapagesan (2005) [145] argue that the order book is informative and specialists better use the book information to their advantage than the traders who routinely place limit orders. In this chapter we discuss the feature of limit order book and various levels of information that have been made available to the investors; the focus is on developing models for studying the dynamics on the book and how the models could be used for high frequency trading.

An excellent survey paper by Parlour and Seppi (2008) [221] reviews the issues related to limit order markets. A limit order for any asset ' $j$ ' is an *ex ante* pre-commitment  $(t, x, p)$  made at time ' $t$ ' to trade ' $x$ ' units at a pre-specified limit price, ' $p$ '. The order remains valid until it is filled or cancelled. With multiple exchanges, the interactions are quite complex; the dynamic nature of state of the order book and the choice of actions in anticipation of the future states make the limit order placement decisions,  $(t, x, p)$  challenging. The key issues in the study of LOB are nicely summarized in Parlour and Seppi (2008) [221]. They relate to, Price Formulation (How limit order markets differ from dynamic dealer markets?), Liquidity (Implication of the quotes posted at various depths in LOB), Dynamics (Dependencies of Various Trading Decisions), Information Aggregation (How forward-looking is the activity in LOB about future order flow in aggregate?) and Inter-Market Competition (How effective are limit order markets to induce competition and how they handle non-market frictions?). How multiple exchanges can sustain competitive pressure to attract postings to their venues? How do limit order markets help reduce market frictions? These issues are addressed in this chapter but our focus is always on how the information flow through LOB dynamics helps trading decisions. The tools we review here in this chapter can be useful to that end.

## 8.1 Limit Order Book (LOB): Basic Description and Characteristics

In automated continuous double auction trading system that have become prevalent in the financial markets, buyers and sellers submit limit orders electronically; orders are then matched and executed as per time and price priorities. The unmatched orders are stored in the limit order book (LOB). The market orders are orders seeking the best available price and they are execute immediately. Thus, one way or other all executed orders can be taken as market orders.

The state of the order book changes quite rapidly due to high frequency program trading. There are six types of events, three on either side that can alter the state of the order book:

- Limit order submission: A limit order is added to the queue at the specified price level.
- Limit order cancellation: An outstanding limit order is expired or cancelled and therefore is removed from the LOB.
- Market order submission: Outstanding limit order at the best price are executed against the market order and thus removed from the book.

Orders on the buy side are called “bids” while those on the sell side are called “asks.” The above events illustrated in the following diagrams. When a market order

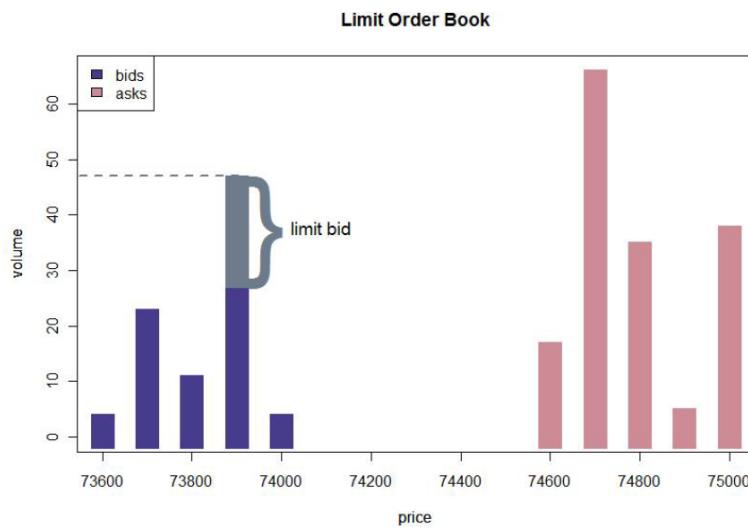


Figure 8.3: Limit Order Book—Limit Bid.

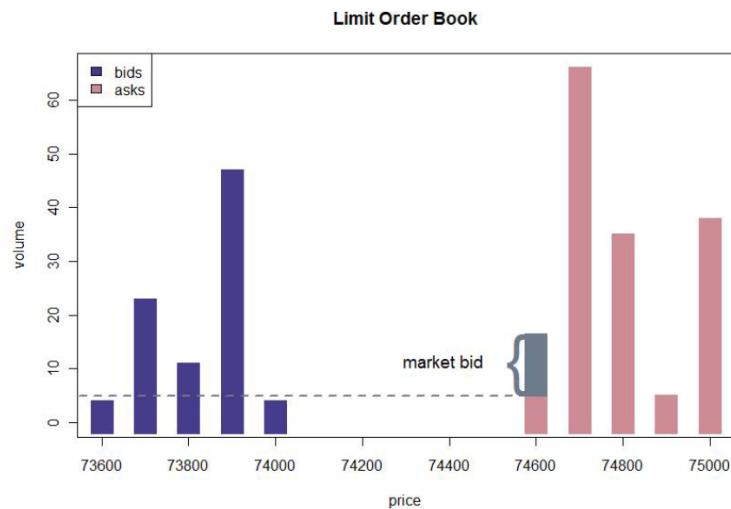


Figure 8.4: Limit Order Book—Market Bid.

is submitted, it decreases the number of outstanding orders at the opposite best price. For example, if a market bid order arrives, it will decrease the number of outstanding asks at the best price. All unexecuted limit orders can be cancelled. When a cancel-

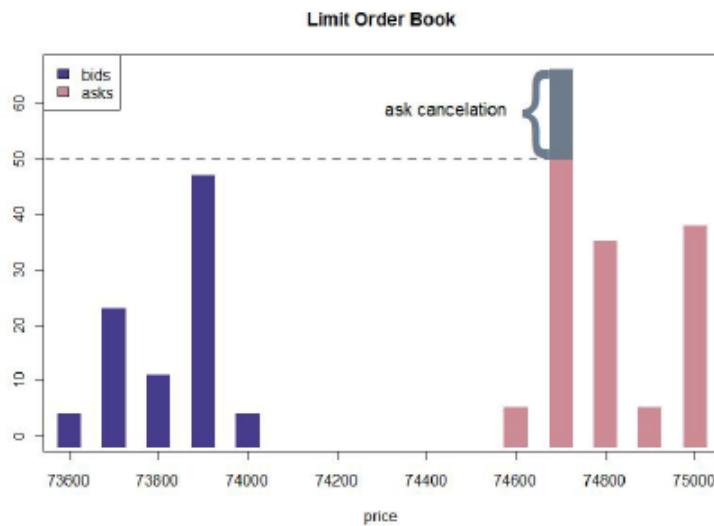


Figure 8.5: Limit Order Book—Ask Cancellation.

lation occurs, it will decrease the number of outstanding orders at the specified price level.

The limit orders constitute a very significant percentage (70%) of stock market trading activity. The main advantage of a limit order is that there is no price risk, that is when the order is executed the limit price is attained. But the execution is not guaranteed and the time to get an order executed depends on various market factors. The trade off between the limit order and the market order depends on the investors need for liquidity and the probability of execution of a limit order. The execution of limit orders does affect how the quotes are posted and are updated. A market order will be filled at or close to National Best Bid and Offer (NBBO). If the size of the order exceeds the number of shares available, it is usually split and is executed at different prices until the order is filled. The market orders are usually restricted to be filled within a day and order placed after the markets close are entered the next day at 9:30 am EST. Other types include stop-limit orders and trailing stop orders. The former type is typically used to trade a security at a specified limit price once it is traded through a specified stop price. If the stock declines in value and traders below or at the stop price, the order will become a limit order rather than a market order. The trailing stop order is in which the specified price will trail by a certain percentage from the best bid or best ask. These rules are handled in the brokers IT technology. In addition to time and duration specifications, other conditions that can be added are: fill or kill (the order gets filled in its entirety or not filled at all) or minimum quantity (if the specified minimum is not available, the order does not get executed).

---

## 8.2 Data from LOB and Relevant Questions

Initially researchers studying the dynamic of LOB had to content with Trades and Quotes (TAQ) data which provide a time-stamped sequence of trades (Market Orders) and updates in the price and depth of bid and ask quotes. No other information beyond the best bid and the best offer are made available. Moreover, the updates contain changes at the two quotes and were aggregated. The aggregation was clearly a limiting factor as one could not discern the kind of events that led to the changing status of the book. The level II data that became available is a time-stamped sequence of trades and quotes for the top 5-levels of the LOB. This data is little deeper than TAQ data, but was still in aggregated form. But recently made available level III data contains time-stamped sequence of all (except for hidden orders) events that occur in the LOB. Every order is identified by a unique order-id and thus it can be tracked to its lifetime.

**Description of Level III data:** The dataset consist of activities during the entire trading session, including early and late hours from 6:00 A.M. to 8:00 P.M. EST in a trading day. It provides intraday depth book activity for 12000 tickers from major national exchanges: Nasdaq, Direct Edge, NYSE, ARCA and BATS (exact number of

exchanges depends on the historical period). All messages are consolidated into one file ordered by timestamp. This data allows to build a full national depth book at any moment intraday. The description of the level III data is given in Table 8.1.

Variable	Description
Timestamp:	Number of milliseconds after the midnight.
Ticker:	Equity symbol (up to 8 characters)
Order:	Unique order ID.
T	Message type. Allowed values: <ul style="list-style-type: none"> <li>• “B”—Add buy order</li> <li>• “S”—Add sell order</li> <li>• “E”—Execute outstanding order in part</li> <li>• “C”—Cancel outstanding order in part</li> <li>• “F”—Execute outstanding order in full</li> <li>• “D”—Delete outstanding order in full</li> <li>• “X”—Bulk volume for the cross event</li> <li>• “T”—Execute non-displayed order</li> </ul>
Shares	Order quantity for the “B”, “S”, “E”, “X”, “C”, “T” messages. Zero for “F” and “D” messages.
Price	Order price, available for the “B”, “S”, “X” and “T” messages. Zero for cancellations and executions. The last 4 digits are decimal digits. The decimal portion is padded on the right with zeros. The decimal point is implied by position; it does not appear inside the price field. Divide by 10000 to convert into currency value.
MPID	Market Participant ID associated with the transaction (4 characters)
MCID	Market Center Code (originating exchange—1 character)

Table 8.1: Level III data

**Special types of orders:** While the display and issuance of new ID to the modified order varies from exchange to exchange, we provide a brief description of these orders below:

1. Order subject to price sliding: the execution price could be one cent worse than the display price at NASDAQ; it is ranked at the locking price as a hidden order, and displayed at the price one minimum price variation (normally 1 cent) inferior to the locking price. New order ID will be used if the order is replaced as a display order. At other exchanges the old order ID will be used.
2. Pegged order: Based on NBBO, not routable, new timestamp upon re-pricing, display rules vary over exchanges.
3. Mid-point peg order: Non-displayed, can result in half-penny execution.
4. Reserve order: displayed size is ranked as a displayed limit order and the reserve size is behind non-displayed orders and pegged orders in priority. The minimum display quantity is 100 and this amount is replenished from the reserve size when it falls below 100 shares. A new timestamp is created and the displayed size will be re-ranked upon replenishing.
5. Discretionary order: displayed at one price while passively trading at a more aggressive discretionary price. The order becomes active when shares are available within the discretionary price range. Ranked last in priority. The execution price could be worse than the display price.
6. Intermarket sweep order: order that can be executed without the need of checking the prevailing NBBO.

Some key measures of the LOB are: relative position of an order from the top position of the side (buy or sell), the depth which is the total number of shares at a given price and the depth profile of the book reflects the depth of both prices. Large market orders can shift the bid and ask prices. If the limit price of a new order crosses the best price on the opposite side of the book, it is executed immediately and these are termed as crossing or marketable limit orders. These are used at times to explore the existence of hidden orders.

The limit order book data provides a rich source of information to better understand the market microstructure. The information is also useful for developing trading strategies. Some key issues that may be explored are:

- The pattern of inter-arrival times of various events.
- Arrival and cancellation rates as a function of distance from nearest touch price.
- Arrival and cancellation rates as a function of other available information, such as in the queue on either side of the book.

The key issues from the trading point of view are:

- What is the impact of market order on the limit order book?
- What are the chances for a limit order to move up the queue from a given entry position?

- What is the probability of making the spread?
  - What is the direction of the price movement in short duration?
- 

### 8.3 Models for LOB dynamics, based on Level I and Level II data

Before we formally present the models, we want to provide a review of how the analysis of LOB data has been approached. Biais, Hillion and Spatt (1995) [27] study the trading activity in Paris Bourse (a fully automated limit order market), the dynamics of order flow and how the order flow varies with the state of the order book and marketplace events relevant to the asset. It is found that the conditional probability of a limit order placement is larger when the bid-ask spread is larger and when the order book is not deep. If a market order has been placed the chance that the next order posted provides liquidity is generally higher. The placement of orders also follows a pattern with new orders coming in the morning when the price discovery occurs and cancellations and large orders occur in the evenings. The durations between trades do indicate that the intensity of trade varies during the course of a day. The main tool used in these calculations is contingency table which provides both marginal and conditional probabilities.

**Slope of LOB:** To begin with we can consider the slope of the order book both on the demand side and on the supply side using the quotes on both sides various depths (see Figure 8.6). This provides information on order book imbalance that can be helpful for the trader to decide on the optimal time to enter or to exist the market (see Figure 8.6). It is observed from Figure 8.6 that the bid-ask spread is at least twice the difference between quotes at successive depths. Thus the slope of the book is steeper closer to the best quotes.

**Order Flow:** The orders can be characterized as buyer or seller initiated and how aggressive they are. The aggressiveness is quantified by the size of the order. Most of the orders are small indicating that they could be part of a large parent orders. Both placement and cancellation of orders seem to decrease away from the best quotes. As the intra-day trading pattern follows typical diurnal pattern, both the depth and order flow must be adjusted for the time of the day. Some observations (stylized facts) worth noting are:

- Large (small) trades on one side of the market tend to be followed by large (small) trades on the same side. The same thing could be said about the new limit orders and cancellations as well. It is speculated that this may be due to traders reacting similarly to the same events or due to parent order splitting and automatic placement of child orders.

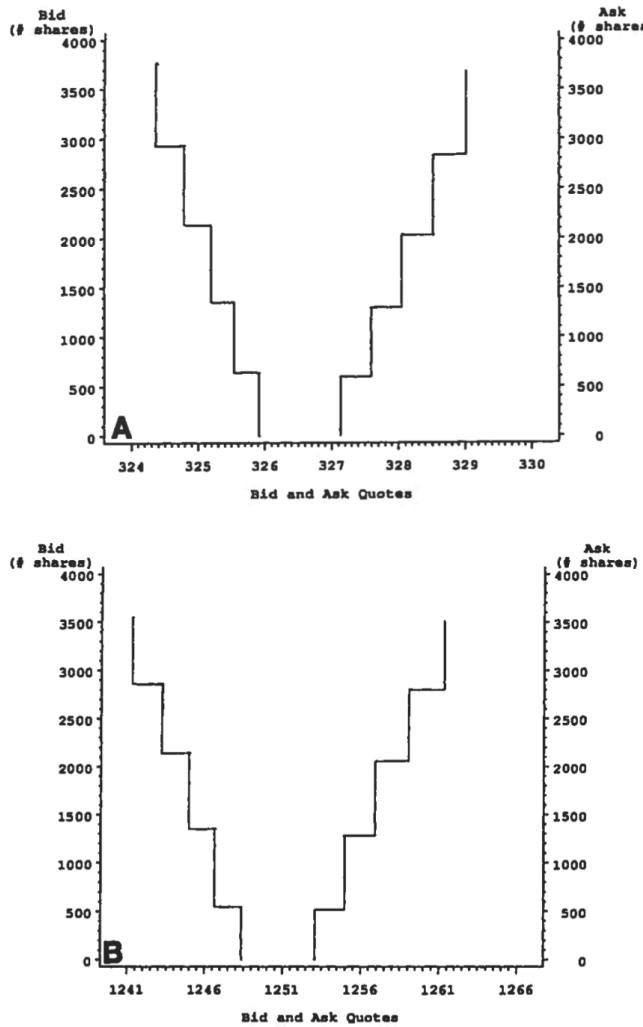


Figure 8.6: The Limit Order Book for Stocks. (A) Plots the cross-sectional average (across with tick=FF 0.1) of the time-series averages of the five best ask and bid quotes and their associated depth. (B) plots the cross-sectional average (across stocks with tick=FF 1) of the time-series averages of the five best ask and bid quotes and their associated depth.

- Cancellations on the buy side of the book are more frequent after market buy orders and the same for sell orders. This may be due to the fact that the large sell orders tend to convey negative information about the stock while the large buy orders convey positive information. Cancelations may also occur because of placement of

some market orders are intended to probe the presence of any hidden orders.

- Conditioning on the state of the book by size of bid-ask spread and the depth, it is observed that trades are more frequent when the spread is tight but the new orders inside the quotes are more frequent when the spread is large. The changes in spread are mainly due to liquidity shocks.
- Frequency of trades tends to be clustered because competing traders who monitor the market closely are likely to place the orders when they read the market in their favor. It is observed that the trading activity is more intense with the information flow. Also observed are the following stylized facts: when the spread is large after liquidity shocks, traders place their orders quickly to take advantage of time priority. Thus the spread reveals back to original level. The expected time interval is generally lower after large trades than after other trades whether the spread is large or not.

Thus the study by Biasi et al. (1995) [27] provides a set of concrete measures that could be used to study the dynamics of the limit order book. Now we will delve into model formulations.

One of the earlier models to appear in the econometric literature is by Lo, MacKinlay and Zhang (2002) [199]; they used the data from Investment Technology Group (ITG) that contained time stamped information on each limit order from submission to cancellation or execution. With three possible paths for each submitted order, the associated (a) time-to-cancellation/modification, (b) time-to-first fill and (c) time-to-completion are tracked and models are developed for each path on both sides of the order. Because ITG largely deals with the institutional investors, the results apply only to a limited segment of the traders. The modeling approach is to get the execution times as the first-passage time to the limit price assuming that price follows a geometric Brownian model with a drift (Model 1.1):

$$dP(t) = \alpha \cdot P(t) dt + \sigma \cdot P(t) dW(t) \quad (8.1)$$

where  $W(t)$  is a standard Brownian motion. A buy limit order with a price ' $P_l$ ' is executed in the time interval  $[t_0, t_0 + t]$  only if  $P_{\min} \leq P_l$  and the probability is

$$P_r(P_{\min} \leq P_l | P(t_0) = P_0) = 1 - \left( 1 - \left( \frac{P_l}{P_0} \right)^{2\mu/\sigma^2} \right) \Phi \left[ \frac{\ln(P_0/P_l) + \mu t}{\sigma \sqrt{t}} \right] \quad (8.2)$$

where  $\mu = \alpha - \frac{1}{2}\sigma^2$  and  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal.

Note (8.2) can be alternatively stated as, if 'T' denotes the limit order execution time,  $F(t) = P(T \leq t | P_0)$  for buy orders and similarly for sell orders,  $F(t) = p_r(P_{\max} \geq P_l)$ . To see if the model is appropriate for the data at hand compare the theoretical c.d.f. of  $F(t)$  to empirical c.d.f. using the well-known result that these are

uniformly distributed. Fix ‘ $\tau$ ’ as the fixed sampling interval and with  $r_t = \ln(P_t) - \ln(P_{t-1})$ , the estimates of  $\mu$  and  $\sigma^2$  are:

$$\hat{\mu} = \frac{1}{N\tau} \sum_{j=1}^N r_j, \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{j=1}^N \frac{(r_j - \hat{\mu}\tau)^2}{\tau} \quad (8.3)$$

where ‘ $N$ ’ is the number of observations in the sample.

The above First Passage Time (FPT) model was used for filled orders without taking into account that other orders that were cancelled or modified in that duration. The model’s other limitations include not accounting for time priority, not including other relevant variables such as price volatility, spreads, etc.. The model’s performance using the empirical cumulative distribution function is shown to be lacking. The model (8.2) is expanded as

$$F(t) = P_r(T_k \leq t | X_k, P_{lk}, S_k, I_k) \quad (8.4)$$

where  $T_k$  is the execution of time of  $k$  orders,  $P_{lk}$  is the limit order price,  $S_k$  is the size and  $I_k$  is the side indicator. An alternative function, the hazard rate (briefly discussed in Chapter 2)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \quad (8.5)$$

is found to be useful to operationalize and explain the modeling of survival rate of time-based events; Note  $S(t)$  is called the survivor function. The censoring information ( $\delta_i = 1$  if observation ‘ $i$ ’ is censored) can be incorporated with  $t_i$ ,  $i$ th realization of the random variable,  $T$ . It is assumed that the censoring mechanism is independent of the likelihood the limit order is executed. Using the generalized Gamma (defined in Chapter 2) as  $f(t)$ , the models are estimated.

The following variables are used as explaining variables:

- Distance between mid-quote and limit price.
- Prior trade indicator for buyer or seller initiated.
- Measures of liquidity.
- Measures of market depth.
- Measures that reflect change in trading activities.

It is shown that this model does fare better than FPT model to capture the execution times. But an important limitation is “...censoring as a result of prices moving away from the limit price would be a violation of the underlying assumption since prices at the time of censoring are not included in  $X_i$ .” We show in our analysis of Level III data, a key determinant for cancellation of an order is that price moving away from the limit price that the trader has desired.

There are a number of studies that have looked into order book dynamics using models that arise from Point Processes. Bouchaud, Mezard and Potters (2002) [37]

investigate the interaction between order flow and liquidity; it is observed that the distribution of incoming limit order prices follow a power law around the current price and the overall shape of the book in terms of volume on both sides is rather systematic. Smith, Farmer, Gillemot and Krishnamunthy (2003) [242] and Farmer et al. (2004) [] study how the order flow influences price formation; the price impact of market orders is a function of touch price depth and spread size. Hewlett (2006) [] focuses on using LOB dynamics to find optimal liquidation strategies in foreign currency markets.

To have a more complete and realistic picture of how the LOB evolves, we must consider ‘hidden liquidity’. Almost all exchanges allow traders to hide all or portions of their orders. The so called, “iceberg order” is split into several smaller parts and is queued along with the other orders, but only the displayed quantity is visible and is part of the market depth. When the order reaches the front of the queue only the displayed is executed. Several reports could be as high as 26%. In the models discussed below, we will not address the hidden liquidity issue, which will be taken up in a later section.

Modeling of LOB dynamics has drawn on ideas from economics, physics, statistics and psychology. The approach taken in economics literature is to focus on the behavior of the traders and the dynamics is modeled as sequential games (Foucault (1999) [], Parloys (1998) [] and Rosu (2009) []). Others, such as researchers from physics, have treated order flows as random and statistical mechanics techniques are used to study the dynamics. Other studies, such as Parlour and Seppi (2008) [221] and Bouchaud et al. (2009) are useful references for studies that come from economics point of view. We will review three models that are of recent origin.

**Cont, Stoikov and Talreja (2010) [68]:** In this model of the limit order book, the number of limit orders at each price level in the book is taken to be a continuous-time Markov chain. The evolution of the order book through market orders, limit orders and cancellations is represented as a counting process. The model replicates the evolution of the events using independent Poisson processes. The study views LOB as a system of queues subject to order book events whose occurrences are modeled as a multidimensional point process. Before we state the specific questions and the models implications, we define certain useful quantities.

It is taken that the price grid as  $P(n) = \{1, 2, \dots, n\}$  multiples of a price tick. The number of outstanding orders in the book  $|X_t^i|$  at price  $i$ , collectively,  $X_t \equiv (X_t^1, \dots, X_t^n)$  is taken to be continuous-time Markov chain; if  $X_t^i < 0$ , then there are  $-X_t^i$  bid orders at price ‘ $i$ ’ and if  $X_t^i > 0$ , then there are  $X_t^i$  ask orders at price  $i$ . Now we have,

$$\begin{aligned} \text{Best Ask Price: } p_A(t) &= \inf \{i : X_t^i > 0\} \\ \text{Best Bid Price: } p_B(t) &= \sup \{i : X_t^i < 0\} \\ \text{Mid Price: } p_M(t) &= (p_A(t) + p_B(t))/2 \\ \text{Spread: } p_S(t) &= p_A(t) - p_B(t) \end{aligned} \tag{8.6}$$

The depth of the order book is stated relative to best bid and best ask on either side of the book. This way the model can be applied to order book for any equity subject

to ever changing, best bid and best ask. Define

$$Q_i^B(t) = \begin{cases} X_{p_A(t)-i}(t), & 0 < i < p_A(t) \\ 0, & p_A(t) \leq i < n \end{cases} \quad (8.7)$$

and

$$Q_i^A(t) = \begin{cases} X_{p_B(t)-i}(t), & 0 < i < n - p_B(t) \\ 0, & n - p_B(t) \leq i < n \end{cases}$$

Thus,  $Q_i^B(t)$  and  $Q_i^A(t)$  denote the number of buy orders at a distance ‘ $i$ ’ from ask and the number of sell orders at a distance ‘ $i$ ’ from bid, respectively. This representation highlights the shape (or depth) of the book relative to the best quotes on either side.

The following assumptions are made on the order arrivals: market buy or sell orders arrive at independent exponential times with rate,  $\mu$ . Limit buy or sell orders arrive at a distance of ‘ $i$ ’ ticks from the opposite best quote at independent exponential times with rate  $\lambda(i)$ . The cancellations of limit orders occur at a rate proportional to the number of outstanding orders,  $\theta(i)X$ . All these events are assumed to be mutually independent.

How the limit order is updated with the inflow of above events can be described as follows: for a state  $x \in \mathbb{Z}^n$  and  $1 \leq i \leq n$ , let  $x^{i\pm 1} = x \pm (0, \dots, 1, \dots, 0)$ , where ‘1’ denotes the change in the  $i$ th component. For example, a limit buy order at price level,  $i < p_A(t)$  increases the quantity at level ‘ $i$ ’, from  $x$  to  $x^{i-1}$  with rate  $\lambda(p_A(t) - i)$ . Similarly, limit sell order arrived which changes the order book from  $x$  to  $x^{i+1}$  with rate  $\lambda(i - p_B(t))$  for  $i > p_B(t)$ . For market buy order decreases at the ask price  $i$  and hence  $x \rightarrow x^{p_A(t)+1}$  with rate ‘ $\mu$ ’ and the sell order at the bid price, ‘ $i$ ’ will change the book status,  $x \rightarrow x^{p_A(t)+1}$  with rate  $\mu$ . Cancellation of a buy order at price level, ‘ $i$ ’, will decrease the quantity at the rate  $\theta(p_A(t) - i)|X_p|$  for  $i < p_A(t)$  and the sell order will decrease the quantity at the rate of  $\theta(i - p_B(t))|X_p|$  for  $i > p_B(t)$ .

It is assumed that the limit order arrival rate  $\lambda(i)$  follows a power law

$$\lambda(i) = \frac{\kappa}{i^\alpha} \quad (8.8)$$

which is confirmed by several empirical studies (Farmer (2002) [] for example). The empirical estimates of the parameter are based on the following quantities:  $s_m$  is the average size of the market orders,  $s_l$  is the average size of the limit orders and  $s_c$  is the average size of the cancelled orders. Also let  $N_l(i)$  be the total number of limit orders that arrived at a distance ‘ $i$ ’ from the opposite best quote and  $T_*$  is the total trading time in minutes and  $N_m$  is the number of market orders during the same time. Then

$$\hat{\lambda}(i) = \frac{N_l(i)}{T_*}, \quad 1 \leq i \leq 5 \quad (8.9)$$

where  $\hat{\lambda}(i)$  is extrapolated beyond five positions using the power law in (8.8) simply by minimizing  $\sum_{i=1}^5 (\hat{\lambda}(i) - \frac{\kappa}{i^\alpha})^2$ , over ‘ $\kappa$ ’ and ‘ $\alpha$ ’. The arrival rate is estimated as:

$$\hat{\mu} = \frac{N_m}{T_*} \cdot \frac{s_m}{s_l} \quad (8.10)$$

The cancellation rate as noted earlier is defined as proportional to the number of order at that price level,

$$\hat{\theta}(i) = \begin{cases} \frac{N_l(i)}{T_* Q_i} \cdot \frac{s_c}{s_i}, & i \leq 5 \\ \hat{\theta}(5), & i > 5 \end{cases} \quad (8.11)$$

It is understood that the cancellation is not due to executing market orders.

While these descriptives are easy to compute and are intuitive, but the main interest in modeling high frequency dynamics of order book is to predict short-term behavior of various key quantities that were identical earlier that may help in algorithmic trade executions. Some relevant questions are as stated before, given the state of the order book, what is the probability that mid-price will move up, what is the probability of executing both buy and sell order at the best quotes before the price changes etc.. These conditional probabilities are evaluated using Laplace transforms.

We present some key results in Cont et al. (2010) [68]. The probability of queue going up when there are no orders in the queue, for  $1 \leq d \leq 5$ , given that the best quotes are not changing is

$$p_{\text{up}}^d(m) = \begin{cases} \frac{\hat{\lambda}(d)}{\hat{\theta}(d)m + \hat{\lambda}(d) + \hat{\mu}}, & d = 1 \\ \frac{\hat{\lambda}(d)}{\hat{\theta}(d)m + \hat{\lambda}(d)}, & d > 1. \end{cases} \quad (8.12)$$

Other questions such as the probability that the mid price goes up when the spread  $\geq 1$  etc. are based on the following quantities:

- Probability of a market buy order

$$\frac{\mu^a}{\mu^b + \mu^a + \sum_j (\lambda_B(j) + \lambda_A(j) + \theta(j)Q_j^A(t) + \theta(j)Q_j^B(t))} \quad (8.13)$$

- Probability of a limit buy order  $d$  ticks away from the best ask

$$\frac{\lambda_B(d)}{\mu^b + \mu^a + \sum_j (\lambda_B(j) + \lambda_A(j) + \theta(j)Q_j^A(t) + \theta(j)Q_j^B(t))} \quad (8.14)$$

- Probability of a cancel buy order  $d$  ticks away from the best ask

$$\frac{\theta(d)Q_d^B(t)}{\mu^b + \mu^a + \sum_j (\lambda_B(j) + \lambda_A(j) + \theta(j)Q_j^A(t) + \theta(j)Q_j^B(t))} \quad (8.15)$$

The model was evaluated using data sets from Tokyo Stock Exchange. The model is shown to capture realistic features of the order book profile.

The use of Poisson processes to model the flows of limit orders, market orders

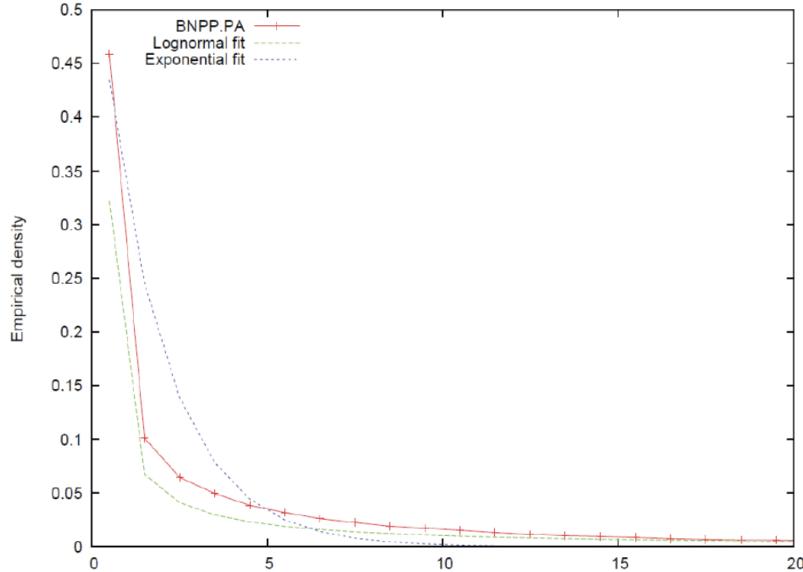


Figure 8.7: Distribution of inter-arrival times of market orders for stock BNPP.PA (Chakraborti et al. 2011).

and cancellations makes the method analytically tractable. The model captures the steady-state shape of the order book (although there may be other distributions that may be fit better besides Poisson, see Figure 8.7) but it may not be useful for predicting short-term behavior that may be more relevant for the traders. The assumption of Poisson process results in that the intensity of arrivals and cancellations does not depend on the state of the order book which is somewhat unrealistic given the feedback loop relationship between the behavior of the market participants and the history of the order book. Recent works, by Zhao (2010) [] and Abergel (2010) [] show that the memory-less property is not empirically valid. Orders tend to be clustered due to dependencies between liquidity taking and liquidity providing and also due to algorithmic execution of parent order that is split into numerous child orders. Also, the assumption that all orders are of the same size is too simplistic.

Zhao (2010) [] proposes a model based on marked point process when the mark represents order types, price, size etc.. The joint modeling also assumes that the order flow is self-exciting, i.e. new orders or cancellations occur in an orderly point processes with an intensity function that depends on the history of the limit order book. The model is somewhat difficult to estimate and does not fully capture the empirical results such as volatility clustering etc..

**Hawkes Process:** The features of volatility clustering and the significant autocorrelations in durations between order arrivals and significant cross-correlation of arrival rates across various event types are better captured by multi-dimensional Hawkes pro-

cess. For a given  $M$ -dimensional point process, let  $N_t = (N_t^1, \dots, N_t^M)$  denote the associated counting process and the Hawkes process is characterized by intensities,  $\lambda^m(t)$ ,  $m = 1, \dots, M$  as

$$\begin{aligned}\lambda^m(t) &= \lambda_0^m(t) + \sum_{n=1}^M \int_0^t \sum_{j=1}^P \alpha_j^{mn} e^{-\beta_j^{mn}(t-s)} dN_s^n \\ &= \lambda_0(t) + \sum_{n=1}^M \sum_{t_j < t} \sum_{j=1}^P \alpha_j^{mn} e^{-\beta_j^{mn}(t_j - t_i^n)}\end{aligned}\quad (8.16)$$

where the number of exponent kernels,  $P$ , is fixed and  $t_i^n$  is the  $i$ th jumping time of the  $n$ th variate. The scale and decay parameters,  $\alpha^{mn}$  and  $\beta^{mn}$  express the influence of the past events  $t_i^n$  of type ' $n$ '. The baseline model in (8.16) does not incorporate the effect of bid-ask spread on order flow. It is shown in Anderson, Cont and Vinkovskaya (2010) [] that  $\alpha$  and  $\beta$  parameters do depend on the bid-ask spread. A simulated Hawkes process is given in Figure 8.8. The main characteristic of the Hawkes process is that intensity goes up at each event and decays exponentially between events.

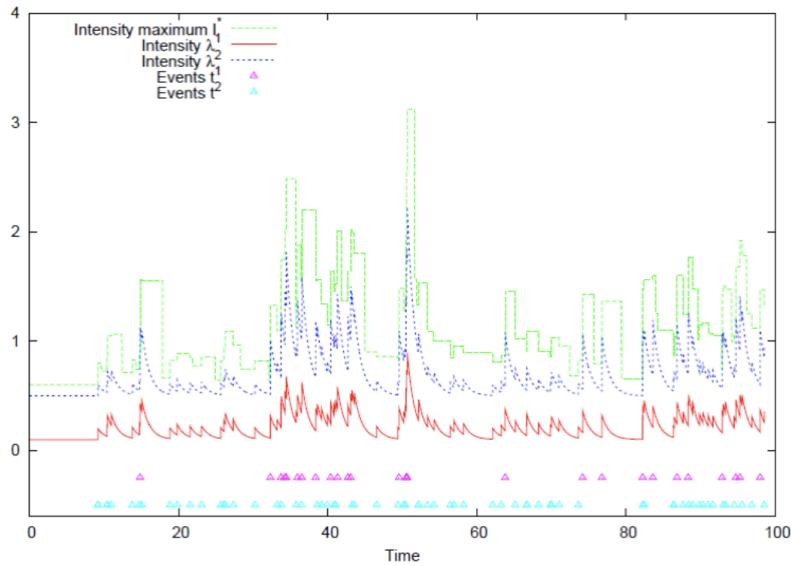


Figure 8.8: Simulated two-dimensional Hawkes process (Toke 2011).

The evolution of the order book is driven by six different intensity functions, but the model changes with the value of the spread. The base intensity and the decay rates are all assumed to be functions of the spread. The model is quite complex and requires the estimation of a large number of parameters, which in turn requires vast amount of data for calibration.

The mathematical framework was developed in Hawkes (1971) [] and was expanded in Hawkes and Oakes (1974) []. The application of Hawkes process in finance

is of recent origin: see Toke (2011) [] or Bacry et al. (2011) []. We will briefly discuss the likelihood estimates of the model in (8.16). First, we want to observe for a Poisson process, the intensity function  $\lambda(t) = \lambda_0$ , a constant. The log-likelihood of the model,

$$\ln \mathcal{L}(\{t_i\}_{i=1,2,\dots,N}) = \sum_{m=1}^M \ln \mathcal{L}^m(\{t_i\}) \quad (8.17)$$

where

$$\ln \mathcal{L}^m(\{t_i\}) = T - \sum_{i=1}^N \sum_{n=1}^M \frac{\alpha^{mn}}{\beta^{mn}} (1 - e^{-\beta^{mn}(T-t_i)}) + \sum_{t_i^m} \ln [\lambda_0^m(t_i^m) + \sum_{n=1}^M \alpha^{mn} R^{mn}(l)]$$

and

$$R^{mn}(l) = \sum_{t_k^n t_l^m} e^{-\beta^{mn}(t_l^m - t_k^n)}$$

is the cumulative decay function. Using the Multiplicative Random Time Change Theorem (see Meyer (1971) []) that states that with a certain transformation of the variables, the Hawkes process can be changed to a unit rate homogeneous Poisson process. Transforming the time variable ‘ $t$ ’ into ‘ $\mathcal{T}$ ’, where

$$\mathcal{T} = \int_0^t \lambda(s) ds \quad (8.18)$$

Hawkes process becomes a Poisson process with unit rate. Thus

$$\mathcal{T}_i - \mathcal{T}_{i-1} = \Lambda(t_{i-1}, t_i) = \int_{t_{i-1}}^{t_i} \lambda(s) ds \quad (8.19)$$

is exponentially distributed.

More generally in the ‘ $M$ ’ dimensional multivariate case,

$$\begin{aligned} \Lambda^m(t_{i-1}^m, t_i^m) &= \int_{t_{i-1}^m}^{t_i^m} \lambda_0^m(s) ds + \int_{t_{i-1}^m}^{t_i^m} \sum_{n=1}^M \sum_{t^n < t_{i-1}^m} \alpha^{mn} e^{-\beta^{mn}(s-t^n)} ds \\ &\quad + \int_{t_{i-1}^m}^{t_i^m} \sum_{n=1}^M \sum_{t_{i-1}^m < t^n < s} \alpha^{mn} e^{-\beta^{mn}(s-t^n)} ds \end{aligned}$$

will follow exponential distribution. The  $\Lambda^m(t_{i-1}^m, t_i^m)$  are known as compensators and can be verified empirically if it follows exponential distribution.

### 8.3.1 Application of a Hawkes process

We want to illustrate the model with an application. We will first look at the statistical properties of the data and then look at the results of fitting one and two dimensional Hawkes processes to it. Our findings show that a two dimensional model

based on Hawkes processes performs considerably better than a model based on two independent Poisson Processes. This provides strong evidence of the fact that the two key features of Hawkes processes (path dependency of the order flow and the possibility of modeling the interaction between the dimensions) are capable of capturing and replicating some of the key features of the empirical data. Moreover, there is evidence of significant and asymmetric interplay between the buy and the sell sides for the Limit Orders and evidence of limited interplay between the two sides when it comes to Market Orders. Finally, a study of the statistical properties of the orders showed a complex of alternating symmetric and asymmetric behavior of the LOB.

The data at hand is INET data for XOM (Exxon Mobil) for the 1st of September 2010. The first step was to look at the submission frequency of the Limit Orders during the trading day. Figure 8.9 shows such frequency plot and we can see how in the first hour a half and in the last half an hour of the trading day there is a much higher submission frequency (with peaks of 25/30 orders submitted per second) than during the rest of the day (around 5–10 orders per second).

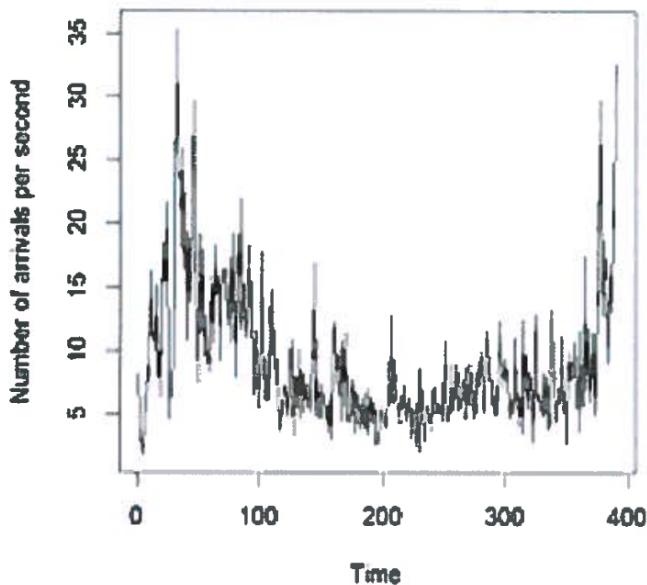


Figure 8.9: Frequency of Limit Order submission for 09/01/2010 (time in minutes after market opening).

This means that if we choose to look at the Limit Order flow with a model that doesn't allow for a varying baseline intensity, we will have to drop these two periods of increased trading activity and keep the data for only the middle portion of 6 hours of the trading day.

We then looked at the data for the central portion of 6 hours of the trading day and found that market orders are just a function of Limit Orders. There were almost

Table 8.2: XOM data for 09/01/2010

	Limit Orders	Market Orders	Total
Buy	48,541	2,107	50,648
Sell	41,194	1,882	43,076
Total	89,735	3,989	93,724

Table 8.3: LO executions and percentage breakdown by number of fills required

Number of LO	% at least partially fill	% completed in one fill	% completely fill
89,735	6.57	5.35	6.15
<hr/>			
	1	2	3
	87.14	10.83	1.34
		4	5
		0.37	0.13
		6	≥ 7
		0.06	0.13

90,000 Limit Orders submitted during that period of the day, compared to the only about 4,000 Market Orders. Similarly, there is an imbalance in the side of the order submission, with around 42,000 Limit Orders submitted on the Sell side compared to the 48,000 submitted on the Buy side (about 18% more). This side-imbalance is also present in the Market Order submission although not that strongly: 1,882 Market Orders submitted on the Sell side compared to the 2,107 orders submitted on the Buy side (only about 12% more). This shows how, on this trading day, most of the “action” was occurring on the Buy Side of the book which could mean that the markets were confident in an increase of the price of XOM (in line with the historical performance of the stock for that day).

Our next step was to re-construct the Limit Order book from the order flow so to look at the life of the Limit Orders after their submission. We first looked into how many of the submitted Limit Orders get executed and, when they do, in how many fills. We found Table 8.3 that only about 6.5% of all submitted LOs get at least partially filled with the percentage dropping to only about 6% for those LO that actually get completely executed. Also, out of the LO that are completely filled nearly 90% of them is filled in just one execution. This allows us to say that the magnitude of the LOs is comparable to that of the Market Orders, given that nearly all of them require only one Market Order to fill it completely. It is also interesting to note how few of the LOs that are submitted actually get executed.

This means that either very few LO are actually submitted close to the market (hence have a chance to get executed) or that many of them don’t stay in the book long enough to get executed. This led us to look into the life-time of the LOs and, more specifically, at their cancellation and execution times. We focused on how long, on average, for an order to be either not filled or partially filled LO stayed in the book before getting cancelled and at how long did it take for a LO to be either filled for the first time or completely filled. Results are given in Table 8.4.

In this particular case, LOs stay in the book for much longer if they are on the sell side than if they are on the buy side. Orders that have never been filled are cancelled after 250 seconds if they are on the sell side compared to 177 seconds if they are

Table 8.4: Mean LO cancellation and execution times (in seconds)

No Fills		Partial Fills	
Buy Orders	Sell Orders	Buy Orders	Sell Orders
177.8	250.5	46.1	63.0
First Fill		Completion	
25.3	56.2	25.6	57.6

on the buy side. Similarly, a partially filled LO on the sell side is cancelled after 63 seconds compared to 46 seconds if it's on the other side of the book. Also, if we look at the time it takes for a LO to be filled for the first time we see that that happens after about 56 seconds on the sell side compared to only 25 seconds on the buy side. Finally, for the LO that do get completely filled, it takes on average 1 minute for that to happen if the order is on the sell side compared to only 25 seconds if the order is on the buy side. This life-time difference for LOs on the two opposite sides of the book is consistent with what we saw earlier in the imbalance of the LO and MO submission. In fact, in this case as well we see how the buy side appears to be the more active side of the book. The shorter execution times for the orders on the buy side could be caused by either more aggressive pricing or by the price movements of the markets while the shorter lifespans and the higher number of orders on that side seem to indicate the use of more active trading strategies. However, there is another interesting observation that can be made. We see that on average a sell LO that is filled for the first time (but that doesn't get completely filled), is cancelled after about 7 seconds, compared to the 20 seconds that it would take had the order been on the buy side. This means that on average it takes three times longer for a partially (and never completely) filled order to be cancelled if it is on the buy side (which is the more active side) than if it is on the sell side (which is the less reactive side). The asymmetry between buy and sell side at the aggregated level has been noted in the literature.

The next step was to observe changes in the behavior of the book when moving further away from the market. We looked into the order submission and cancellation frequencies as a function of distance from the market price. Figure 8.12 and Figure 8.14 show that most LO are submitted and cancelled at the market (respectively, 40% and almost 30%) and that the rates drop sharply with each price level further away from it. This supports the conclusion that the very small percentage of executed LO is not caused by the fact that most LO are submitted far away from the market and hence don't get the chance to get executed. We can also see how approximately 70% (96%) of all Limit Orders is placed in the top 5 (20) levels of the book and that 69% (96%) of all cancellations also occurs in the top 5 (20) levels. This a first indication of the fact that Level III data offers more information than Level II since with the top 5 levels of the book we were only able to capture around 70% of all the occurring events. However, this also indicates that there is no need to use all the information provided in Level III data since by looking at only the top 20 levels of the book we are still able to capture nearly all of the occurring events. Both plots show a very similar

behavior for the buy and sell side which led us to assume an almost symmetric behavior of the book. This assumption was later tested and confirmed at a 99% confidence level with a two-sided Kolmogorov-Smirnov Test.

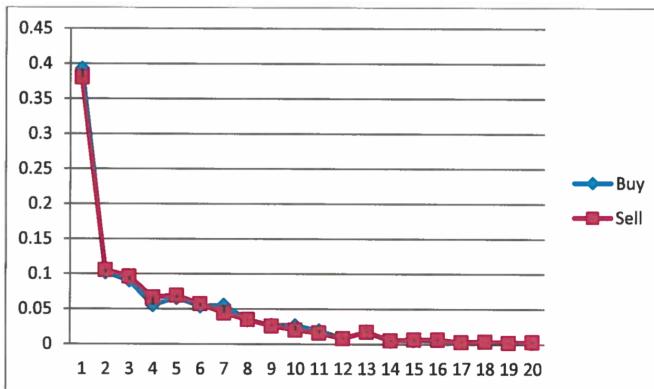


Figure 8.10: Submission frequencies–Distance near touch price  
Kolmogorov-Smirnov Test Statistics= 0.61 → fail to reject null at 1% sig. level

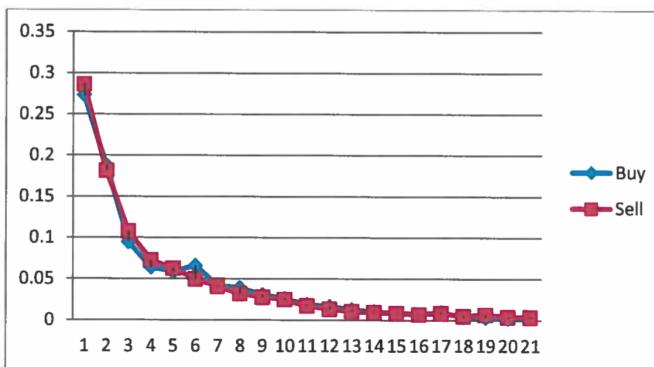


Figure 8.11: Cancellation Frequencies–Distance from near touch price  
Kolmogorov-Smirnov Test Statistic= 0.83 → fail to reject null at 1% sig. level

Better understanding what happens to the LOs once they enter the book is key in developing a more accurate way of choosing when to submit a LO or a MO. For this reason we also looked at the cancellation frequencies as a function of position in the queue and we found that (Figure 8.12) most of the cancellations occur when the LO are in 4th or 5th position in the queue. In this case as well, we carried out a test of the assumption of a symmetric behavior of the book which confirmed previous similar findings. However, it would be more interesting to look at the cancellation frequencies as a function of overall distance from the market (as a measure of distance from possible execution).

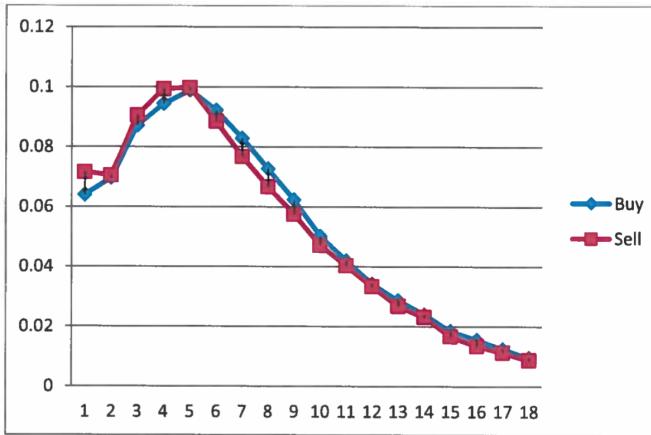


Figure 8.12: Cancellation Frequencies—position in queue.

With two more trading days (3rd and 9th of September 2013) data available, we are able to study the robustness of the above results. The submission and cancellation frequencies as a function of relative price which were symmetric for the 1st of September were not symmetric anymore on the 3rd of September. In this trading day, Figure 8.14, the peak of activity on the buy side occurred 12 price levels away from the market. This seems to suggest that on the 3rd of September the markets were betting that the price of XOM would drop in the foreseeable future, possibly reacting to some new event. A closer investigation of that time period revealed in the previous days the announcement by Exxon Mobil of the plan of closing a refinery which then led to a drop in the price shortly after 09/03/2010.

A second interesting difference was in the amount of volume on the two sides of the book (Figure 8.13). Across the 3 days there was a significant volume imbalance with considerably more volume on the sell side than on the buy side but the amount of imbalance between the two sides changes significantly across the three days. Overall there is no explanation as to why there would be such a strong volume imbalance in favor of the sell side other than some kind of temporary anomaly. These findings seem to point out the fact that the shape and behavior of the LOB change dramatically across different days and that, in order to capture some kind of more general/average behavior of the book, many more days of data are required.

### 8.3.2 One-Dimensional Models

We started by building a one dimensional model for the Limit Orders submitted on the buy side and one for those submitted on the sell side. Similarly, we build a one dimensional model for the Market Orders submitted on each side of the book. In the one dimensional case, the intensity function driving the Hawkes process is of the

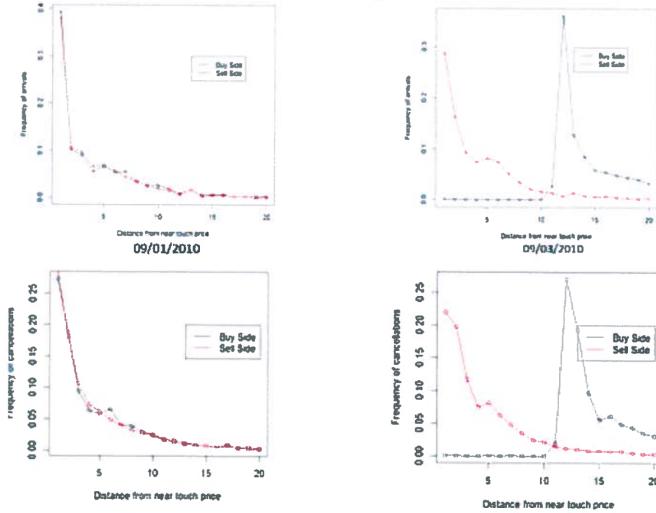


Figure 8.13: LO submission and cancellation frequencies.

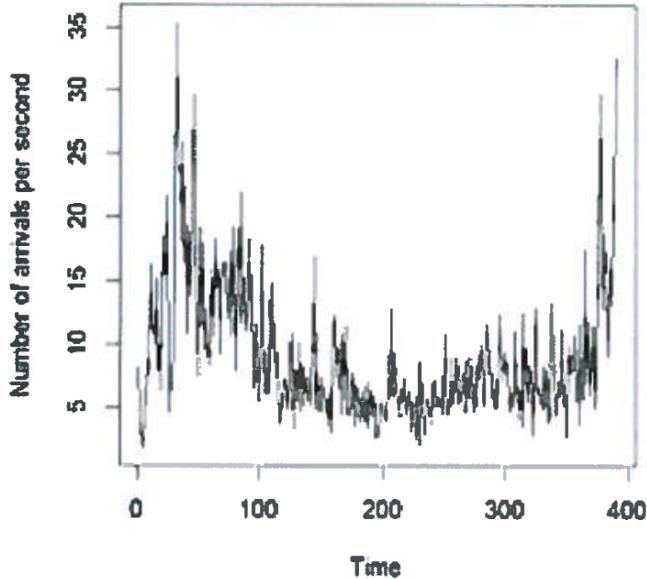


Figure 8.14: Frequency of Limit Order submission for 09/01/2010 (time in minutes after market opening).

form

$$\lambda(t) = \lambda_0 + \alpha \sum_{t_i < t} e^{-\beta(t-t_i)} \quad (8.20)$$

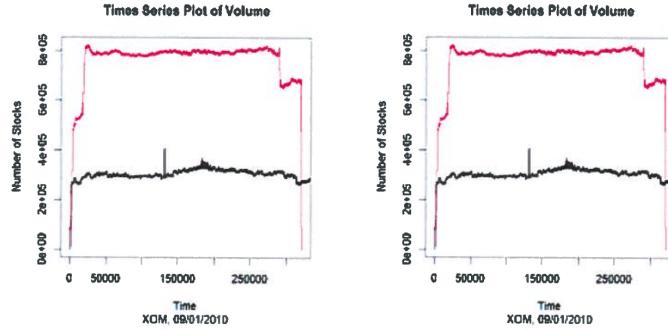


Figure 8.15: Frequency of Limit Order submission for 09/01/2010 (time in minutes after market opening).

Table 8.5: 1-D Limit Order Model (buy side)

	$\lambda$	$\alpha$	$\beta$
09/01/2010	0.008108431	1.421767060	1.435094466
09/03/2010	0.006127003	1.290012415	1.299564704
09/09/2010	0.002341542	1.023410145	1.026757239

with  $\lambda_0$  indicating the baseline intensity,  $\alpha$  indicating the jump in the intensity after the occurrence of an event and  $\beta$  indicating the rate of exponential decay of the intensity after the occurrence of an event.

Table 8.5 and Table 8.6 show the results of the MLE estimation of the parameters of the one dimensional models for the Limit Orders on the two sides of the book. The difference in the values of the estimates of the parameters for the two sides of the book (strongest on the 9th and weakest on the 1st of September) indicates that there is a level of asymmetry between these two sides. Also, even though the magnitude of the values of the parameters is the same for all three days (on each side of the book) there is still some variation between them, especially when we look at the model for the buy side. This seems to suggest that the behavior of the book can change considerably from day to day and that a considerable amount of data might be necessary in order to capture some kind of general behavior.

One of the advantages of having chosen an exponential decay for the intensity

Table 8.6: 1-D Limit Order Model (sell side)

	$\lambda$	$\alpha$	$\beta$
09/01/2010	0.002055786	1.071071139	1.073989940
09/03/2010	0.002342963	0.983678355	0.986999404
09/09/2010	0.006123845	0.946420813	0.955774394

function of the process, becomes apparent when we analyze the meaning of these parameters. Using a well-known property of exponentially distributed variables, we can tell the half-life of the excitation effect caused by the occurrence of an event. In the specific, in our one-dimensional models we find that, on average, half of the excitation effect caused by the occurrence of an event is dissipated after about 0.7 seconds ( $\frac{\ln 2}{\beta} \approx$  half-life). Also, the small estimated values for  $\alpha$  indicate that the occurrence of an event has a very small impact on the likelihood of another event occurring. This seems to suggest that there is a weak self-excitation effect in the LO process when modeling it with a one-dimensional model.

We then build, in a similar fashion, the two one-dimensional models for the flow of buy and sell Market Orders for all three days and obtained the following parameter estimation: The first thing that is obvious is the asymmetry between the buy and the

Table 8.7: 1-D Market Order Model (buy side)

	$\lambda$	$\alpha$	$\beta$
09/01/2010	0.0238743	7.3187478	19.8477514
09/03/2010	0.01450873	3.38168999	7.70581048
09/09/2010	0.0267254	4.8818834	13.5427254

Table 8.8: 1-D Market Order Model (sell side)

	$\lambda$	$\alpha$	$\beta$
09/01/2010	0.02083999	9.38842281	24.55777564
09/03/2010	0.01616096	6.64370599	16.68090519
09/09/2010	0.0239992	8.2180945	25.0115980

sell sides. Nevertheless, for the Market Orders they asymmetry appears to be much more significant than it was for the Limit Orders, given the much bigger differences in the values of the estimated parameters. Also, for Market Orders we see that the occurrence of each event seems to have a much stronger impact on the likelihood of occurrence of future events than it was the case for Limit Orders. In fact, the values of  $\alpha$  in Table 8.7 and Table 8.8 are considerably higher than those we saw in Table 8.5 and Table 8.6. Similarly, the values of  $\beta$  are also much larger in the Market Order models than they were in the Limit Order models which means that the half-life of the excitation effect is much shorter for Market Orders (0.03 seconds vs. 0.7 seconds) than for Limit Orders. These differences in the results of the estimations are not that surprising if we consider what we saw in Table 8.2 for order submissions. Market Orders are submitted in much fewer numbers than Limit Orders which explains why even though the occurrence, such self-excitation has a very short lifetime. The significant increase in the chances of another Market Order occurring given that one has just occurred (even through for a very short time), might seem contradicting the fact that very few Market Orders actually occur. However, what can be inferred is that

Market Orders depend strongly from the occurrence of other Market Orders and that they tend to cluster around submissions.

### 8.3.3 An extension

An important limitation of this basic Hawkes Process model is that each event has the same impact on the intensity function. This implies that each event carries the same amount of information in the process. However, it is easy to imagine that a large LO would carry much more information about future price changes than a small one and this should be accounted for in the model. We include order size in the intensity function by scaling the size of the jump  $\alpha$  in the value of the intensity by the ratio  $\frac{w_i}{\bar{w}_l}$ , where  $w_i$  is the size of the LO and  $\bar{w}_l$  is the average LO size up to that point. The intensity function for this “extended” one-dimensional model becomes

$$\lambda(t) = \lambda_0 + \alpha \sum_{t_i < t} \frac{w_i}{\bar{w}_l} e^{-\beta(t-t_i)}$$

We then compared the performance of this “size-adjusted” model with the “basic” one by using the before mentioned *Multivariate Random Time Change* Theorem and looking at the Q-Q plot of the compensators for the two models. Figure 8.16 and Figure 8.17 show how in both case (buy and sell Limit Orders) adjusting for size doesn’t improve the fit of the model. This result is somewhat counter-intuitive given that more than 70% of all Limit Orders are of size 100 shares and that this should imply that any deviation from this size explanation is that trading strategies are successful in concealing the true intension of the traders and that there is no significant information left in the size of the LO once that the strategies are implemented and the orders are submitted.

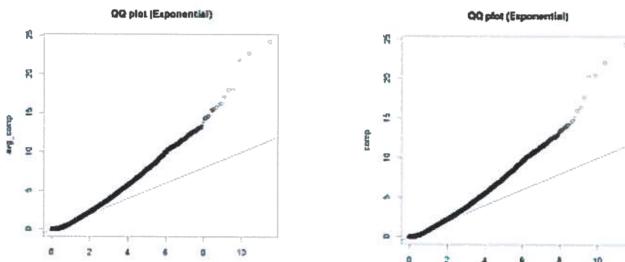


Figure 8.16: LO-buy side with size adjustment and without.

### 8.3.4 Two-Dimensional models

Our next step was to build two-dimensional models that would simultaneously model the buy and sell sides of the book and with two separate (yet interacting) in-

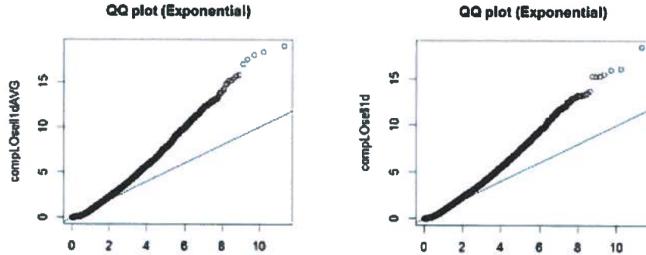


Figure 8.17: LO–sell side with size adjustment and without.

Table 8.9: 2-Dimensional LO model

$\lambda^B = 2.480 \cdot 10^{-5}$	$\alpha^{BB} = 14.386$	$\beta^{BB} = 31.517$	$\alpha^{BS} = 0.303$	$\beta^{BS} = 0.464$
$\lambda^S = 2.677 \cdot 10^{-5}$	$\lambda^{SS} = 15.657$	$\beta^{SS} = 37.299$	$\alpha^{SB} = 0.149$	$\beta^{SB} = 0.312$

tensity functions. The intensity functions for this kind of model become

$$\begin{aligned}\lambda^B(t) &= \lambda_0^B + \int_0^t \alpha^{BB} e^{-\beta^{BB}(t-u)} dN_u^B + \int_0^t \alpha^{BS} e^{-\beta^{BS}(t-u)} dN_u^S \\ \lambda^S(t) &= \lambda_0^S + \int_0^t \alpha^{SB} e^{-\beta^{SB}(t-u)} dN_u^B + \int_0^t \alpha^{SS} e^{-\beta^{SS}(t-u)} dN_i^S\end{aligned}$$

and we have 5 parameters for each dimension of the model for a total of 10 parameters across the two dimensions (buy and sell). The two additional parameters in each dimension account for the cross-excitation effect between the two dimensions. In our case,  $\alpha^{BS}$  tells us the increase in the likeliness of a LO occurring on the buy side after the occurrence of a LO on the sell side, while  $\beta^{BS}$  tells us the rate of decay of the buy side intensity after the occurrence of a LO on the sell side (and vice-versa for  $\alpha^{SB}$  and  $\beta^{SB}$ ). We then used MLE to estimate the parameters of the model and we can make several remarks about the results (Table 8.9).

There appears to be some asymmetry in the cross-excitation effects with the occurrence of an event on the sell side having a stronger impact on the likeliness of occurrence of an event on the buy side rather than the other way round ( $\alpha^{BS} > \alpha^{SB}$ ). On the other hand though, since  $\beta^{BS} < \beta^{SB}$ , the effect on the sell side of an occurrence on the buy side is more persistent than that of an event on the buy side on the sell side (2.22 seconds vs. 1.49 seconds). The self-excitation components of the model are rather similar for both dimensions and it is interesting to point out the very short persistence of the self-excitation effect in both cases (about 0.02 seconds).

In order to evaluate whether the two key features of a model based on Hawkes processes (path dependency of the order flow and the possibility of modeling the relation between the various dimensions) are really helpful in improving the performance of the model, we decided to compare the performance of our model to that of a bi-dimensional model based on two independent Poisson Processes (which by

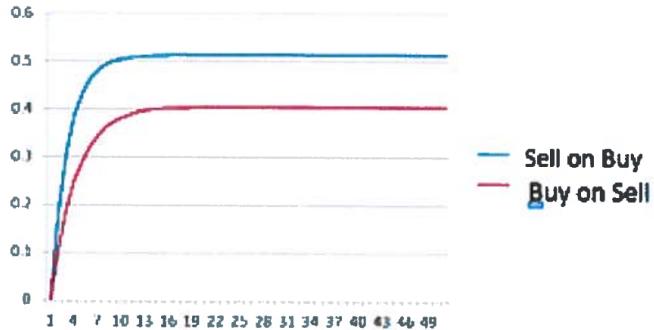


Figure 8.18: Asymmetry in cross-excitation effect.

definition of Poisson process cannot account for any dependency in the order flow and by construction assume independence between the two dimensions). The results proved to be very encouraging with a remarkably better fit to the empirical data for the Hawkes Process based model (Figure 8.19).

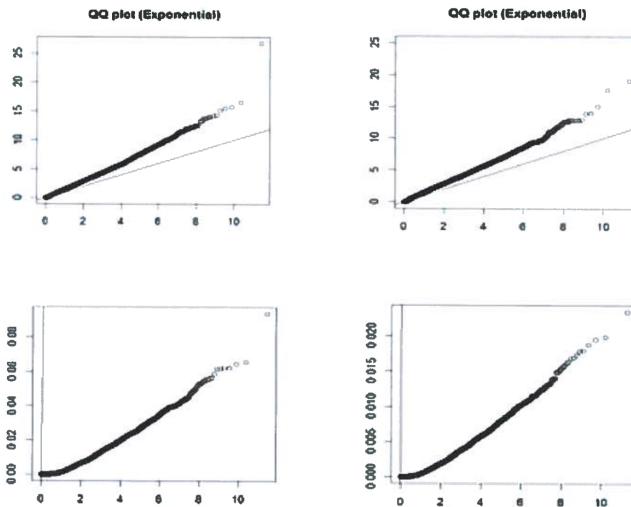


Figure 8.19: Top–Fit of Hawkes Process (buy &amp; sell); Bottom–Fit of Poisson Process (buy &amp; sell).

Following a similar procedure, we build a bi-dimensional model for the Market Orders and performed MLE to estimate the parameters. The results (Table 8.10) showed very weak and symmetric cross-excitation between the dimensions which, however, turns out to be very persistent (141 seconds for the half-life of the effect of buy on sell and 215 seconds for that of sell on buy). The self-excitation effect is

Table 8.10: 2-Dimensional MO model

$\lambda^B = 2.701 \cdot 10^{-5}$	$\alpha^{BB} = 11.793$	$\beta^{BB} = 36.153$	$\alpha^{BS} = 0.0037$	$\beta^{BS} = 0.0049$
$\lambda^S = 3.451 \cdot 10^{-5}$	$\alpha^{SS} = 12.454$	$\beta^{SS} = 35.063$	$\alpha^{SB} = 0.0017$	$\beta^{SB} = 0.0032$

comparable across the two dimensions and is also very short lived (with a half-life of about 0.02 seconds).

Finally, we compared the performance of our model to that of the bi-dimensional Poisson Process. Similarly to the LO case, the model at hand significantly outperforms the Poisson Process one for both dimensions.

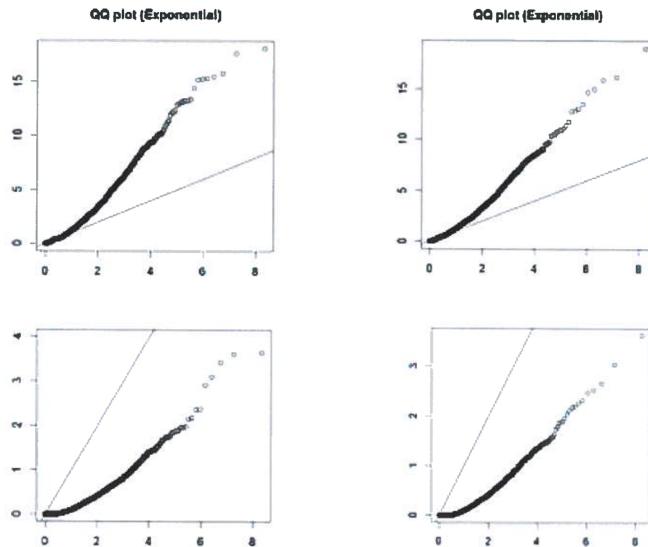


Figure 8.20: Top–Fit of Hawkes Process (buy & sell); Bottom–Fit of Poisson Process (buy & sell).

### 8.3.5 Limitations and findings

There are some important considerations to be made based on our analysis. The first issue is related to the cleaning of the data that was necessary in order to adapt it to our needs. In fact, Hawkes processes are “simple” processes which means that they don’t allow for more events to occur simultaneously. Unfortunately, in the INET data it is quite common to encounter clusters of 10/20 events reported with the same timestamp. This is typical of high frequency data due to the approximation with which events are often reported and to the obvious limitation of the chosen time scale (e.g., milliseconds). In our case we decided to keep only the first event of each cluster, dropping all the others. A possible alternative used in the literature is to simply uni-

formly distribute the events from the clusters across the time interval given by the occurrence of the cluster and that of the next event. However, this approach could introduce bias in the parameter estimation phase. In fact, the clustering of events by itself is of interest.

Also, an important limitation of the bi-dimensional Hawkes Process model is that it is very computationally intense. Evaluating the 5 parameters for one of the two intensity function using data from only one trading day took almost 48 hours. It becomes clear that in order to evaluate the parameters of a model using data for a longer time period or if evaluating the parameters for a model with higher dimension (hence with more parameters) will require considerably more computational power.

---

#### **8.4 Models for Level II and Level III data**

---

---

#### **8.5 Estimation of Hidden Orders**

---

---

#### **8.6 Volume Imbalance and Other Indicators**

---

---

#### **8.7 Hidden Markov Models (HMM) for dynamic market microstructure: Estimation of Parameter and States**

---

---

#### **8.8 Appendix: Adaptive Particle Filters in HMM**

---

---

#### **8.9 Supplements and Problems**

---



# 9

---

## *Algorithmic Trading In Electronic Platforms*

---

### CONTENTS

9.1	Execution Algorithms: History and Review .....	244
9.1.1	Bertsimas and Lo Model .....	245
9.1.2	Almgren and Chriss Model .....	246
9.1.3	Obizhaeva and Wang Model .....	249
9.1.4	Easley, De Prado and O'Hara Model .....	252
9.2	Multiple Exchanges: Smart Order Routing .....	253
9.3	Market Impact Models: Permanent and Temporary Impacts .....	257
9.4	Microstructure Noise Models for Multiple Stocks: Issues of Synchronizations, etc. ....	274
9.5	Simultaneous Execution Algorithms for Multiple Instruments .....	275
9.6	A Stochastic Adaptive Control Approach to Optimal Execution .....	275
9.7	Role of Market Makers in Limit Order Markets .....	275
9.8	Open Issues .....	275
9.9	Supplements and Problems .....	275

While algorithmic trading techniques are useful in both low and high frequency settings, their importance has become more pronounced in the age of electronic trading because of the speed in execution and the necessity for efficient matching of demand and supply sides in the limit order books. Added to this is that modern equity markets where order submission and cancellations are automated are also highly fragmented with dozens of exchanges and about forty alternative trading systems where investors can choose to trade. The trading algorithms differ over various types of market participants, but all of them try to optimize dynamically where, how often and at what price to trade. Typically, the execution layers can be broadly described as follows (see Gatheral (2012)[]):

- **The Macrotrader:** Given a meta-order (parent trade) translate the objective function of a strategy into a trading schedule; when the algorithm should trade, in what size (child order) and for how long.
- **The Microtrader:** Given a slice (child order) of the meta-order to trade, this layer of algorithm decides whether to place it as a market or as a limit order and at what price level.

- **The Smart Order Router:** For a given order (quantity and price level) to which venue should the order be sent.

In this chapter we discuss these and other execution issues in the high frequency trading context. Generally the objectives of equity trading are to minimize the probability of making errors—minimizing the trading impact and avoiding adverse selection. In order to achieve high completion rates, the trading should operate with speed and necessary risk control. It has to be noted that the average trade size has come down over time; thus the parent order are increasingly decimated and are sent to the exchanges as small child orders. Handling this with large parent order sizes with the focus on short-term alphas requires a great deal of trade timing insights.

## 9.1 Execution Algorithms: History and Review

The primary goal of most of the execution algorithms is to minimize the transaction costs which are broadly divided into direct costs and into indirect costs. The direct costs such as commissions, taxes etc are easy to measure and primarily depend on quantity of trading. The indirect costs such as market impact and opportunity costs are hard to measure and depend upon trading strategies. Because many mutual fund and hedge fund companies engage in portfolio rebalancing periodically, they trade a large block of shares in a fixed duration. There is a trade-off between trading a large market order that is likely to affect the price adversely and splitting the parent order into many child orders and let the market recover after each trade. The modeling elements generally involve representing the permanent impact of a parent order and the temporary impact of trading a child order when the price normally returns back to equilibrium, short time after the trade, drift in price (alpha) and the volatility of the stock. The optimizing function balances the cost-loss in revenue due to price impact and the risk-uncertainty due to volatility of the underlying stock. If the trade occurs fast, we incur high cost and low risk but if the trade is slow, we incur low cost and high risk.

The execution algorithms are generally evaluated through some benchmark prices. Some examples are:

- $p_0$ : Arrival price, that is the market mid-quote at the start of parent order execution at  $t = 0$ .
- $p_{VWAP}$ : Volume weighted average market price over the period of execution,  $t \in (0, T)$ .

If the parent order of size,  $X$  is split into ‘ $n$ ’ child orders with size,  $x_k$ ,  $k = 1, \dots, n$ , then

- $p_{\text{exec}}$ : Volume weighted average price of the parent order’s execution.

Two related quantities that the traders examine in their evaluation strategies are

- $p_{\text{bench}} - p_{\text{exec}}$ : Slippage =  $x$ ;
- $p_0 - p_{\text{VWAP}}$ : Market Impact

The execution cost is taken to be the slippage and the execution risk is its standard deviation.

The influential paper by Kyle (1985) [182] argues that traders split the parent order to hide private information. The splitting of the order and how they are sent to the market—the strategies—depend on the evolution of the stock price. These are broadly classified into static, one determined in advance of trading and into dynamic, one that depends on the state of the market. The static strategy under some conditions can also be dynamically optimal. Few commonly used strategies are:

$$\text{VWAP-Strategy: } x_k = x \cdot \frac{V_k}{V}$$

Implementation Shortfall Strategy: Minimize  $E(x) + \lambda \text{Var}(x)$

$$\text{Percentage of Volume Strategy: Minimize } \left( \frac{\text{Volume}_{\text{exec}}}{\text{Volume}_{\text{market}}} - v_{\text{target}} \right).$$

These strategies are static and they are based on a priori models for the evolution price and volume traded over different times of the day.

We present below select results in this area. The quantities that the trader can determine are the number of child orders ( $n$ ) and their sizes ( $x_k$ ) and the duration for the trading ( $T$ ). The relevance of this problem to the practitioners is discussed in Chan and Lakonishok (1997) [56] and Keim and Madhavan (1997) [178]. Bertsimas and Lo (1998) [25] develop a model that suggests splitting the parent order to reduce the average trading costs. Almgren and Chriss (2000) [5] split the parent order to minimize the mean-variance of trading costs similar to Markowitz's portfolio theorem. We describe in the following sections, these strategies which are mainly static.

### 9.1.1 Bertsimas and Lo Model

The basis of the model discussed here is that trading affects not only the current prices but also the price dynamics that affects the future trading costs. Thus the problem involves dynamic optimization but the solution turns out to be static. Assume that an investor wants to sell a parent order of  $X$ -shares in  $(0, T)$ , where  $T$  is arbitrary, and let  $x_k$  be the number of shares sold (child orders) with the realized price,  $\tilde{p}_k$ . The cost of trading, also known as slippage is:

$$x = Xp_0 - \sum_{k=1}^n x_k \tilde{p}_k \quad (9.1)$$

where  $p_0$  is the arrival price. The model proposed in Bertsimas and Lo (1998) [25] can be stated as:

$$\min_{x_k} E_1 \left[ \sum_{k=1}^n x_k p_k \right] \text{ such that } \sum_{k=1}^n x_k = X \quad (9.2)$$

The criteria in (9.2) call for determining the sizes of child orders a priori. To solve (9.2), the model for price, where the price impact and price dynamics are separated, is taken to be,

$$p_k = p_{k-1} - \theta x_k + \varepsilon_k \quad (9.3)$$

where  $\theta > 0$  representing the price impact and when  $\theta = 0$ , the model is simply a random walk model discussed in Chapter 2. Under (9.3) the best strategy for splitting the parent order,  $x$ , is shown to be

$$x_1^* = x_2^* = \dots = x_n^* = \frac{X}{n} \quad (9.4)$$

the equal split. What the trader needs to decide a priori is the number of splits,  $n$ . The allocation strategy in (9.4) is termed as the time weighted average prices (TWAP) strategy in industry.

The dynamic programming solution to (9.2) via the so called Bellmann equation is elegant and simple because the price impact ( $\theta x_k$ ) does not depend upon the prevailing price ( $p_{k-1}$ ) nor on the size of the unexecuted order ( $X - \sum_{j=1}^{k-1} x_j$ ). The indirect assumptions here are that the volume curve is even over different times of the day and the volatility profile are flat. Bestimas and Lo (1998) [25] also consider extensions of the model (9.3), where a serially correlated state variable such as market condition (S&P 500) is added. In this case it is shown that the best execution strategy at any point is a function of the state variable and the remaining unexecuted size.

As a point of comparison to the equal allocation strategy observe that VWAP-strategy is similar in the sense that it does not also depend on any price movement. Simply the allocation is made as  $\frac{x_k}{X} = \frac{V_k}{V}$ , which leads to

$$x_k = X \frac{V_k}{V} \quad (9.5)$$

where  $V_k$  is the market volume during the  $k$ th time interval. If the traders are restricted to only ' $n$ ' time intervals, it is important to have a good forecast for the quantities,  $V_k$  and thus,  $V$ .

### 9.1.2 Almgren and Chriss Model

The seminal paper by Almgren and Chriss (2000) [5] and Almgren (2003) [] consider the same problem taking into account the stock characteristics such as volatility via the mean-variance formulation of minimizing execution costs. The formulation treats the execution as a tradeoff between risk and cost; the faster the execution, higher the cost but lower the risk. The objective again is to liquidate,  $X$  units before time

‘ $T$ ’ at minimum cost, in time steps,  $t_k = k\tau$  for  $k = 0, 1, \dots, n$  where  $\tau = T/n$  is the average length of interval between successive child order submission times. If  $X_k$  is the number of units held at  $t_k$ , the goal is to decide on the trading trajectory,  $X_0, X_1, \dots, X_n$  with  $X_0 = X$  and  $X_n = 0$ . The number of units liquidated  $x_k$  between  $t_{k-1}$  and  $t_k$  is called the trading schedule, with  $x_k = X_{k-1} - X_k$ . The trading strategy requires determining ‘ $x_k$ ’ in terms of information available at  $t_{k-1}$ . The static strategies determine the trading schedule in advance of starting trade time,  $t_0$  and the dynamic strategies depend on information available up to  $t_{k-1}$ . Almgren and Chriss (2000) [5] find trading trajectories that minimize  $E(x) + \lambda \text{Var}(x)$  for various values of  $\lambda$ .

The model for price impact is same as in (9.3) but stated as

$$p_k = p_{k-1} - (\tau\theta)\frac{x_k}{\tau} + \sigma\tau^{1/2}\varepsilon_k \quad (9.6)$$

where  $v_k = x_k/\tau$  is the average rate of trading in  $[t_{k-1}, t_k]$  and  $\varepsilon_k$  are independent with mean zero and unit variance and ‘ $\sigma$ ’ is the volatility of the stock. The temporary impact represented by price per share received for sale in  $[t_{k-1}, t_k]$  is

$$\begin{aligned} \tilde{p}_k &= p_{k-1} - \eta \left( \frac{x_k}{\tau} \right) \\ &= p_{k-1} - c \operatorname{sgn}(x_k) + \frac{\eta}{\tau} \cdot x_k \end{aligned} \quad (9.7)$$

where  $c$  can be taken to be fixed cost of selling which is the sum of mid-spread and fees. The temporary impact component is not present in the next period. Note (9.5) can be written as,

$$p_k = p_0 + \sigma\tau^{\frac{1}{2}} \sum_{j=1}^k \varepsilon_j - \theta(X - X_k) \quad (9.8)$$

and thus the shortfall,

$$x = Xp_0 - \sum_{k=1}^N x_k \tilde{p}_k = \sum_{k=1}^N \left( \frac{x_k}{\tau} \right) \eta - \sum_{k=1}^N [\sigma\tau^{1/2}\varepsilon_k - \theta x_k] X_k \quad (9.9)$$

is a function of the market volatility and the trade schedule, ‘ $x_k$ ’. It captures the difference between the initial book value and the weighted average of realized prices. From (9.9) it follows that

$$\begin{aligned} E(x) &= \theta \sum_{k=1}^N x_k X_k + \eta \sum_{k=1}^N \left( \frac{x_k}{\tau} \right) \\ \text{Var}(x) &= \sigma^2 \sum_{k=1}^N \tau X_k^2. \end{aligned} \quad (9.10)$$

If the trajectory is to sell at a constant rate,  $x_k = \frac{X}{n}$  and  $X_k = (n-k)\frac{X}{n}$ ,  $k = 1, \dots, N$ ,

$$\begin{aligned} E(x) &= \frac{1}{2}\theta X^2 + \varepsilon X + \left(\eta - \frac{1}{2}\theta\tau\right) \frac{X^2}{T} \\ \text{Var}(x) &= \frac{1}{3}\sigma^2 X^2 T \left(1 - \frac{1}{N}\right) \left(1 - \frac{1}{2N}\right) \end{aligned} \quad (9.11)$$

The trajectory minimizes expected cost but variance increases with  $T$ . On the other hand if we liquidate entire  $X$  in the first time period,

$$\begin{aligned} E(x) &= cX + \eta \frac{X^2}{\tau} \\ \text{Var}(x) &= 0 \end{aligned} \quad (9.12)$$

If  $n$  is large, hence  $\tau$  small,  $E(x)$  can be very large significant resulting in impact price.

With these illustrations of extremes, the optimization problem can be restated as,

$$\min_x (E(x) + \lambda V(x)) \quad (9.13)$$

and if  $\lambda > 0$ ,  $E(x) + \lambda V(x)$  is strictly convex and thus has a unique solution. The solution to (9.13) can be written as a combination of the exponentials. The results of trading trajectory and the trade list are given below:

$$\begin{aligned} X_j &= \frac{\sinh(\kappa(T - t_j))}{\sinh(\kappa(T))} \cdot X, \quad j = 0, \dots, N \\ x_j &= \frac{2 \sinh(\frac{1}{2}\kappa\tau)}{\sinh(\kappa(T))} \cdot \cosh(\kappa(T - t_{j-\frac{1}{2}})) X, \quad j = 1, \dots, n \end{aligned} \quad (9.14)$$

Here  $t_{j-\frac{1}{2}} = (j - \frac{1}{2})\tau$  and  $\kappa = \sqrt{\lambda\sigma^2/\eta}$ , that depends on the volatility and the risk aversion coefficient, ‘ $\lambda$ ’ is termed as urgency parameter. The inverse of ‘ $\kappa$ ’ is termed as the trade’s, “half-life”; the larger its value the more rapid the depletion rate of the trade list. The solution given in (9.14) is shown to be adaptive at any point in time, ‘ $j$ ’. Thus the initial optimal solution over the entire interval is also optimal over each subinterval. A fixed trajectory as in (9.14) was previously constructed by Grinold and Kahn (2000) [138].

The value of ‘ $\lambda$ ’, the price of risk depends on the traders utility function. The suggested optimal strategy is just volume weighted average pricing (VWAP), that is trading at a constant rate, when  $\lambda \rightarrow 0$ , that is if the trader is risk-neutral. The trade schedule for various value of ‘ $\lambda$ ’ is given in Figure 9.1 and the projected execution schedule is given in Figure 9.2. Various extensions of the basic model (9.6) are considered by Almgren (2008) [4]. The effect of a short term drift in prices and the serial correlation in the errors which is observed in high frequency etc can be incorporated easily in the model (9.8). Lorenz and Almgren (2011) [201] provide a Bayesian approach to the optimization problem. While the methodology yields elegant solutions,

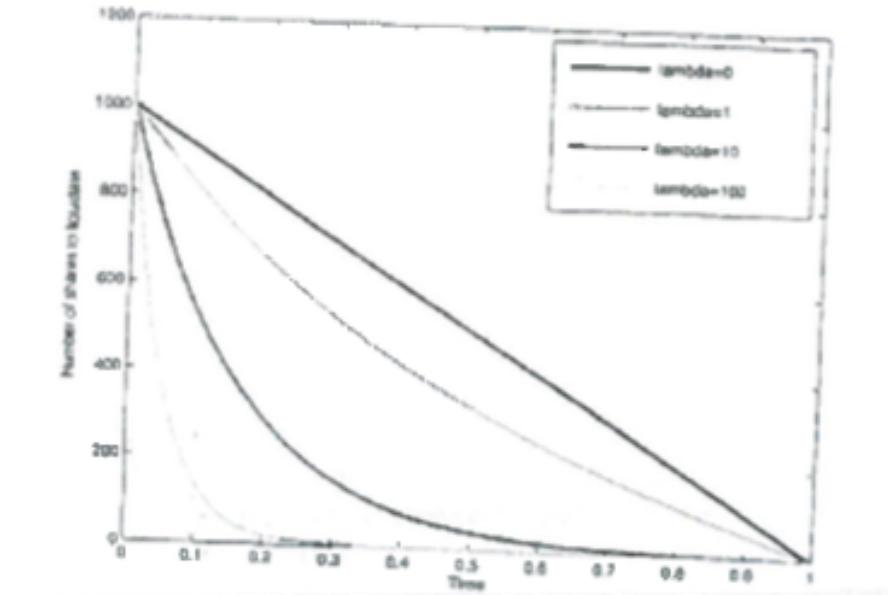


Figure 9.1: Trade Schedule: The cross validation errors of lasso models with different levels of interactions.

empirical use of these results require a close monitoring of order book dynamics as the changes in demand and supply sides are quite rapid.

The above mean-variance formulation of execution cost is combined with the traditional mean-variance portfolio optimization in Engle and Ferstenberg (2007) [97] and Engle, Ferstenberg and Russell (2012) [100]. We will discuss this in the context of portfolio rebalancing and the need for minimizing the transaction costs associated with all trades.

### 9.1.3 Obizhaeva and Wang Model

The models discussed in the previous sections assume price impact function that is not sensitive to liquidity dynamics. Obizhaeva and Wang (2013) [215] develop a model that accounts for the dynamic properties of supply and demand as represented in the limit order book. The model captures ‘resilience’; that is, how the current trade affects the status of the future limit order book. The resulting strategy consists of a large trade initially aimed at moving the order book away from its steady state followed by a number of small trades that will utilize the inflow of new liquidity providers. The speed at which the book replenishes itself is an important part of execution cost. The level of resilience depends on the level of hidden liquidity in the market. The price impact model considered here takes into account the dynamics of the limit order book.

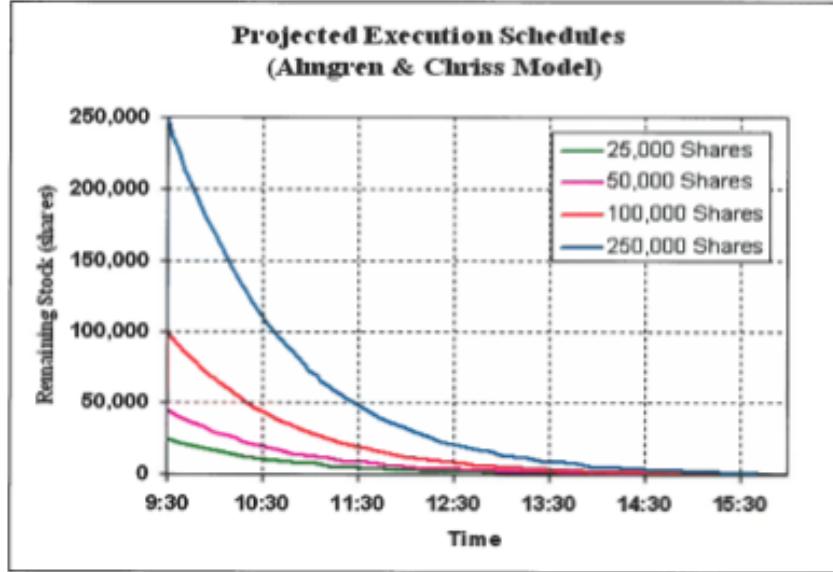


Figure 9.2: Projected Execution Schedules

It is assumed that the parent order  $X$  will be traded in the interval  $(0, T)$ , ' $n$ ' times at  $t_1, t_2, \dots, t_n$ . To model the execution of a large order in the limit order book, it is assumed that the price of execution ( $p_t$ ) depends on the true value of the equity ( $p_t^*$ ) and the state umicsles ( $z_t$ ) such as past trades that may affect the book. Let  $q(p_t^*, P_t; z_t)$  be the density of the limit orders in the other side of the book. The mid-quote ( $V_t$ ) is generally taken to reflect the true price,  $p_t^*$ . If the trade is on the buy side, the initial large transaction,  $x_1$ , can push the ask price to a higher level,  $p_1^* + (s/2) + (x_1/q)$ , where 's' is the spread and the average execution price is then equal to  $p_1^* + (s/2) + x_1/(2q)$ . After the initial execution the price converges to a new steady state,  $p_t^* + s/2 + \lambda x_1$ . Assuming that the limit-order book converges to its steady-state exponentially and at time ' $t$ ',

$$q_t(p) = q \cdot I(p > A_t) \quad (9.15)$$

and

$$A_t = V_t + \frac{s}{2} + x_1 \cdot \kappa e^{-\rho t}$$

where  $A_t$  is the asking price,  $\kappa = 1/q - \lambda$  and ' $\rho$ ' measures the resilience of the limit order book. If more the current ask price ( $A_t$ ) deviates from steady state level,  $V_t + s/2$ , the new ask limit orders will come into the book as the rate of  $\rho q(A_t - V_t - s/2)$ .

Given the above description of the limit order book dynamics, the optimal execu-

tion problem can be restated as follows:

$$\min_{x_k} E_1 \left( \sum_{k=1}^n [A_k + x_k(2q)]x_k \right) \quad (9.16)$$

such that

$$A_k = p_k^* + \lambda(X_1 - X_n) + \frac{s}{2} + \sum_{i=1}^{k-1} x_i \cdot \kappa e^{-\rho\tau(n-i)}$$

where  $p_k^*$ , the true price is taken to follow a random walk.

The solution to this dynamic programming problem is given in Obizhaeva and Wang (2013, p.14, Proposition 1) [215] and is somewhat involved. Instead we state, the strategy when  $n \rightarrow \infty$  which is of practical interest as many parent orders are divided into many, many child orders. The market impact study that is presented later in this chapter will attest to that. The solution to (9.16) as  $n \rightarrow \infty$  are:

$$\begin{aligned} \text{First and final orders : } x_1 &= \frac{X}{\rho_{T+2}} = x_n \\ \text{Spread of in-between trading : } x_t &= \frac{\rho x}{\rho_{T+2}} \end{aligned} \quad (9.17)$$

The expected cost is determined as

$$\text{Expected cost} = \left( p_0^* + \frac{s}{2} \right) X_t + \lambda X_0 X_t + \alpha_t X_t^2 + \beta_t X_t D_t + \gamma_t D_t^2 \quad (9.18)$$

where  $\alpha_t = \frac{\kappa}{\rho(T-t)+2} - \frac{\lambda}{2}$ ,  $\beta_t = \frac{2}{\rho(T-t)+2}$  and  $\gamma_t = -\frac{\rho(T-t)}{2\kappa[\rho(T-t)+2]}$ . The initial and final orders are discrete in nature and the orders in-between are continuous. They make use of incoming orders at desirable prices.

Some comments are worth noting. Note the solution given in (9.17) does not depend on the market depth, ‘ $q$ ’ and the price impact, ‘ $\lambda$ ’. It is shown that the price impact is not a factor if the trade times are determined optimally and if they are not set apriori as in the other strategies. The reason for ‘ $q$ ’ not being a factor is due to the fact that the depth is taken to be constant at all times. This implies that there is enough liquidity in the market and the book gets replenished, albeit at a constant rate. The two factors that play important roles are the resiliency factor, ‘ $\rho$ ’ and the trading horizon, ‘ $T$ ’. When  $\rho = 0$  the execution costs are strategy dependent and when  $\rho \rightarrow \infty$ , the order book rebuilds itself faster. When ‘ $T$ ’ increases, the size of the first and final order decreases; if there is more time to trade, the trades are spread out to manage the execution cost. The net cost of this strategy

$$\text{Net cost} = \frac{\lambda}{2} \cdot X^2 + \left( \frac{\kappa}{\rho_{T+2}} \right)^2 X^2 \quad (9.19)$$

and the above is shown to be smaller than the cost incurred with the strategy of constant rate trading.

Obizhaeva and Wang (2013) [215] consider the extension of the optimization criterion in (9.16) to include the risk aversion as well, as in Almgreva and Chriss (2000) [5]. Interested readers should refer to their paper, Section 8. For a practical implementation of these methods, it is necessary to consider the trading that happens in multiple exchanges and how the replenishment patterns can differ over different exchanges where ‘liquidity’ and the fee structure have become major considerations.

#### 9.1.4 Easley, De Prado and O’Hara Model

In the models described so far in this section, assume that the number of child orders ‘ $n$ ’ and the execution horizon, ‘ $T$ ’, are decided exogenously. Also the impact of a trade is modeled through the modified random walk model for the price with additional terms reflecting the permanent or temporary impact (Equation (9.3), (9.6) and (9.7)). The process of how price impact arises due to friction in the liquidity access is an important part of market microstructure theory and this needs to be taken into account in determining the optimal execution. For example, a buyer in a seller’s market can incur a lower cost of trading than a seller in a similar market. Obizhaeva and Wang model is based on the arrival dynamics to the order book. The approach taken by Easley, De Prado and O’Hara (2015) [90] is based on an asymmetric information model of the market maker’s behavior. The key measure is the probability of information-based (PIN) trading that is estimated by the order book imbalance (see Easley, De Prado and O’Hara (2012) [89]). Selling a large order in a market already imbalanced toward sell, will reinforce adverse selection from the other side and thus widen the bid-ask spread resulting in higher market impact. The optimal execution horizon (OEH) model in Easley et al (2015) provide a framework for determining the trading horizon, ‘ $T$ ’, and does complement other earlier studies on execution strategies that minimize the price impact.

The PIN that was developed in a series of papers (see Easley, Kiefer, O’Hara and Paperman (1996) [93] and the references therein) views trading as a sequential game between liquidity providers and liquidity takers, repeated over the trading duration. If the information about the asset occurs with probability  $\alpha$  and the chance that the information is good, is denoted by,  $(1 - \delta)$  and further assume that the informed traders, who knew the terminal value of the asset under good news ( $\bar{S}$ ) and under bad news ( $S$ ) arrive at the rate,  $\mu$ , and the noise traders arrive at the rate of ‘ $\epsilon$ ’. The PIN is approximated as (assuming  $\delta = 1/2$ )

$$\text{PIN} = \frac{\alpha\mu}{\alpha\mu + 2\epsilon} \sim E[\text{OI}] \quad (9.20)$$

where order importance (OI),  $\text{OI} = \frac{V^B - V^S}{V}$ , with  $V$  denoting the volume. An aggressive buy order of size ‘ $m$ ’ can affect the order imbalance as

$$\text{OI} = \left(2 \cdot \frac{V^B}{V} - 1\right) \left(1 - \frac{m}{V}\right) + \frac{m}{V} \quad (9.21)$$

and note that  $\left(2 \cdot \frac{V^B}{V} - 1\right)$  is the order imbalance, when  $m = 0$ , that is without our

large trade. When ‘ $m$ ’ is small, the OI likely to be perturbed too much and  $m \rightarrow V$ , that is the buy order will take all the available liquidity, then  $\text{OI} \rightarrow 1$ . This also indirectly quantifies the amount of leakage or signaling that occurs with each trade.

An important factor that is associated with PIN is the range of liquidity that may exist in the market in the presence of both informed and noise traders:

$$\Sigma = \text{PIN} \cdot [\bar{S} - \underline{S}] \quad (9.22)$$

slicing a large order into small orders as suggested by earlier execution strategies does have timing risk and the asset price is assumed to follow a random-walk:

$$\Delta S = \sigma \sqrt{\frac{V}{V_\sigma}} \xi \quad (9.23)$$

where  $\xi \sim N(0, 1)$ ,  $V_\sigma$  is the volume in the mid-price range. With ‘ $\lambda$ ’ as the probability of accepting a loss greater than  $z_\lambda \cdot \sigma \cdot \sqrt{V/V_\sigma}$ , the loss from trade size, ‘ $m$ ’, can be shown to be bounded by,

$$P \left[ \frac{m \cdot \Delta S}{\hat{\sigma} \sqrt{\frac{V}{V_\sigma}}} > z_\lambda \right] = 1 - \lambda \quad (9.24)$$

and thus ‘ $\lambda$ ’ can be interpreted as a ‘risk aversion’ parameter. The OEH’s goal is to determine the optimal trading volume,  $V$ , that can hide the purported trade,  $m$ , with minimum timing risk. The probabilistic loss function  $\Pi$  that incorporates both liquidity and timing risk component is defined as:

$$\Pi = \left| \varphi(m) \cdot \text{OI} + (1 - \varphi(m))(2V^B - 1)(\bar{S} - \underline{S}) \right| - z_\lambda \sqrt{\frac{V}{V_\sigma}} \sigma \quad (9.25)$$

Here  $\varphi(m)$  is a monotonic function that maps into the range  $(0, 1)$ . If ‘ $V$ ’ is greater, its impact is smaller in OI and larger in the trading range. The optimization results are given in Easley et al (2015) [90].

The key quantities in determining the OEH are the order imbalance and the trade size/side. Figure 9.3, reproduced from Easley et al demonstrates how selling in a buyer’s market allows for shorter horizons and in seller’s market leads to longer horizons.

## 9.2 Multiple Exchanges: Smart Order Routing

In the USA there are over ten lit venues and close to forty dark venues and thus equity markets are highly fragmented. Each venue functions as an electronic limit order book, where orders are prioritized first based on their prices and then at a given

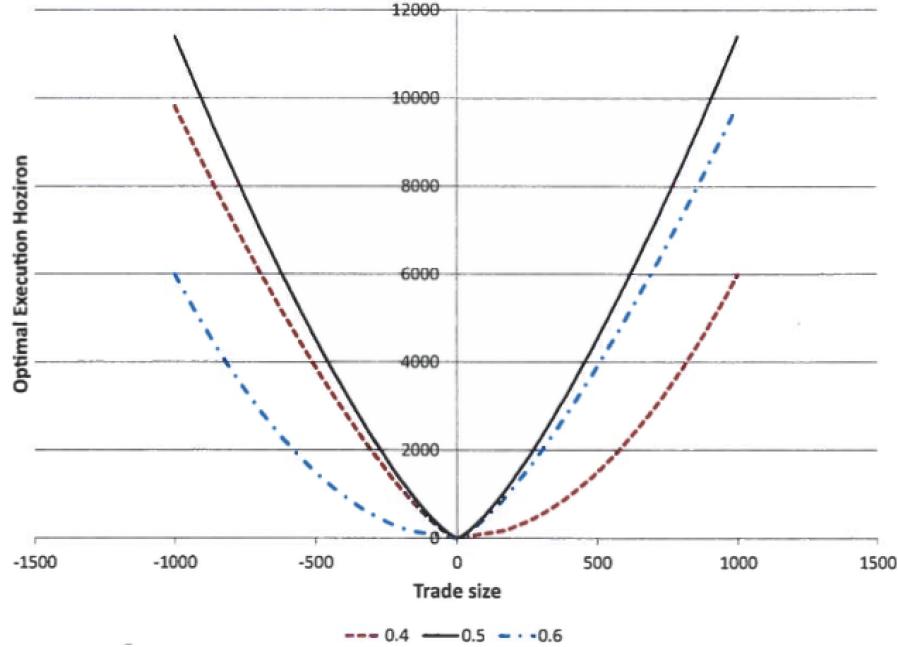


Figure 9.3: Optimal execution horizons for various order imbalances and trade sizes/-sides. Combining alternative trade sizes and sides with our three scenarios ( $v^B = 0.4$ ,  $v^B = \frac{1}{2}$ ,  $v^B = 0.6$ ) results in the optimal execution horizons displayed in the figure above,  $\hat{\sigma} = 1,000$ ,  $V_\sigma = 10,000$ ,  $m = 1,000$ ,  $[\bar{S} - \underline{S}] = 10,000$ ,  $\lambda = 0.05$  and  $\varphi[|m|]$  linear.

price level, according to their time of arrival. Exchanges publish information for each security in real-time and the information can change rapidly due to cancellations. The exchanges may differ with respect to best bid and offer price levels, the market depth at various prices etc. At any point in time, the highest bid and the lowest offer among all exchanges comprise the National Best Bid and Offer (NBBO). The exchanges also differ in their fee structures. Under the maker-taker pricing, exchanges offer rebates to liquidity providers and charge fees to takers of liquidity. These fees can range from  $-\$0.001$  to  $\$0.0030$  and since the typical bid-ask spread is  $\$0.01$ , the fees and rebates are a fairly significant fraction of the trading costs. Thus market participants must decide where child orders should be sent. It may be expensive to route to all the exchanges and by not routing to the right venue, it is possible to miss liquidity and hence incur greater market impact. Most traders use the smart order routing algorithms with goal to buy or sell the maximum number of shares in the shortest possible time with the least market impact possible. But a key issue is that many of the lit venues have hidden orders (12-45%). How to estimate the size of the hidden orders and how to

make routing decisions in the presence of hidden orders has been a focus of recent research studies.

While exchanges compete along many fronts, for example, payment for order flow, transparency and execution speed, the key variables driving the efficiency are liquidity and price improvement. Thus the design of market structure is considered by market participants and regulators as the key determinant of flow of liquidity. The coexistence of multiple exchanges as mentioned in Parlour and Seppi (2003) [220] raises specific questions stated below:

- a) Do liquidity and trading concentrate in a few exchanges?
- b) Do some market designs provide greater liquidity than others?
- c) Is the fragmentation of order flow desirable from a policy point of view?
- d) What is the constructive role for regulators to enhance liquidity?

Because exchanges operate under rules governing certain market designs, the performance and efficiency issues are related to the types of markets, such as pure limit order and a hybrid market (limit order book plus a specialist). In the hybrid market a specialist can provide supplementary liquidity after a market order has arrived. Parlour and Seppi (2003) [220] conclude only under some conditions such as enforcement of time priority, the efficiency is possible. Merely increasing the number of trading venues may result in degradation of market quality, if the enforcement of time priority is not followed; as it might discourage traders from posting the limit orders.

Foucault and Menkeld (2008) [115] consider the effect of fragmentation on two limit order markets. They examine the competition between Euronext and London Stock Exchange (LSE) in the Dutch market. The consolidated limit order book is found to be deeper after entry of the LSE. General conclusions from the above study and others such as Hendershott, Jones and Menkeld (2011) [153] and O'Hara and Ye (2011) [144] are that fragmentation of order flows improves the liquidity supply and protecting orders against trade-throughs is important. Thus multiple exchanges are here to stay and smart order routing algorithms that can work well with the increased number of exchanges would be favorably sought by the market participants.

The smart order router (SOR), generally work as follows. A user can customize the use of the strategy by establishing a set of rules in splitting the parent order into child orders and the smart order routers are designed to ensure that the orders are routed to the venue with the best price and the orders are filled according to the trading strategies that may be pre-determined or may be adaptive. To be successful, SORs must be able to handle a variety of trading strategies and multiple venues. In addition they must deal with vast amount of incoming and historical market data. The publicly available information on the SORs used by major traders states that the order placement objective in general is to minimize client all-in shortfall with or without venue fees. The all-in shortfall consists of the shortfall on the filled shares and the cost of the clean-up trade for the unfilled shares. The latter is taken to be an important component

of order placement optimization as it enables to quantify the effect of venue differences in fill rates, (see Street Smart, Issue 42, Jan 14, 2011). Other programs used in industry chose to optimize the expected time to execute the client's overall order; expected queue speeds at various venues are predicted from recent trading observations. Small orders tend to be placed in a single venue to minimize the queueing time whereas large orders will be placed in multiple venues to access the maximum liquidity. In this model forecasting trading rates at different venues is crucial for successful implementation of the program. The rates are modeled as a function of lagged rates, thus capturing the momentum, and imbalances between supply and demand sides of the order book. Thus order placement in a fragmented market is not a trivial task.

A recent empirical study by Battalio, Corwin and Jennings (2016) [] confirms that brokers use both limit and marketable orders to execute trades. Past studies provide evidence that market orders are sent to venues with lower trading costs and the trading fees and rebates generally affect consolidated market depth. Moallemi, Maglaras and Zheng (2012) [] show that limit orders are submitted to exchanges with high rebates and lower waiting time for execution while market orders are sent to venues that have lower fees and larger posted quote sizes. Cont and Kukanov (2017) [66] develop a model that is somewhat more realistic to the current practice by decoupling the 'order placement' decision from the scheduling decision and more importantly consider the option of placing limit orders on several exchanges, simultaneously. The model also accounts for the execution risk, the risk of not filling an order. Filling the unfilled portion may be costly and the allocation may shift toward market orders or toward overbooking, that is placing more orders than needed to refill.

To formalize the order placement problem, we assume a size of order  $S$  is to be filled in the duration  $(0, T)$ . The decision is to split this order into a market order  $M$  and ' $K$ ' limit orders  $L_1, \dots, L_K$  with the same size to be placed in ' $K$ ' exchanges with queue sizes  $Q_1, \dots, Q_K$ ; thus the order allocation is summarized by the elements of the vector  $X = (M, L_1, \dots, L_K)$  that need to be optimally determined. If order cancellations in the duration  $(0, T)$  in exchange ' $K$ ' is represented by  $\xi_K$ , then the number of shares transacted can be written as

$$A(X, \xi) = M + \sum_{k=1}^K [(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] \quad (9.26)$$

where the terms in the parenthesis refers to the initial position and the final position of the queue outflows. The execution cost must account for fee ( $f$ ) and rebate structures (discussed in Chapter 6) in each exchange and the cost of adverse selection ( $r_k$ ) and can be stated as follows:

$$C(X, \xi) = (h + f)M - \sum_{k=1}^K (h + r_k) [(\xi_k - Q_k)_+ - (\xi_k - Q_k - L_k)_+] \quad (9.27)$$

Here  $h$  is one-half of bid-ask spread.

The cost function in (9.27) can be modified to account for the cost of unfilled limit orders which may be filled through market orders of higher cost or it is possible that the prices have decreased resulting in additional adverse selection cost with penalty

$(\lambda_u)$  for falling behind and penalty  $(\lambda_0)$  for exceeding the target; Thus, the execution risk can be written as,

$$\text{ER} = \lambda_u(S - A(X, \xi))_+ + \lambda_0(A(X, \xi) - S) \quad (9.28)$$

To be more realistic, the market impact function can be considered as follows:

$$\text{MI} = \theta \left[ M + \sum_{k=1}^K L_k + (S - A(X, \xi))_+ \right] \quad (9.29)$$

The total cost function that includes both implicit and explicit costs can be states as:

$$V(X, \xi) = C(X, \xi) + \text{ER} + \text{MI} \quad (9.30)$$

The random variable in (9.30) is  $\xi$ , the cancellations that occur in various exchanges and the minimization function is  $E[V(X, \xi)]$  under some distribution  $F$  of  $\xi$ . Under some reasonable assumptions such as that the trader will not execute more than the target  $S$  and market orders in the beginning of the duration  $(0, T)$  are less expensive than at the end when unfilled orders are converted to market orders, an optimal solution is shown to exist (see Cont and Kukanov (2017) [66] for details).

The analytical solution minimizing  $V(X, \xi)$  is not easily tractable due to dimensionality issues. A numerical solution via gradient method is proposed. Random samples of  $\xi$  are obtained and averaged to approximate  $E(\xi)$ . Let  $q(X, \xi) = \Delta V(X, \xi)$  be the gradient of  $V$ . The following iterative algorithm is shown and converges:

- Start with  $X_0$  and for  $n = 1, 2, \dots, N$  do
- $X_n = X_{n-1} - \gamma_N g(x_{n-1}, \xi^n)$
- End;  $X_N^* = \frac{1}{N} \sum_{n=1}^N X_n$

The step size,

$$\gamma_N = \sqrt{k} S \left( N(h + f + \theta + \lambda_u + \lambda_0)^2 + N \sum_{k=1}^K (h + r_k + \theta + \lambda_u + \lambda_0)^2 \right)^{-1/2} \quad (9.31)$$

can be seen as a function of all the costs associated with the execution of the order.

To summarize, recall that the algorithm needs the following input:

- Trading Costs:
  - (a) One half of bid-ask spread ( $h$ )
  - (b) Market order fee ( $f$ )
  - (c) Effective limit order risks ( $r_k$ )
  - (d) Market impact coefficient ( $\theta$ )
  - (e) Penalties for overfilling or underfilling ( $\lambda_0, \lambda_u$ )
- Market Variables: Number of exchanges ( $K$ ) and limit order queues ( $Q_k$ ).
- Execution Variables: Time horizon ( $T$ ) and target quantity ( $S$ ).

While many of these quantities can be estimated using past transactions, the limit order queues ( $Q_k$ ) and the cancellation ( $\xi_k$ ) are to be obtained at the time of execution.

### 9.3 Market Impact Models: Permanent and Temporary Impacts

One of the main focuses of Algorithmic Trading is minimizing costs of trade execution. When a large order is executed in the market, it causes market impact (MI), which is commonly defined as the deviation of the post-trade market price from the market price that could have prevailed had the trade not occurred (Fabozzi, Focardi and Kolm (2006) [106]). To alleviate MI and the implied transaction cost (TC), a trader may use an algorithmic trading system to slice a large “parent” order into a sequence of “child” orders. While a slicing strategy is typically motivated by the liquidity constraint associated with executing a large order in one print, it is usually balanced against the risk and opportunity costs associated with the execution of a sequence of child orders over a longer duration as discussed earlier.

In asset management, with the competitive focus of the industry on cost management and diminishing returns of crowded strategies, understanding and managing MI has become a crucial component of a systematic investment process. For an estimate of its size among the transaction costs that include explicit costs such as brokerage commissions and fees, see Table 9.1 from Berkovec and Heidle (2010) [33]. Managing MI has important implications on three aspects of algorithmic trading: pre-trade TC estimation, post-trade TC analysis and the design of optimal trading strategies (Almgren (2008) [4]).

Table 9.1: Empirical Transaction Cost Estimates<sup>1</sup>

Type	Fixed	Variable
Explicit	Commissions (7.4bp)	Bid-ask spreads (1.9bp)
	Fees	Taxes
Implicit		Delay Cost (9.5bp)
		Price Movement Risk & Market Impact Costs* (35.6bp)
		Timing Risk & Opportunity Costs (11.7bp)

\*Major Component

1: Readers should note changes in the market that have taken place since these numbers were reported. The commissions are quite expensive for current trading. Most developed markets have commissions less than 1 bps, while E.M. markets are in the range of 2–5 bps for e-trading. The bid-ask spread makes sense if it is half a spread of liquid U.S. stocks. For other markets, it can be much higher.

The study of price impact dates back to Bagehot (1971) [] who postulated that the market consists of heterogeneously informed traders. The specialist cannot distinguish between informed and uninformed (liquidity) traders and fixes a spread that can balance the trading cost. The asymmetric information models based on this notion have been extensively studied in finance literature. The traders do convey information and Hasbrouck (1991) [148] provides empirical evidence that the market maker can infer information from the characteristics of the sequence of trades. The seminal work

by Kyle (1985) [182] and Huberman and Stanzl (2004) [163] indicate that an order of large size (parent order) must be sliced into a number of small order (child orders) so that the volume per se will not signal informed trading. The hypothesis that time between the trades could convey the information flow (Easley and O’Hara (1992) [91]) and its impact on price is studied in Dufour and Engle (2000) [88]. Long durations are usually associated with no new news and the variations in trading intensity are taken to be positively correlated with the behavior of the informed traders.

**Origins of Market Impact:** Most conventional definitions of the MI of a trade equate it with the deviation of the post-trade market price from the market price that would have prevailed had the trade not occurred. Figure 9.4 shows an idealized MI picture of  $n$ -share sell order from Fabozzi, Focardi and Kolm (2006) [106] but Figure 9.5 is augmented to include other information: (1) the state of the order book establishes a pre-trade equilibrium, (2) this equilibrium is disturbed by a market order to sell  $n$  shares, (3) as the sell order depletes the bid order book, it obtains an increasingly lower trade price and results in the trade print, and, (4) over time the price gradually moves up back to recover some of the price drop. Accordingly, the difference between (4) and (3) is called temporary impact, whereas the difference between (4) and (1) is called permanent impact. If we are looking to model the MI effect of sequential trades, as shown in Figure 9.4, the temporary and permanent impact will be superpositions of those of individual trades. The permanent impact is often assumed to be immediate and linear in the parent trade size, and the post-trade equilibrium to be the same for the order of  $n$  shares or a sequence of  $m$  orders of  $n/m$  shares each (Almgren et al. (2005) [6]; Fabozzi et al., (2006) [106]). By contrast, the temporary impact is a function of how the “parent” trade is split into smaller “child” trades. The

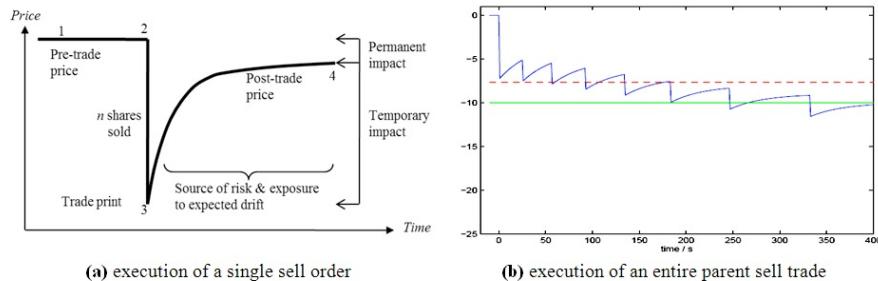


Figure 9.4: Idealized MI Model: single sell trade (left) and entire parent sell trade (right).

mechanisms responsible for the permanent and temporary impacts have been a subject of significant interest. One common view links price changes accompanying large trades to the information cost and liquidity cost (Chan and Lakonishok (1995) [57]; Holthausen et al. (1990) [159]) potentially attributed to any transaction. In terms of information cost, the market response to the fact that a market participant has decided to sell (buy) shares of a particular stock could be perceived as conveying new (pri-

vate) information about the fundamentals of a security (e.g., firm's management, debt conditions) and thus its future prices. Consequently, it results in a permanent price change with no price reversals. In terms of liquidity cost, the market response to the fact that liquidity was removed out of the order book can be perceived as the effect of distorting the liquidity demand/supply equilibrium. The consequence is a temporary price impact that dissipates soon after the trading occurs, at a speed that depends on the market ability to absorb liquidity demand.

Independent of the market microstructure responsible for MI, two central concerns to much theoretical and empirical work are the functional form (shape) of the MI and the key determinants of this form. As seen in Figure 9.5, MI is influenced by trade-related factors (e.g., relative trade size, time of trade), asset-specific factors (e.g., market capitalization, shares outstanding, bid-ask spread) and exchange- and market-related factors (e.g., liquidity, trading volume, institutional features). Models that incorporate these factors tend to be more complex but with better predictive power. Numerous stylized MI models offer some insight into the possible shapes of

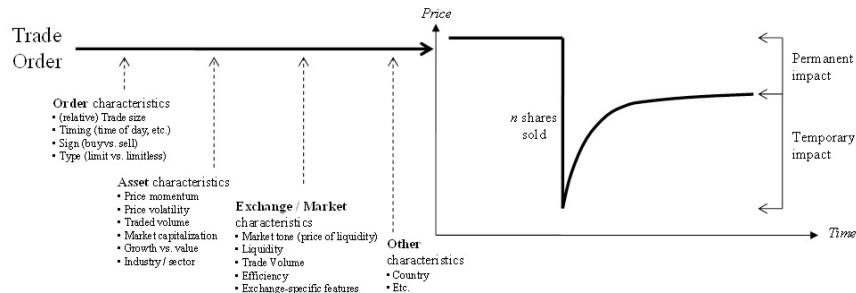


Figure 9.5: Determinants of Market Impact.

the permanent and temporary impact functions. Some of these models focus only on the permanent impact, while some models also attempt to capture the effect of temporary impact.

**Equilibrium Permanent Impact Models:** Models of one class, including Kyle's (1985) [182] and Huberman and Stanzl's (2004) [163], rest on the efficient market hypothesis positing that all available information is included in prices. These models use a time-independent framework to establish an equilibrium price after a trade is executed. This means that trades are assumed to have only a permanent price impact (Huberman and Stanzl, 2004, p. 1260) [163]

Under these settings, Huberman and Stanzl (2004) [163] show that, if no quasi-arbitrage possibilities exist (to ensure viable markets), the permanent impact must be linear in the trade quantity and symmetric between buys and sells. Linearity in trade size has been established by Kyle (1985) [182] as well. In fact, Huberman and Stanzl (2004) [163] add that "Nonlinear, time-dependent price-update functions may assume "chaotic" shapes without giving rise to price manipulation. Only when additional

assumptions are made on the shapes of these functions does the analysis become meaningful."

Almgren et al. (2005) [6] develop a different market impact model, in the context of finding an optimal execution strategy for a large trade [see the discussion of Almgren and Chriss model in Section 7.1.2] that is sliced into several smaller trades. They describe the random process satisfied by the stock price using an arithmetic Brownian process that depends directly on a permanent impact function that is linear in trade size and therefore has no effect on the optimal execution strategy. However, they recognize that if the total number of units traded is sufficiently large, the execution price may steadily change between trades, in part because the supply of liquidity is exhausted at each successive price level. They assume that this effect is short-lived because liquidity returns after each period and a new equilibrium price is established. They model this effect using a temporary price impact function that is non-linear in trade size. They then postulate that both the permanent and temporary impact functions follow power laws based on theirs' and others' empirical results. (Loeb (1983) [200]; Lillo, Farmer and Mantegna (2003) [193]).

**Hybrid Impact Models:** Models of another class assume that trading agents have zero intelligence (instead of being fully rational) and take random decisions to buy or to sell, but that their action is interpreted by all the other agents as containing some potential information. The mere fact of buying, or selling typically leads to a change of the ask  $a$ , or bid  $b$ , price and hence to a change of the midpoint  $p = \frac{a+b}{2}$ . The new midprice is also expected to follow a random walk (at least for sufficient large times), if it is immediately adopted by all other market participants as the new reference price around which new orders are launched.

Madhavan et al. (1997) [202] develop a price formation model postulating that the (bid-ask) mid price ' $p$ ' changes because of unpredictable public information shocks (news) and microstructure effects that include statistical effects of order flow fluctuations (e.g., autocorrelation) and trading frictions (e.g., asymmetric information, dealer costs). This postulate automatically removes any predictability in the price returns and ensures market efficiency. If all trades have the same volume and the surprise component of the order flow at the  $k$ th trade is given by  $\epsilon_k - \rho \epsilon_{k-1}$ , where the signs of trades  $\epsilon_n$  are generated by a Markov process with correlation  $\rho$ ,<sup>1</sup> one writes the following evolution equation for the midprice as:

$$\Delta p_{k+1} = p_{k+1} - p_k = \theta[\epsilon_k - \rho \epsilon_{k-1}] + \eta_k \quad (9.32)$$

where  $\eta$  is the shock component and the constant  $\theta$  measures the size of trade impact. Empirical estimation of their model are not aimed at investigating the shape of impact functions.<sup>2</sup>

---

<sup>1</sup>This means that the expected value of  $\epsilon_k$  conditioned on the past only depends on  $\epsilon_{k-1}$ , given by:  $E(\epsilon_k | \epsilon_{k-1}) = \rho \epsilon_{k-1}$ .

<sup>2</sup>First, both information flows and trading frictions are important factors in explaining intraday price

Others extend the model of Madhavan et al. (1997) to gain insight into the shape of impact functions. Bouchaud et al. (2009) [35] use (9.32) to compute several important quantities. Principally, they write the lagged return impact function (for time points  $n$  and  $n + l$ ), denoted  $R_l$ , as:

$$R_l = p_{n+l} - p_n = \theta \sum_{j=n}^{n+l-1} [\epsilon_j - \rho \epsilon_{j-1}] + \sum_{j=n}^{n+l-1} \eta_j \quad (9.33)$$

and then show that the lagged impact function is constant and equal to:

$$R_l = \theta(1 - \rho^2), \forall l \quad (9.34)$$

Now define the “bare” impact of a single trade taken at time  $l$ , denoted  $G_0(l)$ , which measures the influence of a trade at time  $n - l$  on the bid-ask midprice at time  $n$ . Written in terms of  $G_0(l)$ , the midpoint process is expressed as:

$$p_n = \sum_{j=-\infty}^{n-1} G_0(n - j - 1) \epsilon_j + \sum_{j=-\infty}^{n-1} \eta_j \quad (9.35)$$

Then it can be shown that  $G_0(0) = \theta$  and  $G_0(l) = \theta(l - \rho)$  for  $l > 0$ . Bouchaud et al. (2009) [35] also observe that “The part  $\theta\rho$  of the impact instantaneously decays to zero after the first trade, whereas the rest of the impact is permanent. The instantaneous drop of part of the impact compensates the sign correlation of the trades.”

Bouchaud et al. (2004) [36] have developed an earlier price evolution model, similar to (9.33), where the price at time  $n$  is written as a sum over all past trades of the impact of one given trade propagated up to time  $n$ :

$$p_n = \sum_{n' < n} G_0(n - n') \epsilon_{n'} \ln V_{n'} + \sum_{n' < n} \eta_{n'} \quad (9.36)$$

where  $V_n$  denotes the volume of a trade at time  $n$  and  $G_0(0)$  is assumed to be a fixed non-random function that only depends on time differences. The  $\eta_n$ 's are also random variables, assumed to be independent from  $\epsilon_n$ , and are used to model all sources of price changes not described by the direct impact of the trades. The authors then consider constraints imposed on the shape of response function,  $G_0$ , by three empirical results they discuss: (a) the midprice price process is close to being purely diffusive, even at the trade-by-trade level; (b) the temporal structure of the impact function first increases and reaches a maximum after some number (100 to 1000) of trades, before decreasing back with a rather limited overall variation; and the sign of the trades shows surprisingly long-range, power-law correlations. These empirical results are reconciled with (9.36) by assuming that the response function  $G_0()$  must instead also

---

volatility in individual stocks. Second, information asymmetry decreases steadily throughout the day, however, dealer costs increase over the day (possibly reflecting the costs of carrying inventory overnight) so that bid-ask spreads exhibit the U-shaped pattern commonly noted in previous research studies.

decay as a power-law in time, with an exponent precisely tuned to ensure simultaneously that prices are nearly diffusive and that the response function is nearly constant. Gatheral (2010) [129], assuming that the trading costs should be non-negative, demonstrates that the exponential decay of market impact is compatible only when the market impact is taken to be linear. Thus the debate on the functional form of MI is ongoing.

It has been noted that the models developed to study market impact generally do not consider situations that may arise from strategies where price impact may depend upon the state of the order book. The second generation models consider the behavior of the limit order book. Hautsch and Huang (2012) [149] consider a time-aggregated model. The activities in the book are captured in Table 9.2 and the market depth is kept to the three best quotes on both sides of the market. Let  $Y_t$  be a  $K$ -dimensional

Table 9.2: Model (9.37) Variable Definitions

Variable	Description
$p_t^a$	Logarithm of the best ask after the $t$ -th event.
$p_t^b$	Logarithm of the best bid after the $t$ -th event.
$v_t^{a,l}$	Logarithm of the market depth at the $l$ -th best ask after the $t$ -th event.
$v_t^{b,l}$	Logarithm of market depth at the $l$ -th best bid after the $t$ -th event.
$\text{BUY}_t$	Dummy equal to one if the $t$ -th event is a buyer-initiated trade.
$\text{SELL}_t$	Dummy equal to one if the $t$ -th event is a seller-initiated trade.

listing the variables in Table 9.2; the co-integration model in Chapter 2 is fit for the data:

$$\Delta Y_t = \mu + \alpha \beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + u_t \quad (9.37)$$

The co-integrating  $(K \times r)$  matrix,  $\beta$  has first two columns restricted to have entries as zeros except for buy and sell indicators. To study the market impact, the model is written in the reduced-form to capture the impulse-response coefficients as:

$$Y_t = \mu + \sum_{i=1}^p A_i Y_{t-i} + U_t \quad (9.38)$$

where  $A_1 = I_k + \alpha \beta' + \Gamma_1$ ,  $A_i = \Gamma_i - \Gamma_{i-1}$  for  $1 < i < p$  and  $A_p = -\Gamma_{p-1}$ . The VAR( $p$ ) model in (9.38) can then be written in the form of VAR(1) as

$$Y_t^* = \mu^* + A^* Y_{t-1} + U_t^* \quad (9.39)$$

where  $\mu^* = (\mu', 0', \dots, 0')'$ ,  $Y_t^* = [Y_t', Y_{t-1}', \dots, Y_{t-p+1}']$ ,  $U_t^* = [U_t', 0', \dots, 0']'$  and

$A^* = \begin{bmatrix} A_1 \cdots A_{p-1} & | & A_p \\ \hline I_{K(p-1)} & | & 0 \end{bmatrix}$ . It can be shown that

$$Y_t = J M_t + \sum_{i=0}^{t-1} J A^i J' U_{t-i} \quad (9.40)$$

where  $J = [I_k, 0, \dots, 0]$  is a  $k \times kp$  selection matrix and  $M_t = A^{*T} Y_0^* + \sum_{i=0}^t A^{*i} U_{t-i}^*$ .

Recall that the market impact measures the impact of current trade on the future trades; it is measured by the impulse response function

$$f(h, \delta_h) = E[y_{t+h} | y_t + \delta_y, t_{t-1}, \dots] - E[y_{t+h} | y_t, y_{t-1}, \dots] = JA^{*h} J' \delta_y \quad (9.41)$$

where  $\delta_y$  is the change to  $Y_t$ , due to the current trade. The long-run effect,  $f(\delta_y) = C\delta_y$ , where  $C = \beta_\perp (\alpha'_\perp (I_K - \sum_{i=1}^{p-1} \Gamma_i) \beta_\perp)^{-1} \alpha'_\perp$ . Here  $\alpha_\perp$  and  $\beta_\perp$  are orthogonal to  $\alpha$  and  $\beta$  in (9.37).

Hautsch and Huang (2012) [149] estimate the market impact using the data from Euronext, Amsterdam for heavily traded stocks. The main findings are, besides the existence of co-integrating relationships between quotes and depths, limit orders have long-term effects on quotes, the price movement is influenced by the order size and the decrease of spreads after the arrival of a limit order is reverted back asymmetrically soon after. The between-stock variations in market impact are shown to be related to trading frequency of the stocks.

Cont, Kukanov and Stoikov (2014) [67] consider the dynamics of market liquidity that is a function of limit orders, market orders and cancellations to model the price impact function. The outstanding limit orders provide a measure of depth in the market on both buy and sell sides and thus also provide a measure of order flow imbalance (OFI).

By setting the number of shares (depth) or price levels beyond the best bid and ask is equal to  $D$ , define the price change in the interval  $(t_{k-1}, t_k)$  as,

$$\begin{aligned} \Delta P_k^b &= \delta \left[ \frac{L_k^b - C_k^b - M_k^s}{D} \right] \\ \Delta P_k^s &= -\delta \left[ \frac{L_k^s - C_k^s - M_k^b}{D} \right] \end{aligned} \quad (9.42)$$

where  $L_k^i$ ,  $C_k^i$  and  $M_k^i$  for  $i = b, s$  denote arrival of the limit orders, cancellations and market orders in the defined interval. Here  $\delta$  is the tick size. If  $P_k = \frac{P_k^b + P_k^s}{2\delta}$ , that is mid-price normalized by tick size, for any *snd*-interval,  $(t_{k-1,i}, t_{k,i})$ , it is postulated that

$$\Delta P_{k,i} = \beta_i \text{OFI}_{k,i} + \varepsilon_{k,i} \quad (9.43)$$

where  $\text{OFI}_k = (L_k^b - C_k^b - M_k^s) - (L_k^s - C_k^s - M_k^b)$  and  $\Delta P_k = \frac{\text{OFI}_k}{2D} + \varepsilon_k$ . Here it is important to note that ' $\beta_i$ ' is taken to be price impact coefficient and is taken to be inversely proportional to market depth:

$$\beta_i = \frac{C}{D_i^\lambda} + v_i \quad (9.44)$$

The models in (9.43) and (9.44) can be regarded to represent instantaneous price

impact. The model as it can be seen from the set-up assumes that all activities in the limit order book have an average linear impact,  $\beta_i$ , in the  $i$ th interval.

Cont et al. (2014) [67] using the TAQ data for a month in 2010 for fifty stocks estimate the model in (9.43) over ten-second intervals. The model is generally confirmed by the data with  $\hat{\beta}_i$  is significant in most of the samples and  $\hat{\alpha}_i$  is only significant in less than twenty percent of the samples. It must be noted that there is a significant number of trades posted gets cancelled within a few seconds of submission. Various theories that are postulated (see Hasbrook and Saar (2009) [147]) are yet to be fully empirically tested. Hence instead of using OFI, trade imbalance ( $TF_k = M_k^b - M_k^s$ ) is used in equation (9.43). Results are still reasonable, but not as strong as using OFI in the model. Finally the relationship between  $\hat{\beta}_i$  and  $D_i$  (price impact coefficient and depth) is studied; the results indicate for model (9.44),  $\hat{C} \sim 0.5$  and  $\hat{\lambda} \sim 1$ . Some interesting conclusions that could be useful for trading are drawn. The impact coefficient varies more with the amount of liquidity provision (OFI) than the trade imbalance (TI). It can be also used to monitor adverse selection. If a limit order is filled after a positive OFI, price is likely to go up and for a limit sell order, a positive change would imply that the order was executed at a loss. These conclusions need to be firmed up with further studies in this area.

**Determinants of Empirical relationship:** A key aspect of MI modeling is to provide information about expected pre-trade costs. The studies that have been reviewed on empirical analysis thus far can help us to understand how the price formation and the market impact function evolve during a trading period. But most clients who want to trade large quantities of equities want to know the trading costs *a priori*. Here we review the relevant studies and augment with our own research in this area. First we review the factors that are found to influence these costs in block trades. Although block trades differ in fundamental ways to algorithmic, they can be treated as all equivalent to a parent trade except that no splitting into child orders occurs.

- **Trade Size.** Many studies observe that the (relative) size of trades positively relates to the magnitude of MI (Keim and Madhavan (1996) [178]), consistent with the prediction of stylized model (Kyle (1985)[182]; Huberman and Stanzl (2004) [163]).
- **Market Capitalization.** Keim and Madhavan (1996) [177] find that the asymmetry of trading costs between buys and sells is more pronounced among smaller stocks. This is consistent with what is broadly known to be the effects of market capitalization on MI (Lillo et al. (2003) [193]; Chan and Lakonishok (1997) [58]). From a practical point of view, market cap itself may not be a true factor but it can be a proxy for some liquidity characteristics.
- **Price Volatility.** Price volatility has been found to be positively related to MI (Chiyachantana et al. (2004) [61]). One explanation is that elevated volatility is associated with greater dispersion in beliefs, with risk averse traders less inclined to participate in markets, thereby resulting in greater price concessions or MI (Domowitz et al. (2001) [86])

- **Trading Activity.** Dufour and Engle (2000, p. 2467) [88] find that the price impact of trades increases as the time duration between transactions decreases, and more broadly that the price impact is especially large due to increased trading activity, as measured by high trading rates. In their view, “times when markets are most active are times when there is an increased presence of informed traders; we interpret such markets as having reduced liquidity.” Likewise, Yang (2011, p. 91) [263] finds that “a trade shortly after the previous trade results in higher price impact than one after a long period.” He also offers “evidence that increased trading activity (as measured by short durations between trades) is associated with larger price impact, therefore implying a higher degree of information-based trading.”
- **Time of Day.** It is reported that the MI for buy trades decreases as time passes during the day, with the highest impact in the first trading hours (Frino et al. (2007) [122]).

We will now summarize briefly the models that are used in the industry. For many years, traders have used the simple sigma-root-liquidity model described for example by Grinold and Kahn (1994) [138]. The model takes the form,

$$\Delta P - \text{Spread Cost} = \alpha \cdot \sigma \sqrt{\frac{X}{V}} \quad (9.45)$$

where  $X$  is the size of the parent order,  $V$  is the average daily volume and ‘ $\sigma$ ’ is the volatility of the stock. Thus the relative size (RS),  $\frac{X}{V}$  is the key factor in the pre-trade cost estimation and the impact is proportional to volatility. The square root of relative size is expected because risk capital should be proportional to the square-root of the holding period. In Toth et al. (2011) [251], the authors present an argument that if latent supply and demand are linear in price over some plausible range of prices, which is a reasonable assumption, market impact should be square-root.

The square-root formula as stated in (9.45) refers only to the size of the trade relative to daily volume. It does not refer to for example, the rate of trading, how the trade is executed and the market capitalization of the stock. If the trading is quite aggressive, the square-root formula tends to break down. Moro et al (2009) [213] observe that the price path during the execution follows a power law:

$$p_t - p_0 = \alpha \cdot \left(\frac{t}{T}\right)^{2/3} \quad (9.46)$$

where  $T$  is the duration of the parent-order. Immediately after completion of a parent-order, the price begins to revert. Other models used in industry are mostly functions of the determinants identified above.

**Select Empirical Studies:** While there is some discussion about the theoretical models for market impact, published studies based on large scale data are somewhat rare especially when it comes to models useful for pre-trade cost estimation, Almgren et al (2005) [6] took US stock trade orders processed by Citigroup for Dec 2001 - June

2003 and estimated a model of the market impact as given in Table 9.3. For two select stocks, IBM and DB, the power law model for market impact are estimated. The temporary impact is defined as  $\tilde{p}_k - p_{k-1}$ , for  $k^{th}$  execution. The model depends on relative size, inverse turnover ratio, which is the number of the shares outstanding to average daily volume and the volatility.

While the square root formula in (9.45) is quite robust and holds even beyond equity markets, recent empirical studies confirm the need for more general power law functions; the quotient of  $RS = \frac{X}{V}$  varies depending on other factors that get included in the model (refer to the estimates in Table 9.3 for RS). Zarinelli, Treccani, Farmer and Lillo (2015) [265] show that logarithmic functional form fits the data better for the large order data that they consider.

Table 9.3: Example of impact costs (Almgren et al. (2005) [6]

		IBM			DRI		
Average daily volume	$V$	6.561	m		1.929m		
Shares outstanding	$\theta$	1728	m		168 m		
Inverse turnover	$\theta/V$	263			87		
Daily volatility (%)	$\sigma$	1.57			2.26		
Normalised trade size	$X/V$	0.1			0.1		
Normalised permanent	$I/\sigma$	0.126			0.096		
Perm. price impact (bp)	$I$	20			22		
Trade duration (days)	$T$	0.1	0.2	0.5	0.1	0.2	0.5
Normalised temporary	$K/\sigma$	0.142	0.094	0.054	0.142	0.094	0.054
Temp. Impact cost (bp)	$K$	22	15	8	32	21	12
Realised cost (bp)	$J$	32	25	18	43	32	23

Note: examples of permanent and temporary impact costs are shown, for a purchase of 10% of the day's average volume, in two different large-cap stocks. The permanent cost is independent of time of execution. The temporary cost depends on the time, but across different assets it is the same fraction of daily volatility (from Almgren et al. (2005) [6]).

Limited access to broker-proprietary algorithmic trading data has stifled detailed academic inquiry in this area. While the impact of algorithmic trading is extensively discussed in the finance literature, formal empirical results are rare. The few exceptions we are aware of are: Engle et al (2012) [100], who use proprietary algorithmic trading data from Morgan Stanley to study the risk-return tradeoff associated with increasing the number of child trades and duration of parent trade; Domowitz and Yegerman(2005) [87], who use ITG's Transaction Cost Analysis Peer Group Database to compare execution costs for a set of 40 buy-side clients; and Hendershott and Riordan (2011) [155], who use algorithmic trading data from firms listed on the Deutsche Boerse DAX to measure the contribution of algorithmic trading to price discovery. A few academic studies resort to using proxies based on publicly-available TAQ data. Two examples are: Hendershott et al. (2011) [154], who use NYSE electronic message traffic data as a proxy for algorithmic liquidity supply to study how algorithmic trading effects liquidity and bid-ask spreads; and Chaboud et al. (2014) [53], who use FX quote data (highest bid and lowest ask) from an electronic

limit order book platform called EBS to construct mid-quote series for studying the effect of algorithmic trading on FX rates' volatility. The use of mid-point prices, in the lack of actual arrival price and completion price for a parent trade, has been criticized for failing to capture the possible asymmetry of the bid and ask price impacts of trading, which has been observed to exist for block trades. In any event, none of these studies deals with the estimation of MI models per-se.

Velu, Gretchka, Benaroch, Nehren and Kuber (2015) [256] address these issues using a large algorithmic trading dataset provided by J.P. Morgan Chase. The data contain "parent" algorithmic executions, including the initial arrival price and fill price as well as the child trades' average fill price (computed as the share-weighted average executing price of the child orders). The algorithmic executions are generated using Volume Weighted Average Price (VWAP) and Percent of Volume (POV) strategies. Recall the volume weighted average price (VWAP) strategy schedules to trade some constant percentage of the expected market volume regularly throughout the day, where the proportion traded is adjusted based on historical market volume profiles and the predict volume. Hence, in VWAP, the trading schedule is predetermined based on a prediction of the daily market volume distribution, although the timing of executions may be randomized so as to make the trading pattern less predictable. By contrast, the percent of volume (POV) strategy trades a fixed percentage of the current market volume where the trading schedule is dynamically determined based on a historical analysis of volume profiles used to anticipate trading volumes. Unlike implementation shortfall strategies, VWAP and POV do not explicitly take into account the expected MI in generating their algorithmic executions. Moreover, although firms typically implement proprietary versions of these strategies with parameters reflecting their experience and clients' profiles, VWAP and POV nonetheless generate relatively standardized algorithmic executions. In other words, from the perspective of MI, these two strategies can be safely considered "plain vanilla" in the sense that firm-to-firm differences in their behavior and performance are too insignificant. The measures employed, therefore, do not depend on any unique traits of the strategies. The unique aspect of the data is that if the meta order is on the buy side or sell side is explicitly provided.

The dataset allows us to study the effect of parent characteristics on the market impact and the implied transaction cost (TC). Specifically, we estimate the parameters of a family of power-law models, investigate the explanatory power of different determinants of market impact and TC, and develop insights useable for pre-trade cost estimation. Our model estimation effort is informed by existing stylized market impact models and empirical findings about the market impact of block trades. These stylized models are reviewed earlier in this section. In addition to studying the effects of buy and sell trades, order size, price volatility and market cap, we examine the effect of the number of child trades and duration of the parent trade. Lastly, we also summarize two extensions of our model estimation effort that pertain to the effect of the bid-ask spread on market impact and to the market impact of ETFs traded via algorithmic trading.

This study has some important findings. A central novel insight that emerges from the analysis is that market impact can be substantially decreased by reducing the num-

ber of child order trades. There could be several explanations, but a simple plausible one is that reducing the number of child trades reduces the information leakage about trading intentions. This explanation is in line with the main premise of Farmer et al. (2012) [113] that, in an efficient market, market impact is determined by the information that is disseminated through trades. It is also consistent with studies showing that the price impact is larger as the duration between trades decreases and there is increased trading activity (Dufour and Engle 2000 [88]; Yang 2011 [263]), and that the number of trades may drive prices more than the actual trade size (Jones et al. 1994 [175]).

Another finding is the nuanced difference in market impact and TC that is related to buy and sell trades. On the overall level, this difference is statistically significant, and this presumably contradicts the buy-sell market impact asymmetry commonly reported for block trades. However, a deeper examination reveals a more nuanced difference between buy and sell trades. Visual inspection of the TC vs. trades' relative size and the TC vs stocks market cap shows that in some sub-ranges there is a difference between buys and sells. After dividing trades into 24 bins based on their relative size and testing the TC difference within each bin we observe that the difference is statistically significant in six bins; six out of 24 exceeds the proportion expected when testing multiple hypotheses at a 5% significance level. Consequently, separate market impact and TC models for buy and the sell trades are estimated, and observed the following about the difference between the coefficients of buys and sells. This difference is significant in models containing a subset of the explanatory factors (e.g., relative size alone, or relative size and volatility), but it is within a two standard errors threshold in models containing all the factors. Our findings notwithstanding, it is necessary to investigate further whether the underlying causes may be linked to the specifics of trade flow, the logic behind many of the trading algorithms.

It is important to mention two caveats . One relates to the role of permanent versus temporary impact in our analysis. The data set does not contain information about the price dynamics after the completion of the parent trades, that is, only completion price for the last child order is available. Therefore, it was not possible to determine what portion of the measured market impact is permanent and what portion of it is temporary. Another caveat relates to the high variability in the data. To obtain a better signal out of the noisy data, the data was smoothed through binning and extraction of key descriptives. The models are based on these descriptives. Nevertheless, for robustness testing all models were estimated using the raw unsmoothed data and the results are qualitatively similar, albeit the *R*-square values are understandably much lower.

**Empirical Details:** The market impact is modeled as a non-linear function of several stock-related and trade-related factors. The power law form is generally recommended for non-linear shape. The data analyzed in Velu et al. (2015) [256] consists of over 250,000 parent orders fully executed over a one-year period, Oct 1, 2009–Sept 30, 2010. All trades were completed within a single trading day. Each parent trade was sliced into multiple child trades; each child trade was executed in one or several partial fills. Only parent trades executed through widely used VWAP (volume-weighted average price) or POV (percent of volume) strategies (Brandes et al. (2007) [41]). This

ensures that the participation rate is well-defined and stays roughly constant over the duration of the parent trade. We define

$$\text{Market Impact (MI) in bps} = \text{side} \cdot \left( \frac{\text{Completion Price} - \text{Arrival Price}}{\text{Arrival Price}} \right) \cdot 10000$$

$$\text{Transaction Cost (TC) in bps} = \text{side} \cdot \left( \frac{\text{Average filled Price} - \text{Arrival Price}}{\text{Arrival Price}} \right) \cdot 10000$$

where the arrival price is the price when the first child order is executed and completion price is the price when the last child order is executed. Every trade in this data is explicitly identified as buy or sell. This contrasts with studies that use TAQ data that do not have the same explicit identification. Using Lee and Ready (1991) [189] heuristic rule to determine the trade direction can be sometimes inaccurate (Asquith, Oman and Safaya (2011) [16] and Chakrabarty, Moulton and Shkilko (2012) [54]). To make the analysis robust, we binned the explanatory variables and obtained smoothed averages; the analysis is based on binned data.

Results start with a visual inspection of the relationships. The graphs clearly indicate that the relationship between MI and RS is non-linear, and it remains non-linear with binning of other factors as well. The relationship between MI and relative size is monotonic till the relative size reaches 20% or so (Figure 9.6). Trades with longer durations clearly have a lower MI (Figure 9.7). Trades for higher volatility stocks, too, can be read to have a higher MI (Figure 9.8). Trades on larger market cap stocks have a smaller MI (Figure 9.9). Lastly, MI becomes higher as the number of child trades increases (Figure 9.10). This last relationship is important and deserves special attention particularly when designing execution strategies. There is a clear trade-off between low “immediate” liquidity impact of small child orders and long term information leverage of the parent order level when hitting the market too frequently. But the power law model is more general and can accommodate interaction effects as well. These graphs for various factors besides RS can be collapsed into a master curve, with appropriate normalization, as given in Lillo et al. (2003) [193].

Table 9.4 shows the results of power-law models ; the standard errors are shown in parentheses below the estimated regression coefficients. All coefficients in the models presented in Table 9.4 are statistically significant. This is consistent with Almgren et al. (2005) [6], except that in their model the market-cap-to-volume ratio is not significant. It should be noted that the values of *R*-square are significantly higher than other published studies, mainly due to binning and smoothing. The response variable is not an individual parent trade cost, but rather the average cost for trades in a bin representing a combination of various levels of predictors. Thus, the models developed here are intended to predict the average MI for a combination of predictors.

Several relevant observations can be made about the coefficients in the estimated models. For example, Model 3, which is traditionally well cited in the literature, can be written as:

$$\text{MI(bps)} = 15.9 \cdot \sigma^{0.92} \cdot (RS)^{0.56} \cdot T^{-0.24}$$

Note that the volatility coefficient is approximately ‘one’ and the coefficient of RS is about 0.5 has indeed been observed in the literature. It is consistent with the folklore that it takes one day’s volatility to trade one day’s volume. The duration effect

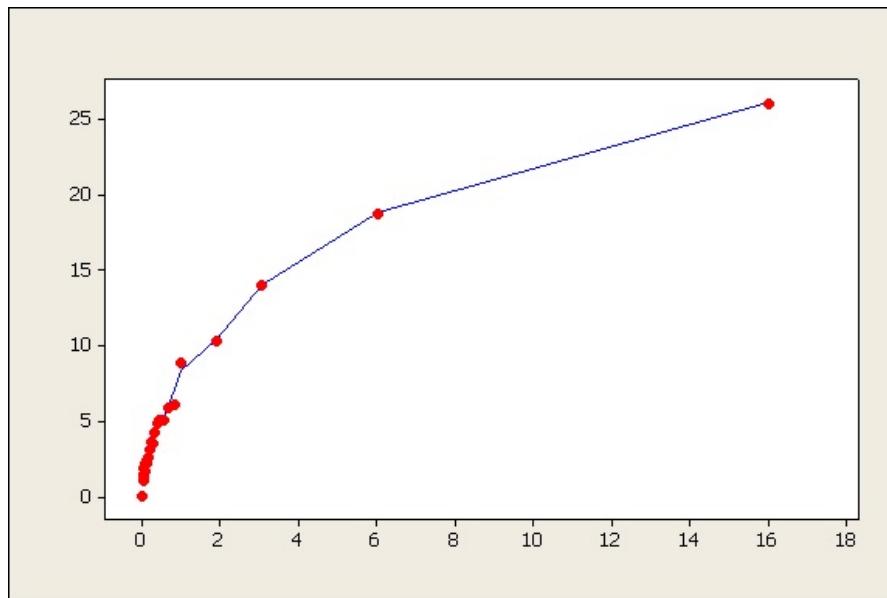


Figure 9.6: MI vs. RS.

Table 9.4: Response Variable is MI (bps) ( $\hat{\sigma}^2 = 225, n = 2,211$  bins)

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6*
Rel. Size	0.53 (0.02)	0.51 (0.02)	0.56 (0.02)	0.63 (0.02)	0.43 (0.02)	
Volatility		1.12 (0.09)	0.92 (0.09)	1.17 (0.11)	0.53 (0.1)	
Duration			-0.24 (0.02)	-0.25 (0.02)	-0.47 (0.3)	-0.27 (0.1)
Market Cap				-0.09 (0.02)	-0.12 (0.02)	-0.04 (0.07)
No. of Children					0.53 (0.04)	0.55 (0.16)
Constant	9.6 (0.34)	19.6 (1.07)	15.9 (0.9)	9.4 (1.3)	2.7 (0.46)	1.3 (1.3)
$\hat{\sigma}_{\text{error}}^2$	143	136	125	124	114	N/A
$R^2$	0.37	0.40	0.45	0.45	0.50	0.03

$$* \text{Response Variable} = \frac{\text{MI(bps)}}{(\text{Vol})(\text{RS})^{0.6}}$$

is also consistent as most of the parent trading occurs in less than two hours in the data set, not always in real life. This has led us to consider an additional model where  $\frac{\text{MI(bps)}}{(\text{Vol})(\text{RS})^{0.6}}$  is used as a normalized response variable. This variable can be taken as the multiplicative residual term of Model (2) in Table 9.4. The regression results of

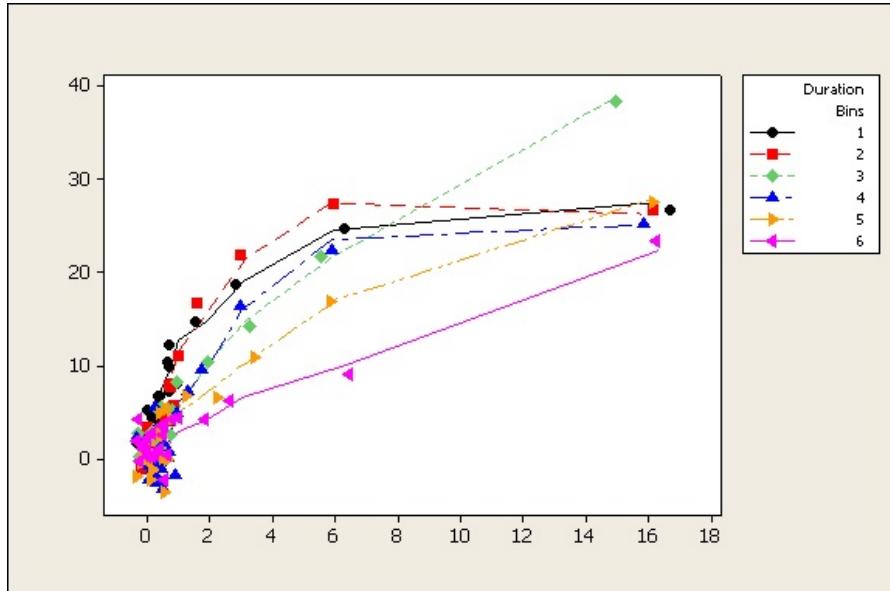


Figure 9.7: MI vs. RS and Duration.

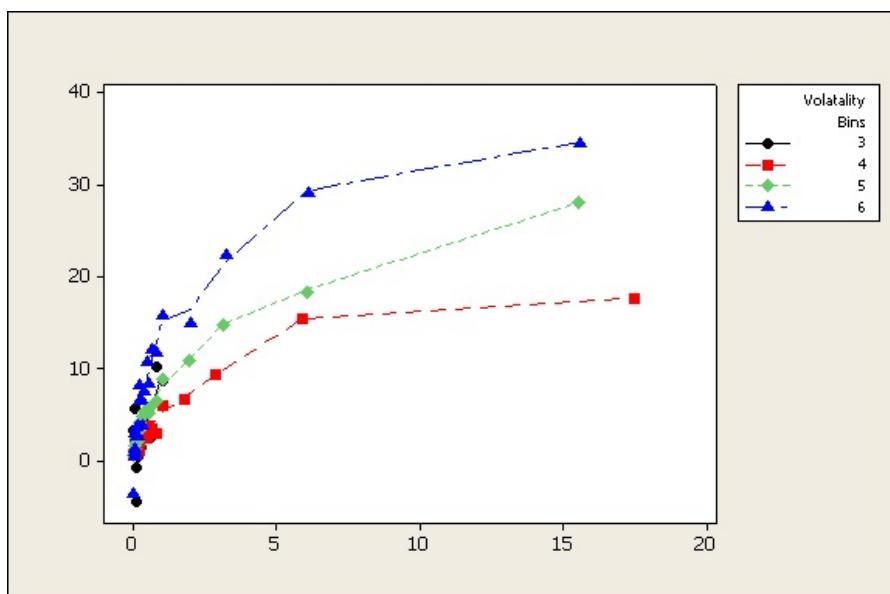


Figure 9.8: MI vs. RS and Volatility.

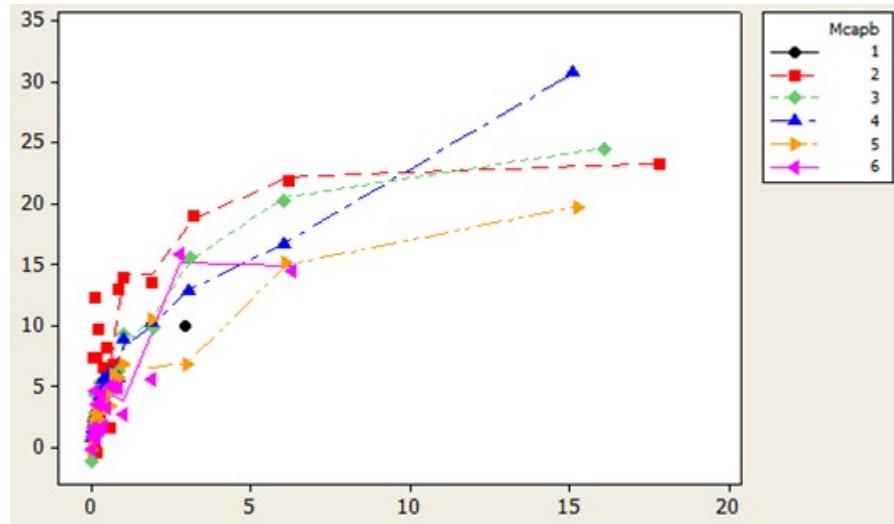


Figure 9.9: MI vs. RS and Market Cap.

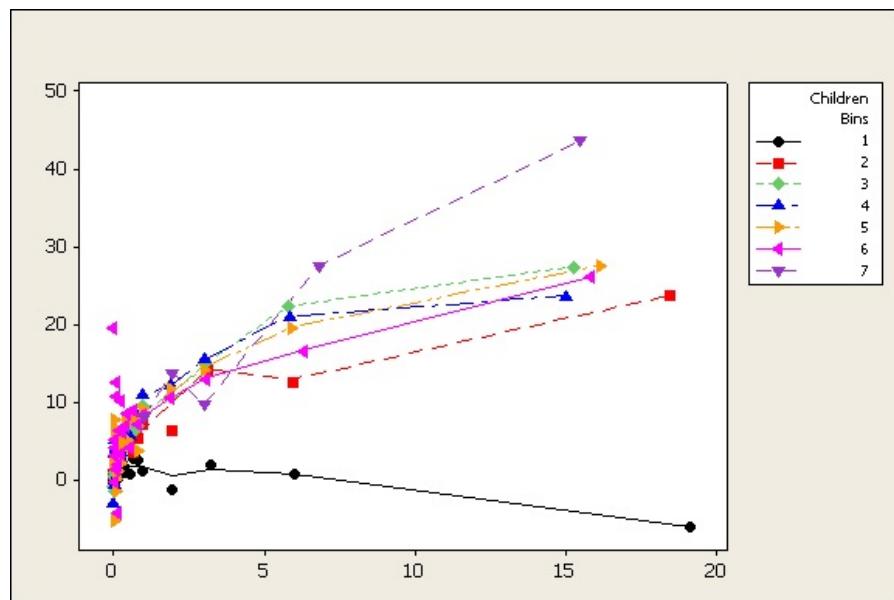


Figure 9.10: MI vs. RS and No. of Children.

this variable on other variables are given under Model (6\*). Two observations can be made: the duration and the number of child trades are still significant factors. However, the increase in *R*-square is somewhat smaller than the unrestricted versions of the same model, labeled as Model (5) in Table 9.4.

Other key findings can be summarized as follows:

- Relative size and volatility consistently exhibit positive influences on MI; their impact varies depending on other factors in the model.
- Duration has an inverse effect. Its use in the model must be carefully calibrated as the execution of parent trades using child trades can be highly concentrated.
- Stocks with higher market capitalizations tend to have smaller price impact for the same relative size.
- The number of child trades, which is usually determined by the trading algorithm, does reveal a positive effect on MI.

There are several practical considerations that may limit the use of these models for pre-trade cost estimation. These challenges in MI calibration are due to the following facts:

- Not enough data for certain combinations of the factors involved.
- MI is a noisy measure which is hard to calibrate ex-post.
- Sometimes, market moves at a much larger magnitude than MI.
- Use of raw data results in a poor predictive model; thus a need for aggregation. But aggregation obviously ignores the wide variation within the aggregated unit.

---

**9.4 Microstructure Noise Models for Multiple Stocks: Issues of Synchronizations, etc.**

---

**9.5 Simultaneous Execution Algorithms for Multiple Instruments**

---

**9.6 A Stochastic Adaptive Control Approach to Optimal Execution**

---

**9.7 Role of Market Makers in Limit Order Markets**

---

**9.8 Open Issues**

---

**9.9 Supplements and Problems**



# 10

---

## *Algorithmic Trading: Some Potential Research Issues*

---

### CONTENTS

10.1	Extending the Algorithms to Derivative markets .....	277
10.2	Developing a Trading Platforms for Simultaneous Execution of Assets .....	277
10.3	Future of Algorithmic Trading .....	277

---

### **10.1 Extending the Algorithms to Derivative markets**

---

### **10.2 Developing a Trading Platforms for Simultaneous Execution of Assets**

---

### **10.3 Future of Algorithmic Trading**



— | — | —

## **Part IV**

# **Informatics and Regulatory Issues**

— | — | —



# 11

---

## *Informatics and Computation for Algorithmic Trading*

---

### CONTENTS

11.1	Model for electronic Trading Platforms .....	281
11.1.1	Trading Rules .....	281
11.1.2	Access to Order Book; Information Flow and the Infrastructure .....	281
11.1.3	High Frequency vs. Low Frequency Set Up .....	281
11.2	Current Practices and Alternative Trading Systems .....	281
11.2.1	Computer Networks and Broadcasting .....	281
11.2.2	Data Center and Co-location Service .....	281
11.2.3	Dark Pools .....	281
11.3	Developing Modern Trading Platforms .....	281

---

### **11.1 Model for electronic Trading Platforms**

#### **11.1.1 Trading Rules**

#### **11.1.2 Access to Order Book; Information Flow and the Infrastructure**

#### **11.1.3 High Frequency vs. Low Frequency Set Up**

---

### **11.2 Current Practices and Alternative Trading Systems**

#### **11.2.1 Computer Networks and Broadcasting**

#### **11.2.2 Data Center and Co-location Service**

#### **11.2.3 Dark Pools**

---

### **11.3 Developing Modern Trading Platforms**



# 12

---

## *Risk Management and Regulatory Issues*

---

### CONTENTS

12.1	Operational Risk and Information Technology Risk .....	283
12.2	Risk Management: Pre-trade and Post-trade .....	283
12.3	Circuit Breakers .....	283
12.4	Order-to-trade Ratio and Quote Stuffing .....	283
12.5	Regulation of Parallel Processes .....	283
12.6	Information Access and Fairness Issues .....	283

---

### **12.1 Operational Risk and Information Technology Risk**

---

### **12.2 Risk Management: Pre-trade and Post-trade**

---

### **12.3 Circuit Breakers**

---

### **12.4 Order-to-trade Ratio and Quote Stuffing**

---

### **12.5 Regulation of Parallel Processes**

---

### **12.6 Information Access and Fairness Issues**

---



---

## Bibliography

---

- [1] A. R. Admati and P. Pfleiderer. A theory of intraday patterns: Volume and price variability. *Review of Financial studies*, 1(1):3–40, 1988.
- [2] Aït-Sahalia, Y. and Mykland, P. A. and Zhang, L. How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial studies*, 18(2):351–416, 2005.
- [3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974.
- [4] R. Almgren. Execution costs. *Encyclopedia of Quantitative Finance*, pages 1–5, 2008.
- [5] R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *The Journal of Risk*, 3:5–39, 2000.
- [6] R. Almgren, C. Thum, E. Hauptmann, and H. Li. Equity market impact. *Risk*, 18(7):57–62, 2005.
- [7] S. Anatolyev and A. Gerko. A Trading Approach to Testing for Predictability. *The Journal of Bunsincer Economics Statistics*, 23:455–461, 2005.
- [8] S. Anatolyev and A. Gerko. Modeling Financial Return Dynamics via Decomposition. *The Journal of Bunsincer Economics Statistics*, 28:232–245, 2010.
- [9] T. G. Andersen. Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility. *The Journal of Finance*, 51:116–204, 1996.
- [10] T. G. Andersen and T. Bollerslev. Answering the Skeptics: Yes, Standard Volatility Models do provide Accurate Forecasts. *International Economic Review*, 39:885–905, 1998.
- [11] T.W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22:327–351, 1951.
- [12] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Second Edition. New York, Wiley, 1984.

- [13] T.W. Anderson and A.M. Walker. On the asymptotic distribution of the auto-correlations of a sample from a linear stochastic process. *Annals of Mathematical Statistics*, 35:1296–1303, 1964.
- [14] A. Ang and A. Timmermann. Regime Changes and Financial Markets. *Annual Review of Finance and Economics*, 4:313–337, 2012.
- [15] W. Antweiler and M. Z. Frank. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.
- [16] P. Asquith, R. Oman, and C. Safaya. Short sales and trade classification algorithms. *The Journal of Financial Markets*, 13:157–173, 2010.
- [17] P. Bajgrowicz and O. Scaillet. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *The Journal of Financial Economics*, 106(3):473–491, 2012.
- [18] M. Baker and J. Wurgler. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006.
- [19] M. Baker and J. Wurgler. Investor sentiment in the stock market. *The Journal of Economic Perspectives*, 21(2):129–151, 2007.
- [20] F.M. Bandi and J.R. Russell. Separating Microstructure Noise from Volatility. *The Journal of Financial Economics*, 79:655–692, 2006.
- [21] N. Barberis, A. Shleifer, and R. Vishny. A model of investor sentiment. *The Journal of Financial Economics*, 49:307–343, 1998.
- [22] O. E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *The Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280, 2002.
- [23] L. Barras, O. Scaillet, and R. Wermers. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance*, 65(1):179–216, 2010.
- [24] P. Bertrand and C. Protopopescu. The statistics of the information ratio. *International Journal of Business*, 15:71–86, 2010.
- [25] D. Bertsimas and A. W. Lo. Optimal Control of Executions Costs. *The Journal of Financial Markets*, 1:1–50, 1998.
- [26] R. Bewley and M. Yang. Tests for cointegration based on canonical correlation analysis. *The Journal of the American Statistical Association*, 90:990–996, 1995.

- [27] B. Biais, P. Hillion, and C. Spatt. An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse. *The Journal of Finance*, 50:1655–1689, 1995.
- [28] F. Black. Towards a Fully Automated Exchange, Part I. *Financial Analysts Journal*, 27:29–34, 1971.
- [29] F. Black. Capital market equilibrium with restricted borrowing. *The Journal of Business*, 45:444–454, 1972.
- [30] F. Black and R. Litterman. Global Portfolio Optimization. *Financial Analysts Journal*, 48 No. 5:28–43, 1992.
- [31] L. Blume, D. Easley, and M. O’Hara. Market Statistics and Technical Analysis: The role of volume. *The Journal of Finance*, 49:153–181, 1994.
- [32] T. Bollerslev. Generalized Autoregressive Conditional hetero skedasticity. *The Journal of Econometrics*, 31:307–327, 1986.
- [33] M. Borkovec and H. G. Heidle. Building and Evaluating a Transaction Cost Model: A Primer. *The Journal of Trading*, 5:57–77, 2010.
- [34] P. Bossaerts. Common nonstationary components of asset prices. *The Journal of Economic Dynamics and Control*, 12(2):347–364, 1988.
- [35] J. P. Bouchaud, J. D. Farmer, and F. Lillo. *How markets digest supply and demand and slowly incorporate information into prices*. Academic Press, 2009.
- [36] J. P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart. Fluctuations and response in financial markets: The subtle nature of "random" price changes. *Quantitative Finance*, 4:176–190, 2004.
- [37] J-P. Bouchaud, M. Mezard, and M. Potters. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2:251–256, 2002.
- [38] D. Bowen, M. C. Hutchinson, and N. O’Sullivan. High-Frequency Equity Pairs Trading: Transaction Costs, Speed of Execution and Patterns in Returns. *The Journal of Trading, Summer*, pages 31–38, 2010.
- [39] G.E.P. Box and G.C. Tiao. A canonical analysis of multiple time series. *Biometrika*, 64:355–365, 1977.
- [40] P. Boyle, L. Garlappi, R. Uppal, and T. Wang. Keynes Meets Markowitz: The Trade-off between familiarity and diversification. *Management Science*, 58:253–272, 2012.
- [41] Y. Brandes, I. Domowitz, B. Jiu, and H. Yegerman. *Algorithms, trading costs and order size, Algorithmic Trading: A Buy side Handbook*, 2nd ed. The Trade (London), 2007.

- [42] M.W. Brandt and P. Santa-Clara. Dynamic Portfolio Selection by Augmenting the Asset Space. *The Journal of Finance*, 61:2187–2217, 2006.
- [43] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [44] L. Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11(7):1493–1517, 1999.
- [45] D.R. Brillinger. *Time Series: Data Analysis and Theory*. expanded edition. San Francisco: Holden-Day, 1981.
- [46] W. Brock, J. Lakonishok, and B. LeBaron. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance*, 47:1731–1764, 1992.
- [47] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Lace. Sparse and Stable Markowitz Portfolios. *Proceedings of the National Academy of Sciences*, 106:12267–12272, 2009.
- [48] C. Brownless, F. Cipollini, and G.M. Gallo. Intra-daily volume modeling and prediction for algorithmic trading. *The Journal of Financial Econometrics*, 9:489–518, 2011.
- [49] B. Bruder, N. Gasssel, J-C. Richard, and T. Roncalli. Regularization of Portfolio Allocation. *White Paper Issue #10*, 2013.
- [50] J.Y. Campbell, S.J. Grossman, and J. Wang. Trading Volume and Serial Correlation in Stock Returns. *The Quarterly Journal of Economics*, 108:905–939, 1993.
- [51] J.Y. Campbell, A.W. Lo, and A.C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, New Jersey, 1996.
- [52] M. M. Carhart. On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82, 1997.
- [53] A.P. Chaboud, B. Chiquoine, E. Hjalmarsson, and C. Vega. Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, 69:2045–2084, 2014.
- [54] B. Chakrabarty, P. C. Moulton, and A. Shkilko. Short sales, long sales, and the lee-ready trade classification algorithm revisited. *The Journal of Financial Markets*, 15(4):467–491, 2012.
- [55] K. Chan and W-M. Fong. Trade Size, Order imbalance and the Volatility-Volume Relation. *The Journal of Financial Economics*, 57:247–273, 2000.
- [56] L. Chan and J. Lakonishok. institutional Equity Trading Costs, NYSE versus Nasdaq. *The Journal of Finance*, 52:713–735, 1997.

- [57] L. K. C. Chan and J. Lakonishok. The behavior of stock prices around institutional trades. *The Journal of Finance*, 50:1147–1174, 1995.
- [58] L. K. C. Chan and J. Lakonishok. Institutional Equity Trading Costs: NYSE Versus Nasdaq. *The Journal of Finance*, 52(2):176–190, 1997.
- [59] N.F. Chen, R. Roll, and S.A. Ross. Economic forces and the stock market. *The Journal of Business*, 59:383–403, 1986.
- [60] S. Chib. Estimation and comparison of multiple change-point models. *The Journal of Econometrics*, 86:221–241, 1998.
- [61] C. Chiyachantana, P. K. Jain, C. Jiang, and R. A. Wood. International Evidence on Institutional Trading Behavior and Price Impact. *The Journal of Finance*, 59:869–898, 2004.
- [62] T. Chordia, R. Roll, and A. Subrahmanyam. Order imbalance, liquidity, and market returns. *The Journal of Financial Economics*, 65:111–130, 2002.
- [63] P. K. Clark. A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. *Econometrica*, 41:135–155, 1973.
- [64] J. Conrad and G. Kaul. An Anatomy of Trading Strategies. *Review of Financial Studies*, 11:489–519, 1998.
- [65] J.S. Conrad, A. Hameed, and C. Niden. Volume and Autocovariances in Short-horizon individual security returns. 49:1305–1329, 1994.
- [66] R. Cont and A. Kukanov. Optimal order placement in limit order markets. *Quantitative Finance*, 17:21–39, 2017.
- [67] R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. *The Journal of Financial Econometrics*, 12:47–88, 2014.
- [68] R. Cont, S. Stoikov, and R. Talreja. A Stochastic Model for Order Book Dynamics. *Operations Research*, 58:549–563, 2010.
- [69] M. Cooper. Filter values based on price and volume in individual security overreaction. *Review of Financial Studies*, 12:901–935, 1999.
- [70] A. Cowles. Can Stock Market Forecasters Forecast. *Econometrica*, 1:309–324, 1933.
- [71] A. Cowles. Stock Market Forecasting. *Econometrica*, 12:206–214, 1944.
- [72] D. R. Cox and P. A. W. Lewis. The Statistical Analysis of Series of Events. *Chapman and Hall, London*, 1966.
- [73] D. M. Cutler, J. M. Poterba, and L. H. Summers. *What moves stock prices?*, volume 15. National Bureau of Economic Research Cambridge, Mass., USA, 1989.

- [74] S. Da, J. Engelberg, and P. Gao. In search of attention. *The Journal of Finance*, 66(5):1461–1499, 2011.
- [75] Z. Da, J. Engelberg, and P. Gao. The sum of all fears investor sentiment and asset prices. *Review of Financial Studies*, 28(1):1–32, 2015.
- [76] R. Dahlhaus and S. Subba Rao. Statistical Inference for Time-Varying ARCH Processes. *The Annals of Statistics*, 34:1075–1114, 2006.
- [77] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Springer, New York, 2003.
- [78] H. E. Daniels. Autocorrelation between first differences of mid-ranges. *Econometrica: Journal of the Econometric Society*, pages 215–219, 1966.
- [79] Das, S.R., and Chen, M.Y. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53:1375–1388, 2007.
- [80] DeLong, B.J., and Shleifer, A., and Summers, L.H., and Walmann, R.J. Noise trader risk in financial markets. *The Journal of Political Economy*, 98:703–738, 1990.
- [81] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naïve diversification: How inefficient is the  $1/N$  portfolio strategy? *The Review of Financial Studies*, 22:1915–1953, 2009.
- [82] A.P. Dempster, N-M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *The Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- [83] D. A. Dickey, W. R. Bell, and R. B. Miller. Unit roots in time series models: tests and implications. *American Statistician*, 40:12–26, 1986.
- [84] D. A. Dickey and W. A. Fuller. Distribution of the estimates for autoregressive time series with a unit root. *The Journal of the American Statistical Association*, 74:427–431, 1979.
- [85] B. Do and R. Faff. Does simple pairs trading still work? *The Financial Analysts Journal*, 66(4):83–95, 2010.
- [86] I. Domowitz, J. Glen, and A. Madhavan. Liquidity, Volatility and Equity Trading costs across countries and over time. *International Finance*, 4:221–255, 2001.
- [87] I. Domowitz and H. Yegerman. The cost of algorithmic trading: A first look at comparative performance. *Algorithmic Trading: Precision, Control, Execution*, 2005.
- [88] A. Dufour and R. Engle. Time and the price impact of a trade. *The Journal of Finance*, 55(2):467–498, 2000.

- [89] D. Easley, M.L. De Prado, and M. O’Hara. Flow Toxicity and Liquidity in a High Frequency World. *Review of Financial Studies*, 25:1457–1493, 2012.
- [90] D. Easley, M.L. De Prado, and M. O’Hara. Optimal Execution Horizon. *Mathematical Finance*, 25:640–672, 2015.
- [91] D. Easley and M. O’ Hara. Price, trade size, and information in security markets. *The Journal of Financial Economics*, 19:69–90, 1987.
- [92] D. Easley and M. O’ Hara. Time and the Process of Security Price Adjustment. *The Journal of Finance*, 19:69–90, 1992.
- [93] D. Easley, N. Kiefer, M. O’Hara, and J. Paperman. Liquidity, Information and Infrequently traded stocks. *The Journal of Finance*, 51:1405–1536, 1996.
- [94] D. Easley, M. Lopez de Prado, and M. (Ed.) O’Hara. *High-frequency Trading: New Realities for Traders, Markets and Regulators*. Risk Books, New York, 2 edition, 2014.
- [95] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Annals of Statistics*, 32:407–499, 2004.
- [96] J. Engelberg, P. Gao, and R. Jagannathan. An anatomy of pairs trading: the role of idiosyncratic news, common information and liquidity. In *Third Singapore International Conference on Finance*, 2009.
- [97] R. Engle and R. Ferstenberg. Execution Risk. *The Journal of Portfolio Management*, 33(4):34–44, 2007.
- [98] R. Engle and A. Lunde. Trades and Quotes: A Bivariate Point Process. *The Journal Financial Economics*, 1:159–188, 2003.
- [99] R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations. *Econometrica*, 50:987–1007, 1982.
- [100] R. F. Engle, R. Ferstenberg, and J. R. Russell. Measuring and modeling execution cost and risk. *The Journal of Portfolio Management*, 38(2):14–28, 2012.
- [101] R. F. Engle and C. W. J. Granger. Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55:251–276, 1987.
- [102] R. F. Engle and D. Kraft. *Multiperiod forecast error variances of inflation estimated from ARCH models*, in: A. Zellner, ed: *Applied time series analysis of economic data*. Bureau of the census, Washington, DC, 1983.
- [103] R. F. Engle and J. R. Russell. Autoregressive Conditioned Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrics*, 66:1127–1162, 1998.
- [104] R.F. Engle and S. Kozicki. Testing for Common Features. *The Journal of Business and Economic Statistics*, 11:369–380, 1993.

- [105] T. W. Epps and M. L. Epps. The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distribution Hypothesis. *Econometrica*, 44:305–321, 1976.
- [106] F. J. Fabozzi, S. M. Focardi, and P. N. Kolm. *Quantitative Equity Investing: Techniques and Strategies*. Wiley and Sons, 2006.
- [107] E. F. Fama and M. E. Blume. Filter Rules and Stock - Market Trading. *The Journal of Business*, 39:226–241, 1966.
- [108] E. F. Fama and K. R. French. A five-factor after pricing model. *The Journal of Financial Economics*, 116:1–22, 2015.
- [109] E. F. Fama and K. R. French. International tests of a five-factor asset pricing model. *Fama-Miller Working Paper*, 2015.
- [110] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *The Journal of Econometrics*, 147:187–197, 2008.
- [111] J. Fan, F. Han, and B. Vickers. Robust inference of risks of large portfolios. *The Journal of Econometrics*, 194:298–308, 2016.
- [112] J. Fan, J. Zhang, and K. Yu. Vast Portfolio Selection with Cross-Exposure Constraints. *The Journal of the American Statistical Association*, 107:592–606, 2012.
- [113] J. D. Farmer, A. Gerig, F. Lillo, and H. Waelbroeck. How Efficiency Shapes Market Impact. <http://arxiv.org/abs/11025457>, 2012.
- [114] W.E. Ferson and A.F. Siegel. The use of conditioning information in portfolios. *The Journal of Finance*, 56:967–982, 2001.
- [115] T. Foucault and A. J. Menkveld. Competition for Order Flow and Smart Order Routing Systems. *The Journal of Finance*, pages 119–157, 2008.
- [116] A. Frazzini and L.H. Pedersen. Betting against beta. *The Journal of Financial Economics*, 111:1–25, 2014.
- [117] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37. Springer, 1995.
- [118] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, volume 96, pages 148–156. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1996.
- [119] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences*, volume 55, pages 119–139, 1997.

- [120] J. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [121] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000.
- [122] A. Frino, E. Jarnebic, and A. Lepone. The determinants of price impact of block trades: Further evidence. *Abacus*, 43(2):94–106, 2007.
- [123] K. A. Froot and E. M. Dabora. How are stock prices affected by the location of trade? *The Journal of Financial Economics*, 53(2):189–216, 1999.
- [124] P. Fryzlewicz, T. Sapatinas, and S. Subba Rao. Normalized Least-Squared Estimation in Time-Varying ARCH Models. *The Annals of Statistics*, 36:742–786, 2008.
- [125] W. A. Fuller. *Introduction to Statistical Time Series, Second Edition*. John Wiley, New York, 1996.
- [126] A. R. Gallant, P. E. Rossi, and G. Tauchen. Nonlinear Dynamic Structures. *Econometrica*, 61:871–908, 1993.
- [127] P.H. Garthwaite. An interpretation of partial least squares. *The Journal of the American Statistical Association*, 89:122–127, 1994.
- [128] E. Gatev, W. N. Goetzmann, and R. G. Rouwenhorst. Pairs Trading: Performance of a Relative Value Arbitrage Rule. *The Review of Financial Studies*, 19:797–827, 2006.
- [129] J. Gatheral. No-Dynamic-Arbitrage and Market Impact. *Quantitative Finance*, 10:769–759, 2010.
- [130] R. Gencay. The predictability of security returns with simple technical trading rules. *The Journal of Empirical Finance*, 5:347–359, 1998.
- [131] S. Gervais, R. Kaniel, and D. H. Mingelgrin. The high-volume return premium. *The Journal of Finance*, 56(3):877–919, 2001.
- [132] S. Gervais, R. Kaniel, and D.H. Mingelgrin. The High-Volume Return Premium. *The Journal of Finance*, 56:877–919, 2001.
- [133] E. Ghysels, C. Gouriéroux, and J. Jasiak. Stochastic volatility duration models. *The Journal of Econometrics*, 119:413–433, 2004.
- [134] M.R. Gibbons, S.A. Ross, and J. Shanken. A test of the efficiency of a given portfolio. *Econometrica*, 57:1121–1152, 1989.
- [135] L. Glosten. Is the Electronic Open Limit Order Book Inevitable. *The Journal of Finance*, 49:1127–1161, 1994.

- [136] L. R. Glosten and P. R. Milgrom. Bid, ask and transaction prices in a specialist market with hetero-geneously informed traders. *The Journal of Financial Economics*, 14:71–100, 1985.
- [137] J. Gonzalo and C.W.J. Granger. Estimation of common long-term memory components in cointegrated systems. *The Journal of Business and Economic Statistics*, 13:27–35, 1995.
- [138] R. C. Grinold and R. N. Kahn. *Active Portfolio Management, Second Edition*. McGraw - Hill, 2000.
- [139] Groß-Klussmann, A., and Hautsch, N. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *The Journal of Empirical Finance*, 18:321–340, 2011.
- [140] J.D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.
- [141] J.D. Hamilton. Analysis of Time Series subject to changes in regime. *The Journal of Econometrics*, 45:39–70, 1990.
- [142] J.D. Hamilton. Macroeconomic regimes and Regime Shifts, Handbook of Macroeconomics. 2:163–201, 2016.
- [143] P. R. Hansen. A test for superior predictive ability. *The Journal of Business and Economic Statistics*, 23:365–380, 2005.
- [144] M. O’ Hara and M. Ye. Is market fragmentation harming market quality. *The Journal of Financial Economics*, pages 454–474, 2011.
- [145] L.E. Harris and V. Panchapagesan. The information content of the limit order book: evidence from NYSE specialist trading decisions. *The Journal of Financial Markets*, 8:25–67.
- [146] A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, 1989.
- [147] J. Hasbrook and G. Saar. Technology and liquidity provision: The blurring of traditional definitions. *The Journal of Financial Markets*, 12:143–172, 2009.
- [148] J. Hasbrouk. Measuring the information content of stock trades. *The Journal of Financial Economics*, 46:179–207, 1991.
- [149] N. Hautsch and R. Huang. The market impact of a limit order. *The Journal of Economic Dynamics and Control*, 36:501–522, 2012.
- [150] I.S. Helland. On the structure of partial least squares regression. *Communications in Statistics, Simulation and Computation*, B17:581–607, 1988.
- [151] I.S. Helland. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 17:97–114, 1990.

- [152] T. Hendershott, J. Brogaard, and R. Riordon. High Frequency Trading and Price Discovery. *Review of Financial Studies*, 27:2267–2306, 2014.
- [153] T. Hendershott, C. M. Jones, and A. J. Menkveld. Does Algorithmic Trading Improve Liquidity? *The Journal of Finance*, pages 1–3, 2011.
- [154] T. Hendershott, C. M. Jones, and A. J. Menkveld. Does Algorithmic Trading Improve Liquidity? *The Journal of Finance*, 66(1):1–33, 2011.
- [155] T. Hendershott and R. Riordan. Algorithmic Trading and Information. <http://faculty.haas.berkeley.edu/hender/ATInformation.pdf>, 2011.
- [156] T. Hendershott and M. Seasholes. Market Maker Inventories and Stock Prices. *American Economic Review*, 97:210–214, 2007.
- [157] T. Ho and H.R. Stoll. Optimal dealer pricing under transactions and return uncertainty. *The Journal of Financial Economics*, 9:47–73, 1981.
- [158] S. Hogan, R. Jarrow, M. Teo, and M. Warachka. Testing market efficiency using statistical arbitrage with application to momentum and value strategies. *The Journal of Financial Economics*, 73:525–565, 2004.
- [159] R. W. Holthausen, R. W. Leftwich, and D. Mayers. Large-block transactions, the speed of response, and temporary and permanent stock-price effect. *The Journal of Financial Economics*, 26:71–95, 1990.
- [160] P. H. Hsu, Y. C. Hsu, and C. M. Kuan. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *The Journal of Empirical Finance*, 17:471–484, 2010.
- [161] Huang, D., and Jiang, F., and Tu, J., and Zhou, G. Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies*, 28:791–837, 2015.
- [162] G. Huberman. Familiarity breeds investment. *Review of Financial Studies*, 14:659–680, 2001.
- [163] G. Huberman and W. Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 72:1247–1275, 2004.
- [164] S. Hvidkjaer. A Trade-Based Analysis of Momentum. *Review of Financial Studies*, 119:457–491, 2006.
- [165] R. Jagannathan and T. Ma. Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *The Journal of Finance*, 58:1651–1683, 2003.
- [166] C.M. Jarque and A.K. Beru. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economic Letters*, 6:255–259, 1980.

- [167] N. Jegadeesh. Discussion of LMW (2000). *The Journal of Finance*, pages 1765–1770, 2000.
- [168] N. Jegadeesh and S. Titman. Returns to buying winners and selling losers: Implications for Stock Market Efficiency. *The Journal of Finance*, 48:65–91, 1993.
- [169] N. Jegadeesh and S. Titman. Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of Finance*, 56:699–720, 2001.
- [170] N. Jegadeesh and S. Titman. Cross-sectional Time Series determinants of momentum returns. *Review of Financial Studies*, 15:143–157, 2002.
- [171] W. Jiang, L. Shu, and D.W. Apley. Adaptive CUSUM procedures with EWMA-based shift estimators. *IIE Transactions*, 40:992–1003, 2008.
- [172] J.D. Jobson and B. Korkie. Estimation for Markowitz Efficient Portfolios. *The Journal of American Statistical Association*, 75:544–554, 1980.
- [173] S. Johansen. Statistical analysis of cointegration vectors. *The Journal of economic dynamics and control*, 12(2):231–254, 1988.
- [174] S. Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, pages 1551–1580, 1991.
- [175] C. Jones, G. Kaul, and M. Lipson. Information, trading and volatility. *The Journal of Financial Economics*, 36:127–154, 1994.
- [176] R.N. Kahn and M. Lemmon. Smart Beta: The Owner’s Manual. *The Journal of Portfolio Management*, 41(2):76–83, 2015.
- [177] D. B. Keim and A. Madhavan. The upstairs market for large-block transactions: Analysis and measurement of price effects. *Review of Financial Studies*, 9:1–36, 1996.
- [178] D.B Keim and A. Madhavan. Transaction Costs and Investment style: an inter-exchange analysis of institutional equity trades. *The Journal of Financial Economics*, 46:265–292, 1997.
- [179] J. M. Keynes. The general theory of employment. *The Quarterly Journal of Economics*, pages 209–223, 1937.
- [180] A. Korajczyk and R. Sadka. Are Momentum Profits Robust to Trading Costs? *The Journal of Finance*, LIX:1039–1081, 2004.
- [181] A. Kourtis. On the distribution and estimation of trading costs. *The Journal of Empirical Finance*, 29:230–245, 2014.
- [182] A. Kyle. Continuous Time Auctions and Insider Trading. *Econometrics*, 53:1315–1336, 1985.

- [183] T. L. Lai and H. Xing. *Statistical Models and Methods for Financial Markets*. Springer, 2008.
- [184] T.L. Lai and H. Xing. Stochastic Change-Point ARX GARCH Models and their applications to econometric times series. *Statistica Sinica*, 23:1573–1594, 2013.
- [185] T.L. Lai, H. Xing, and Z. Chen. Mean-Variance Portfolio Optimization when Means and Covariances are Unknown. *The Annals of Applied Statistics*, 5:798–823, 2011.
- [186] J. Lakonishok, A. Shleifer, and R. W. Vishny. Contrarian investment, extrapolation, and risk. *The Journal of Finance*, 49(5):1541–1578, 1994.
- [187] O. Ledoit and M. Wolf. Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection. *The Journal of Empirical Finance*, 10:603–621, 2003.
- [188] C. M. Lee and B. Swaminathan. Price Momentum and Trading Volume. *The Journal of Finance*, LV:2017–2069, 2000.
- [189] C.M. Lee and M. Ready. Inferring trade direction from intradaily data. *The Journal of Finance*, 46:733–746, 1991.
- [190] Y.S. Lee. Some results on the distribution of Wilks’ likelihood-ratio criterion. *Biometrika*, 59:649–664, 1972.
- [191] J. Lewellen. Momentum and Autocorrelation in Stock Returns. *Review of Financial Studies*, 15:65–91, 2002.
- [192] Liew, J.K-S., and Guo, S., and Zhang, T. Tweet sentiments and crowd-sourced earnings estimates as valuable sources of information around earnings releases. *The Journal of Alternative Investments*, Winter Issue:1–20, 2017.
- [193] F. Lillo, J. D. Farmer, and R. Mantegna. Master curve for price impact function. *Nature*, 421:129–130, 2003.
- [194] J. Lintner. The valuation of risky assets and the selection of risky investment in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47:13–37, 1965.
- [195] G. Llorente, R. Michaely, G. Saar, and J. Wang. Dynamic volume-return relation of individual stocks. 15:1005–1047, 2002.
- [196] A. W. Lo, H. Mamaysky, and J. Wang. Foundation of Technical Analysis: Computational Algorithms, Statistical Inference and Empirical Implementation. *The Journal of Finance*, LV(4):1705–1765, 2000.
- [197] A.W. Lo. The statistics of sharpe ratios. *Financial Analysis Journal*, pages 36–52, 2002.

- [198] A.W. Lo and A. C. MacKinlay. When are Contrarian Profits due to Stock Market Overreaction? *Review of Financial Studies*, 3:175–205, 1990.
- [199] A.W. Lo, A.C. MacKinlay, and J. Zhang. Econometric models of limit-order executions. *The Journal of Financial Economics*, 65:31–71, 2002.
- [200] T. F. Loeb. Trading Cost: The Critical Link between Investment Information and Results. *Financial Analyst Journal*, 39(3):39–44, 1983.
- [201] J. Lorenz and R. Almgren. Mean-Variance Optimal Adaptive Execution. *Applied Mathematical Finance*, 18(4):395–422, 2011.
- [202] A. Madhavan, M. Richardson, and M. Roomans. Why do security prices change? A transaction-level analysis of NYSE stocks. *Review of Financial Studies*, 10(4):1035–1064, 1997.
- [203] B. G. Malkiel. *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing, 10th Edition*. Norton, 2012.
- [204] H. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Wiley: New York, 1959.
- [205] A. I. McLeod and W. K. Li. Diagnostic checking ARMA time series models using squared-residual autocorrelations. *The Journal of Time Series Analysis*, 4:269–273, 1983.
- [206] L. Menkhoff, L. Sarno, M. Schmeling, and A. Schrimpf. Currency Momentum Strategies. *The Journal of Financial Economics*, 106:660–684, 2012.
- [207] L. Menkhoff and M. P. Taylor. The Obstinate Passion of Foreign Exchange Professionals: Technical Analysis. *The Journal of Economic Literature*, XLV:936–972, 2007.
- [208] R.C. Merton. On estimating the expected return on the market. *The Journal of Financial Economics*, 8:323–336, 1980.
- [209] R.T. Merton. An intertemporal capital asset pricing model. *Econometrica*, 41:867–887, 1973.
- [210] R.O. Michaud. *Efficient Asset Management*. Harvard Business School Press, Boston, 1989.
- [211] G. Mitra and L. Mitra. *The Handbook of New Analytics in Finance (Ed)*. Wiley Finance, 2011.
- [212] T. Moorman. An empirical investigation of methods to reduce transaction costs. *The Journal of Empirical Finance*, 29:230–245, 2014.
- [213] E. Moro, J. Vicente, L. G. Moyano, A. Gerig, J. D. Farmer, G. Vaglica, F. Lillo, and R. N. Mantegna. Market impact and trading profile of hidden orders in stock markets. *Physical Review E*, 80(6):066102, 2009.

- [214] T. J. Moskowitz, Y. H. Ooi, and L. H. Pedersen. Time Series Momentum. *The Journal of Financial Economics*, 104:228–250, 2012.
- [215] A.A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *The Journal of Financial Markets*, 16:1–32, 2013.
- [216] M. O’Hara. Presidential Address: Liquidity and Price Discovery. *The Journal of Finance*, 58:1335–1354, 2003.
- [217] J. Okunev and D. White. Do Momentum - Based Strategies Still Work in Foreign Currency Markets? *The Journal of Financial and Quantitative Analysis*, 38:425–447, 2003.
- [218] J. K. Ord, A. B. Koehler, and R. D. Snyder. Estimation and prediction for a class of dynamic nonlinear statistical models. *The Journal of the American Statistical Association*, 92(440):1621–1629, 1997.
- [219] D.A. Pachamanova and F.J. Fabozzi. Recent Trends in Equity Portfolio Construction Analytics. *The Journal of Portfolio Management*, 40:137–151, 2014.
- [220] C. Parlour and D. Seppi. Liquidity - Based Competition for Order Flow. *Review of Financial Studies*, 16:301–343, 2003.
- [221] C.A. Parlour and D.J. Seppi. *Limit Order Markets – A Survey*, In *Handbook of Financial Intermediation and Banking*, Edited by A. Thakor and A. Boot. Elsevier, Amsterdam, 2008.
- [222] L. Pastor and R.F. Stambaugh. The Equity Premium and Structural Breaks. *The Journal of Finance*, 56:1207–1239, 2001.
- [223] A.J. Patton and A. Timmermann. Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and the portfolio sorts. *The Journal of Financial Economics*, 98:605–625, 2010.
- [224] R.L. Peterson. *The Power of Minds over Markets*. Wiley Finance, 2016.
- [225] K.C. Pillai and A.K. Gupta. On the exact distribution of Wilks’ criterion. *Biometrika*, 56:109–118, 1969.
- [226] C.R. Rao. *Linear Statistical inference and its applications*. Wiley-Interscience, New York, 2 edition, 1973.
- [227] M. J. Ready. Profits from Technical Trading Rules. *Financial Management*, Autumn:43–61, 2002.
- [228] G. C. Reinsel and S. K. Ahn. Vector autoregressive models with unit roots and reduced rank structure: estimation, likelihood ratio test, and forecasting. *The Journal of Time Series Analysis*, 13(4):353–375, 1992.
- [229] G.C. Reinsel. *Element of Multivariate Time Series Analysis, Second Edition*. Springer-Verlag, New York, 2002.

- [230] G.C. Reinsel and R. Velu. *Multivariate Reduced-Rank Regression, Theory and Application*. Springer-Verlag, New York, 1998.
- [231] J. P. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73:1237–1282, 2005.
- [232] S.A. Ross. The arbitrage theory of capital asset pricing. *The Journal of Economic Theory*, 13:341–360, 1976.
- [233] T. H. Rydberg and N. Shephard. Dynamic Trade - by - Trade Price Movement: Decomposition and Models. *The Journal of Financial Econometrics*, 1:2–25, 2003.
- [234] S. Satchell and A. Snowcraft. A demystification of the Black-Litterman model: Managing quantitative and traditional portfolio construction. *The Journal of Asset Management*, 1:138–150, 2000.
- [235] V. Satish, A. Saxena, and M. Palmer. Predicting intraday trading volume and volume percentages. *The Journal of Trading*, 9:15–25, 2014.
- [236] M. Schatzoff. Exact distribution of Wilks' likelihood ratio criterion. *Biometrika*, 53:347–358, 1966.
- [237] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:401–404, 1978.
- [238] J. T. Scruggs. Noise Trader Risk: Evidence from the Siamese Twins. *The Journal of Financial Markets*, 10:76–105, 2007.
- [239] D. Seppi. Liquidity Provision with Limit Orders and a Strategic Specialist. *Review of Financial Studies*, 10:103–150, 1997.
- [240] W.F. Sharpe. Capital asset prices: a theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19:425–442, 1964.
- [241] R. H. Shumway and D. S. Stoffer. Arima models. In *Time Series Analysis and Its Applications*, pages 83–171. Springer, 2011.
- [242] E. Smith, D.J. Farmer, L. Gillemot, and S. Krishnamunthy. Statistical theory of the continuous double auction. *Quantitative Finance*, 3:481–514, 2003.
- [243] M.S. Srivastava and C.G. Khatri. *An Introduction to Multivariate Statistics*. New York: North Holland, 1979.
- [244] Stambaugh, R., and Yu, J., and Yuan, Y. The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, 104:288–302, 2012.
- [245] M. Statman, S. Thorley, and K. Vorkink. Investor overconfidence and trading volume. *Review of Financial Studies*, 19(4):1531–1565, 2006.

- [246] H.R. Stoll. Electronic Trading in Stock Markets. *The Journal of Economic Perspectives*, 20:153–174, 2006.
- [247] R. Sullivan, A. Timmermann, and H. White. Data snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance*, 54:1647–1691, 1999.
- [248] R. J. Sweeney. Some New Filter Rule Tests: Methods and Results. *The Journal of Financial and Quantitative Analysis*, 20(3):285–300, 1988.
- [249] G. E. Tauchen and M. Pitts. The Price Variability-Volume Relationship on Speculative Markets. *Econometrica*, 51:485–505, 1983.
- [250] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.
- [251] B. Toth, Y. Lemperiere, C. Deremble, J. De Lataillade, J. Kockelkoren, and J-P Bouchaud. Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2):021006, 2011.
- [252] R. Tsay. *Analysis of Financial Time Series, Third Edition*. Wiley, 2010.
- [253] J. Tu and G. Zhou. Markowitz meets Talmud: A Combination of Sophisticated and Naïve Diversification. *The Journal of Financial Economics*, 99:204–215, 2011.
- [254] G.E. Uhlenbeck and L.S. Ornstein. On the Theory of Brownian Motion. *Physical Review*, 36:823–841, 1930.
- [255] F. Vahid and R.F. Engle. Common Trends and common cycles. *The Journal of Applied Econometrics*, 8:341–360, 1993.
- [256] R. Velu, A. Gretchika, M. Benaroch, D. Nehren, and K. Kuber. Market Impact: To Trade Small or to Trade Seldom? Evidence from Algorithmic Execution Data. *Unpublished*, 2015.
- [257] R.P. Velu, D.W. Wichern, and G.C. Reinsel. A note on non-stationarity and canonical analysis of multiple time series models. *The Journal of Time Series Analysis*, 8:479–487, 1987.
- [258] A. A. Weiss. ARMA models with ARCH errors. *The Journal of Time Series Analysis*, 5:129–143, 1984.
- [259] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models, Second Edition*. Springer-Verlag, New York, 1997.
- [260] H. White. A reality check for data snooping. *Econometrica*, 68:1097–1126, 1999.
- [261] H. Wold. *PLS regression*, volume 6. Eds. N.L. Johnson and S. Kotz, 1984.

- [262] H. Working. Note on the correlation of first differences of averages in a random chain. *Econometrica: Journal of the Econometric Society*, pages 916–918, 1960.
- [263] J. W. Yang. Transaction duration and asymmetric price impact of trades - Evidence from Australia. *The Journal of Empirical Finance*, 18:91–102, 2011.
- [264] Yu, J., and Yuan, Y. Investor sentiment and the mean–variance relation. *The Journal of Financial Economics*, 100:367–381, 2011.
- [265] E. Zarinelli, M. Treccani, J.D. Farmer, and F. Lollo. Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate. *Market Microstructure and Liquidity*, 1:1–31, 2015.