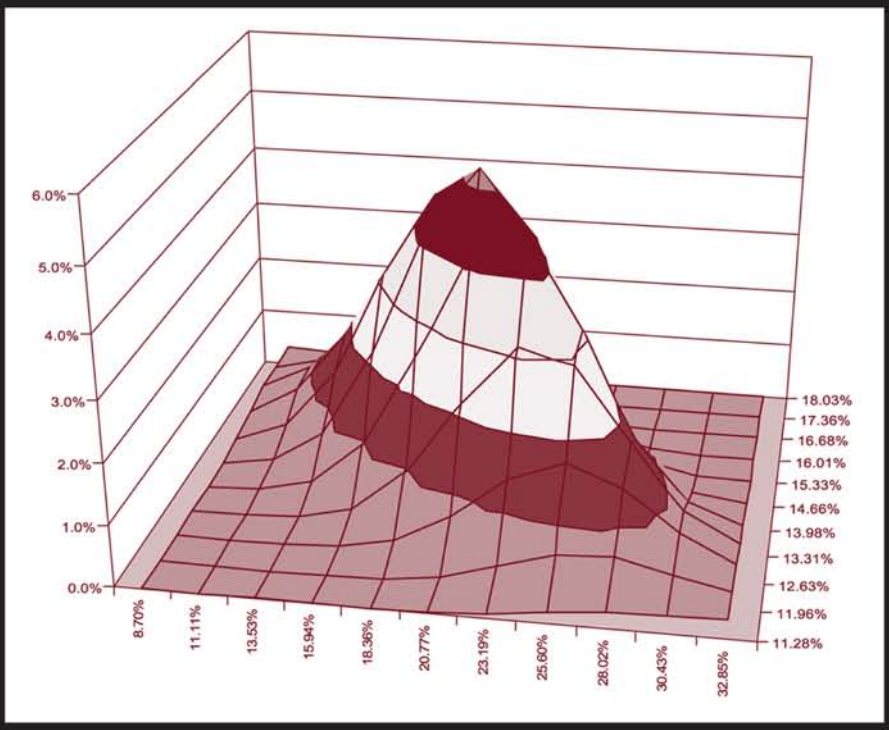
Edward E. Qian, Ronald H. Hua,
and Eric H. Sorensen

Quantitative Equity Portfolio Management

Modern Techniques and Applications



Chapman & Hall/CRC FINANCIAL MATHEMATICS SERIES

CHAPMAN & HALL/CRC FINANCIAL MATHEMATICS SERIES

Quantitative Equity Portfolio Management

Modern Techniques and Applications

CHAPMAN & HALL/CRC
Financial Mathematics Series

Aims and scope:
The field of financial mathematics forms an ever-expanding slice of the financial sector. This series aims to capture new developments and summarize what is known over the whole spectrum of this field. It will include a broad range of textbooks, reference works, and handbooks that are meant to appeal to both academics and practitioners. This series encourages the inclusion of numerical code and concrete real-world examples.

Published Titles

American-Style Derivatives; Valuation and Computation, *Jerome Detemple*

Financial Modelling with Jump Processes, *Rama Cont and Peter Tankov*

An Introduction to Credit Risk Modeling, *Christian Bluhm, Ludger Overbeck, and Christoph Wagner*

Portfolio Optimization and Performance Analysis, *Jean-Luc Prigent*

Quantitative Equity Portfolio Management: Modern Techniques and Applications, *Edward E. Qian, Ronald H. Hua, and Eric H. Sorensen*

Robust Libor Modelling and Pricing of Derivative Products, *John Schoenmakers*

Structured Credit Portfolio Analysis, Baskets & CDOs, *Christian Bluhm and Ludger Overbeck*

Proposals for the series should be submitted to one of the series editors above or directly to:
CRC Press, Taylor and Francis Group
24-25 Blades Court
Deodar Road
London SW15 2NU
UK

CHAPMAN & HALL/CRC FINANCIAL MATHEMATICS SERIES

Quantitative Equity Portfolio Management

Modern Techniques and Applications

Edward E. Qian, Ronald H. Hua, and Eric H. Sorensen

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20110720

International Standard Book Number-13: 978-1-4200-1079-4 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

Over the last 40 years, academic researchers have made major breakthroughs in advancing modern practice in finance. These include portfolio theory, corporate finance, financial engineering of derivative instruments, and many other applications pertaining to financial markets overall. Formal portfolio theory research saw major advances in the context of normative choice modeling, including how to form an optimal portfolio, beginning with Harry Markowitz. Parallel with this, we saw new advances in capital market theory in the context of descriptive equilibrium propositions in terms of the risk/return tradeoff, beginning with Bill Sharpe and the Capital Asset Pricing Model (CAPM). Many related academic developments provided rich portfolio management insight, including Arbitrage Pricing Theory (APT), market efficiency proposition, market anomalies, and behavioral finance.

Against this backdrop, it is therefore not surprising, over the past two decades, that modernizing portfolio management has been the ambition of hundreds of professional investment management practitioners as well as fiduciaries. Driven by market demand and the search of higher returns, a new breed of investment professionals has emerged — quants, i.e., quantitative professions with advanced degrees in science and economic/finance, seeking to exploit market anomalies with increasing success.

As a result, quantitative equity investment strategies have been gaining acceptance and popularity in the investment community. They are deployed in many forms, from enhanced products that aim to beat market indices while limiting the amount of risk, to absolute return strategies (long-short hedge funds) that strive to produce positive return regardless of the overall market condition.

Quantitative equity portfolio management combines theories and advanced techniques from several disciplines, including financial economics, accounting, mathematics, and operational research. Although many books are devoted to these disciplines, few deal with quantitative equity

investing in a systematic and mathematical framework that is suitable for quantitative investment professionals and students with interests in quantitative equity investing.

The motivation for this book is to provide a self-contained overview and detailed mathematical treatment of various topics that serve collectively as the foundation of quantitative equity portfolio management. In many cases, we frame related problems in this field in mathematical terms and solve these problems with mathematical rigor while establishing an analytical framework. We also illustrate the mathematical concepts and solutions with numerical and empirical examples. In the process, we provide a review of quantitative investment strategies or factors accompanied by their academic origins.

This book serves as a guide for practitioners in the field who are frustrated with certain naïve treatments of many common modeling issues and wish to gain in-depth insights from mathematical analysis. We hope that the book will also serve as a text and reference for students in computational and quantitative finance programs interested in quantitative equity investing out of pure curiosity or in search of employment opportunities. As practitioners, we feel strongly that current curriculum of many such programs is often light on portfolio theory and portfolio management, and long on option pricing theory and various microscopic views of market efficiency (or lack thereof).

As practitioners and active researchers in the field, we have selected topics essential to quantitative equity portfolio management, from theoretical foundation to recently developed techniques. Due to our variety of topics, we adopt a flexible style: we employ theoretical, numerical, and empirical approaches, when appropriate, for specific subjects within the book.

Many people have helped us in making this book possible. We are grateful to Joe Joseph of Putnam Investments who is responsible for many ideas developed in Chapter 6. We thank Dan diBartolomeo of Northfield and participants of Northfield research conferences for feedbacks to several research presentations that have made their way into the book. Frank Fabozzi and Gifford Fong also deserve credit in recognizing the value of our research and publishing it in the *Journal of Portfolio Management* and the *Journal of Investment Management*, respectively. We also thank our colleagues at PanAgora and Putnam for helpful comments. Betty Anne Case, Craig Nolder, and Alec Kercheval of Florida State University provided encouragement and academic perspective for our effort. Others who provided feedback to us include Artemiza Woodgate and Fred Copper. Last, but not least, we are very grateful to Jennifer Crotty for editorial assistance. Any errors, however, remain entirely ours.

Abstract

This book provides a self-contained overview, empirical examination, and detailed mathematical treatment of various topics from financial economics/accounting, mathematics, and operational research that serve collectively as the foundation of quantitative equity portfolio management. In the process, we review quantitative investment strategies or factors that are commonly used in practice, including value, momentum, and quality, accompanied by their academic origins. We present advanced techniques and applications in return forecasting models, risk management, portfolio construction, and portfolio implementation. Examples include optimal multifactor models, contextual and nonlinear models, factor timing techniques, portfolio turnover control, Monte Carlo valuation of firm values, and optimal trading.

We frame and solve related problems in mathematical terms and also illustrate the mathematical concepts and solutions with numerical and empirical examples. This book serves as a guide for practitioners in the field who wish to gain in-depth insights from mathematical analysis. We hope that the book will also serve as a text and reference for students in finance/economics, computational, and quantitative finance programs, interested in quantitative equity investing, out of pure curiosity, or in search of employment opportunities.

About the Authors

Edward E. Qian, Ph.D., C.F.A., is a director and head of research of the Macro Strategies Group at PanAgora Asset Management in Boston. Previously, Dr. Qian spent 6 years at Putnam Investments as a senior quantitative analyst. He also worked at Back Bay Advisors as a fixed income quantitative analyst and at 2100 Capital Group as a quantitative analyst of hedge fund strategies. Dr. Qian has published numerous research articles in the areas of asset allocation and quantitative equity. Prior to 1996, Dr. Qian was a postdoctoral researcher in applied mathematics at the University of Leiden in the Netherlands and at the Massachusetts Institute of Technology. He received his Ph.D. in applied mathematics from Florida State University in 1993 and a bachelor's degree in mathematics from Peking University in 1986. He was a National Science Foundation postdoctoral research fellow at MIT from 1994 to 1996.

Ronald Hua, C.F.A., is a director of Equity Group and head of equity research at PanAgora Asset Management in Boston. Prior to joining PanAgora in 2004, he was a senior vice president at Putnam Investments for 5 years and was responsible for quantitative equity research and equity portfolio management. Between 1994 and 1999, he worked at Fidelity Management and Research Company as a quantitative research analyst. He received his M.B.A. degree from Stern School of Business at New York University in 1994 and M.S. degree from Courant Institute of Mathematical Sciences at New York University in 1991.

Eric H. Sorensen, Ph.D., is the president and CEO of PanAgora Asset Management in Boston. PanAgora specializes in quantitative money management solutions implementing modern portfolio strategies totaling approximately $24 billion in institutional assets. He and his team are industry leaders in the field of quantitative investment strategies and focus

on a variety of innovative bottom-up equity as well as multialpha macro strategies. Between 2000 and 2004, Dr. Sorensen was the global head of Quantitative Research and a chief investment officer at Putnam Investments. Prior to joining Putnam Investments, Dr. Sorensen was global head of quantitative research at Salomon Brothers (now Citigroup). During his 14 years on Wall Street, he published extensively, and consulted with institutional investor clients around the world and developed a global reputation as a leader in the field of modern investment strategies. Prior to Wall Street, he was a professor of finance and department head at the University of Arizona. He has published over 50 academic journal articles and has served on the editorial boards of several academic finance journals. His background also includes a tour as a U.S. Air Force officer and jet pilot from 1969–1974. Dr. Sorensen received his Ph.D. in finance from University of Oregon in 1976 and M.B.A. from Oklahoma City University in 1974.

CHAPTER 1

Introduction: Beliefs, Risk, and Process

THIS BOOK IS ABOUT QUANTITATIVE EQUITY INVESTMENT STRATEGIES, focusing on modern techniques and applications. Three fundamental activities form the basis of a modern investment practice: in order to be successful, the investment team must have (1) a strong philosophy based on commitment to a set of beliefs, (2) a clear approach in translating uncertainty into an appropriate risk/return trade-off, and (3) a comprehensive investment process from beginning to end.

1.1 BELIEFS

What do markets give us, and how do we believe we can go after it? This two-part question is essential to a portfolio manager's belief system. In the premodern 1950s world of fundamental stock picking, the analysis focused exclusively on the second part of the question — go for the "best" stocks and enjoy the results. Inherent in this belief is that one has sufficient skill and is significantly blessed above others who compete in the same game. Across a diverse spectrum of stock-picking techniques, there certainly have been (and are) some that win more than others. However, over the years, formal academic research and practitioner experience converge on the conclusion that it is difficult to win consistently if we account for the proper risks. With consideration of the risks, we should think of the game as well worth winning but not necessarily worth playing.

As for the first part of the question, there has been a common evolution of beliefs. What does the opportunity set look like? How do the distributions of relative stock returns behave? Are these return differences exploitable? In the 1960s, there began a tension surrounding the true value of past price and volume information in security returns — "technical analysis." A well-accepted investment approach was to study the pattern of past price returns in order to forecast future returns. As we will see in later chapters, the same underlying price data may be also relevant today, though in the context of a modern, comprehensive process.

As academics began to formally study return distributions, they gravitated to a concept of "random walk." They increasingly came to the conclusion that "price has no memory" (Lorie and Hamilton 1973). If the investor's technique is conditioned on some *ad hoc* price configuration, there will be little value added because a random walk stock will give us no profitable clues about future prices.

It was Fama (1970) who artfully formed and expanded the notion of random walk into what he popularized as the efficient market hypothesis (EMH). In summary, it is hard (if not impossible) to beat the market depending on the investors' information set. Past price data does not cut it. Taken to an extreme, a very strong EMH belief is that all information, both public and private, is not sufficient to beat the market, after consideration of appropriate costs and proper risk specifications.

By the 1970s, variations of efficient markets beliefs were firmly implanted in the brains of many financial economists. In fact, it was quite difficult for a bright assistant professor of finance to publish any empirical findings that disproved the EMH. However, by the early 1980s, the ambitious and persistent academic empiricists found a way — just call it something else! In the 1980s, there came a volume of formal literature that discovered inefficiencies that could lead to abnormal returns if rigorously applied. The list includes size effect, January effect, value irregularities, momentum effect, etc. We called them anomalies[1] and reverently acknowledged in the conclusion that these discoveries (1) were likely not repeatable in the future (now that we know them), (2) may be inconclusive because of potential "risk misspecification," or (3) were lacking the proper allocation of costs in the strategy. In a modern quantitative process we call these anomalies "factors," which are an in-depth topic of later chapters.

What are our beliefs? What are the principles underlying our book? We choose rather safe ones that are explained in many of the subsequent chapters. First, skill and return dispersion are the key drivers of opportunity.

Second, the market is not efficient, which, in many cases, is attributable to investors' irrational behavior described by "behavioral finance." Third, the variables or factors we use to predict return must be grounded in financial theory and reflect logical cause and effect. (Sunspots do not cut it.) Fourth, true alpha-generation is available to practitioners who creatively combine modern tools — econometrics, mathematics, investment theory, financial accounting, psychology, operations research, and computer science. Fifth, objective discipline is essential in the implementation of strategies. This is not to say subjective judgment is lacking in the world of quantitative management — but it lies in perfecting the comprehensive portfolio system, rather than in comprehending the perfect stock selection.

This comprehensive system is the core of quantitative investment process. Active investment is about the processing of information. One must have the best information as well as the best way to process and implement them in a portfolio. With the advent of the information age, advance of financial markets, and increasing computing power, quantitative investment process provides a way of unifying all these together to deliver consistent returns. In a way, this is analogous to combining the best machinery with the best operators. In the late 1960s, there was a common belief in the U.S. Air Force that advances in aeronautical engineering would obviate any role for the human pilot. On the contrary, air superiority today resides with the force that combines the best equipment with the best-trained pilots. The best equipment is not knowable without design inputs from the best pilots.

1.2 RISK

The quantification of uncertainty is also one of the evolutionary breakthroughs in the theory of investment during the last century. Frank Knight (1921) laid the groundwork with a quite intuitive definitional distinction between uncertainty and risk: (1) decision makers crudely operate in a world of random uncertainty, and (2) risk is a condition in which the decision maker assigns formal mathematical probabilities to specify the uncertainty. Later, Von Neumann and Morgenstern (1944) formalized the specification of risk into microeconomic theory, laying a foundation for rational decision making under uncertainty with the concept of expected utility.[2]

It was Markowitz (1952) who inaugurated the vast body of literature we know as modern portfolio theory (MPT). Markowitz combined the notion that when a rational investor is faced with a set of security choices that follow a normal distribution, he or she will seek to maximize expected utility by formally trading off expected return with risk measured by variance.

In a world characterized by diminishing marginal utility for wealth, the optimal portfolio is specified and the security weights are solved using the mean and variance of the portfolio return distribution (see Chapter 2 for a complete treatment).

Bill Sharpe's article in 1964 took the normative mean–variance portfolio concept to the next level by developing an equilibrium pricing model to describe the first formal capital market pricing of risk framework — the capital asset pricing model (CAPM).[3] For this, he later received the Nobel Prize, as did Harry Markowitz. Assuming frictionless markets and homogeneous expectations of investors, the pricing relationship is depicted in terms of expected returns. The expected return of a security (or a portfolio) consists of two parts: (1) market price of time — the risk-free rate and (2) market price of risk — beta times the market excess return.

For investors, CAPM concludes that the market provides a fair risk premium — take systematic or market (beta) risk and be rewarded. As such, prudent investments should be combinations of two passively managed portfolios — the market portfolio and the risk-free portfolio; the precise combination is governed by the risk tolerance of a particular investor.

In theoretical equilibrium, beta is the elasticity of the portfolio return with the market and presents a linear trade-off between risk and return in the long run, i.e., capital market line (CML). However, can't we do better in practice? Isn't what this book and myriads of writings before are about? How can we generate alpha — the return above the CML that is in excess of the risk? It takes positive skill!

1.2.1 Beta, Benchmarks, and Risk

Risk-adjusted positive skill is the true goal of the game. The development of risk and capital market theory from the 1950s, and for 30 years thereafter, ushered in a host of phenomena and participants to the game. Three stand out. First, beginning in the 1980s, the attraction of indexing to a benchmark — index such as the S&P 500 — exploded. Entrepreneurs at Wells Fargo (BGI today), Mellon, and later, Vanguard and State Street, offered passive zero alpha index funds with an efficient beta of 1 and low fees. It was as if the new risk tools combined with the now acceptable belief in market efficiency to produce a powerful antidote to those that had been stung by underdelivered promises of traditional active return managers.

Second, a new player category entered the fray in the 1980s. Managers who promised active strategies (positive alpha) found themselves increasingly exposed to benchmark comparisons by a new labor force

— the influential pension plan consultants. Within the consulting firms emerged armies of analysts equipped with MPT devices to conduct manager research, evaluating them against designated benchmarks (growth/value, large/small, domestic/international, developed/emerging, etc.). Their objective was to provide service to institutional investors and the ability to "separate alpha from beta" by performing scientific attribution of active managers, as well as to pronounce an active strategy dead or alive. The game was still worth "winning" but now had more talented officials evaluating the "playing."

Third, enter hedge fund managers who got away with no benchmarks. Hedge fund is not a new phenomenon — combining subjective long and short positions (asset classes of securities) goes back to the 1960s. For example, equity hedge funds are long-short — buy securities as well as sell borrowed ones — but they are not necessarily market beta neutral. It is often hard, if not impossible, to disentangle what is alpha and what is beta. For a long time, nobody cared because most of the investors in the hedge funds were high-net-worth individuals who had their eyes on the absolute returns, not abstract geeks. Today, the situation has changed dramatically. Equity market neutral managers (mostly quants) manage zero-beta funds with refined risk management systems, and often deliver pure alpha. Institutional investors are increasingly pursuing and paying handsomly for alpha, but are unwilling to pay excessively for beta management. Hence, we have the rise of market-neutral hedge funds with a new benchmark — cash.

1.3 QUANTITATIVE INVESTMENT PROCESS

What steps characterize a quantitative investment process? What are the instruments in the toolbox of quantitative investment professionals? There are at least five essential components.

Alpha model: First and foremost is an alpha model that forecasts excess return of stocks. If return distribution is characterized by the expected return and the standard deviation, it is often the expected return that determines whether we buy or sell, overweight or underweight, and the standard deviation that determines the size of the portfolio allocations. It is easier to find random factors that represent non-compensated market risks than to find alpha factors that represent incremental rewards. The alpha model is often proprietary and highly guarded, reflecting creativity as well as superior systems. It is the most important differentiator within the investment firm.

Risk models: Good quantitative investment processes require sophisticated risk tools that embody many "drivers' of risk beyond the one-factor CAPM — plain vanilla beta. Today, commercial risk models such as BARRA serve to isolate and control stock specific factors that measure unwanted risk, such as size, value and the like. However, some BARRA factors, first estimated in the mid-1980's, overlap with potential stock-specific alpha factors. Ross and Roll (1976, 1977) introduced the arbitrage pricing model (APT), and estimated it with a set of four purely macroeconomic time-series factors, such as the cycle of long-term interest rates. Later others developed more complete specifications of macro models using such phenomenon as economic growth, term structure of rates, inflation, oil and so on. Salomon Brothers quantitative team first estimated a set of macroeconomic risk systems for local and global equity markets in the late 1980's Similarly, the Northfield Company delivered a portfolio optimization package using a macro risk model in the 1990's.

Portfolio optimization: The normative machinery that calculates the tradeoff between alpha factors (wanted risk) with risk factors (unwanted risk) formally is the optimization tool. Effectively, portfolio optimization formally combines both proprietary alpha with exogenous risk to create the ex ante optimum set of portfolio weights, subject to the risk appetite of the manager. Managers can optimize active portfolios versus a benchmark such as S&P 500 index, or against cash for market-neutral long/short portfolios. These tools allow managers to dissect the ex ante risks, and place their exposures with their alphas. However, there is a tendency to be overconfident in risk model outputs. As we will see later, there is alpha model risk also, and it must be modeled to achieve the best portfolio results.

Portfolio implementation: Risks and alphas change. The complete process requires trading — turnover. Relatively high-turnover active portfolios demand close attention to transaction costs. Since the 1970's, market maker competition and computer networking technology influenced and drove down the costs of trading — both commissions as well as market pricing impact proportional to volume. Nevertheless, trading costs are positive and less subject to randomness than are security prices (and alphas). The modern implementation process, therefore, includes a risk/return framework to address the portfolio implementation. Asset management

firms and brokerage firms are increasingly relying on proprietary or commercial models to implement trades with the goal of minimizing implementation shortfall under uncertainty.

Performance attribution: Well, in the end does this all work? If so, how much is *working* and how much is *random*? Modern managers perform attributions regularly to ascribe ex post returns to ex ante factor exposures. It is increasingly imperative for active managers to identify their skill vis-a-vis ex ante alpha efficacy, and to attribute ex post results to maintaining exposure of these alpha sources. Here quantitative mangers possess a clear advantage over pure fundamental managers.

Successful investment firms would find a way to integrate these five components together and constantly search for improvements in all of them to stay ahead of the market and the competitors.

1.3.1 Quantitative vs. Fundamental

It is inaccurate to say that fundamental managers dig deep at the solo stock level, but have no models or disciplines. It is also unfair to say that quantitative managers apply skills to so broad a set of stocks that the process is superficial at the fundamental level, and often labeled black-box, data-mining nerds. This is a misrepresentation. Many quantitative investment strategies rely on factors that are based on not only solid economic principles, but also on sound fundamental intuition (more on this in Chapters 5 and 6). At the same time, fundamental managers all use models. These may be rules-of-thumb or heuristics, and not subject to rigorous testing, but the deep implementation of the *model* into the security makes up for the lack of breadth. To repeat, quantitative management — lies in broadly perfecting the comprehensive portfolio system, whereas, fundamental management lies in deeply comprehending the perfect stock selection.

In many instances, the underlying principles of quantitative investment are no different from traditional fundamental research. At a basic level, all investment strategies seek to buy low and sell high — requiring a measured valuation methodology. John Burr Williams [1938] developed the first modern expression for the fundamental valuation of intrinsic value — that a company's stock should achieve a market price that quantifies the present value of all future potentially profitable operations of the firm that accrue to shareholders. This is the forerunner of the now common dividend discount model (DDM) and a variety of related cash

flow valuation expressions. This valuation framework is indispensable to fundamental analysis. Who can say it is not quantitative analysis — do we value bonds, even those with embedded options, similarly?

Notably, Benjamin Graham (1934, 1949) laid the foundation of fundamental investing, which deemphasizes movements of market prices and focus on a firm's intrinsic value and fundamental analysis. Warren Buffet is perhaps the best-known disciple of Graham and offers at least an implicit process firmly founded on the original valuation principals. Can quantitative investing have a much closer affinity and be kindred spirit to the Ben Graham principles? We provide some answers to this question in the book.

Perhaps, some of the misperception about quantitative investing is self-inflicted. After all, we are quants — as some would assume all it takes is a brainy nerd and a fast computer, right? Many become easily get excited about mean–variance optimization and Monte Carlo simulation but are bored with balance sheet and cash-flow analysis. This is the wrong attitude, perhaps. Some of the most valuable information, quantitative or fundamental, is only garnered through painstaking analysis of financial statements. We hope readers would agree with this after reading the book.

1.4 INFORMATION CAPTURE

Investing without true information is just speculation. How do we know we have true information that can predict security returns? On one level, predicting a market crash is not enough, even if you are correct once. In the same vein, neither is finding the correct target prices for a couple of stocks a proof of skill. The key to investment success is consistency in forecasting (skill) applied repeatedly (breadth).

We have Grinold and Kahn (2000) to thank for introducing the fundamental law of active management (FLAM). It has become an important framework for evaluating skills in active management. In their framework, the skill is measured by the information coefficient (IC) — the cross-sectional correlation coefficient between forecasts and subsequent returns. Consistency is measured by the information ratio (IR) — the ratio of average excess return to the standard deviation of excess return. Under a host of assumptions, FLAM combines skill and opportunity set together into a convenient expression for IR:

$$\mathrm{IR} = \mathrm{IC}\sqrt{N}$$

where N is the number of independent securities.

Although FLAM represents a milestone in active portfolio management theory, important practical extensions have gone in two directions. First, we can reexamine FLAM and modify for portfolios with real world constraints. For instance, Grinold and Kahn (2000) compare the IR of long-only portfolios with long-short portfolios. Clarke et al. (2002) generalize FLAM introducing the concept of transfer coefficient to approximate the loss of information due to constraints. These studies highlight the dampening effect of overly stringent constraints on investment performance. This awareness across the investment community has created increased receptivity to long-short portfolios, either "pure" or constrained, in the search of more consistent alpha (see Chapter 11).

The second extension, more subtle but arguably more significant, is a multiperiod version of IR. Unknown to many, FLAM is a result for a single period — the expected excess return to the targeted tracking error. Qian and Hua (2004) first pointed out that, in a multiperiod framework, the standard deviation of IC plays an important role in determining the *ex post* tracking error, which is not necessarily the same as the *ex ante* tracking error. This insight is further extended in Sorensen et al. (2004), using an alternative expression for IR to combine multiple alpha factors with optimal factor weights that achieves maximum IR (Chapter 4 and Chapter 7).

Multiperiod portfolio management is dynamic in nature. This dynamic link is amplified by portfolio turnover constraints (Sneddon 2005; Grinold 2006). The turnover constraint, while controlling transaction costs, inhibits information transfer to the portfolio. However, its impact varies across alpha factors with differing information horizon (Chapter 8 and Chapter 12). Such recent research raises the awareness of important normative implications of the fundamental law and proposed various methods to modify it for practical use.

Quality information is the most precious substance in the investment business. Simple yet naïve models that are unconditional and one-size-fits-all do not capture all the information available. These simple models fall short in two ways. First, stocks are idiosyncratic in nature. A one-size-fits-all model assumes that all stocks respond to the factor exposure in the same way all the time. Practitioners know this is not true, and are beginning to analyze factor significance within this context. How do we systemize this approach? Second, the market is inherently dynamic due to influences from macroeconomic factors and the changing behavior of players — firms, investors, etc. As a result, the efficacy of alpha factors does

not necessarily remain stable as the market environment changes. There is a growing list of academic literatures covering conditional CAPM. For practical purposes, how do we build a forecasting model that is adaptive to allow its factor combination to change over time? We cover this topic in the book.

Much of this book goes deep into the elements of FLAM. Our purpose is to enrich this framework to highlight key elements of a modern process. It will be apparent that our approach is part art, part science, part quantitative, and part fundamental. These steps may not be the ultimate way to capture all the information, but they represent considerable improvement in our journey to build the perfect comprehensive portfolio system.

1.4.1 Alpha

True risk-adjusted alpha has always been scarce. Some refer to the search for alpha as a zero-sum game. To win the game — using a baseball analogy — a team must play well by having a high batting average, similar to a high average IC. Skill combined with many times at bat is tantamount to a high average IC. Great batters can't win if the game is rained out. Poor batters can't win no matter how many times they get to the plate. To win more games than its opponents, a team must play consistently throughout the year by not having prolonged slumps, analogous to a low standard deviation of IC. In order to do this, the players must complement each other: when some are not playing well, others are there to pick up the slack, similar to a diversifying set of alpha factors. To win a division title, a team must play a lot of games, and players' time at the plate is high. The best team is expected to always win the division, but the play-off could be a toss-up in a seven-game series.

Alpha can also be allusive, and today's alpha could be gone tomorrow or reclassified as beta in the future. However, one thing is constant: investors such as institutional fiduciaries, pension funds, endowments, and the like, will continue to pursue risk-adjusted alpha through active equity management. It might be that the latest surge of formal quantitative investing has, in part, ushered in better metrics for "separating alpha from beta" and therefore led to a higher level of general understanding of the difference. It is our hope that this book can contribute to that pursuit by presenting investors and researchers the best practice of quantitative equity investing and what it takes to be successful in the search for alpha.

1.5 THE CHAPTERS

The rest of the book consists of 3 parts with 11 chapters. Part I lays the basics of MPT framework. We present the modern portfolio theory from Markowitz through the CAPM and introduce some applications in Chapter 2. In Chapter 3, we develop modern risk models to include APT, fundamental factor models, and macroeconomic risk models, with emphasis on how these are used in quantitative portfolio management.

In Part II, we have 4 chapters devoted to the development and implementation of quantitative factors that form the bases for security selection. Chapter 4 introduces the typical objective functions of IR and Sharpe ratio, with a focus on cross-sectional estimation of the predictive power of factors, represented by average information coefficient, and the inherent risks of alpha strategies, represented by the standard deviation of IC. Chapter 5 focuses on the broad set of factors that academics and practitioners have researched over the last decade. We outline their economic and behavior intuition and analyze their efficacy through the framework developed in Chapter 4. Chapter 6 devotes attention to firm valuation based on the discount cash flow method. It extends the one-path-one-value approach to a multipath approach, which gives rise to measures of confidence around the fair-value estimation. Lastly, Chapter 7 presents mathematical frameworks for constructing multifactor models, with a focus on exploiting the diversification benefit among factors and maximizing information ratio.

Part III, the final section, puts it all together with a series of advanced implementation issues. These include Chapter 8, portfolio turnover and alpha integration; Chapter 9, advanced alpha modeling techniques to account for security context and nonlinear patterns; Chapter 10, dynamic factor timing; Chapter 11, dealing with real-world portfolio constraints optimally; and lastly, Chapter 12, incorporating transactions costs in the comprehensive optimal strategy.

Although we have tried to blend theoretical analyses and empirical examinations throughout the book, each chapter tends to have either a theoretical or empirical focus. Chapters with more analytical focus are 2, 3, 4, 7, 8, 11, and 12. Chapters with more empirical emphasis are 5, 6, 9, and 10.

APPENDIX: PSYCHOLOGY AND BEHAVIOR FINANCE

The literature on behavior finance has exploded in recent years, much of it goes beyond the scope of the book. However, it is important for readers

to have some basic understanding of its tenets, which will provide some insight into materials in the later chapters.

A1.1 ADVANCES IN PSYCHOLOGY

In the 1960s, cognitive psychology began to describe the brain as an information processing device, as opposed to a stimulus–response machine. Psychologists such as Ward Edwards, Duncan Luce, Amos Tversky, and Daniel Kahneman began to explore cognitive models of decision-making under uncertainty and to benchmark their models against neoclassical economic models of rational behavior. Their works had far-reaching impact on finance as well as many other fields, such as economics, political science, and consumer behavior. Kahneman and Tversky (1979) wrote the seminal paper, "Prospect theory: Decision making under risk," which detailed an alternative model of choice under uncertainty — prospect theory — in contrast to the expected utility theory from Von Neumann and Morgenstern (1944). Prospect theory provided explanations for a number of documented anomalies beyond the capabilities of the expected utility theory. They also articulated the difference between a normative model, such as the expected utility theory, and a descriptive model such as their prospect theory. Kahneman and Tversky (1984) noted, "The normative analysis is concerned with the nature of rationality and the logic of decision making. The descriptive analysis, in contrast, is concerned with people's beliefs and preferences as they are, not as they should be." Their later work regarded the framing of decisions. Kahneman and Tversky (1986) articulated four normative rules underlying the expected utility theory: cancellation, transitivity, dominance, and invariance. They noted, "Because these rules are normatively essential but descriptively invalid, no theory of choice can be both normatively adequate and descriptively accurate."

A1.2 BEHAVIORAL FINANCE

Behavioral finance flourished in the 1990s. Its research integrates insights from psychology with neoclassical economic theory, with a foundation rooted in alternative views that question the assumption of rational agents (homo-economicus) and the notion of riskless arbitrage. Historically, fundamental equity investing came into vogue in the last half century. Demand for fundamental research attracted interests in three research areas within the accounting discipline, including fundamental analysis, accounting-based valuation, and value relevance of financial reporting.

After years of unsatisfactory efforts to explain market anomalies by efficient market theorists, behavioral economists took an alternative approach to challenge two key tenets of equilibrium pricing models: (1) arbitrage activity eliminates pricing discrepancies completely and (2) investors behave rationally. A series of papers, known as "Limits to Arbitrage," showed that irrationality can have a substantial and long-lived impact on prices, and they provided a differing view from Friedman's (1953) classical arbitrage argument. In essence, this literature argued that the arbitrage strategy designed to correct mispricing can be both risky and costly, rendering it unattractive. On an intuitive level, risk simply comes from the imperfection of the substitution, thus exposing the arbitrageur to fundamental risk. On a more sophisticated level, the arbitrageur also faces the noise trader risk. Shleifer (2000) argued that irrationality is to some extent unpredictable, and it is plausible for today's mispricing to become even more extreme tomorrow. In other words, convergence of price dislocation is not a certainty. Hirshleifer (2001) argued that pricing equilibrium reflects the beliefs of both rational and irrational traders. Because each group has a risk-bearing capacity, both influence security prices. The years of 1999 and 2000 are salient reminders, as many value shops went out of business when the market became more and more irrational. Experimental psychology documented a long list of behavioral biases of investors when making decisions under risk. Hirshleifer (2001) argued that heuristic simplification, self-deception, and emotional loss of control provide a unified explanation for most biases.

Heuristic simplification: Kahneman and Riepe (1998) dubbed heuristic simplification as biases of preference. The premise of this bias lies in the fact that humans have limited time, attention, memory, and processing capacity in tackling information and making decisions. As such, problem solving is simplified to a rules-of-thumb or heuristic approach. Commonly cited behavioral anomalies include narrow framing, mental accounting, loss aversion, and representativeness heuristic.

Self-deception: Kahneman and Riepe (1998) referred to it as biases of judgment. Overconfidence, optimism, and biased self-attribution are the three major cognitive illusions, wherein perceptions deviate, sometimes significantly, from reality. Overconfidence relates to the observation that humans are poor judges of probability and that their predictions tend to fail more often than they expect.

Optimism means that people display unrealistically rosy views of their own abilities and underestimate the likelihood of bad outcomes over which they have no control. Biased self-attribution is that phenomenon in which people attribute success to skill and failure to bad luck. Kahneman and Riepe (1998) noted, "The combination of overconfidence and optimism is a potent brew, which causes people to overestimate their knowledge, underestimate risks, and exaggerate their ability to control events."

Emotions and self-control: Hirshleifer (2001) posited that emotion could overpower reason. For example, people who are in good moods are more optimistic in their choices.

A1.3 BEHAVIORAL MODELS

Three behavioral models, shown in Table 1.1, provide an integrated explanation of several cross-sectional pricing anomalies, including short-term price momentum (Jegadeesh 1993), long-term reversal of price momentum (DeBondt and Thaler 1985), excess volatility (Shiller 1981), earnings announcement drift (Ball and Brown 1968), earnings revision (Givoly and Lakonishok 1979), analyst recommendations (Womack 1996), and the value premium.

1. Daniel, Hirshleifer, and Subrahmanyam (DHS) (1998) assume that investors are overconfident about their private information, and their overconfidence increases gradually with the arrival of public information with biased self-attribution. The pattern of increased confidence leads to a prediction of the return pattern, manifested in short-run positive autocorrelation and long-run negative autocorrelation. Specifically, overconfidence induces overreaction, which pushes prices beyond the underlying fundamentals when information is positive, and below the fundamentals when negative. Such over- or underpricing is eventually eliminated as price reverts back to fundamental, thus resulting in long-term return reversal. Short-term return continuation is traced to the progressive nature of the increased overconfidence, largely due to biased self-attribution. As an investor becomes more and more overconfident, he pushes the stock price further and further away from its fair value, thus giving rise to short-term momentum continuation.

TABLE 1.1 Summary of Behavioral Models

Models	Departure from EMH Assumptions	Short-Term Momentum Continuation	Long-Run Momentum Reversal	Representative Agents
HS	1. Investors are boundely rational with limited computational capacity 2. Information diffuses slowly across the population	Underreaction	Overreaction	1. News-watchers 2. Momentum traders
DHS	1. Informed investors are overconfident about their private information 2. Their overconfidence increase progressively due to biased self-attribution	Overreaction	More overreaction	1. The informed and the risk-neutral price setter 2. The uninformed and the risk-averse price taker
BSV	Investors exhibit two biases in updating their prior beliefs: conservatism and representativeness	Underreaction	Overreaction	A risk-averse investor who shifts his or her belief between two regimes: trending or reverting

2. Hong and Stein (HS) (1999) make two assumptions: (1) investors are bounded rational, meaning that they have limited intellectual capacity and that they are rational in processing only a small subset of the available information; and (2) information diffuses slowly across the population. They specify two bounded rational agents — news-watchers and momentum traders. Both are risk-averse, and their interactions set security prices. On the one hand, news-watchers exhibit similar behavior to a typical fundamental manager in practice, observe some private information, and ignore information in past and current prices. On the other hand, momentum traders condition their forecasts *only* on past price changes, and their forecast method is simple. The slow diffusion of information among news-watchers induces underreactions in the short-horizon. Underreaction leads to

positively autocorrelated returns — momentum continuation. Upon observing this predictable return pattern, momentum traders condition their forecast only on past price changes and arbitrage the profit opportunity. Arbitrage activity eventually leads to overreaction in the long-horizon, creating dislocation between price and fundamentals. The reversion of price back to fundamental is the source of long-term momentum reversal.

3. Barberis, Shleifer, and Vishny (BVS) (1998) suggest that investors exhibit two biases in updating their prior beliefs with public information: conservatism and representativeness. Conservatism (Edwards 1968) states that investors are slow to change their beliefs in the face of new evidence; representativeness heuristic (Tevrsky and Kahneman 1974) involves assessing the probability of an event by finding a "similar known" event and assuming that the probabilities will be similar, i.e., "if it walks like a duck and quacks like a duck, it must be a duck." Conservatism underweights new information and causes underreaction. For example, after a positive earnings surprise, conservatism means that the investor reacts insufficiently, creating a positive postannouncement drift. In contrast, after a series of positive surprises, representativeness causes people to extrapolate and overreact, pushing price beyond the fundamental value. This eventually results in long-term momentum reversal.

REFERENCES

Ball, R. and Brown, P., An empirical evaluation of accounting income numbers, *Journal of Accounting Research,* 159–178, 1968.

Banz, R., The relationship between return and market value of common stock, *Journal of Financial Economics*, Vol. 9, 3–18, 1981.

Barberis, N., Schleifer, A., and Vishny, R., A model of investor sentiment, *Journal of Financial Economics*, Vol. 49, No. 3, 307–343, September 1998.

Clarke, R., de Silva, H., and Thorley, S., Portfolio constraints and the fundamental law of active management, *Financial Analyst Journal,* Vol. 58, No. 5, 48–66, September–October 2002.

DeBondt, W.F.M. and Thaler, R., Does the stock market overact? *Journal of Finance*, Vol. 40, 739–805, 1985.

Edwards, W., Conservatism in human information processing, *Formal Representation of Human Judgment*, John Wiley and Sons, In Kleinmutz, B. (Ed.), New York, 1968, pp. 17–52.

Fama, E.F., Efficient capital markets: A review of theory and empirical work, *Journal of Finance*, Vol. 25, 383–417, 1970.

Friedman, M., The case for flexible exchange rates, *Essays in Positive Economics*, University of Chicago Press, 1953, pp. 157–203.

Givoly, D. and Lakonishok, J., Financial analysts' forecasts of earnings: Their value to investors, *Journal of Banking and Finance*, 1980.

Givoly, D. and Lakonishok, J., The information content of financial analysts' forecasts of earnings: Some evidence on semi-strong inefficiency, *Journal of Accounting and Economics,* 1979.

Graham, B., *The Intelligent Investor*, Harper Business, 1949.

Graham, B. and Dodd, D., *Security Analysis*, McGraw-Hill, 1934.

Griffin, D. and Tversky, A., The weighting of evidence and the determinants of confidence, *Cognitive Psychology*, 411–435, 1992.

Grinold, R., A dynamic model of portfolio management, *Journal of Investment Management*, Vol. 4, No. 2, 2006.

Grinold, R.C. and Kahn, R.N., *Active Portfolio Management*, McGraw-Hill, New York, 2000.

Hirshleifer, D., Investor psychology and asset pricing, *Journal of Finance*, Vol. 56, 1533–1597, 2001.

Jegadeesh, N. and Titman, S., Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance*, Vol. 48, 65–91, 1993.

Kahneman D. and Riepe M.W., Aspects of investor psychology, *Journal of Portfolio Management,* Summer 1998.

Kahneman, D. and Tversky, A., Prospect Theory: An Analysis of Decision under Risk, *Econometrica,* XVLII, 263–291, 1979.

Kahneman, D. and Tversky A., Choices, Values, and Frames, *America Psychologist,* Vol. 39, No. 4, 341–50, 1984.

Keim, D.B., Size-related anomalies and stock return seasonality: Further empirical evidence, *Journal of Financial Economics*, 13–32, 1983.

Knight, F.H., *Risk, Uncertainty, and Profit*. Hart, Schaffner, and Marx Prize Essays, No. 31, Houghton Mifflin, Boston and New York, 1921.

Lo, A.W. and Mackinlay, A.C., Data-snooping biases in tests of financial asset pricing models, *Review of Financial Studies*, Vol. 3, 431–467, 1990.

Lorie, J.H. and Hamilton, M.T., *The Stock Market: Theories and Evidence*, Richard D. Irwin, Homewood, IL, 1973.

Markowitz, H.M., Portfolio selection, *Journal of Finance*, Vol. 7, 77–91, 1952.

Modigliani, F. and Miller, M., The cost of capital, corporation finance and the theory of investment, *American Economic Review*, June 1958.

Qian, E.E. and Hua, R., Active risk and information ratio, *Journal of Investment Management*, Vol. 2, Third Quarter, 2004.

Reinganum, M.R., Misspecification of capital asset pricing: empirical anomalies based on earnings' yields and market values, *Journal of Financial Economics*, 19–46, 1981.

Rosenberg, B., Reid, K., and Lanstein, R., Persuasive evidence of market inefficiency, *Journal of Portfolio Management*, Vol. 11, No. 3, 9, Spring 1985.

Ross, S.A. and Roll, R., The arbitrage theory of capital asset pricing, *Journal of Economic Theory*, Vol. 13, 341–360, 1976.

Sharpe, W.F., Capital assets prices: A theory of market equilibrium under conditions of risk, *Journal of Finance*, Vol. 10, 425–442, 1964.

Shiller, R.J., Do stock prices move too much to be justified by subsequent changes in dividends?, *The American Economic Review*, Vol. 71, No. 3, 421–436, June 1981.

Shleifer, A., *Inefficient Markets: An Introduction to Behavioral Finance*, Oxford University Press, 2000.

Sneddon, L., The dynamics of active portfolios, *Proceeding of Northfield Research Conference*, 2005.

Sorensen, E.H., Qian, E.E., Hua, R., and Schoen, R., Multiple alpha sources and active management, *Journal of Portfolio Management*, Vol. 31, No. 2, 39–45, Winter 2004.

Tobin, J., Liquidity preference as behavior towards risk, *The Review of Economic Studies*, Vol. 25, 65–86, 1958.

Tversky, A. and Kahneman, D., Judgment under uncertainty: Heuristics and biases, *Science*, 1974, pp.1124–1131.

Von Neumann, J. and Morgenstern, O., *Theory of Games and Economic Behavior*, Princeton University Press, 1944.

Williams, J.B., *The Theory of Investment Value*, Harvard University Press, 1938.

Womack, K.L., Do brokerage analysts' recommendations have investment value? *Journal of Finance*, Vol. 51, No. 1, 137–167, March 1996.

ENDNOTES

1. Anomalies: Pricing anomalies began to appear in the literature in the 1980s. An early example is firm size. Banz (1981) and Reinganum (1981) concluded that small capitalization stocks earned higher average return than the CAPM might predict. Keim (1983) showed that much of the abnormal return to small stocks occurs in January (the "January Effect"). Similarly, the abnormal returns to cheap (value) stocks also received significant attention, starting with Basu (1983), who documented that high-earnings-yield (E/P) firms delivered positive abnormal returns. Rosenberg (1985) further showed that stocks with high book-to-market ratios outperform others as a group. In the realm of technical analysis, new momentum strategies emerged. DeBondt and Thaler (1985) identified long-term reversals of returns to both winner and loser portfolios. Jegadeesh and Titman (1993) further documented a short-term reversal (1st month after portfolio formation) and an intermediate-term momentum continuation (2nd to 12th month after portfolio formation). Ball and Brown (1968) were the first to document the postearnings-announcement drift, in which the market appears to underreact to earnings news. Givoly and Lakonishok (1979) concluded that market reaction to analysts' earnings revisions was relatively slow.
2. This work ushered in a series of other important pieces: Arrow and Debreu (1954), Savage (1954), and Samuelson (1969).

3. Academic literature also examines the effect of relaxing the assumptions of the CAPM: (1) different riskless lending and borrowing rates, (2) the inclusion of personal taxes, (3) existence of nonmarketable assets such as human capital, and (4) heterogeneity of expectations. These research projects typically examine CAPM's assumptions one at a time. The intertemporal CAPM (ICAPM) was devised to extend CAPM into multiperiod to discover other sources of risk that may be priced in the equilibrium. They included aggregate consumption growth (Breeden 1979), inflation risk (Friend 1976), or other sources of risk concerning investors in general (Merton 1971, 1973) beyond the movement of the market portfolio, such as default risk or term structure risk that are generally related to business cycles.

Part I

CHAPTER 2

Portfolio Theory

THE TRADITIONAL OBJECTIVE OF ACTIVE PORTFOLIO MANAGEMENT is to consistently deliver excess return against a benchmark index with a given amount of risk. The benchmark in question could be one of the traditional market indices, such as the Standard & Poor's (S&P) 500 Index and the Russell 2000 Index, or a cash return, such as Treasury bill rate, or LIBOR, in the case of market-neutral hedge funds. To be successful, quantitative equity managers must rely on four key components to their investment process. First and foremost on the list is an alpha model, which predicts the relative returns of stocks within a specified investment. The second component is a risk model that estimates the risks of individual stocks and the return correlations among different stocks. The third piece is a portfolio construction methodology to combine both return forecasts and risk forecasts to form an optimal portfolio. Lastly, one must have the portfolio implementation process in place to execute the trades. We present the portfolio construction methodology in this chapter. Risk models, alpha models, and portfolio implementations are introduced in later chapters.

Ever since the seminal work by Markowitz (1959), the mean–variance optimization has served as the workhorse for many areas of quantitative finance, including asset allocation, equity, and fixed income portfolio management. It finds the appropriate portfolio weights by solving an optimization problem. There could be several versions of this optimization: one to maximize expected portfolio return for a given level of risk, and another to minimize portfolio variance for a required expected return. Yet another version is to maximize an objective function, that is, the expected portfolio return minus a multiple (risk-aversion parameter) of

the portfolio variance. Despite some of its shortcomings, one of them being the sensitivity of optimal weights to the inputs (noted by practitioners over the years), and many variants of portfolio construction methods aimed to overcome these shortcomings, the mean–variance optimization remains a core tenet of modern portfolio management. A firm understanding of the method and its intuition is thus essential to the understanding and successful implementation of quantitative investment strategies.

We shall first introduce the basic assumptions in the mean–variance optimization. We then present the mathematical analysis for the procedure, deriving the optimal portfolio and analyzing its implications. We shall form the portfolio with minimal constraints in order to derive an analytic solution, allowing us to develop insights and intuitions that might otherwise be obscured in numerical simulations. We analyze two versions of the mean–variance optimization: one for total risk and total return, and the other for active risk and active return. The latter version can be used for both an active portfolio managed against a traditional benchmark and long-short hedge funds.

In this chapter, we also introduce the capital asset pricing model (CAPM) as a risk model and consider optimal portfolios with a beta-neutral constraint as well as a dollar neutral constraint. These portfolios can be obtained by solving a constrained mean–variance optimization or by finding a linear combination of characteristic portfolios.

2.1 DISTRIBUTIONS OF INVESTMENT RETURNS

Return and risk are two inherent characteristics of any investment. The limiting case being cash, which is risk free — devoid of uncertainty — in the short term. The return of an uncertain investment is best described by a probability distribution. One of the most challenging tasks in quantitative finance is to select a type of distribution function that adequately models a given investment instrument and yet is amendable to mathematical analysis. For stocks, the simplest choice is either a normal or lognormal distribution, both of which have their advantages and disadvantages.

A normal distribution, describing the return of a stock over the next time period, can be denoted by $r \sim N\left(\mu, \sigma^2\right)$, where μ is the average or expected return and σ is the standard deviation. The term σ^2 is the variance. The most attractive feature of modeling security return with normal distribution is that the return distribution of a portfolio investing in a number of stocks would also be normal. First, we denote the joint return distribution of multiple stocks as a multivariate normal distribution

$\mathbf{r} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{r} = (r_1, \cdots, r_N)'$ is the return vector, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_N)'$ is the expected return vector, and $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^{N}$ is the covariance matrix among returns of different stocks. The covariance matrix is symmetric with $\sigma_{ij} = \sigma_{ji}$ and positive definite. If we denote the portfolio weights by the weight vector $\mathbf{w} = (w_1, \cdots, w_N)'$, then the portfolio returns distribution is

$$r_p \sim N(\mathbf{w}' \cdot \boldsymbol{\mu}, \mathbf{w}' \boldsymbol{\Sigma} \mathbf{w}). \tag{2.1}$$

Therefore, the portfolio expected return is a weighted average of individual expected returns, and the portfolio return variance is a quadratic function of the weight vector.

Several features of the normal distribution are undesirable or unrealistic when it is used to model stock returns. First, a stock investor has only limited liability — he could not lose more than what he invested in. Therefore, the return of a stock over any time horizon should never be less than –100%. But a normal distribution assigns nonzero probability to losses of any size, even those exceeding –100%. Second, if we assume that a single-period return for a stock is normal, the compound return over multiple periods is no longer normal. This can be illustrated with an example for just two periods. If the return for the first period is r_1 and for the second period is r_2, the compound return over the two periods is $r = (1 + r_1)(1 + r_2) - 1 = r_1 + r_2 + r_1 r_2$. The compound return consists of the sum of two individual period returns and their product. Because the product of two normal variables is not normal, the compound return is not normal. However, note the following remark:

- There are other drawbacks in using a normal distribution to model stocks and returns. The normal distribution is symmetric, whereas in reality, returns exhibit skewness and often have fatter tails (higher probabilities of a large loss or gain) than a normal distribution.

Some of these issues are negated if we use a lognormal distribution for stock returns, i.e., $\ln(1 + r)$ obeys a normal distribution function. The lognormal distribution not only eliminates the possibility of return being less then –100% but also assures that the compound return over multiple time periods is also lognormal. Unfortunately, we know that a linear combination of lognormal variables is not lognormal. Therefore, portfolio returns will not be lognormal even if individual stock returns are. This makes it

difficult for us to use lognormal distributions in portfolio analysis. Therefore, although we are aware of some of its limitations, we will use the normal distribution function to model stock returns throughout this book.

2.1.1 Correlation Coefficient and Diversification

The concept of diversification refers to the fact that the total risk of a portfolio is often less than the sum of all its parts. Diversification arises when the returns among different stocks are not perfectly correlated.

The correlation coefficient between two stocks relates to their covariance and standard deviations by

$$\rho_{1,2} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}. \tag{2.2}$$

It is known that $|\rho_{1,2}| \le 1$. When given the covariance matrix $\mathbf{\Sigma} = \left(\sigma_{ij}\right)_{i,j=1}^{N}$, the standard deviations $\left(\sigma_1, \cdots, \sigma_N\right)$ are the square roots of its diagonal elements. The equivalent of (2.2) in the matrix form gives the correlation matrix of N assets:

$$\mathbf{C} = \mathrm{diag}\left(\sigma_1^{-1}, \cdots, \sigma_N^{-1}\right) \mathbf{\Sigma} \mathrm{diag}\left(\sigma_1^{-1}, \cdots, \sigma_N^{-1}\right). \tag{2.3}$$

In Equation 2.3, $\mathrm{diag}\left(\sigma_1^{-1}, \cdots, \sigma_N^{-1}\right)$ denotes a diagonal matrix with $\left(\sigma_1^{-1}, \cdots, \sigma_N^{-1}\right)$ as diagonal elements and zero elsewhere.

Example 2.1

Before we delve into any mathematical analysis, we first consider a simple hypothetical example to illustrate the benefit of diversification. Imagine two stocks A and B, both priced at \$1. Stock A goes up 100% to \$2 in the first month, and then goes down 50% and back to \$1 again in the second month. Stock B does the opposite, down 50% in the first month and then up 100% in the second month. In this hypothetical case, the two stocks have a correlation of –1. Now, if we have invested in either stock, we would have gone nowhere with our investments after two turbulent months. However, if we had invested in both stocks with a 50/50 split and *rebalanced* the mix back to 50/50 after the first month, we would have grown our investment by 56.25% after the 2 months.

It is informative to analyze the diversification benefit of a portfolio of just two stocks. The total portfolio variance is then

$$\sigma_p^2 = w_1^2\sigma_1^2 + 2\rho_{1,2}w_1w_2\sigma_1\sigma_2 + w_2^2\sigma_2^2 . \tag{2.4}$$

It is easy to see that when both weights are nonnegative,

$$\sigma_p = \begin{cases} w_1\sigma_1 + w_2\sigma_2 & \text{if } \rho_{1,2} = 1 \\ \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2} & \text{if } \rho_{1,2} = 0 \\ \left|w_1\sigma_1 - w_2\sigma_2\right| & \text{if } \rho_{1,2} = -1 \end{cases} . \tag{2.5}$$

At one extreme, when the correlation is 1, the portfolio volatility is the weighted sum of two stock volatilities, and there is no diversification benefit. At the other extreme, when the correlation is –1, the portfolio volatility is the absolute difference of the two, and the diversification is at the maximum. When the correlation is 0, the portfolio volatility is between the two extremes. In this case, the variances are additive instead.

Example 2.2

For a portfolio of N stocks, assume each has the same return standard deviation denoted by σ. Further assume the returns are uncorrelated, and the portfolio return standard deviation is then

$$\sigma_p = \sqrt{\sum_{i=1}^{N} w_i^2\sigma^2} = \sigma\sqrt{\sum_{i=1}^{N} w_i^2} . \tag{2.6}$$

For an equally weighted portfolio, $\sigma_p = \sigma/\sqrt{N}$, the risk declines as the square root of *N*.

We have just seen how the portfolio variance changes with the correlation. It is also instructive to see how it changes when the underlying security weights change. Still using the stock example, we require $w_1 + w_2 = 1$. In other words, the portfolio is fully invested in the two risky securities under consideration. Figure 2.1 displays the variance as a function of w_1 with $\sigma_1 = 40\%$, $\sigma_2 = 30\%$, and $\rho_{1,2} = 0.3$. In the plot, we let the weight to be both negative and greater than 100% to allow shorting of both stocks.

The portfolio variance (2.4) is a quadratic function of the weight, and it attains the minimum when

$$w_1 = \frac{\sigma_2^2 - \rho_{1,2}\sigma_1\sigma_2}{\sigma_1^2 - 2\rho_{1,2}\sigma_1\sigma_2 + \sigma_2^2},\ w_2 = 1 - w_1 . \tag{2.7}$$

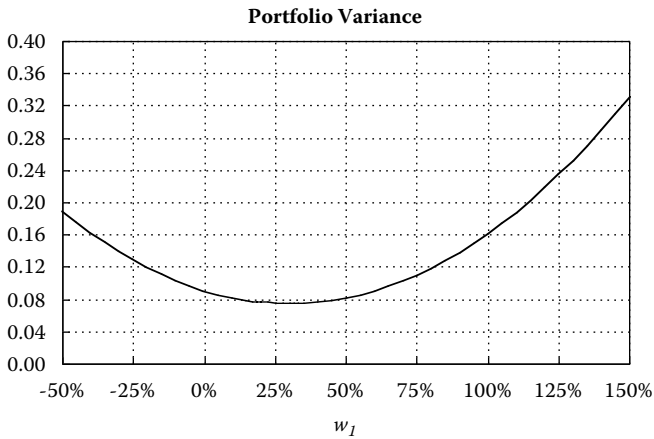


FIGURE 2.1. Portfolio variance as a function of stock weight w_1.

This is the minimum variance portfolio that has the least risk. For parameters used in Figure 2.1, the minimum occurs when w_1 = 30%, and in this case the minimum portfolio volatility is 27%, smaller than either of the individual volatilities.

2.2 OPTIMAL PORTFOLIOS

In this section, we shall derive various optimal portfolios with different objective functions.

2.2.1 Minimum Variance Portfolio

Suppose there are N stocks in the investmentable universe and we have a fully invested portfolio investing 100% of the capital. The covariance matrix is denoted as $\mathbf{\Sigma}$. We are interested in finding the portfolio with minimum variance. An investor choosing this portfolio is only concerned about the risk of the portfolio. Denoting a vector of ones by $\mathbf{i} = (1, \cdots, 1)'$, we have the following optimization problem:

$$\begin{aligned} &\text{Minimize } \frac{1}{2}\mathbf{w}'\mathbf{\Sigma}\mathbf{w} \\ &\text{subject to: } \mathbf{w}' \cdot \mathbf{i} = w_1 + w_2 + \cdots + w_N = 1. \end{aligned} \tag{2.8}$$

The constraint in (2.8) is often referred to as a *budget constraint*. The fraction one half is merely a scaling constant, and the reason for including

it will soon be apparent. The problem can be solved by the method of Lagrangian multipliers. We form a new objective function

$$Q(\mathbf{w},l)=\frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}-l(\mathbf{w}'\cdot\mathbf{i}-1). \tag{2.9}$$

The additional term in (2.9) is the Lagrangian multiplier times a constraint-related term. Taking the partial derivative of the new function with respect to the weight vector and equating it to zero yields the condition for the optimal weight

$$\boldsymbol{\Sigma}\mathbf{w}-l\mathbf{i}=0 \tag{2.10}$$

and solving for the weight vector gives

$$\mathbf{w}=l\boldsymbol{\Sigma}^{-1}\mathbf{i}, \tag{2.11}$$

where $\boldsymbol{\Sigma}^{-1}$ is the inverse matrix of $\boldsymbol{\Sigma}$. To determine the Lagrangian multiplier l, we substitute the weight vector into the constraint in Equation 2.8 to obtain

$$l=\frac{1}{\left(\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}\right)}. \tag{2.12}$$

Finally, substituting Equation 2.12 into Equation 2.11 yields the minimum variance portfolio weight vector

$$\mathbf{w}^{*}_{\min}=\frac{\boldsymbol{\Sigma}^{-1}\mathbf{i}}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}}. \tag{2.13}$$

It is easy to verify that the optimal weight (2.13) satisfies the budget constraint. Finally, the minimum variance is

$$\sigma^2_{\min}=\left(\mathbf{w}^{*}_{\min}\right)'\boldsymbol{\Sigma}\mathbf{w}^{*}_{\min}=\frac{1}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}}, \tag{2.14}$$

equal to the Lagrangian multiplier (2.12).

2.2.2 Mean–Variance Optimal Portfolio with Cash

The minimum variance portfolio focuses solely on the risk and ignores the expected return of the portfolio. Most investors prefer a balance between the two, provided they have return expectation for stocks. The mean–variance optimization serves as the main tool for finding the optimal portfolio with the maximum expected return for a given level of risk. We first consider portfolios that include cash and denote its return by r_f and its weight by w_0. We denote the expected return vector of N stocks by $\mathbf{f} = \left(f_1, \cdots, f_N\right)'$, which is a collection of forecasts generated by investors through investment research. For the time being, we take these forecasted returns as given inputs. In Part II of this book, we will identify some quantitative factors for forecasting stock returns. The mean–variance optimal portfolio with a risk-aversion parameter λ is

$$\begin{aligned} &\text{Maximize } w_0 r_f + \mathbf{w}' \cdot \mathbf{f} - \frac{1}{2}\lambda\left(\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}\right) \\ &\text{subject to: } w_0 + \mathbf{w}' \cdot \mathbf{i} = 1 \end{aligned} \tag{2.15}$$

Note that cash is risk free — it only contributes to return but has no risk, at least for a single-period optimization. The risk-aversion parameter $\lambda > 0$ determines the degree of influence that risk has on the portfolio. If $\lambda = 0$, then the risk term drops out and the problem reduces to maximizing expected return under the assumed budget constraint. The solution is generally unbounded because one can borrow unlimited amount from the low-return asset and invest that sum in the higher return asset. On the other hand, if $\lambda \to \infty$, (meaning the investor is extremely risk averse and), then the optimal portfolio would have 100% in cash and have no risk at all.

The problem (2.15) can be converted into an unconstrained optimization problem for the stock weights by using the constraint in the objective function. Writing the constraint as $w_0 = 1 - \mathbf{w}' \cdot \mathbf{i}$ and substituting it into the objective function yields

$$\text{Maximize } \mathbf{w}' \cdot \mathbf{f}_e - \frac{1}{2}\lambda\left(\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}\right), \quad \text{with } \mathbf{f}_e = \mathbf{f} - r_f \mathbf{i}. \tag{2.16}$$

The vector $\mathbf{f}_e$ represents the stocks' excess returns above cash. The optimal weights are found by equating partial derivatives of the objective function (2.16) to zero. We have

$$\mathbf{w}^* = \frac{1}{\lambda}\mathbf{\Sigma}^{-1}\mathbf{f}_e$$
$$w_0^* = 1 - \mathbf{w}^{*\prime}\cdot\mathbf{i} = 1 - \frac{1}{\lambda}\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{f}_e \tag{2.17}$$

The following examples from solution (2.17) help us gain insights to the mean–variance optimization.

Example 2.3

When the covariance matrix is diagonal, i.e., when the stock returns are uncorrelated, the optimal weight of an individual stock is

$$w_i^* = \frac{1}{\lambda}\frac{f_i - r_f}{\sigma_i^2} = \frac{1}{\lambda}\frac{f_{e,i}}{\sigma_i^2}. \tag{2.18}$$

Therefore, in isolation, the optimal weight of stock is proportional to its own excess return and inversely proportional to its own variance and the risk-aversion parameter. Because of this relationship, the optimal weight is in fact twice as sensitive to the standard deviation as to the expected return on the margin. Mathematically, if the changes in the forecast and standard deviation are small:

$$\frac{\Delta w_i^*}{w_i^*} = \frac{\Delta f_{e,i}}{f_{e,i}} - 2\frac{\Delta\sigma_i}{\sigma_i}. \tag{2.19}$$

Hence, a relative increase in the expected return will bring the same relative increase in the optimal weight. On the other hand, a relative increase in the stock volatility would bring down the optimal weight by a factor of two.

Example 2.4

This example illustrates the effect of the correlation coefficient on the optimal weights. We choose the case of two stocks because the inverse of a 2×2 covariance matrix is readily available. We have

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad \mathbf{\Sigma}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1 c} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}. \tag{2.20}$$

Substituting the inverse matrix into Equation 2.17 yields

$$
\begin{aligned}
w_1^* &= \frac{1}{\lambda\left(1-\rho^2\right)}\left(\frac{f_{e1}}{\sigma_1^2}-\rho\frac{f_{e2}}{\sigma_1\sigma_2}\right) \\
w_2^* &= \frac{1}{\lambda\left(1-\rho^2\right)}\left(\frac{f_{e2}}{\sigma_2^2}-\rho\frac{f_{e1}}{\sigma_1\sigma_2}\right)
\end{aligned}
\tag{2.21}
$$

In contrast to Equation 2.18, the optimal weight of each stock has one additional term that is dependent on the expected return of the other stock. Suppose the correlation coefficient is positive; then the additional term would be negative — a reduction in optimal weight if the expected excess return of the other stock is also positive. On the other hand, if the correlation is negative, then the optimal weight would be increased if the expected excess return of the other stock is positive. This is the essence of diversification at work. With positive correlation, one should reduce the combined weight of the two stocks to reduce overall risk. But with negative correlation, one should increase the combined weight because the risks in two stocks are offsetting each other.

2.2.3 Mean–Variance Optimal Portfolio without Cash

The optimal portfolio with cash might be useful in determining appropriate allocation between stocks and cash but is of little use when an equity portfolio must be fully invested in stocks. Most equity portfolios for mutual fund investors and institutional investors are managed this way. Thus, we must consider the mean–variance optimization for fully invested portfolios. We can formulate the problem by simply setting $w_0 = 0$ in (2.15).

Because the budget constraint is now binding, we must use the method of Lagrangian multipliers to solve the optimization problem (see Problem 2.5). We have

$$
\mathbf{w}^* = \frac{\mathbf{\Sigma}^{-1}\mathbf{i}}{\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}} + \frac{1}{\lambda}\frac{\left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}\right)\mathbf{\Sigma}^{-1}\mathbf{f} - \left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{f}\right)\mathbf{\Sigma}^{-1}\mathbf{i}}{\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}}. \tag{2.22}
$$

The first term in the solution (2.22) is just the minimum variance solution, independent of both the forecast and the risk-aversion parameter. The second term is affected by the forecast and the risk-aversion parameter.

Because cash is excluded, we need not worry about excess return. Note the following remark:

- There are two cases in which the solution (2.22) reduces to the minimum variance weights. The first is when $\lambda \to \infty$ and the second term vanishes. The second case is less obvious, and that is when all the return forecasts are identical, i.e., $\mathbf{f} = k\mathbf{i}$; again, the solution is identical to the minimum variance solution. This is intuitive; when all returns are the same, the portfolio return will be the same as well. Hence, the minimum variance portfolio is the mean–variance optimal portfolio. Consequently, if we increase all the return forecasts by an identical amount, the optimal solution remains unchanged.

The expected return and variance of the optimal portfolio are

$$\begin{aligned} \mu^* &= \mathbf{f}' \cdot \mathbf{w}^* = \frac{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{f}}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}} + \frac{1}{\lambda}\frac{\left(\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\boldsymbol{\Sigma}^{-1}\mathbf{f}\right) - \left(\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{f}\right)^2}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}} \\ \left(\sigma^*\right)^2 &= \mathbf{w}^{*\prime}\boldsymbol{\Sigma}\mathbf{w}^* = \frac{1}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}} + \frac{1}{\lambda^2}\frac{\left(\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\boldsymbol{\Sigma}^{-1}\mathbf{f}\right) - \left(\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{f}\right)^2}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}} \end{aligned} \tag{2.23}$$

The expected return μ^* is the maximum expected return for a given level of risk at σ^*. As we change the risk-aversion parameter, the pair $\left(\sigma^*, \mu^*\right)$ forms a curve called the efficient frontier in the risk/return space.

Example 2.5

The hyperbolic curve in Figure 2.2 depicts such an efficient frontier for portfolios of just three stocks with the following: return forecasts, volatilities (we have written the volatilities into a vector just for simplicity), and correlation matrix.

$$\mathbf{f} = \begin{pmatrix} 10\% \\ 0\% \\ -10\% \end{pmatrix}, \quad \boldsymbol{\sigma} = \begin{pmatrix} 30\% \\ 30\% \\ 30\% \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

The straight line depicts another efficient frontier, which we will discuss next. For this set of inputs, the minimum portfolio ($\lambda = \infty$) is an equally weighted portfolio with zero expected return and volatility of 24%. As the

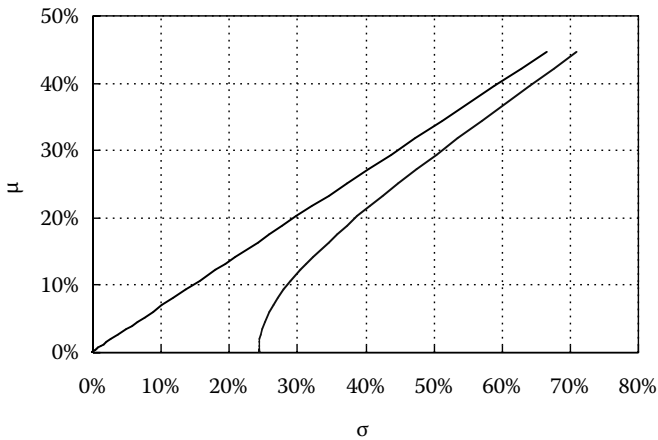


FIGURE 2.2. Efficient frontiers: the curved line is the efficient frontier of a fully invested equity portfolio, and the straight line is the efficient frontier of a long-short dollar neutral portfolio.

risk-aversion parameter descends from infinity, both the expected return and risk of the optimal portfolio increase in a concave shape that is typical of efficient frontiers.

2.2.4 Active Mean–Variance Optimization

In many cases, equity portfolios are managed against a benchmark, such as the S&P 500 index or the Russell 2000 index. The return and risk of these portfolios are measured relative to the benchmark and are called *active return* and *active risk*. An active mean–variance optimal portfolio is one that has the maximum expected active return for a given level of active risk.

We can decompose the portfolio weights into benchmark weights and active weights: $\mathbf{w} = \mathbf{b} + \mathbf{a}$. Because both benchmark and portfolio weights satisfy the budget constraint, the active weights must be dollar neutral, i.e., $\mathbf{a}' \cdot \mathbf{i} = 0$. In other words, overweights ($a_i > 0$) must be perfectly balanced or financed by underweights ($a_i < 0$).

For long-short market-neutral equity hedge funds, the traditional equity benchmarks no longer apply. Instead, a cash benchmark is often used. In this case, the active weights are just the portfolio weights. If the fund is also dollar neutral, then the weights must also satisfy the constraint $\mathbf{a}' \cdot \mathbf{i} = 0$. Dollar neutral is not the same as market neutral. As we shall see later in this chapter and in Chapter 3.

Given the expected return vector **f**, the expected active return is $\mathbf{a}' \cdot \mathbf{f}$. The active risk in variance is $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$. The objective of active mean–variance optimization is to find optimal active weights through

$$\begin{aligned} &\text{Maximize } \mathbf{a}' \cdot \mathbf{f} - \frac{1}{2}\lambda\left(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}\right) \\ &\text{subject to: } \mathbf{a}' \cdot \mathbf{i} = 0 \end{aligned} \tag{2.24}$$

The solution of this mean–variance optimization turns out to be identical to the second term in Equation 2.22. The optimal active weights are

$$\mathbf{a}^* = \frac{1}{\lambda} \frac{\left(\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}\right)\boldsymbol{\Sigma}^{-1}\mathbf{f} - \left(\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{f}\right)\boldsymbol{\Sigma}^{-1}\mathbf{i}}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}}. \tag{2.25}$$

- This solution has several features worth noting. First, it is inversely proportional to the risk-aversion parameter. Therefore, depending on investors' risk appetite, the optimal weights are entirely scalable. Second, it is independent of the benchmark. Consequently, the expected active return or alpha and the active risk are also independent of the benchmark. It is therefore theoretically feasible to utilize or port it on any benchmark. In other words, two active equity portfolios managed against two different equity benchmarks could have the same active weights. For instance, the active weights of an equity portfolio managed against S&P 500 index could be the same as the weights of a long-short market-neutral hedge fund. This is the idea behind the so-called portable alpha strategies, i.e., the alpha or excess return generated from a strategy can be ported onto another different benchmark. In reality, however, this is not entirely possible for most traditional equity portfolios because they must strictly obey the no-shorting rule. We have not included this type of constraint into the mean–variance optimization. We shall see in Chapter 9 that imposing this constraint and various other constraints will alter the optimal active weights greatly.

One alternative form of the optimal active weights (2.25) that provides more insights is the following:

$$\mathbf{a}^* = \frac{1}{\lambda}\boldsymbol{\Sigma}^{-1}\left(\mathbf{f} - l\mathbf{i}\right), \quad \text{with } l = \frac{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{f}}{\mathbf{i}'\boldsymbol{\Sigma}^{-1}\mathbf{i}}. \tag{2.26}$$

This is similar to the unconstrained optimal weights (2.17). There, the forecasted are uniformly adjusted by the risk-free rate. Here, we adjust the forecasts by the Lagrangian multiplier to ensure that the active weights are dollar neutral. The only case in which the adjustment is not needed is when $l = 0$ or when $\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{f} = 0$. This conditionality implies that the original forecasts would give rise to a set of optimal weights $\mathbf{a}^0 = \lambda^{-1}\mathbf{\Sigma}^{-1}\mathbf{f}$ that are already dollar neutral. When it is not satisfied, we must adjust the forecasts according to Equation 2.26.

The expected active return from the optimal weights (2.25) is

$$\alpha^* = \mathbf{f}' \cdot \mathbf{a}^* = \frac{1}{\lambda} \frac{\left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\mathbf{\Sigma}^{-1}\mathbf{f}\right) - \left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{f}\right)^2}{\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}}. \tag{2.27}$$

The active risk in standard deviation, or, as it is often called, the expected tracking error of the portfolio to the benchmark, is

$$\sigma^* = \sqrt{\mathbf{a}^{*\prime}\mathbf{\Sigma}\mathbf{a}^*} = \frac{1}{\lambda}\sqrt{\frac{\left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\mathbf{\Sigma}^{-1}\mathbf{f}\right) - \left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{f}\right)^2}{\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}}}. \tag{2.28}$$

For a long-short dollar neutral hedge fund, these are not relative but absolute return and risk. As both Equation 2.27 and Equation 2.28 have the same dependence on the risk-aversion parameter, the associated efficient frontier is a straight line going through the origin

$$\frac{\alpha^*}{\sigma^*} = \sqrt{\frac{\left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\mathbf{\Sigma}^{-1}\mathbf{f}\right) - \left(\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{f}\right)^2}{\mathbf{i}'\mathbf{\Sigma}^{-1}\mathbf{i}}}. \tag{2.29}$$

- There are two different ways to interpret this efficient frontier: one in active space for traditional portfolios, and the other in absolute space for long-short hedge funds. The ratio represents expected excess return per unit of risk in terms of standard deviation. This is often referred to as *information ratio* (IR) of the portfolio. The portfolios on the efficient frontier offer the maximum information ratio among all portfolios with the same level of risks. Because we are only concerned with the optimal portfolio for one time period, this information ratio is a one-period IR. We shall discuss multiple-period IR later in the book.

In Figure 2.2, we graph this efficient frontier together with the efficient frontier for a fully invested portfolio with the same inputs. By comparing the two frontiers, the graph makes it possible to compare a fully invested portfolio with a long-short hedge fund in absolute risk/return space.

Therefore, there are several features in Figure 2.2 worth noting. First, the efficient frontier of the long-short hedge fund always lies on top of the efficient frontier of the fully invested portfolio. This indicates that, for the same amount of risk, i.e., above 24%, one can expect higher return from the hedge fund than from the fully invested portfolio. This is reasonable because the average stock return in our input is 0%. Thus, fully invested portfolios take additional risk with no additional return. The second and perhaps less obvious feature is that, whereas the risk of fully invested portfolios has a minimum (24% in this case), the hedge fund risk can be targeted at any level without a minimum or maximum. In our example, if an investor's risk preference is below 24%, the hedge fund is the only available investment choice.

Third, the relative placement of two efficient frontiers can be quite different if any of the inputs to mean–variance optimization changes. For example, if the expected returns are increased by 10% for each stock and the covariance matrix remain the same, the efficient frontier of the fully invested portfolio is lifted and becomes a better choice for most of the risk spectrum than the hedge fund. The expected returns of hedge fund portfolios remain unaffected because they depend on the relative differences in returns, not the absolute level. This is shown in Figure 2.3.

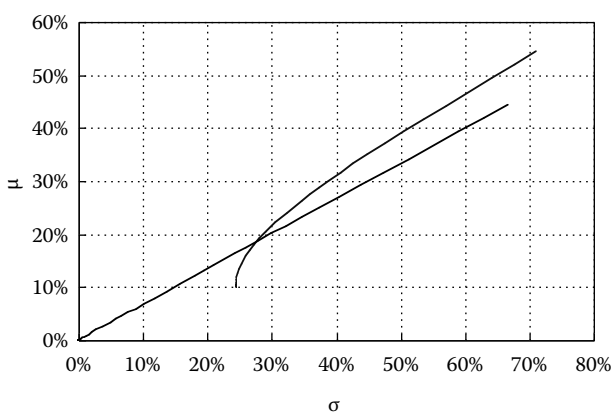


FIGURE 2.3. Efficient frontiers similar to those in Figure 2.2, except for the change in the expected returns, which are 10% higher for each stock.

2.3 CAPITAL ASSET PRICING MODEL

At least two inputs are required in order to use mean–variance optimization for portfolio construction. They are expected return forecasts and return covariance matrix. Additional inputs are practical constraints that are required for realistic portfolios, i.e., limits on stock holdings and/or sector weights. Forecasting returns and portfolio constraints will be discussed extensively in Part II and Part III of this book. For the remainder of this chapter and the next, we focus on the covariance matrix.

So far, we have left the covariance matrix rather arbitrary in mean–variance analysis. For a portfolio of N stocks, there are $N(N+1)/2$ variances and covariances. For the stock market as a whole, or portfolios with thousands of stocks, the estimation of so many parameters proves to be an impossible task. CAPM, developed by Sharpe (1964), Tobin (1958), and Lintner (1965), provides a particular simple structure for the covariance matrix.

Denoting the return of the overall market by r_{M}, CAPM stipulates that individual stocks' returns r_i is the sum of systematic return and specific return

$$r_i = r_f + \beta_i \left(r_{\mathrm{M}} - r_f \right) + \varepsilon_i , \tag{2.30}$$

where r_f is the risk-free rate. The systematic return is a function of beta that measures the sensitivity of individual stocks' returns to the market return. It is given as the regression coefficient of r_i vs. the market return r_{M}

$$\beta_i = \frac{\mathrm{cov}\left(r_i, r_{\mathrm{M}}\right)}{\mathrm{cov}\left(r_i, r_{\mathrm{M}}\right)} = \frac{\rho_{i,\mathrm{M}} \sigma_i \sigma_{\mathrm{M}}}{\sigma_{\mathrm{M}}^2} = \frac{\rho_{i,\mathrm{M}} \sigma_i}{\sigma_{\mathrm{M}}}. \tag{2.31}$$

In Equation 2.31, $\rho_{i,m}$ denotes the correlation coefficient between r_i and r_{M}, and σ_{M} denotes the volatility of market returns. The last term in (2.30) is the specific return component and is a normal random variable with zero mean:

$$\varepsilon_i \sim N\left(0, \theta_i^2\right). \tag{2.32}$$

The volatility of the specific return θ_i is often referred to as the *specific risk*.

CAPM assumes that, for an individual stock, the systematic return and the specific return are independent of each other. Furthermore, the specific returns of different stocks are also independent of one another.

Essentially, the portfolio covariance structure maps each security's pairwise covariance into its linkage through beta.

It is worth noting that even when CAPM is not applicable, we can still define beta as in (2.31). If $\mathbf{\Sigma}$ is a general covariance matrix and $\mathbf{b}$ is the weight vector of the market or a benchmark portfolio, then the beta vector $\boldsymbol{\beta} = \left(\beta_1, \cdots, \beta_N\right)'$ is given as

$$\boldsymbol{\beta} = \frac{\mathbf{\Sigma b}}{\mathbf{b}'\mathbf{\Sigma b}}. \tag{2.33}$$

It is easy to show that, under CAPM, the covariance matrix is

$$\begin{aligned}\mathbf{\Sigma} &= \boldsymbol{\beta}\boldsymbol{\beta}'\sigma_{\mathrm{M}}^2 + \mathrm{diag}\left(\theta_1^2, \cdots, \theta_N^2\right) \\ &= \boldsymbol{\beta}\boldsymbol{\beta}'\sigma_{\mathrm{M}}^2 + \mathbf{S} \\ &= \begin{pmatrix} \beta_1\beta_1 & \cdots & \beta_1\beta_N \\ \vdots & \ddots & \vdots \\ \beta_N\beta_1 & \cdots & \beta_N\beta_N \end{pmatrix}\sigma_{\mathrm{M}}^2 + \begin{pmatrix} \theta_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \theta_N^2 \end{pmatrix}\end{aligned} \tag{2.34}$$

We have used $\mathbf{S}$ for the diagonal matrix consisting of specific variances.

For a portfolio with weight vector $\mathbf{w}$, the portfolio beta is then the weighted average of stock betas $\beta_p = \mathbf{w}' \cdot \boldsymbol{\beta}$. The portfolio variance can be separated into systematic variance and specific variance:

$$\sigma_p^2 = \beta_p^2\sigma_{\mathrm{M}}^2 + \sum_{i=1}^{N} w_i^2\theta_i^2 . \tag{2.35}$$

This shows that the portfolio has two sources of risk, one systematic and the other specific. Although we leave the detailed discussion of risk contribution until the next chapter, we provide a few remarks regarding the relative importance of the two sources of risk.

- We notice the specific risk of a portfolio can be diversified away with increasing number of stocks. For simplicity, suppose all stock-specific risks are the same for all stocks; an equally weighted portfolio would have the specific variance of θ_0^2/N. The corresponding specific volatility is $\theta_0/\sqrt{N}$. Suppose the specific risk is 30%; then, the

portfolio specific risk would be 3% with 100 stocks and 1.5% with 400 stocks. The systematic risk, on the other hand, does not depend explicitly on the number of stocks; it is solely a function of portfolio beta and market risk. Suppose the market volatility is around 15%. A portfolio with unit beta would have 15% systematic volatility. Therefore, a traditional long-only portfolio would have most of its risk in the market risk. However, a zero beta portfolio, typically a long-short market-neutral portfolio, would have no systematic or market risk. All its risk is specific risk. Of course, this depends heavily on the accuracy of beta estimation.

2.3.1 Optimal Portfolios under CAPM

We now have the special form of the covariance matrix (2.34) under CAPM and will study the mean–variance optimization solution under it. In order to do so, we first must find the inverse of the covariance matrix. Using the result from Problem 2.10, we obtain

$$\boldsymbol{\Sigma}^{-1} = \mathbf{S}^{-1} - \frac{\sigma_{\mathrm{M}}^2}{1+\kappa}\boldsymbol{\beta}_s\boldsymbol{\beta}_s', \tag{2.36}$$

where

$$\kappa = \sum_{i=1}^{N} \frac{\sigma_{\mathrm{M}}^2 \beta_i^2}{\theta_i^2}, \quad \boldsymbol{\beta}_s = \left(\frac{\beta_1}{\theta_1^2}, \cdots, \frac{\beta_N}{\theta_N^2}\right)'. \tag{2.37}$$

In the sum κ, each term is the ratio of systematic variance to specific variance for an individual stock. In the vector $\boldsymbol{\beta}_s$, the components are beta scaled by the specific variance.

Example 2.6

We can get a sense of the magnitude of κ by considering a stock with beta 1 and the specific risk of 30%. Assuming $\sigma_{\mathrm{M}} = 15\%$, we obtain $\sigma_{\mathrm{M}}^2\beta_i^2 / \theta_i^2 = 1/4$. Hence, a rough estimate of κ would be $\kappa \approx N/4$.

To understand how mean–variance optimal weights behave under CAPM, we once again consider the case of optimal portfolios including cash in which the weights of the risky assets is given by an unconstrained optimization. According to (2.17), it is the inverse of the covariance matrix times the excess return vector:

$$\mathbf{w}^* = \frac{1}{\lambda}\boldsymbol{\Sigma}^{-1}\mathbf{f}_e = \frac{1}{\lambda}\left(\mathbf{S}^{-1}\mathbf{f}_e - \frac{\sigma_{\mathrm{M}}^2}{1+\kappa}\boldsymbol{\beta}_s\boldsymbol{\beta}_s' \cdot \mathbf{f}_e\right). \tag{2.38}$$

Let us denote

$$\mathbf{w}_0^* = \frac{1}{\lambda}\mathbf{S}^{-1}\mathbf{f}_e \tag{2.39}$$

as "the partial solution" given by the specific covariance matrix and the forecasts. Then we see that

$$\frac{1}{\lambda}\boldsymbol{\beta}_s' \cdot \mathbf{f}_e = \frac{1}{\lambda}\sum_{i=1}^{N}\frac{\beta_i f_{e,i}}{\theta_i^2} = \sum_{i=1}^{N}\beta_i w_{0,i}^* = \beta_0 . \tag{2.40}$$

It is the portfolio beta given by the partial solution (2.39). Combining Equation 2.39 and Equation 2.40, we rewrite the optimal solution as in Equation 2.38 as

$$\mathbf{w}^* = \mathbf{w}_0^* - \frac{\sigma_{\mathrm{M}}^2\beta_0}{1+\kappa}\boldsymbol{\beta}_s . \tag{2.41}$$

In terms of weight of a single stock, we have

$$w_i^* = w_{0,i}^* - \frac{1}{1+\kappa}\frac{\sigma_{\mathrm{M}}^2\beta_0\beta_i}{\theta_i^2} = w_{0,i}^* - \frac{1}{1+\kappa}\frac{\mathrm{cov}\left(r_i, r_{\mathbf{w}_0^*}\right)}{\theta_i^2} . \tag{2.42}$$

In other words, the optimal weight of a stock is the partial weight less the ratio of its covariance with the partial portfolio to its specific variance times a scalar. Note the following remarks:

- If the excess return forecasts adjusted by specific variances are uncorrelated with the stocks' beta estimates, then $\beta_0 = 0$. In this special case, the optimal weights are identical to the partial solution (2.39).
- In general, we can derive the optimal weights in two steps. In the first step, we simply derive the partial weights based only on the specific risks. In the second step, we modify the partial weights by the covariance term. Note that, if $\beta_i\beta_0 > 0$, i.e., the stock beta and the partial solution beta are of the same sign, we reduce the partial weight. On the other hand, if $\beta_i\beta_0 < 0$, i.e., the stock beta and the partial solution beta are of the opposite sign, we increase the partial weight. The net effect is to reduce the beta of the partial solution.

The beta of the optimal portfolio is

$$\begin{aligned}\beta^* &= \sum_{i=1}^{N} w_i^* \beta_i = \sum_{i=1}^{N} w_{0,i}^* \beta_i - \frac{\beta_0}{1+\kappa} \sum_{i=1}^{N} \frac{\sigma_M^2 \beta_i}{\theta_i^2} \\ &= \beta_0 - \frac{\beta_0 \kappa}{1+\kappa} = \frac{\beta_0}{1+\kappa}\end{aligned} . \tag{2.43}$$

We have used the definition (2.37) in the derivation. Because the parameter κ is proportional to N, we conclude that, for a portfolio of reasonable size, the beta of the optimal portfolio should be significantly less than β_0.

We next derive the specific risk of the optimal portfolio:

$$\begin{aligned}\sum_{i=1}^{N} \left(w_i^* \theta_i\right)^2 &= \sum_{i=1}^{N} \left(w_{0,i}^* \theta_i - \frac{1}{1+\kappa} \frac{\sigma_M^2 \beta_0 \beta_i}{\theta_i} \right)^2 \\ &= \sum_{i=1}^{N} \left[\left(w_{0,i}^* \theta_i\right)^2 + \frac{1}{\left(1+\kappa\right)^2} \frac{\sigma_M^4 \beta_0^2 \beta_i^2}{\theta_i^2} - \frac{2\sigma_M^2 \beta_0}{1+\kappa} w_{0,i}^* \beta_i \right] \\ &= \sum_{i=1}^{N} \left(w_{0,i}^* \theta_i\right)^2 + \frac{\sigma_M^2 \beta_0^2 \kappa}{\left(1+\kappa\right)^2} - \frac{2\sigma_M^2 \beta_0^2}{1+\kappa} \\ &= \sum_{i=1}^{N} \left(w_{0,i}^* \theta_i\right)^2 - \sigma_M^2 \beta_0^2 \left[\frac{1}{1+\kappa} + \frac{1}{\left(1+\kappa\right)^2} \right]\end{aligned} . \tag{2.44}$$

This shows that the specific variance of the optimal portfolio is the specific variance of the partial solution minus a correction term that is proportional to the beta of the partial solution. The total risk of the optimal portfolio is then

$$\sigma_p^2 = \sum_{i=1}^{N} \left(w_i^* \theta_i\right)^2 + \left(\beta^*\right)^2 \sigma_M^2 = \sum_{i=1}^{N} \left(w_{0,i}^* \theta_i\right)^2 - \frac{\sigma_M^2 \beta_0^2}{1+\kappa} . \tag{2.45}$$

Example 2.7

We shall consider an example with three stocks and an optimal portfolio. Table 2.1 lists their relevant attributes. The betas are 1.5, 1.0, and 0.5,

TABLE 2.1 Optimal Portfolios with Three Stocks

Stock	Beta	Systematic Risk	Specific Risk	Total Risk	Forecast	$\mathbf{w}_0^*$	$\mathbf{w}^*$
1	1.5	23%	30%	38%	10%	44%	36%
2	1.0	15%	30%	34%	0%	0%	–6%
3	0.5	8%	30%	31%	–10%	–44%	–47%

respectively. Assuming a market risk of 15%, the stocks' systematic risks are 23%, 15%, and 8%, respectively. The stocks' specific risks are the same at 30%. Combining the systematic and specific risks yields the total risk of 38%, 34%, and 31%, respectively.

With expected return of 10%, 0%, and –10%, the average forecast is 0%. We have chosen $\lambda = 2.5$ for the optimal portfolio. The partial solution using only forecast and specific risk is 44%, 0%, and –44%, respectively. The beta for this portfolio is 0.44. The optimal weight is 36%, –6%, and –47%, respectively, with a beta of 0.23. As the partial solution has a positive beta, 0.44, and all stocks also have positive beta, the optimal weights are all less than the partial solution in order to reduce beta exposure. The optimal portfolio has a systematic risk of 3.6%, a specific risk of 17.9%, and a total risk of 18.2%. The majority of the total risk is attributed to the specific risk, at 96%.

2.3.2 Beta-Neutral Portfolios

As we have seen from the last section, an active mean–variance optimal portfolio in general will have some beta exposure. For a long-only portfolio managed against a benchmark, the active portfolio will have a beta bias, affecting its relative return against the benchmark. For instance, suppose the active portfolio is low beta, at 0.9. Then a market return of 5% will cause an underperformance of 0.5% (= $0.1 \cdot 5\%$) or 50 basis points by the portfolio. For a long-short market-neutral portfolio, this translates to a pure loss of 50 basis points. Therefore, an unintended beta exposure is a source of market risk. One way to eliminate it is to force the active portfolio to have zero beta exposure, i.e., $\mathbf{w}' \cdot \boldsymbol{\beta} = 0$. We shall derive beta-neutral optimal portfolios in this section.

A mean–variance optimization with beta-neutral constraint under CAPM is surprisingly simple. As the optimal portfolio will be beta neutral, its risk will consist entirely of specific risk. We can reformulate the optimization problem with the diagonal matrix $\mathbf{S}$ in (2.34) as

$$\begin{aligned} &\text{Maximize } \mathbf{w}' \cdot \mathbf{f} - \frac{1}{2}\lambda\left(\mathbf{w}'\mathbf{S}\mathbf{w}\right) \\ &\text{subject to: } \mathbf{w}' \cdot \boldsymbol{\beta} = 0 \end{aligned} . \tag{2.46}$$

We find the solution by using the Lagrangian multiplier method:

$$\mathbf{w}^* = \frac{1}{\lambda}\mathbf{S}^{-1}\left(\mathbf{f} - l\boldsymbol{\beta}\right), \quad \text{with } l = \frac{\mathbf{f}'\mathbf{S}^{-1}\boldsymbol{\beta}}{\boldsymbol{\beta}'\mathbf{S}^{-1}\boldsymbol{\beta}} . \tag{2.47}$$

As **S** is a diagonal matrix, we can write the weights explicitly as in

$$w_i^* = \frac{1}{\lambda}\frac{f_i - l\beta_i}{\theta_i^2}, \quad i = 1, \cdots, N; \text{ with } l = \sum_{i=1}^{N}\frac{f_i\beta_i}{\theta_i^2} \Big/ \sum_{i=1}^{N}\frac{\beta_i^2}{\theta_i^2}. \tag{2.48}$$

We note that the solution in this case resembles optimal weights (2.18) in which the covariance matrix was diagonal. By requiring beta neutrality, we have effectively eliminated the market risk from the covariance matrix. What remains is the specific risk. However, instead of the original forecast, we now use a beta-adjusted forecast in (2.48).

- If the forecasts and betas are such that $\sum_{i=1}^{N}\frac{f_i\beta_i}{\theta_i^2} = \mathbf{f}'\mathbf{S}^{-1}\boldsymbol{\beta} = 0$, i.e., they are orthogonal with respect to the matrix $\mathbf{S}^{-1}$, then no beta adjustment is needed.

In addition to market-neutral portfolios, many long-short hedge funds also adhere to a dollar neutral constraint, $\mathbf{w}' \cdot \mathbf{i} = 0$. The solution for the optimal weights with both constraints takes on the same form as in (2.48). However, instead of adjusting the forecasts just for the beta constraint, we now need an additional adjustment for the dollar neutral constraint. We cite the following results and leave the derivation as an exercise. We have

$$w_i^* = \frac{1}{\lambda}\frac{f_i - l_1 - l_2\beta_i}{\theta_i^2}, \quad i = 1, \cdots, N \tag{2.49}$$

where

$$
\begin{aligned}
l_1 &= \frac{\left(\sum_{i=1}^{N}\frac{\beta_i^2}{\theta_i^2}\right)\left(\sum_{i=1}^{N}\frac{f_i}{\theta_i^2}\right)-\left(\sum_{i=1}^{N}\frac{\beta_i}{\theta_i^2}\right)\left(\sum_{i=1}^{N}\frac{f_i\beta_i}{\theta_i^2}\right)}{\left(\sum_{i=1}^{N}\frac{\beta_i^2}{\theta_i^2}\right)\left(\sum_{i=1}^{N}\frac{1}{\theta_i^2}\right)-\left(\sum_{i=1}^{N}\frac{f_i\beta_i}{\theta_i^2}\right)^2} \\
&= \frac{\left(\boldsymbol{\beta}'\mathbf{S}^{-1}\boldsymbol{\beta}\right)\left(\mathbf{f}'\mathbf{S}^{-1}\mathbf{i}\right)-\left(\boldsymbol{\beta}'\mathbf{S}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\mathbf{S}^{-1}\boldsymbol{\beta}\right)}{\left(\boldsymbol{\beta}'\mathbf{S}^{-1}\boldsymbol{\beta}\right)\left(\mathbf{i}'\mathbf{S}^{-1}\mathbf{i}\right)-\left(\mathbf{f}'\mathbf{S}^{-1}\boldsymbol{\beta}\right)^2} \\
l_2 &= \frac{\left(\sum_{i=1}^{N}\frac{1}{\theta_i^2}\right)\left(\sum_{i=1}^{N}\frac{f_i\beta_i}{\theta_i^2}\right)-\left(\sum_{i=1}^{N}\frac{\beta_i}{\theta_i^2}\right)\left(\sum_{i=1}^{N}\frac{f_i}{\theta_i^2}\right)}{\left(\sum_{i=1}^{N}\frac{\beta_i^2}{\theta_i^2}\right)\left(\sum_{i=1}^{N}\frac{1}{\theta_i^2}\right)-\left(\sum_{i=1}^{N}\frac{f_i\beta_i}{\theta_i^2}\right)^2} \\
&= \frac{\left(\mathbf{i}'\mathbf{S}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\mathbf{S}^{-1}\boldsymbol{\beta}\right)-\left(\boldsymbol{\beta}'\mathbf{S}^{-1}\mathbf{i}\right)\left(\mathbf{f}'\mathbf{S}^{-1}\mathbf{i}\right)}{\left(\boldsymbol{\beta}'\mathbf{S}^{-1}\boldsymbol{\beta}\right)\left(\mathbf{i}'\mathbf{S}^{-1}\mathbf{i}\right)-\left(\mathbf{f}'\mathbf{S}^{-1}\boldsymbol{\beta}\right)^2}
\end{aligned}
\tag{2.50}
$$

2.4 CHARACTERISTIC PORTFOLIOS

So far in this chapter, we have been using the method of Lagrangian multipliers to find optimal portfolios with various objective functions (variance only for minimum variance portfolio and quadratic utility function for optimal portfolio with forecasts) and portfolio constraints (dollar neutral and beta neutral). The form of these solutions is

$$
\begin{aligned}
\mathbf{w}^* &= \frac{1}{\lambda}\boldsymbol{\Sigma}^{-1}\left(\mathbf{f}-l_1\mathbf{i}-l_2\boldsymbol{\beta}\right), \text{ or} \\
\mathbf{w}^* &= c_1\boldsymbol{\Sigma}^{-1}\mathbf{f}+c_2\boldsymbol{\Sigma}^{-1}\mathbf{i}+c_3\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}.
\end{aligned}
\tag{2.51}
$$

This suggests that the optimal weights are a linear combination of a generic expression — the inverse of the covariance matrix times a vector of attributes. Equation 2.51 contains three examples of attributes: expected return forecasts represented by **f**, the membership in the portfolio by **i**, and the beta by $\boldsymbol{\beta}$. Other examples of attributes can be additional risk

factors and alpha factors, which appear later in the book. This motivates us to define characteristic portfolios for each attribute and express a set of general optimal weights as a combination of them.

For a given attribute **t**, we define the *characteristic portfolio* as the portfolio that has unit exposure to **t** and has the minimum variance. Finding the characteristic portfolio is not hard (Problem 2.12). We have

$$\mathbf{w}_t = \frac{\mathbf{\Sigma}^{-1}\mathbf{t}}{\mathbf{t}'\mathbf{\Sigma}^{-1}\mathbf{t}}. \tag{2.52}$$

There are two special characteristic portfolios. First, if the attribute is 1, then the characteristic portfolio is the minimum variance portfolio of (2.13). Second, if the attribute is beta, then the characteristic portfolio is

$$\mathbf{w}_\beta = \frac{\mathbf{\Sigma}^{-1}\boldsymbol{\beta}}{\boldsymbol{\beta}'\mathbf{\Sigma}^{-1}\boldsymbol{\beta}}. \tag{2.53}$$

According to (2.33), beta is related to the benchmark by

$$\boldsymbol{\beta} = \frac{\mathbf{\Sigma}\mathbf{b}}{\mathbf{b}'\mathbf{\Sigma}\mathbf{b}}.$$

Hence, (2.53) reduces to the benchmark weights **b**. This makes intuitive sense (e.g., Grinold and Kahn, 2000) because all $\beta = 1$ portfolios have the same systematic risk according to CAPM, and only the benchmark portfolio has zero residual risk. Therefore, it has the least total risk among all $\beta = 1$ portfolios.

By definition, a characteristic portfolio has unit exposure in its own attributes. We can also calculate its exposures in other attributes. For instance, the beta exposure for the characteristic portfolio of f is $\boldsymbol{\beta}' \cdot \mathbf{w}_f$, and the percentage invested for the characteristic portfolio of f is $\mathbf{i}' \cdot \mathbf{w}_f$. Using these exposures, we can form optimal weights with desired exposures to various attributes.

Example 2.8

Let us first find the optimal portfolio with unit exposure to f and zero exposure to beta. It is easy to show that $\mathbf{w}_f - \left(\boldsymbol{\beta}' \cdot \mathbf{w}_f\right)\mathbf{w}_\beta$ has zero beta exposure, and its exposure to f is $1 - \left(\boldsymbol{\beta}' \cdot \mathbf{w}_f\right)\left(\mathbf{f}' \cdot \mathbf{w}_\beta\right)$. Therefore, the optimal weights we are looking for are

$$\mathbf{w}^* = \frac{1}{1-\left(\boldsymbol{\beta}'\cdot\mathbf{w}_f\right)\left(\mathbf{f}'\cdot\mathbf{w}_\beta\right)}\left[\mathbf{w}_f - \left(\boldsymbol{\beta}'\cdot\mathbf{w}_f\right)\mathbf{w}_\beta\right]. \tag{2.54}$$

By combining characteristic portfolios of f, beta, and membership, we can find the optimal portfolio with unit exposure to f with both beta neutral and dollar neutral. As we noted above, the solution will be a linear combination of three characteristic portfolios:

$$\mathbf{w}^* = c_1\mathbf{w}_f + c_2\mathbf{w}_\beta + c_3\mathbf{w}_1. \tag{2.55}$$

Imposing exposure constraints leads to a system of linear equations for the unknown coefficients

$$\begin{aligned} c_1 + c_2\left(\mathbf{f}'\cdot\mathbf{w}_\beta\right) + c_3\left(\mathbf{f}'\cdot\mathbf{w}_1\right) &= 1 \\ c_1\left(\boldsymbol{\beta}'\cdot\mathbf{w}_f\right) + c_2 + c_3\left(\boldsymbol{\beta}'\cdot\mathbf{w}_1\right) &= 0 \\ c_1\left(\mathbf{i}'\cdot\mathbf{w}_f\right) + c_2\left(\mathbf{i}'\cdot\mathbf{w}_\beta\right) + c_3 &= 0 \end{aligned} \tag{2.56}$$

The coefficients $c's$ can be found as

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{f}'\cdot\mathbf{w}_\beta & \mathbf{f}'\cdot\mathbf{w}_1 \\ \boldsymbol{\beta}'\cdot\mathbf{w}_f & 1 & \boldsymbol{\beta}'\cdot\mathbf{w}_1 \\ \mathbf{i}'\cdot\mathbf{w}_f & \mathbf{i}'\cdot\mathbf{w}_\beta & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \tag{2.57}$$

provided the inverse matrix exists (Problem 2.13).

Both optimal weights (2.55) and (2.54) have unit exposure to the forecast. Normally, we need to scale these weights by a risk-aversion parameter so that the final optimal portfolios have the targeted level of risk.

PROBLEMS

2.1 Derive the weight (2.7) that gives the minimum variance.

2.2 (Geometric vs. arithmetic average.) Given L periods investment return $r_1, \cdots, r_L$, define arithmetic average return as

$$\mu_a = \frac{1}{L}\sum_{i=1}^{L} r_i .$$

Define geometric average as

$$\mu_g = \left[(1+r_1)(1+r_2)\cdots(1+r_L)\right]^{\frac{1}{L}} - 1 = \left[\prod_{i=1}^{L}(1+r_i)\right]^{\frac{1}{L}} - 1 .$$

(a) Prove $\mu_g \le \mu_a$.

(b) Suppose $r_i = \mu + \sigma\varepsilon_i$, where ε_i's are independent standard normal variables. Prove that as $L \to \infty$, $\mu_g \approx \mu - \frac{1}{2}\sigma^2$.

2.3 (Annualized volatility) It is customary in the financial industry to quote financial statistics on an annualized basis. For example, monthly statistics have to be annualized. Suppose the average monthly return is μ and the monthly standard deviation is σ.

(a) When the individual monthly returns are independent, prove that the annualized average return is

$$\mu_{year} = (1+\mu)^{12} - 1 .$$

the annualized volatility is

$$\sigma_{year} = \sqrt{\left[(1+\mu)^2 + \sigma^2\right]^{12} - (1+\mu)^{24}} .$$

and, when σ is small

$$\sigma_{year} \approx \sigma\sqrt{12}(1+\mu)^{11} .$$

(b) When the individual monthly returns are not independent, we denote the autocorrelation of monthly returns by

$$\operatorname{corr}\left(r_i, r_{i+h}\right) = \rho(h).$$

Show that, when σ is small,

$$\sigma_{\text{year}} \approx \sigma(1+\mu)^{11}\sqrt{12+22\rho(1)+20\rho(2)+\cdots+2\rho(11)}$$
$$= \sigma(1+\mu)^{11}\sqrt{12+2\sum_{i=1}^{11}(12-i)\rho(i)}$$

2.4 Given two random variables r_1, r_2 with volatility σ_1, σ_2 and correlation ρ, define two vectors on a plane, $\overrightarrow{OA}, \overrightarrow{OB}$, with lengths equal to σ_1, σ_2 and the angle between the two vectors given by

$$\cos\theta = \rho.$$

Show that the volatility of $r_1 + r_2$ equals the length of vector $\overrightarrow{AB}$.

2.5 Derive the mean–variance optimal weight (2.22) for a fully invested portfolio.

2.6 Derive the active optimal weight (2.25).

2.7 Prove that the expected return (2.27) of a dollar neutral, long-short portfolio is always nonnegative. When is it zero?

2.8 (Implied correlation.) When option contracts are available both as an index and its underlying stocks, one can use implied volatilities to derive an implied stock correlation, assuming it is the same for all stocks.

(a) Derive an analytic formula for the implied correlation using stock weights in the index, implied stock volatilities, and implied index volatility.

(b) It seems unrealistic to assume all pairwise correlations are the same. Is there another interpretation for the implied correlation?

(c) The covariance matrix of stocks with identical pairwise correlation is of the form

$$\mathbf{\Sigma} = \text{diag}\left(\sigma_1, \cdots, \sigma_N\right) \cdot \mathbf{C} \cdot \text{diag}\left(\sigma_1, \cdots, \sigma_N\right)$$

$$\mathbf{C} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

Show that the inverse of the correlation matrix is

$$\mathbf{C}^{-1} = \frac{1}{(1-\rho)\left[1+(N-1)\rho\right]} \begin{pmatrix} 1+(N-2)\rho & -\rho & \cdots & -\rho \\ -\rho & 1+(N-2)\rho & \cdots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \cdots & 1+(N-2)\rho \end{pmatrix}.$$

(d) For $N=3$, $\rho=0.5$, risk-aversion parameter $\lambda=100$, forecasts of excess return as $\vec{f}=(2\%, 1\%, -3\%)'$, and volatilities as $\vec{\sigma}=(40\%, 30\%, 20\%)'$, calculate optimal portfolio weights in the three stocks and cash using the inverse of the covariance matrix in part (c).

2.9 The beta of a stock or a portfolio depends on what we choose as the market. In fact, it is common to choose an index such as S&P 500 or Russell 3000 as the market in calculating beta. Suppose we first choose S&P 500 as the market and find that Russell 3000 index's beta is 0.9. Next, we choose Russell 3000 as the market instead and find that S&P 500 index's beta is 0.95. Therefore, both beta of one index vs. the other is less than 1. Can this be true?

2.10 (a) Given **I**, an $N \times N$ identity matrix, and a vector **a** of length N, prove that

$$\left(\mathbf{I}+\mathbf{a}\mathbf{a}'\right)^{-1} = \mathbf{I} - \frac{\mathbf{a}\mathbf{a}'}{1+\mathbf{a}'\cdot\mathbf{a}}.$$

(b) Prove the inverse matrix of (2.34) is (2.36).

2.11 Derive the optimal portfolio weights (2.49) and (2.50) by solving the optimization problem

$$\begin{aligned} &\text{Maximize } \mathbf{w}' \cdot \mathbf{f} - \frac{1}{2}\lambda\left(\mathbf{w}'\mathbf{S}\mathbf{w}\right) \\ &\text{subject to: } \mathbf{w}' \cdot \boldsymbol{\beta} = 0 \text{ and } \mathbf{w}' \cdot \mathbf{i} = 0 \end{aligned}.$$

2.12 Find the weights of a characteristic portfolio with minimum variance and unit exposure to stock attribute *t*.

2.13 Prove that the inverse in (2.57) exists when the vectors **f**, **β**, and **i** are not linearly dependent.

REFERENCES

Grinold, R.C. and Kahn, R.N., *Active Portfolio Management*, McGraw-Hill, New York, 2000.

Jacobs, B.I., Levy, K.N., and Starer, D., On the optimality of long-short strategies, *Financial Analyst Journal,* Vol. 52, No. 5, 81–85, 1996.

Lintner, J., The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics,* February, 13–37, 1965.

Markowitz, H.M., *Portfolio Selection: Efficient Diversification of Investments*, Cowles Foundation Monograph 16, Yale University Press, New Haven, CT, 1959.

Sharpe, W.F., Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance,* Vol. 19, No. 3, 425–442, 1964.

Tobin, J., Liquidity preference as a behavior toward risk, *Review of Economic Studies,* February, 65–86, 1958.

CHAPTER 3

Risk Models and Risk Analysis

THE CAPITAL ASSET PRICING MODEL (CAPM), discussed in the previous chapter, was originally developed as an equilibrium pricing model and not as a risk model *per se*. As a pricing model, its function is to provide return expectations of individual stocks given their betas vs. a market portfolio and expected excess return of the market, that is,

$$\mathrm{E}\left(r_i - r_f\right) = \beta_i \left[E\left(r_M\right) - r_f \right]. \tag{3.1}$$

In essence, CAPM states that the market should set prices of stocks in a way such that their expected returns are proportional to their systematic risks measured by beta. Specific risks, on the other hand, can be diversified away by holding portfolios of stocks and therefore shall not be rewarded with excess returns.

Readers may have noticed this is not the way we used CAPM in the previous chapter. There we used it as a risk model, i.e., the total risk of a stock or a portfolio consists of systematic risk measured by beta and stock-specific risk, while leaving the expected returns aside. From a statistical standpoint it can be argued that both models originate from the same equation; however, the pricing model interprets the equation by expectation, but the risk model interprets the equation by variance.

This subtle yet obvious difference seems to reflect how academics and industry practitioners view and construct asset-pricing models differently. For example, after a long list of pricing anomalies was discovered contradicting CAPM prediction, variants of alternative asset-pricing

models were proposed in the academia to describe how assets are priced in the equilibrium. For example, Fama and French (1992) proposed a three-factor model with beta, market capitalization, and book-to-price ratio to describe prices. But from the practitioners' point of view, this simply indicates that there exist other priced factors in addition to the market beta. Still, risk models should encompass more. Specifically, some factors may not be priced or rewarded unconditionally through time, but they do differentiate cross-sectional security returns. In other words, it is conceivable to assume that there are nonpriced risk factors whose returns exhibit a low unconditional mean but high unconditional variance. Finding other priced factors would improve the descriptive accuracy of CAPM as a pricing model, but it would carry little implication for risk modeling. As a consequence, many practitioners use arbitrage pricing theory (APT) to model risk models by incorporating a set of nonpriced risk factors in addition to priced factors, thereby constructing risk-adjusted portfolios and managing portfolio risk in general. As readers shall discover later in the book, many alpha models take on the same form as the risk models.

This is the approach we take in this book. In this chapter, we will introduce multifactor risk models that are based on APT. We first briefly describe the APT model. Then, we outline three different variants of multifactor models: macroeconomic factor model, fundamental factor model, and statistical factor model. We also present concepts of risk contribution, which are important in risk management practice.

3.1 ARBITRAGE PRICING THEORY AND APT MODELS

APT has two main ingredients. The first is an assumption regarding the security-return-generating process, and the second is the law of one price — two identical items must have the same price. The return-generating process requires that returns of any stocks be linearly related to a set of factors or indices

$$r_i = b_{i0} + b_{i1} I_1 + \cdots + b_{iK} I_K + \varepsilon_i . \tag{3.2}$$

In this case, there are K factors, $I_1, \cdots, I_K$, and b_{ij} is the sensitivity or exposure of the i-th stock to the j-th factor. The last term ε_i is the stock-specific return with zero mean. It is assumed that all specific returns are uncorrelated with each other, as well as all the factors.

Note that Equation 3.2 is remarkably similar to Equation 2.34 in that it is a generalization of a single-factor model. The covariance matrix of stock returns given by (3.2) is then

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_I\mathbf{B}' + \mathbf{S}. \tag{3.3}$$

The matrix **B** is the exposure matrix given by

$$\mathbf{B} = \begin{pmatrix} b_{11} & \cdots & b_{1K} \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{NK} \end{pmatrix}_{N\times K} = \left(\mathbf{b}_1, \cdots, \mathbf{b}_K\right). \tag{3.4}$$

The vector $\mathbf{b}_k$ consists of stocks' exposures to the k-th factor. The matrix $\mathbf{\Sigma}_I$ is the factor return covariance matrix

$$\mathbf{\Sigma}_I = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1K} \\ \vdots & \ddots & \vdots \\ \sigma_{K1} & \cdots & \sigma_{KK} \end{pmatrix}_{K\times K}. \tag{3.5}$$

Finally, similar to CAPM model, the matrix **S** is the diagonal of specific risks.

However, there are important differences between the CAPM risk model and APT risk model. On the one hand, in a CAPM model, the factor is explicitly prescribed as the market return, and the exposure of a stock to the factor is defined as the *beta of the stock*. On the other hand, APT is very general. In an APT model, we do not know what the underlying factors are or the number of factors. Furthermore, APT does not specify how to measure stocks' exposure to the factors.

The lack of a definitive form for APT models has several consequences. First, it is challenging to test the theory empirically, both in terms of the return-generating process and the pricing mechanism. Second, its flexibility also provides multiple approaches to the empirical investigation of stock returns. As a result of extensive interest and research effort from both the academic and investment communities, there are several competing versions of multifactor risk models.

In general, we classify the multifactor risk models into three categories: macroeconomic factor models, fundamental factor models, and statistical factor models. This classification, to a large extent, is based on how each model selects the factors. Macroeconomic factor models are the most intuitive. Cyclical phenomena such as movements in interest rates create risk for stocks. The first commercial risk model (BARRA) is a so-called

fundamental approach. In the early 1980's many portfolio mangers found the concept of beta too academic. So, the fundamental risk model evolved to capture some of a stock's (portfolio's) risk by modeling well understood stock attributes. These fundamentals include value (price ratios), dividend policy, earnings variability, firm size and so on. Statistical factor models are based on factors that are derived by statistical techniques such as principal component analysis. We shall cover them each in detail. But once the factors are selected, all three model approaches use the same method to derive factor returns and their covariance matrix. For a comparison study, see Connor (1995).

3.1.1 Macroeconomic Factor Models

The fact that stock prices are sensitive to macroeconomic factors, such as interest rate, inflation, and growth of the economy, should not come as a surprise (Table 3.1). It is quite intuitive and based squarely in valuation theory. In a straightforward discounted cash flow model, stock price is the present value of future payments received by shareholders (examples are the dividend discount model and the earnings cash flow model). Thus, macroeconomic factors that affect both company earnings and the required rate of return by investors would impact stock prices and do so differently.

For example, when the Federal Reserve cuts the interest rate, the stock market as a whole generally responds favorably, because lower interest rates not only stimulate the economy resulting in greater aggregate earning growth, but also reduce the required rate of return by shareholders. That is, stocks have positive durations — like bonds (Leibowitz et al. 1989). This effect is often stronger for companies with poorer investment quality because of financial or operational leverages. Another example of a macro factor is the oil price. In general, a higher oil price exerts a drag on the economy and, therefore, has a negative impact on the stock market (akin to a tax). But the impact would be different for an airline where oil price is an input cost, an oil producer where oil price reflects the selling price, and a software company that is relatively insulated from the oscillations of oil price.

In the 1980's Salomon Brothers (now Citigroup) developed a comprehensive macro-based risk model for US stocks (and later global stocks). This was the original application of the four-factor macro APT model posited by Chen, Ross and Roll (1986). During the same time period, the Northfield Company also began to produce a macroeconomic portfolio risk tool. The original Salomon Brothers model estimates stock sensitivities (betas)

TABLE 3.1 Commonly Used Macroeconomic Factors

	Macroeconomic Factor
1	Market return
2	Change in short-term interest rate
3	Change in industrial production
4	Change in inflation
5	Term spread
6	Default spread
7	Change in oil price

to a set of factors: economic growth, long-term rates, short-term rates, risky bond spreads (credit), inflation, exchange rate movements, small cap premia and an overall market factor (CAPM beta). Effectively, this type of macro-based risk model "decomposes" the simple one-factor CAPM approach into several other cyclical variables. However, this creates an econometric problem due to a multicolinearity of factors. For example, interest rates and the overall market are linked in themselves. Also, as credit spreads fall or small cap stocks rise, other things being equal, the overall market also reacts. Thus, Citigroup researchers and others use advanced econometric procedures to iteratively purge some macro-factors from the influence of others (Sorensen et al. 1998). The goal is to specify the model so that each factor is additive and statistically significant.

With the selection and refined specification of these macroeconomic factors, one then proceeds to estimate the exposures of each stock to the select factors through a time-series regression

$$r_{it} - r_{ft} = \alpha_i + \beta_i \left(r_{Mt} - r_{ft} \right) + \sum_{k=1}^{K} b_{ki} I_{kt} + \varepsilon_{it}. \tag{3.6}$$

The index i is for stocks, the index t for time periods, and k for factors. The regression finds the alpha for the stock, its beta exposure to the market, and the exposures to the factors. It is typically carried out with a rolling window of many months. When the regression is completed for each stock, we obtain the exposure matrix in the form of (3.4). The historical macroeconomic factor covariance matrix is the factor return covariance matrix, and the standard error of each regression gives rise to specific risk of each stock.

In a macroeconomic factor model such as (3.6), because factor values are predetermined, the cross-sectional return variation associated with

a factor depends on the cross-sectional variation of the factor exposures. For instance, the cross-sectional variation associated with the market factor for the time period t is

$$\mathrm{var}(\boldsymbol{\beta})(r_{Mt} - r_{ft})^2. \tag{3.7}$$

The vector $\boldsymbol{\beta}$ consists of betas for all stocks. Therefore, if the market excess return for the time period is minimal, the model would imply it would contribute little to the cross-sectional variation of stock returns. The same is true for other macroeconomic factors when there are little economic shocks. What else can explain the cross-sectional variability of stock returns that seems to be pervasive in the stock market?

3.1.2 Fundamental Factor Models

Return and risk are often inseparable. If we are looking for the sources of cross-sectional return variability, we need to look no further than places where investors search for excess returns. So how do investors search for excess returns? One way is doing fundamental research, in which analysts first carry out an industry analysis, and then follow it by a fundamental analysis of companies, along the lines of valuation, quality, and investor expectations, among other things. In essence, fundamental research aims to forecast stock returns by analyzing the stocks' fundamental attributes. Fundamental factor models follow a similar path in using the stocks' fundamental attributes to explain the return difference between stocks.

Using BARRA's (1998) U.S. Equity model as an example, there are two groups of fundamental factors: industry factors and style factors. (The latter are also referred to as risk indices. Industry factors are based on industry classification of stocks.) Borrowing from our earlier example, one would naturally expect an airline stock and a software stock to behave differently because they belong to different industries. The source of this return difference might well be the oil price, but it could also be some other underlying economic factors. In this case, the airline stock has an exposure of one to the airline industry and zero to all other industries. Similarly, the software company only has exposure to the software industry. In most fundamental factor models, the exposure is identical and is equal for all stocks in the same industry. For conglomerates that operate in multiple businesses, they can have fractional exposures to multiple industries. All together, there are between 50 and 60 industry factors.

TABLE 3.2 Commonly Used Fundamental Factors

Category	Fundamental Factor
Industry	Industries
Style	Size
Style	Book-to-price
Style	Earning yield
Style	Dividend yield
Style	Momentum
Style	Growth
Style	Earning variability
Style	Financial leverage
Style	Volatility
Style	Trading activity

The second group of factors relates to the company-specific attributes. Table 3.2 provides a list of commonly used style factors; some are intuitive, whereas others are not. Moreover, many of them are correlated to the simple CAPM beta, leaving some econometric issues as described above for macro models. For example, the size factor is based on the market capitalization of a company. The fact that market participants classify stocks and stock mutual funds into size categories, such as large cap, mid cap, small cap, and even micro cap, reflects different behaviors of these stocks as a source of cross-sectional variability. The next factor book-to-price, also referred to as book-to-market, is the ratio of book value to market value of a company, one of the value measures. To a value investor, a stock with a high book-to-price ratio would appear cheap, whereas a stock with a low book-to-price ratio looks expensive (more on book-to-price as an alpha factor in Chapter 4). However, to a growth investor, a low book-to-price ratio reflects the prospect of high growth expected by the market. A growth investor would be willing to pay for that growth if the expectation is justified. Thus, book-to-price, among a few other factors, defines the line between value stocks and growth stocks.

There have been considerable controversies surrounding the size factor and book-to-price factor. Historically, small cap stocks have outperformed large cap stocks, whereas high book-to-price stocks have done better than low book-to-price stocks. One explanation would be that small and value stocks bear more risk than large and growth stocks; therefore, they have should have high returns. Another explanation is that they represent market inefficiency — the small and value premiums are caused by investors'

behavior that is inconsistent with rational decision-making. We shall return to book-to-price when we discuss alpha factors in the later chapters. For now, we recognize it as a fundamental factor that is capable of explaining cross-sectional return differences among stocks.

The other factors are briefly described in the following text. (For detailed description, see BARRA United States Equity Version 3 Handbook.) The next two factors — earning yield and dividend yield — are also valuation measures. The momentum factor measures price momentum and relative strength. The growth factor represents growth in earning and revenue based on either past history or forward projections provided by the institutional brokers' estimate system (IBES). Earning variability is the historical standard deviation of earning per share. Financial leverage is the debt-to-equity ratio. Volatility is essentially the standard deviation of the residual stock returns. Trading activity is the turnover of shares traded. A stock's exposures to these factors are quite simple: they are simply the values of these attributes. One typically normalizes these factors cross-sectionally so they have mean 0 and standard deviation 1.

Once the fundamental factors are selected and the stocks' normalized exposures to the factors are calculated for a time period, a cross-sectional regression against the actual return of stocks is run to fit cross-sectional returns with cross-sectional factor exposures. The regression coefficients are called *returns on factors* for the time period. This procedure bears resemblance to the second pass of the Fama–MacBeth (1976) regression procedure.

For a given period t, the regression is run for the returns of the subsequent period against the factor exposure known at the time t

$$r_i^{t+1} = b_0^t + b_1^t I_{i,1}^t + \cdots + b_K^t I_{i,K}^t + \varepsilon_i. \tag{3.8}$$

To obtain the covariance matrix of factor returns, one runs the cross-sectional regression for multiple periods and then calculates the covariance matrix based on the times series of factor returns. Note the following remarks:

- There are several practical issues for the model estimation. First, it is important for a risk model to fit a large percentage of market capitalization. This might lead one to use a weighted regression, with weights being the market cap of the stocks. Second, it is reasonable to expect that the most recent factor returns are more informative

to future return variances and covariances. Hence, one can put higher weights on the more recent periods and lower weights on the distant periods. This is typically achieved by a weight scheme that decays in the time, i.e., $\cdots, \omega^{t-T}, \cdots, \omega^2, \omega, 1$, with the weight for the most recent period being 1 and $\omega < 1$. The half-life of the weights is $H = -\ln 2 / \ln \omega$.

- One additional issue is the estimation of stock-specific risks. Ideally, for each stock, one would form a time series of residuals from the Fama–MacBeth regression and use the volatility of the time series as the specific risk. In practice, this is very hard to do. For instance, some newly issued stocks simply have not been around long enough. For this and other reasons, the specific risks are not estimated directly and individually. They are partially estimated based on some of the same fundamental characteristics that go into the factor model.

In summary, although the generic multifactor model provides a clear theoretical foundation, its actual construction is a daunting task. That is why many quantitative managers rely on commercially available risk models and spend most of their time and energy on finding an alpha model for forecasting future returns. In the end, most good risk models could have similar estimates of the total volatility or benchmark-relative risk of a given portfolio.

3.1.3 Statistical Factor Models

Statistical models are another type of multifactor model. Unlike the previous two types, they pay no attention to either the macro or company fundamental data and are purely based on historical returns. The factors in a statistical model are derived from the principal component analysis of returns. The good news is that they literally exploit price information and thus are good at explaining risk. The bad news is that they are merely fitting price data which can be noise, and since they lack any model of economic causality they may be weak at forecasting risk for longterm horizons.

Principal component analysis provides a statistical method to analyze the underlying structure of data sets without any prescribed assumption. Its basic intuition is that it asks what combination of raw data gives rise to the maximum variance among all possible linear combinations. One good example of its application in finance is the term structure of interest rates, which corresponds to yields of bonds with different maturities. For

any given period, yields of all maturities change differently. It would seem we need many factors to describe the change in the yield curve. However, principal component analysis reveals that three components consisting of a linear combination of different points on the curve account for the majority of variation of all the changes along the whole curve. The first component corresponds to the level of yield curve, the second corresponds to the slope, and the third corresponds to the curvature.

Suppose the raw covariance matrix of stock returns is an $N \times N$ symmetric matrix $\mathbf{\Sigma}$, with N being the number of stocks. Then, the principal component analysis would decompose it into

$$\mathbf{\Sigma} = \mathbf{LPL}', \tag{3.9}$$

where $\mathbf{P}$ is a diagonal matrix $\mathbf{P} = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_N\right)$, with $\lambda_1 > \lambda_2 > \cdots > \lambda_N > 0$ being the eigenvalues of matrix $\mathbf{\Sigma}$. The matrix $\mathbf{L}$ is an orthogonal matrix consisting of the eigenvectors

$$\mathbf{LL}' = \mathbf{I}, \tag{3.10}$$

with $\mathbf{I}$ being the identity matrix. We shall denote L_{ij} as the matrix element, $\mathbf{l}_i$ as the row vector, and $\mathbf{L}_j$ as the column vector, i.e.,

$$\mathbf{L} = \begin{pmatrix} L_{11} & \cdots & L_{1N} \\ \vdots & \ddots & \vdots \\ L_{N1} & \cdots & L_{NN} \end{pmatrix} = \left(\mathbf{L}_1, \cdots, \mathbf{L}_N\right) = \begin{pmatrix} \mathbf{l}_1' \\ \vdots \\ \mathbf{l}_N' \end{pmatrix}. \tag{3.11}$$

Then

$$\mathbf{l}_i' \cdot \mathbf{L}_j = \delta_{ij} = \begin{cases} 1, \text{ if } i = j \\ 0, \text{ if } i \neq j \end{cases}. \tag{3.12}$$

Comparing Equation 3.9 to Equation 3.3, we can conclude that Equation 3.9 represents a model of N orthogonal factors, with λ's being their variances and $\mathbf{L}$ being the exposure matrix of each security to the N orthogonal principal component factors. Specifically, the row vector $\mathbf{l}_j'$ is the exposure of the j-th stock to the N factors. They are also called the *factor loadings* for each stock.

To illustrate these relationships mathematically, let us assume that $\mathbf{R} = \left(r_{jt}\right)_{N\times T}$ is an $(N \times T)$ matrix representing returns of N securities during T nonoverlapping periods, and $\mathbf{Q} = \left(q_{jt}\right)_{N\times T}$ is also an $(N \times T)$ matrix reflecting returns of the N orthogonal principal component factors during the same T periods. **R** (security returns) can be expressed as the product of both **L** (factor exposure matrix) and **Q** (factor returns) as follows:

$$\mathbf{R} = \mathbf{LQ}\,. \tag{3.13}$$

Because $\mathbf{LL}' = \mathbf{I}$, the factor return matrix **Q** can be derived by

$$\mathbf{Q} = \mathbf{L}'\mathbf{R}\,. \tag{3.14}$$

Given **Q**, we can now derive the return covariance matrix of the N principal component factors. As shown in the following proof, it is equal to the diagonal matrix of eigenvalues (**P**).

$$\begin{aligned}\hat{\mathbf{\Sigma}} &= \tfrac{1}{T}\cdot\mathbf{QQ}' = \mathbf{L}'(\tfrac{1}{T}\cdot\mathbf{R}\times\mathbf{R}')\mathbf{L}\\ &= \mathbf{L}'\mathbf{\Sigma}\mathbf{L} = \mathbf{L}'\mathbf{L}\mathbf{P}\mathbf{L}'\mathbf{L} = \mathbf{P}\end{aligned} \tag{3.15}$$

- Given **R** = **LQ**, each row vector of **L** corresponds to the factor exposures of each individual security, whereas $\mathbf{Q} = \mathbf{L}'\mathbf{R}$ means that each column vector of **L** represents "security exposures" of each individual orthogonal principal component factor. The reader must be careful not to confuse one with the other.

Comparing Equation 3.9 with Equation 3.3 also reveals a big difference between the two. There is no specific risk in Equation 3.9. The reason for this is obvious: we have the same number of factors as the number of stocks. In reality, the number of factors is much smaller, possibly in single digit (Connor and Korajczyk 1988). If we choose K factors for a statistical factor, then, in theory, the percentage of variance captured by the model is

$$\sum_{i=1}^{K}\lambda_i \Big/ \sum_{i=1}^{N}\lambda_i\,.$$

The selection of principal components as statistical factors is an important step in the modeling process. If there are too few factors, then the

model does not adequately describe systematic risks. If there are too many factors, the model might be overly fit; some of the factors might be noise and lose their significance over subsequent periods. One mathematical tool that offers some help is the theory of random matrix (see, for example, Plerou et al. 1999). By comparing the distribution of eigenvalues with that of a random matrix, one might be able to select only the factors that are statistically significant and leave out other noise factors.

3.2 RISK ANALYSIS

Previously, we presented a general framework of multifactor models and described three different types of multifactor models. The remainder of this chapter is devoted to portfolio risk analysis under this framework. Risk analysis is an integrated part of portfolio management. It serves at least two purposes. First, it reveals where the risks are present in an existing portfolio. An efficient portfolio should have risks in places where we expect excess return, whether it is in sectors, alpha factors, or individual stocks. This can be done by portfolio risk attribution. The second purpose of risk analysis is to see how the portfolio's risk characteristics might change if we were to change the portfolio weights. This is achieved through analyzing marginal contribution to risk (MCR). We discuss the marginal contribution to risk first.

3.2.1 Marginal Contribution to Risk

Given risk models, such as the ones in (3.3), (3.4), and (3.5), and portfolio weights $\mathbf{w} = \left(w_1, \cdots, w_N\right)'$, the total portfolio variance is

$$\sigma^2 = \mathbf{w}'\mathbf{\Sigma}\mathbf{w} = \left(\mathbf{w}'\mathbf{B}\right)\mathbf{\Sigma}_\mathbf{I}\left(\mathbf{B}'\mathbf{w}\right) + \mathbf{w}'\mathbf{S}\mathbf{w} . \tag{3.16}$$

The first term is the systematic risk, with $\mathbf{w}'\mathbf{B}$ being the portfolio exposure to risk indices or factors. The second term represents the specific variance. Equation 3.16 is valid for absolute risk as well as active risks. The standard deviation or tracking error of the portfolio is then

$$\sigma = \sqrt{\left(\mathbf{w}'\mathbf{B}\right)\mathbf{\Sigma}_\mathbf{I}\left(\mathbf{B}'\mathbf{w}\right) + \mathbf{w}'\mathbf{S}\mathbf{w}} . \tag{3.17}$$

The marginal contribution to risk (MCR) from stock i is defined as the *partial derivative* of σ with respect to its weight: $\mathrm{MCR}_i = \partial\sigma / \partial w_i$. It

measures the rate of change in σ, as the weight w_i changes by an infinitesimal amount. We can calculate the vector of MCR as

$$\mathbf{MCR}=\frac{\partial\sigma}{\partial\mathbf{w}}=\frac{\mathbf{B}\boldsymbol{\Sigma}_I\mathbf{B}'\mathbf{w}+\mathbf{S}\mathbf{w}}{\sqrt{\left(\mathbf{w}'\mathbf{B}\right)\boldsymbol{\Sigma}_I\left(\mathbf{B}'\mathbf{w}\right)+\mathbf{w}'\mathbf{S}\mathbf{w}}}=\frac{\mathbf{B}\boldsymbol{\Sigma}_I\mathbf{B}'\mathbf{w}+\mathbf{S}\mathbf{w}}{\sigma}. \tag{3.18}$$

One can similarly define marginal contribution to systematic risk and marginal contribution to specific risk because it is common in practice to look at these two sources of risk separately. Mathematically, we have

$$\mathbf{MCR}_{\text{systematic}}=\frac{\mathbf{B}\boldsymbol{\Sigma}_\mathbf{I}\mathbf{B}'\mathbf{w}}{\sqrt{\left(\mathbf{w}'\mathbf{B}\right)\boldsymbol{\Sigma}_\mathbf{I}\left(\mathbf{B}'\mathbf{w}\right)}}=\frac{\mathbf{B}\boldsymbol{\Sigma}_\mathbf{I}\mathbf{B}'\mathbf{w}}{\sigma_{\text{systematic}}}, \tag{3.19}$$

and

$$\mathbf{MCR}_{\text{specific}}=\frac{\mathbf{S}\mathbf{w}}{\sqrt{\mathbf{w}'\mathbf{S}\mathbf{w}}}=\frac{\mathbf{S}\mathbf{w}}{\sigma_{\text{specific}}}. \tag{3.20}$$

We have defined the portfolio systematic and specific risks. Combining the three definitions yields the relationship between the three:

$$\frac{\sigma_{\text{systematic}}}{\sigma}\mathbf{MCR}_{\text{systematic}}+\frac{\sigma_{\text{specific}}}{\sigma}\mathbf{MCR}_{\text{specific}}=\mathbf{MCR}. \tag{3.21}$$

MCR is a weighted average of systematic MCR and specific MCR, with the weights being the portions of systematic risk and specific risk in the total risk. Note that the weights do not sum to one; instead their squares sum to one.

Example 3.1

The interpretation of MCR is rather straightforward. For instance, suppose MCR_i is 0.1, then an increase of 1% in the weight w_i should increase the portfolio risk by 0.1%, whereas a decrease of the same magnitude would decrease the portfolio risk by the same amount.

3.2.2 Group Marginal Contribution to Risk

We note that this simple interpretation is valid only if the change in the portfolio weight w_i comes at the expense of cash. In most cases, one simply

cannot change the weight of a single security alone. For example, we cannot adjust the weight of a stock in a fully invested long-only portfolio without adjusting the weight of another stock. Or, in a dollar neutral long-short portfolio, if we increase the long of a stock, then we have to either decrease the long of another stock or increase the short of another stock in order to maintain the dollar neutrality.

To have a meaningful interpretation of MCR, it is better to consider it in combination of two or more stocks. For instance,

$$\mathrm{MCR}_{i,j} = \mathrm{MCR}_i - \mathrm{MCR}_j \tag{3.22}$$

measures the marginal contribution of increasing weight w_i and simultaneously decreasing weight w_j by the same amount or, in other words, buying stock *i* and at the same time selling stock *j*. This can be useful in making a trading decision from the risk perspective. For example, if $\mathrm{MCR}_i = 0.1$, $\mathrm{MCR}_j = 0.2$, then $\mathrm{MCR}_{i,j} = -0.1$, implying that a trade of buying 1% of stock *i* and selling 1% of stock *j* would lower the risk by 0.1%.

Trading decision is not necessarily limited to pairs. It can be a group of stocks, as long as the aggregated change of all the weights is zero. The requirement can be achieved by using a vector $\mathbf{t} = (t_1, \cdots, t_N)'$ representing proportions of trading in each stock and letting $\mathbf{t}' \cdot \mathbf{i} = 0$. Recall that $\mathbf{i}$ is the vector of ones. Then, the marginal contribution to risk for the trade vector $\mathbf{t}$ would be

$$\mathrm{MCR}_{\mathbf{t}} = \mathbf{t}' \cdot \mathbf{MCR}\,. \tag{3.23}$$

Example 3.2

For example, a vector $\mathbf{t} = (1, 0.5, -0.75, -0.75, 0, \cdots, 0)'$ would imply buying one unit of stock 1, buying a half unit of stock 2, and selling three quarter units of stock 3 and stock 4. The unit might be 1% or any other trading size. If $\mathbf{MCR} = (0.1, 0.2, 0.3, 0.3, \cdots)'$, then

$$\mathrm{MCR}_{\mathbf{t}} = \mathbf{t}' \cdot \mathbf{MCR} = 1 \cdot 0.1 + 0.5 \cdot 0.2 - 0.75 \cdot 0.3 - 0.75 \cdot 0.3 = -0.25\,.$$

This representation is especially useful when we analyze the marginal contribution of a sector in a long-short portfolio with sector-neutral constraints (see Problem 3.4). Additional constraints may be placed on the trades. For example, if the portfolio is beta neutral and is required to

remain so after the trades, then the vector **t** must also satisfy the equation $\mathbf{t}' \cdot \boldsymbol{\beta} = 0$.

3.2.3 Risk Contribution

Contribution to risk, or simply risk contribution, is a different way to analyze portfolio risk. In contrast to MCR, which is a dynamic concept regarding changes to a portfolio, contribution to risk is a static measure of how the current portfolio risk is allocated among its constituents. For portfolio managers, it is important to understand the makeup of the portfolio risk so they know the bets are placed appropriately. For instance, for a portfolio with a given level of tracking error against a benchmark, we are interested in knowing how much of that tracking error is made up of systematic and specific risks. Alternatively, we might be interested in the contribution to risk from all the sectors. For a long-short portfolio, it is common to ask how much risk is from the long side and how much from the short side. Because contribution to risk adds up to the total risk, the concept is also referred to as risk budgets. When one actively uses risk budgets to construct portfolios instead of passively monitoring portfolio risk contribution, the process is often called *risk budgeting*.

The concept of risk contribution is widely used in both risk management and risk budgeting practices, in the areas of asset allocation as well as active portfolio management (Litterman 1996, Lee and Lam 2001, Wander et al. 2002, Winkelmann 2004). Despite the ubiquitous presence of risks, questions have remained regarding their validity. The questions stem from both the simple belief that risks are nonadditive and a lack of financial intuition behind mathematical definitions of these concepts. In the remainder of the chapter, we shall define risk contribution first and then present a financial interpretation in terms of loss contribution.

The definition of risk contribution is related to the marginal contribution to risk. For contribution to total risk, we have

$$\mathrm{CR}_i = w_i \frac{\partial \sigma}{\partial w_i}. \tag{3.24}$$

The vector form of Equation 3.24, using Equation 3.18, is

$$\mathbf{CR} = \mathbf{w} \otimes \frac{\partial \sigma}{\partial \mathbf{w}} = \frac{\mathbf{w} \otimes \left(\mathbf{B}\boldsymbol{\Sigma}_I \mathbf{B}'\mathbf{w} + \mathbf{S}\mathbf{w} \right)}{\sigma}. \tag{3.25}$$

The operator $\otimes$ denotes element-by-element multiplication, i.e., $(\mathbf{A} \otimes \mathbf{B})_i = \mathbf{A}_i \mathbf{B}_i$ for two vectors of the same length. It is easy to prove (Problem 3.7) that contributions to risk from all stocks add up to the total risk, i.e.,

$$\mathbf{CR}' \cdot \mathbf{i} = \sum_{i=1}^{N} \mathrm{CR}_i = \sum_{i=1}^{N} w_i \frac{\partial \sigma}{\partial w_i} = \sigma . \tag{3.26}$$

Hence, Equation 3.26 constitutes as a risk decomposition of the total risk. We refer to it as the risk budget equation. Dividing it by the total risk σ, we obtain a percentage contribution to risk (PCR) from each stock:

$$\mathrm{PCR}_i = \frac{w_i \dfrac{\partial \sigma}{\partial w_i}}{\sigma}, \quad \sum_{i=1}^{N} \mathrm{PCR}_i = 1. \tag{3.27}$$

Example 3.3

Let us look at a portfolio with two securities and with a covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{3.28}$$

Then, the total risk with $\mathbf{w} = (w_1, w_2)'$ is

$$\sigma = \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2\rho w_1 w_2 \sigma_1 \sigma_2} . \tag{3.29}$$

The risk contribution and PCR are

$$\begin{aligned} \mathrm{CR}_i &= \frac{w_i^2\sigma_i^2 + \rho w_1 w_2 \sigma_1 \sigma_2}{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2\rho w_1 w_2 \sigma_1 \sigma_2}}, \\ \mathrm{PCR}_i &= \frac{w_i^2\sigma_i^2 + \rho w_1 w_2 \sigma_1 \sigma_2}{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2\rho w_1 w_2 \sigma_1 \sigma_2}. \end{aligned} \tag{3.30}$$

Thus, PCR is equivalent to variance decomposition. The denominator is the total variance of the portfolio whereas the numerator is the variance and covariance attributable to each stock. Although it is true that the

volatility or standard deviation is nonadditive, the variance and covariance are.

We can write PCR as

$$\mathrm{PCR}_i = \frac{\mathrm{cov}\left(w_i r_i, w_1 r_1 + w_2 r_2\right)}{\mathrm{cov}\left(w_1 r_1 + w_2 r_2, w_1 r_1 + w_2 r_2\right)} = \beta_{w_i r_i, r}. \tag{3.31}$$

Written this way, the PCR is the ratio of beta of the return component of a stock to the return of the whole portfolio.

3.2.4 Economic Interpretation of Risk Contribution

The interpretation of risk contribution is not as simple as the MCR. First, a mere mathematical decomposition of risk does not necessarily qualify it as risk contribution (Sharpe 2002). Second, because it is mathematically defined through marginal contribution to risk, various authors have attempted to explain it in terms of the latter. For example, Grinold and Kahn (2000) interpret it as "relative marginal contribution to risk." Earlier, Litterman (1996) also interpreted risk contribution in terms of marginal analysis. However, these types of interpretations do not seem to offer anything new beyond a recast of MCR. Because of the difficulty, some expressed critical views toward risk contribution and even suggested abandoning the concept altogether.

Does risk contribution have an independent, intuitive financial interpretation? The answer is yes. The interpretation is loss contribution and percentage contribution to loss. One of the common pressing questions facing portfolio managers in the event of a sizable loss is what underlying components are directly responsible for the disappointing portfolio losses. This question can be addressed by using the theory of conditional expectation.

We present the solution for a two-security portfolio and leave the general case as an exercise (Problem 3.6). Suppose the portfolio suffered a loss of size L; the expected percentage contribution to loss L (PCL) from security i is the conditional expectation divided by the total loss L:

$$\mathrm{PCL}_i = \frac{\mathrm{E}\left(w_i r_i \mid w_1 r_1 + w_2 r_2 = L\right)}{L}. \tag{3.32}$$

According to the theory of conditional distribution, the conditional expectation of a normal variable equals the unconditional mean plus its

beta (which equals PCR, according to [3.31]) to the given variable, in this case, the total portfolio return, times the difference between the given variable and its unconditional mean. We have

$$\begin{aligned} \mathrm{PCL}_1 &= \frac{w_1\mu_1}{L} + \mathrm{PCR}_1\left(1 - \frac{w_1\mu_1}{L} - \frac{w_2\mu_2}{L}\right) = \mathrm{PCR}_1 + \frac{D_1}{L} \\ \mathrm{PCL}_2 &= \frac{w_2\mu_2}{L} + \mathrm{PCR}_2\left(1 - \frac{w_1\mu_1}{L} - \frac{w_2\mu_2}{L}\right) = \mathrm{PCR}_2 + \frac{D_2}{L} \end{aligned} . \tag{3.33}$$

We have defined

$$\begin{aligned} D_1 &= \mathrm{PCR}_2 w_1\mu_1 - \mathrm{PCR}_1 w_2\mu_2 \\ D_2 &= \mathrm{PCR}_1 w_2\mu_2 - \mathrm{PCR}_2 w_1\mu_1 \end{aligned} . \tag{3.34}$$

It is easy to see that $\mathrm{PCL}_1 + \mathrm{PCL}_2 = \mathrm{PCR}_1 + \mathrm{PCR}_2 = 1$ because $D_2 = -D_1$. Equation 3.33 shows that the expected PCL bear close relationship to PCR. In fact, they are identical if $D_2 = -D_1 = 0$. The two are very close otherwise if the loss is large compared to D_1 and D_2. There are three instances in which $D_2 = -D_1 = 0$.

- Case I: First, if μ_1 and μ_2 are both zero, then $D_2 = D_1 = 0$, implying $\mathrm{PCL}_i = \mathrm{PCR}_i$ for any loss L. Therefore, PCR perfectly explains the expected PCL. This case applies to short investment horizons where we can assume the expected returns to be zero. In practice, much risk management analyses are indeed done over one-day or one-week horizons.
- Case II: The second case is when one security has zero weight; therefore, its contribution to risk is zero. Consequently, $D_2 = D_1 = 0$. This is a trivial case in which the remaining security accounts for 100% of the risk as well as 100% of the loss. However, this loss contribution remains approximately true if the security weight is small, and the loss L is relatively large compared to D_1 and D_2.
- Case III: The third and more interesting case arises when $D_1 = \mathrm{PCR}_2 w_1\mu_1 - \mathrm{PCR}_1 w_2\mu_2 = 0$, or equivalently

$$\frac{w_1\mu_1}{\mathrm{PCR}_1} = \frac{w_2\mu_2}{\mathrm{PCR}_2} . \tag{3.35}$$

Equation 3.35 is the first-order condition of marginal utility for an optimal mean–variance portfolio. Therefore, it implies that, for optimal portfolios, PCR is equivalent to expected percentage contribution to the portfolio's total expected return. In other words, risk budgets become the budgets of expected return for mean–variance optimal portfolios.

- Sharpe (2002) discusses this property at length and suggests that "risk-budgeting and risk-monitoring systems are best viewed in terms of a budget of implied expected excess return and deviation from the budget." However, this equivalency is only true for mean–variance optimal portfolios. For a real-world portfolio, which might not be optimal in the mean–variance sense, our interpretation of PCR still allows managers to estimate the likely contribution to a given loss.

In fact, Equation 3.33 allows us to estimate the impact of the portfolios' *suboptimality* measured by D_i's on PCL. For instance, if the allocation to security 1 is more than the mean–variance optimal weight, then $D_1 < 0$. This is because when the weight w_1 increases from the optimal weight, the increase in its risk contribution dominates its increase in the expected return contribution. Therefore, for a given loss L (<0), the percentage contribution to loss PCL_1 will be greater than the percentage contribution to risk PCR_1 because D_1/L is positive.

- We further note that, when the loss L far exceeds the quantity D_i's, then PCL and PCR are approximately the same. This observation is very relevant during financial crises when portfolio losses could be significantly higher than the expected returns. Consequently, loss contribution would be well captured by risk contribution. On the contrary, during quiet periods when portfolio losses are relatively small, loss contribution, or simply *ex post* return attribution, is unlikely to bear any relationship to risk contribution at all!

In summary, contribution to risk can be interpreted as contribution to a given loss of the total portfolio. The two are identical when expected returns are each zero or when the portfolio is mean–variance optimal. In other cases, the interpretation is appropriate when the given loss is large compared to the value of D_i's, which measure the portfolio's deviation from mean–variance optimality. Qian (2006) showed empirically that risk contribution of stock/bond asset allocation portfolios explains the loss contribution. In the context of active equity portfolios, the risk

contribution in terms of systematic risk and specific risk should be a guide for loss contribution from those sources.

3.3 CONTRIBUTION TO VALUE AT RISK

We have shown that risk contribution can be regarded as loss contribution. We based our analysis on the conditional expectation of a multivariate normal distribution, for which analytic formulas are available. However, in reality, few returns follow normal distribution. For returns measuring longer investment horizons, they are log normal at best and often exhibit both skewness and excess kurtosis or fat tails. For nonnormal returns, standard deviation as a risk measure is inadequate. A common substitute for it is value at risk (VaR), which represents loss with a given cumulative probability. We shall now extend our results to VaR contribution.

Let us first define VaR. For a portfolio with normal distribution, VaR is simply the expected return plus a constant multiple of standard deviation. For a nonnormal distribution, a $(1-\alpha)\%$ VaR is defined through the following equation:

$$\text{Prob}(r \le \text{VaR}) = \int_{-\infty}^{\text{VaR}} p(r)\,dr = \alpha\,, \tag{3.36}$$

where $p(r)$ is the probability density of the return distribution and α is the cumulative probability of loss, typically set at 5% or 1%. However, note the following:

- Although it is a more realistic risk measure, VaR does have some drawbacks. One drawback is that analytic expressions rarely exist for VaR as a function of portfolio weights, and one has to resort to numerical simulations to calculate VaR of individual securities or portfolios.

The following equations define the marginal contribution to VaR and contribution to VaR

$$\text{MCV}_i = \frac{\partial \text{VaR}}{\partial w_i}, \quad \text{CV}_i = w_i \frac{\partial \text{VaR}}{\partial w_i}\,. \tag{3.37}$$

As before, the contribution to VaR is a product of weight and the marginal contribution. Because VaR is a linear homogeneous function of weights, it is mathematically true that (Problem 3.7)

$$\mathrm{VaR} = \sum_{i=1}^{N} w_i \frac{\partial \mathrm{VaR}}{\partial w_i} . \tag{3.38}$$

Hence, we have the VaR budget identity.

It turns out that contribution to VaR can also be interpreted as expected contribution to loss, whose size equals VaR. The following proof is due to Hallerbach (2003). Suppose a portfolio suffers a loss of size VaR, i.e.,

$$r_p = w_1 r_1 + \cdots + w_N r_N = \mathrm{VaR} . \tag{3.39}$$

Then, taking expectation of (3.39) with respect to the returns $\left(r_1, \cdots, r_N\right)$ yields

$$\mathrm{E}\left(w_1 r_1 + \cdots + w_N r_N \mid r_p = \mathrm{VaR}\right) = \mathrm{VaR} . \tag{3.40}$$

VaR is simply a constant in this process. Because the weights are regarded as constants in the equation, the expectation on the left side can be written as a linear combination:

$$\sum_{i=1}^{N} w_i E\left(r_i \mid r_p = \mathrm{VaR}\right) = \mathrm{VaR} . \tag{3.41}$$

Comparing Equation 3.38 and Equation 3.41 leads to

$$w_i \mathrm{E}\left(r_i \mid r_p = \mathrm{VaR}\right) = w_i \frac{\partial \mathrm{VaR}}{\partial w_i} . \tag{3.42}$$

Equation 3.42 is the interpretation we have sought — contribution to VaR (on the right-hand side) equals contribution to a loss of the size VaR (on the left-hand side). It further implies that the marginal contribution to VaR equals the expected security return given the portfolio return of VaR. However, note the following:

- Although contributions to risk in terms of both standard deviation and VaR have the same financial interpretation, there are several subtle differences. First, in the case of standard deviation under normality assumption, percentage contributions to risk are independent

of loss size. We have shown that, under some circumstances, they approximate loss contributions with sufficient accuracy regardless of the loss size. However, the interpretation of contribution to VaR is rather restrictive — it only applies to the loss that exactly equals a given VaR. VaR contribution changes when VaR changes. Therefore, for losses of different sizes, one must recalculate its VaR contribution.

- Another difference is the computational complexity. Although risk contribution based on standard deviation is easy to calculate, it is a daunting task to calculate risk contribution to VaR because analytic expressions are rarely available for VaR as functions of weights. Even when there is an analytic expression, calculating its partial derivative with respect to weights can be quite challenging (Chow and Kritzman 2001, Chow et al. 2001). In most instances, one has to resort to Monte Carlo simulations to obtain VaR decomposition as well as VaR itself. One alternative is to use Cornish–Fisher approximation to VaR based on moments of the return distribution (Mina and Ulmer 1999, Jaschke 2000). The approximation gives rise to an algebraic expression of VaR, and it can be used to calculate VaR contribution analytically (Qian 2006).

PROBLEMS

3.1 Suppose decaying weights are $\cdots,\omega^{t-T},\cdots,\omega^2,\omega,1$, with the weight for the most recent period being 1 and $\omega<1$. Prove the half-life of the weights is $H=-\ln 2/\ln\omega$.

3.2 Prove Equation 3.26, i.e., risk contributions add up to the total risk.

3.3 For a long-short portfolio, prove (a) the marginal contribution to specific risk of a long (short) position is positive (negative), and (b) contribution to specific is always positive.

3.4 In a long-only portfolio where all the stock weights are nonnegative, is it possible to have negative MCR?

3.5 In an active portfolio vs. a benchmark or a long-short portfolio, it is typical to impose sector-neutral constraints

$$\sum_{i\in S} w_i = 0 . \tag{3.43}$$

The marginal contribution to risk of the sector S could be defined as

$$\mathrm{MCR}_S = \sum_{i \in S} w_i \mathrm{MCR}_i \,. \tag{3.44}$$

Find an interpretation of MCR_S in terms of the leverage for the sector.

3.6 This problem extends the results for risk contribution to portfolios with N securities whose returns follow a multivariate normal distribution, $\mathbf{r} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Denote portfolio return by $r_p = w_1 r_1 + \cdots + w_N r_N$, and the portfolio expected return by μ_p. Suppose the portfolio had a loss L, prove that:

(a) The PCL is

$$\mathrm{PCL}_i = \mathrm{E}\left(w_i r_i \,|\, r_p = L\right) / L = \mathrm{PCR}_i + \frac{w_i \mu_i - \mathrm{PCR}_i \mu_R}{L} \,. \tag{3.45}$$

(b) The PCL is the same as PCR for all securities if

$$\frac{w_1 \mu_1}{\mathrm{PCR}_1} = \frac{w_2 \mu_2}{\mathrm{PCR}_2} = \cdots = \frac{w_N \mu_N}{\mathrm{PCR}_N} \,. \tag{3.46}$$

(c) For a mean–variance optimal portfolio, Equation 3.46 holds.

(d) The conditional standard deviation of PCL is

$$\mathrm{std}\left(\frac{w_i r_i}{L} \,|\, r_p = L\right) = \frac{\sqrt{w_i^2 \sigma_i^2 - \mathrm{PCR}_i^2 \sigma^2}}{L} \,. \tag{3.47}$$

As the loss L increases, the conditional standard deviation decreases as 1 over L.

3.7 A scalar $f(\vec{w})$ is a linear homogenous function of $\vec{w}$ if $f(c\vec{w}) = cf(\vec{w})$ for any constant c. Prove that

(a) The average return μ_p and the standard deviation σ_p of portfolio returns are linear homogeneous functions of portfolio weights $\vec{w}$.

(b) VaR is a linear homogeneous function of portfolio weights $\vec{w}$.

(c) For any linear homogeneous function $f(\vec{w})$,

$$\sum_{i=1}^{N} w_i \frac{\partial f}{\partial w_i} = f \,. \tag{3.48}$$

3.8 This problem proves the VaR budget identity (3.38) by a direct parametric approach. Without loss of generality, we again assume a portfolio of two securities whose returns have a joint probability distribution $f(r_1, r_2)$. Denote the portfolio return as $r_p = w_1 r_1 + w_2 r_2$.

(a) Prove the probability of r_p being less than the $(1-\alpha)\%$ VaR is

$$\text{Prob}(r_p \le \text{VaR}) = \int_{-\infty}^{\infty} dr_1 \int_{-\infty}^{(\text{VaR}-w_1 r_1)/w_2} f(r_1, r_2)\, dr_2 = \alpha \,. \tag{3.49}$$

(b) Equation 3.49 defines VaR as an implicit function of w_1 and w_2. Prove that the partial derivative of VaR with respect to w_1 equals

$$\frac{\partial \text{VaR}}{\partial w_1} = \frac{\int_{-\infty}^{\infty} r_1 f\left(r_1, \dfrac{\text{VaR} - w_1 r_1}{w_2}\right) dr_1}{\int_{-\infty}^{\infty} f\left(r_1, \dfrac{\text{VaR} - w_1 r_1}{w_2}\right) dr_1} \,. \tag{3.50}$$

(c) Based on (3.50), show that $\dfrac{\partial \text{VaR}}{\partial w_1}$ is the conditional expected return of r_1 given the portfolio return is $r_p = w_1 r_1 + w_2 r_2 = \text{VaR}$.

REFERENCES

Arnott, R.D., Risk budgeting and portable alpha, *Journal of Investing,* Vol. 11, No. 2, 2002.

BARRA, *United States Equity — Version 3, Risk Model Handbook,* BARRA Inc., 1998.

Chen, N.-F., Roll, R., and Ross, S., Economic forces and the stock market, *Journal of Business*, 386–403, July 1986.

Chow, G. and Kritzman, M., Risk budgets, *Journal of Portfolio Management*, Vol. 27, No. 2, Winter 2001.

Chow, G., Kritzman, M., and Van Royen, Anne-Sophie, Risk budgets: Comment, *Journal of Portfolio Management*, Vol. 27, No. 4, Summer 2001.

Connor, G. and Korajczyk, R., Risk and return in an equilibrium APT, *Journal of Financial Economics,* Vol. 21, 255–289, 1988.

Connor, G., The three types of factor models: A comparison of their explanatory power, *Financial Analyst Journal,* May–June 1995.

Fama, E. and French, K., The cross section of expected stock returns, *The Journal of Finance*, 427–466, June 1992.

Fama, E. and MacBeth J.D., Risk, return and equilibrium: Empirical tests, *Journal of Political Economy,* 81, 1973.

Grinold, R.C., and Kahn, R.N., *Active Portfolio Management*, McGraw-Hill, New York, 2000.

Hallerbach, W.G., Decomposing portfolio value-at-risk: a general analysis, *Journal of Risk,* Vol. 5, No. 2, Winter 2003.

Jaschke, S.R., The cornish-fisher expansion in the context of delta-gamma-normal approximation, *Journal of Risk,* Vol. 4, No. 4, Summer 2002.

Kung, E. and Pohlman, L.F., Portable alpha: philosophy, process, and performance, *Journal of Portfolio Management*, Vol. 30, No. 3, Spring 2004.

Lee, W. and Lam, D.Y., Implementing optimal risk budgeting, *Journal of Portfolio Management*, Vol. 28, No. 1, Fall 2001.

Leibowitz, M., Sorenson, E.H., and Arnott, R., A total differential approach to equity duration, *Financial Analysts Journal,* September/October, 1989.

Litterman, R., Hot spots and hedges, *Journal of Portfolio Management*, December 1996.

Mina, J. and Ulmer, A., Delta-Gamma Four Ways, RiskMetrics Group, 1999.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., and Stanley, H.E., Universal and nonuniversal properties of cross correlations in financial time series, *Physical Review Letter*, Vol. 83, No. 7, 1471–1473, 1999.

Qian, E., On the financial interpretation of risk contribution: Risk budgets do add up, *Journal of Investment Management,* Vol. 4, No. 4, Fourth Quarter, 2006.

Sharpe, W.F., Budgeting and monitoring pension fund risk, *Financial Analyst Journal*, Vol. 58, No. 5, September–October 2002.

Sorensen, E.H., Samak, V., Miller, K., U.S. equity risk attribute model (RAM): A new look at a classic model, Salomon Smith Barney, September, 1998.

Wander, B.H., De Silva, H., and Clarke, R.G., Risk allocation versus asset allocation, *Journal of Portfolio Management*, Vol. 29, No. 1, Fall 2002.

Winkelmann, K., Improving portfolio efficiency, *Journal of Portfolio Management*, Vol. 30, No. 2, Winter 2004.

Part II

CHAPTER 4

Evaluation of Alpha Factors

Mean–variance optimization and risk models described in Chapter 2 and Chapter 3 provide the theoretical foundation of quantitative equity portfolio management. In Part II of the book, we dig deeper into the key ingredients of the Modern Portfolio Theory (MPT) paradigm. An important component of any successful investment strategy is forecasting expected returns using alpha models. In this chapter, we consider the process of selecting or evaluating return factors that go into a comprehensive alpha model. In Chapter 5, we consider the typical set of quantitative alpha factors used in practice and their performance. In Chapter 6, we consider the firm valuation approach used in fundamental analysis and retool it for quantitative use. Chapter 7 presents the analytical framework for combining specific return factors into a comprehensive multiple-factor model designed to lead to consistent long-term performance. The essence is to create an expected return/covariance approach to "factor diversification," analogous to classical stock selection methods discussed in Chapter 2. One additional dimension of factor evaluation is its associated portfolio turnover implication. We discuss this important topic in Chapter 8.

4.1 ALPHA PERFORMANCE BENCHMARKS: THE RATIOS

The evaluation of success of most investment strategies requires modern performance measurement, on a long-term basis. For many institutional investors, such as corporate pension plans and university endowments, the investment horizon is infinite, at least in theory. For individual or retail

investors who invest for retirement, the investment horizon can be years or even decades. It is therefore important to have appropriate long-term performance measures to not only build long-term investment strategies but also to evaluate and compare different strategies.

The two common risk/return measures that derive from the CAPM theory in Chapter 2 are the Sharpe ratio (SR) and the information ratio (IR). Both assess the returns of a process (alpha factor or model) conditioned on a dimension of risk. The SR conditions on total risk or volatility of the portfolio, and is the ratio of average excess return to the standard deviation of excess return

$$\mathrm{SR} = \frac{\bar{\mu} - r_f}{\sigma}. \tag{4.1}$$

For example, assume a portfolio of U.S. large cap stocks has an annual volatility of 15% and an excess return of 5% — the SR is 0.33. Intuitively, one can interpret the SR as the accrued returns (benefit) per unit of total risk (cost). In our example, U.S. large cap stocks delivered 33 basis points (bps) of returns per unit of risk.

IR, on the other hand, has an added layer of relativity. It measures the average of an active portfolio return (relative to a passive portfolio), relative to the increased volatility of the active portfolio, also relative to a passive portfolio. The pension consultant community introduced in Chapter 1 makes considerable use of IR. It is particularly important in comparing long-only (no shorting) professional equity managers to (1) other active managers and (2) a passive benchmark that can be mimicked with relatively low cost, like owning the entire S&P 500 index. "Tracking error" is the common term to reference periodic deviation from the passive benchmark (or active risk). Thus, IR compares the average alpha over time to the incremental benchmark-tracking risk (alpha volatility)

$$\mathrm{IR} = \frac{\bar{\alpha}}{\sigma(\alpha)}. \tag{4.2}$$

For long-only portfolios managed against a benchmark, alpha is the portfolio excess return over the benchmark; for long-short market-neutral portfolio, alpha is the excess return over cash, the benchmark for most long-short products. Similar to SR, IR measures the accrued active return per unit of active risk. For a given level of tracking error, it is evident that we prefer a strategy with a higher IR to a strategy with a lower IR. In practice,

long-only managers that achieve an IR above 1 should be considered quite successful. The median IR over the last 20 years for active large cap U.S. investors is considerably less than 1. However, note the following:

- Several remarks should be made about the use of IR in practice. First, it is customary to quote IR on an annualized basis, whereas the alpha stream is often reported on a much shorter horizon such as quarterly or monthly. In these cases, one has to annualize the IR. Second, it is important to emphasize that IR is a multiperiod statistical metric. Although it is straightforward to calculate *ex post* (or realized) IR given a history of periodic excess returns, it is much more difficult to estimate *ex ante* or expected IR. Nevertheless, an *ex ante* IR would be much more useful to investors as a guide for their future investment allocations.

It is useful to note that the IR definition is closely related to the *t*-statistics. Indeed, we can transform the IR into a t-stat that helps measure the consistency of an alpha process as follows:

$$t_\alpha = \frac{\sqrt{T-1}\bar{\alpha}}{\sigma(\alpha)} = \mathrm{IR}\sqrt{T-1}\,, \tag{4.3}$$

where T is the number of sample points. We can use IR to test the hypothesis whether the expected alpha is statistically positive. For example, an IR of .67 derived from 10 years of return history demonstrates statistical significance of value added at the 95% confidence level.

4.2 SINGLE-PERIOD SKILL: INFORMATION COEFFICIENT

The information coefficient (IC) statistic (Grinhold 1989, Grinhold and Kahn 2000) is a key building block in measuring the "alpha power" of a factor or process. We can imagine many ways to associate skill with a predictive factor. For example, we might merely count the success in terms of the number of securities in the portfolio over an interval that outperformed an index-type benchmark. This would be a type of "hit rate." It turns out that a process that can deliver a hit rate of, say, 55 to 60% is exceptional if it can be achieved consistently.

IC is a more formal measure of forecasting alpha power. It is a linear statistic that measures the cross-sectional correlation between the security return forecasts coming from a factor and the subsequent actual

returns for securities. IC is important in evaluating factors because of its translation into IR — our ultimate objective — which is developed later in the chapter through the following equation:

$$IR = \frac{\overline{IC_t}}{\text{std}\left(IC_t\right)}.$$

Other things being equal, the higher the average IC for a factor is over time, the better the reward-to-risk ratio. In addition, the more stable the IC over time, the better the result.

4.2.1 Raw IC

In order to analyze multiperiod IR for a strategy, we need to develop the IC component of the strategy or factor that is embedded in IR. This analysis first entails an extension of the simple one-period "raw IC" for total return correlation to a refined "risk-adjusted IC".

We start from single-period excess return, which is a function of portfolio weights at a given time t and subsequent returns of stocks. Denote active weights by $\mathbf{w} = \left(w_1, \cdots, w_N\right)'$ and subsequent returns by $\mathbf{r} = \left(r_1, \cdots, r_N\right)'$. We have suppressed the time index t for the moment for clarity. The realized excess return for the period is

$$\alpha_t = \sum_{i=1}^{N} w_i r_i = \mathbf{w}' \cdot \mathbf{r}. \tag{4.4}$$

For a dollar-neutral long-short portfolio or a long-only portfolio against a benchmark, we have $\mathbf{w}' \cdot \mathbf{i} = 0$. Therefore Equation 4.4 remains unchanged if we replace returns with relative returns against the cross-sectional average $\overline{r}$

$$\alpha_t = \sum_{i=1}^{N} w_i \left(r_i - \overline{r}\right) = \mathbf{w}' \cdot \left(\mathbf{r} - \overline{r}\,\mathbf{i}\right). \tag{4.5}$$

The summation in (4.5) is related to the covariance between the weight vector and the return vector. Writing the covariance in terms of correlation and cross-sectional dispersion (we reserve the use of standard deviation for time-series measures), we have

$$\alpha_t = \sum_{i=1}^{N} w_i \left(r_i - \bar{r}\right) = \left(N-1\right)\mathrm{corr}\left(\mathbf{w},\mathbf{r}\right)\mathrm{dis}\left(\mathbf{w}\right)\mathrm{dis}\left(\mathbf{r}\right). \tag{4.6}$$

Because both dispersions are positive, the excess return has the same sign as the correlation term. In order to generate positive excess return, we must, in general, overweight stocks with higher returns and simultaneously underweight stocks with lower returns. This is true regardless of the general direction of average return.

Example 4.1

It is easy to observe this in a simple two-stock example. Suppose we have stock 1 and stock 2, and we overweight stock 1 by 5% ($w_1 = 5\%$) and underweight stock 2 by 5% ($w_2 = -5\%$). Consider two return scenarios A and B. In scenario A, stock returns are 10 and 5% for stock 1 and stock 2, respectively. In this case,

$$\alpha = 5\% \cdot 10\% - 5\% \cdot 5\% = 0.25\%,$$

or 25 basis points (bps). In scenario B, stock returns are –5% and –10% for stock 1 and stock 2, respectively. We obtain positive alpha again, because

$$\alpha = 5\% \cdot \left(-5\%\right) - 5\% \cdot \left(-10\%\right) = 0.25\%.$$

To connect excess return in (4.6) with the raw IC, which is the cross-sectional correlation coefficient between the forecasts and the returns, we are forced to make an unrealistic assumption that portfolio weights are proportional to the forecasts, i.e.,

$$\mathbf{w} = c\mathbf{f}, \text{ or } w_i = cf_i, \text{ for all } i. \tag{4.7}$$

Assuming the forecasts have zero cross-sectional mean, we have

$$\begin{gathered} \alpha_t = \sum_{i=1}^{N} w_i \left(r_i - \bar{r}\right) = c\left(N-1\right)\mathrm{IC} \cdot \mathrm{dis}\left(\mathbf{f}\right)\mathrm{dis}\left(\mathbf{r}\right) \\ \mathrm{IC} = \mathrm{corr}\left(\mathbf{f},\mathbf{r}\right) \end{gathered} \tag{4.8}$$

Realized portfolio excess return is decomposed into three intuitive components — IC (skill), dispersion of the forecasts (conviction), and dispersion of actual returns (opportunities). Because both dispersions are always positive, the sign of excess return depends on the sign of the IC. A high positive IC is desired. Typically, an IC of 0.1 or higher on an annual basis is considered quite strong, depending on its time-series volatility. Of course, if a factor f consistently has negative IC, we can just use $-f$ as a factor.

4.2.2 Risk-Adjusted IC

Although the aforementioned IC definition facilitates an intuitive interpretation of portfolio excess return in terms of the three components, it has a serious flaw. The problem arises from the unrealistic assumption of portfolio weights in Equation 4.7. For a quantitative manager, such naïve portfolio weights are mean–variance optimal, only if the risk model consists of a single diagonal matrix with equal diagonal elements, i.e., there is no systematic risk in the market, and all stocks have the same specific risk. From a realistic perspective, systematic risks do exist in the market, and specific risks are uneven across stocks. Therefore, a portfolio constructed by (4.7) is susceptible to unintended systematic risk exposures. In addition, it is inefficient in terms of the distribution of specific risk among the stocks according to Chapter 2. An example is the book-to-price factor. If we have used it in the same manner as in (4.7), the portfolio would have had a low beta bias since high B/P stocks have historically had low beta on average. As a result, the portfolio tends to underperform when the overall market goes up — an unintended beta bet.

The traditional "raw IC," based on raw forecasts and raw returns, is too removed from realistic portfolios to be an effective alpha diagnostic. It might serve as a preliminary check, but its applications are limited. What we need is a new IC, a risk-adjusted IC, which is consistent with a realistic portfolio process, which strips out the systematic bias in the factor, and incorporates uneven levels of specific risks in portfolio weight selection. This new IC is linked directly to a realistic quantitative portfolio process, and therefore serves as a better proxy of how the factor will perform in a portfolio context.

We define a risk-adjusted IC by first solving a mean–variance optimization to get the optimal weights of a market-neutral portfolio; second, we derive the single-period alpha using those weights and subsequent returns and, third, we relate the alpha to a risk-adjusted IC.

Given a forecast vector $\mathbf{f}$, we solve the following mean–variance optimization to obtain portfolio weights $\mathbf{w}$

$$\begin{aligned} &\text{Maximize} \quad \mathbf{f}' \cdot \mathbf{w} - \frac{1}{2}\lambda \cdot (\mathbf{w}' \cdot \mathbf{\Sigma} \cdot \mathbf{w}_t) \\ &\text{subject to} \quad \mathbf{w}' \cdot \mathbf{i} = 0, \text{and } \mathbf{w}' \cdot \mathbf{B} = 0 \end{aligned} . \tag{4.9}$$

The covariance matrix is that of a multifactor model, i.e.,

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_I \mathbf{B}' + \mathbf{S} . \tag{4.10}$$

The active weights are not only dollar neutral but also neutral to all risk factors. Therefore, there will be no systematic risk in the final portfolio. As a result, we can reduce the objective function in (4.9) to the following, provided that we keep all the constraints

$$\mathbf{f}' \cdot \mathbf{w} - \frac{1}{2}\lambda \cdot (\mathbf{w}' \cdot \mathbf{S} \cdot \mathbf{w}) . \tag{4.11}$$

We can now solve the optimization analytically with Lagrangian multipliers. We switch from matrix notation to a summation form. The new objective function including $K + 1$ Lagrangian multipliers (1 for the dollar neutral constraint and K for K risk factors) is:

$$\sum_{i=1}^{N} f_i w_i - \frac{1}{2}\lambda \sum_{i=1}^{N} w_i^2 \sigma_i^2 - l_0 \sum_{i=1}^{N} w_i - l_1 \sum_{i=1}^{N} w_i \beta_{1i} - \cdots - l_K \sum_{i=1}^{N} w_i \beta_{Ki} . \tag{4.12}$$

Taking the partial derivative with respect to w_i and equating it to zero gives

$$w_i = \lambda^{-1} \frac{f_i - l_0 - l_1 \beta_{1i} - \cdots - l_K \beta_{Ki}}{\sigma_i^2} . \tag{4.13}$$

Equation 4.13 states the optimal portfolio weights are the risk-neutral forecasts divided by the specific variances. The values of the Lagrangian multipliers can be determined by the constraints through a system of linear equations. Denote

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}' \cdot \mathbf{S}^{-1} \cdot \mathbf{y} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} . \tag{4.14}$$

The system of equations is

$$\begin{cases} l_0\langle \mathbf{i},\mathbf{i}\rangle + l_1\langle \mathbf{i},\mathbf{b}_1\rangle + \cdots + l_K\langle \mathbf{i},\mathbf{b}_K\rangle = \langle \mathbf{i},\mathbf{f}\rangle \\ l_0\langle \mathbf{b}_1,\mathbf{i}\rangle + l_1\langle \mathbf{b}_1,\mathbf{b}_1\rangle + \cdots + l_K\langle \mathbf{b}_1,\mathbf{b}_K\rangle = \langle \mathbf{b}_1,\mathbf{f}\rangle \\ \vdots \\ l_0\langle \mathbf{b}_K,\mathbf{i}\rangle + l_1\langle \mathbf{b}_K,\mathbf{b}_1\rangle + \cdots + l_K\langle \mathbf{b}_K,\mathbf{b}_K\rangle = \langle \mathbf{b}_K,\mathbf{f}\rangle \end{cases}. \tag{4.15}$$

The solution is given by

$$\begin{pmatrix} l_0 \\ l_1 \\ \vdots \\ l_K \end{pmatrix} = \begin{pmatrix} \langle \mathbf{i},\mathbf{i}\rangle & \langle \mathbf{i},\mathbf{b}_1\rangle & \cdots & \langle \mathbf{i},\mathbf{b}_K\rangle \\ \langle \mathbf{b}_1,\mathbf{i}\rangle & \langle \mathbf{b}_1,\mathbf{b}_1\rangle & \cdots & \langle \mathbf{b}_1,\mathbf{b}_K\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{b}_K,\mathbf{i}\rangle & \langle \mathbf{b}_K,\mathbf{b}_1\rangle & \cdots & \langle \mathbf{b}_K,\mathbf{b}_K\rangle \end{pmatrix}^{-1} \begin{pmatrix} \langle \mathbf{i},\mathbf{f}\rangle \\ \langle \mathbf{b}_1,\mathbf{f}\rangle \\ \vdots \\ \langle \mathbf{b}_K,\mathbf{f}\rangle \end{pmatrix}. \tag{4.16}$$

Given the active weights, the portfolio excess return is the summed product of the active weights and the actual returns

$$\alpha_t = \sum_{i=1}^{N} w_i r_i = \lambda^{-1} \sum_{i=1}^{N} \frac{f_i - l_0 - l_1\beta_{1i} - \cdots - l_K\beta_{Ki}}{\sigma_i^2} r_i . \tag{4.17}$$

We now replace the return r_i by $r_i - m_0 - m_1\beta_{1i} - \cdots - m_K\beta_{Ki}$, where $(m_1,\cdots,m_K)$, which are the returns to K risk factors, derived from the cross-sectional ordinary least square (OLS) regression. We do so to express returns in the same format as the forecast, and it does not change the equation because of the constraints placed on the active weights. We shall see in the following text that this is not just for cosmetic purposes. The value of m_0 is still undetermined but will become clear later. *Risk-adjusted forecast and return* are defined as

$$\begin{aligned} F_i &= \frac{f_i - l_0 - l_1\beta_{1i} - \cdots - l_K\beta_{Ki}}{\sigma_i} \\ R_i &= \frac{r_i - m_0 - m_1\beta_{1i} - \cdots - m_K\beta_{Ki}}{\sigma_i} \end{aligned} \tag{4.18}$$

We have

$$\alpha_t = \sum_{i=1}^{N} w_i r_i = \lambda^{-1} \sum_{i=1}^{N} F_i R_i \ . \tag{4.19}$$

Therefore, excess return is a summed product of risk-adjusted forecasts and risk-adjusted returns, scaled by the risk-aversion parameter.

From this point on, there are two directions to proceed. One common approach is to take the expectation of Equation 4.19 and assume the expected security return is the product of IC, specific risk, and score, which is the standardized forecast (Grinold 1994). Such prescription is useful in practice for translating z-scores into alpha forecasts. It can also lead to an estimate of the single-period IR (Problem 4.3). However, this linearity assumption is not theoretically valid with cross-sectional z-scores. In addition, as we shall see shortly, such prescription is not necessary in deriving the IR.

In the second approach, we make no explicit assumption about the expected return of individual stocks, because the excess return of an active portfolio depends collectively on the cross-sectional correlation between the forecasts and the actual returns. Similar to Equation 4.6, we recast Equation 4.19 in terms of correlation and dispersions

$$\alpha_t = \left(N-1\right)\lambda_t^{-1} \mathrm{corr}\left(\mathbf{F}_t, \mathbf{R}_t\right) \mathrm{dis}\left(\mathbf{F}_t\right) \mathrm{dis}(\mathbf{R}_t) \ , \tag{4.20}$$

provided that the cross-sectional average of $\mathbf{R}_t$ is zero. Thus, we choose m_0 in Equation 4.18 such that

$$\mathrm{avg}\left(\mathbf{R}_t\right) = 0 \ . \tag{4.21}$$

Note we have reinserted the subscript t for all the terms except the number of stocks. The correlation between the risk-adjusted forecasts and the risk-adjusted returns is the *risk-adjusted IC* that we have sought, as it is directly related to the excess return of a risk-managed portfolio. Note that Equation 4.20 is essentially a mathematical identity. Note the following remarks:

- First, it is obvious that for the same alpha factor, the risk-adjusted IC could be quite different from the raw IC. Indeed, in some cases,

they could be of different signs. This difference can lead to serious disparity between the real portfolio performance, which is risk-adjusted, and a naïve model performance, which is not risk-adjusted. This can contribute to the "unexplained" portion (often large and volatile) of a univariate performance attribution, a popular *ex post* attribution tool used by practitioners in decomposing sources of value that are added.

- Second, the neutrality constraints on all risk factors embedded in the risk-adjusted IC are rather restrictive. In practice, many portfolios are constrained to have limited factor exposures, which are not necessarily zero. Therefore, the risk-adjusted IC serves as an approximated performance indicator for these portfolios. Overall, however, it is more indicative of the realistic portfolio performance than the raw IC.

Example 4.2

We use a three-stock example to illustrate the risk-adjusted IC in which the only risk factor is the beta. Table 4.1 first lists the raw forecasts, followed by their betas, risk-adjusted forecasts, actual returns, and risk-adjusted returns. As we can see, the raw forecast f favors the first stock, is neutral on the second stock, and dislikes the third stock. Stock 2 has the best return (r) and is followed by stock 3; stock 1 has the worst return. The raw IC between f and r is −0.24. Therefore, if we overweight stock 1, underweight stock 3, and take no active weight on stock 2, according to f, we would have a negative excess return.

However, stock 1 has a beta of 0.9, whereas stock 3 has a beta of 1.1. The naïve weights above would result in a low-beta bias, which a beta-neutral portfolio would not allow. For a beta-neutral (also dollar neutral) portfolio, the risk-adjusted forecast (F) is the determinant of performance and they are 1.25, 1.25, and −2.50 for the three stocks. In essence, to be dollar neutral and beta neutral, we should overweight both stock 1 and stock 2 by the same amount and offset it by the underweight in stock 3. Because stock

TABLE 4.1 Forecast, Beta, and Return for the Three Stocks

Stock	f	β	F	r	R
1	0.5	0.9	1.25	−5%	8.3%
2	0	1.1	1.25	15%	8.3%
3	−0.5	1	−2.50	0%	−16.7%

Note: The specific risk is the same 20%.

2 returns 15%, this beta-neutral portfolio has a positive excess return. We calculate the risk-adjusted return R and discover the risk-adjusted IC is actually a perfect 1.

4.2.3 Target Tracking Error and the Risk-Aversion Parameter

Because the portfolio above has no systematic risk, the risk-model tracking error (tracking error predicted by a risk model) is computed as the residual variance. The model tracking error is the product of the sum of specific variance and the square of the active weights. Note that we use risk-model tracking error and target tracking error interchangeably. We have

$$\sigma_{\text{model}}^2 = \sum_{i=1}^{N} w_i^2 \sigma_i^2 = \lambda_t^{-2} \sum_{i=1}^{N} F_i^2 . \tag{4.22}$$

The residual variance is therefore the sum of the squares of the risk-adjusted forecasts:

$$\begin{aligned} \sigma_{\text{model}} &= \lambda_t^{-1} \sqrt{\sum_{i=1}^{N} F_i^2} \\ &= \lambda_t^{-1} \sqrt{N-1} \sqrt{\left[\text{dis}\left(\mathbf{F}_t\right)\right]^2 + \left[\text{avg}\left(\mathbf{F}_t\right)\right]^2} . \\ &\approx \lambda_t^{-1} \sqrt{N-1}\,\text{dis}\left(\mathbf{F}_t\right) \end{aligned} \tag{4.23}$$

We assume that $\text{avg}(\mathbf{F}_t) \approx 0$, and this approximation is quite accurate in practice. Solving for the risk-aversion parameter, we have

$$\lambda_t = \frac{\sqrt{N-1}\,\text{dis}\left(\mathbf{F}_t\right)}{\sigma_{\text{model}}} . \tag{4.24}$$

The risk-model tracking error (aka the target tracking error) is proportional to the cross-sectional dispersion of the forecasts (conviction) and square root of the number of stocks (breadth), but inversely proportional to the risk-aversion parameter. Scaling the forecasts and the risk-aversion parameter by the same amount would have no effect on the weights and tracking error at all.

Substituting Equation 4.24 into Equation 4.20, we obtain the main result for the single-period excess return

$$\alpha_t = \mathrm{IC}_t \sqrt{N-1}\sigma_{\mathrm{model}}\mathrm{dis}(\mathbf{R}_t) \approx \mathrm{IC}_t \sqrt{N}\sigma_{\mathrm{model}}\mathrm{dis}(\mathbf{R}_t)\,. \tag{4.25}$$

Therefore, the single-period excess return is the product of the risk-adjusted IC (skill), square root of N (breadth), target tracking error (risk budget), and dispersion of the risk-adjusted returns (opportunity). The IC in the equation is the risk-adjusted IC. We have replaced $N-1$ by N, which is justified when it is large enough.

Example 4.3

If the IC of a forecast is 0.05 for a given year, the number of stocks is 500, the targeted tracking error is 3%, and the dispersion of risk-adjusted returns is 1, then the excess return for the year is $0.05 \cdot \sqrt{500} \cdot 3\% = 3.35\%$.

4.2.4 Dispersion of the Risk-Adjusted Returns

Cross-sectional dispersion of stock returns can be considered as a measure of opportunity that exists in the market. Consider active positions in just two stocks, long 5% in stock 1 and short 5% in stock 2. The result of this pair trading would depend on the difference of the two stocks' realized returns. The larger the return difference, the greater will be the profit or loss. In general, dispersion of raw or unadjusted returns can exhibit great variation over time. The raw returns are influenced by the return to risk factors, which are systematic and subject to macroeconomic and/or profit cycles. What about the risk-adjusted returns defined in Equation 4.18, from which the risk factor returns have been subtracted?

In theory, the dispersion of risk-adjusted return should show little time-series variation, given that the risk model correctly describes the stock returns. To see this, we note that for each stock, the risk-adjusted return is, in fact, the specific return (or residual return) scaled by specific risk. Therefore, each R_i is approximately a standard normal variable. The variance of N such independent variables is a scaled chi-square distribution if their mean is zero. It can be proven that when N is large, the dispersion is close to unity using the approximation of chi-square distribution (Keeping 1995). Thus, when the number of stocks is large, say a few hundred, the cross-sectional dispersion of the risk-adjusted returns is close to one. Under this assumption, the Equation 4.25 is simplified to

$$\alpha_t \approx \mathrm{IC}_t \sqrt{N} \sigma_{\text{model}} . \tag{4.26}$$

Equation 4.26 reveals the real benefit of replacing the raw returns by the risk-adjusted returns in the calculation of excess return. We thus have one less variable to worry about. Note the following remark:

- In practice, the dispersion of risk-adjusted returns is neither exactly unity nor constant over time. There are at least three reasons for the possible bias and variation. First, there could be systematic factors missing from the risk model. In fact, this is almost a certainty if we are to believe there are separate alpha factors. Second, there are systematic estimation errors in the specific risks. Lastly, there is a distinct possibility that a multifactor risk model is simply not adequate.

4.2.5 "Purified Alpha" and Its IC

A similar approach to remove systematic exposures embedded in any alpha factor is to regress it against the risk factors and use only the residual from the regression — purified alpha — as forecasts. In this way, the alpha is "purified" and we can then calculate its IC — the cross-sectional correlation coefficient between the purified alpha and the raw returns. Let us denote the purified alpha by

$$\mathbf{f}_{\text{pure}} = \mathbf{f} - n_0 - n_0 \mathbf{b}_1 - \cdots - n_K \mathbf{b}_K , \tag{4.27}$$

with n's being the regression coefficients, given by

$$\begin{pmatrix} n_0 \\ n_1 \\ \vdots \\ n_K \end{pmatrix} = \left[\begin{pmatrix} \mathbf{i}' \\ \mathbf{b}_1' \\ \vdots \\ \mathbf{b}_K' \end{pmatrix} \begin{pmatrix} \mathbf{i} & \mathbf{b}_1 & \cdots & \mathbf{b}_K \end{pmatrix} \right]^{-1} \begin{pmatrix} \mathbf{i}' \cdot \mathbf{f} \\ \mathbf{b}_1' \cdot \mathbf{f} \\ \vdots \\ \mathbf{b}_K' \cdot \mathbf{f} \end{pmatrix} . \tag{4.28}$$

At the first glance, the purified alpha should not introduce any systematic risk to the IC, and the only weakness is in its dealing with stock-specific risks. This first impression is not correct unless all stock-specific risks are the same. Alternatively, it is only correct if we form portfolios in such a way that portfolio weights are proportional to the purified alpha in the

manner of (4.7). Because this is usually not the case, the purified alpha is not so pure. Although the purified alpha and its IC represent an improvement over the raw forecasts and the raw IC (4.8), it is not free of systematic exposures under the risk model (4.10).

We demonstrate this by showing that the purified alpha is equivalent to the risk-adjusted forecast when we have the following risk model

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_I\mathbf{B}' + s^2\mathbf{I}\,, \tag{4.29}$$

with $\mathbf{I}$ being an identity matrix, i.e., the specific risk is s for all stocks. When this is the case, Equation 4.14 is just proportional to the inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}' \cdot \mathbf{S}^{-1} \cdot \mathbf{y} = \frac{1}{s^2}\sum_{i=1}^{N} x_i y_i = \frac{1}{s^2}\mathbf{x}' \cdot \mathbf{y}\,. \tag{4.30}$$

And the solution of (4.16) for the Lagrangian multipliers reduces to the solution of (4.28) for the regression coefficients. Therefore, the purified alpha and the risk-adjusted forecast are proportional to each other.

When the specific risks are not identical, we can align purified alpha in line with the risk-adjusted forecast by a weighted cross-sectional linear regression, with weight for each stock being the inverse of its specific variance. In such a case, it can be proven the purified alpha equals the risk-neutral forecast — the denominator of (4.13). This is left as an exercise.

4.3 MULTIPERIOD *EX ANTE* INFORMATION RATIO

Equation 4.25 is close to a mathematical identity. Although it is always true *ex post*, we now use it *ex ante* by considering its expectation and standard deviation, i.e., the expected excess return and the expected active risk. Among the four terms affecting the excess return, we assume that the number of stocks does not change over time. We also assume the risk-model tracking error remains constant, implying we target the same level of active risk at each rebalance of the portfolio, a typical practice for many quantitative portfolio managers. There are good reasons for keeping the target tracking error constant. First, varying the tracking error introduces portfolio turnover or trading, purely based on changing risk aversion. Second, and perhaps more importantly, for most quantitative factors, such as value and momentum, the dispersion of the forecasts does not seem to be

correlated with the dispersion of returns. In other words, conviction does not translate into realized opportunity in reality. Then, it is reasonable that one does not benefit from varying active risks.

For the two remaining terms that do change over time, the IC is usually associated with greater variability than the dispersion of the risk-adjusted returns. The latter term, as we discussed earlier, should approximately equal unity, at least in theory. Therefore, as a first approximation, we treat it as a constant.

Assuming $\text{dis}(\mathbf{R}_t)$ is constant and equal to its mean, the expected excess return is

$$\overline{\alpha_t} = \overline{IC_t}\sqrt{N}\sigma_{\text{model}}\overline{\text{dis}(\mathbf{R}_t)}. \tag{4.31}$$

The expected excess return is therefore the product of the average IC (skill), square root of N (breadth), the risk-model tracking error (risk budget), and the dispersion of actual returns (opportunity). The expected active risk is

$$\sigma = \text{std}(IC_t)\sqrt{N}\sigma_{\text{model}}\overline{\text{dis}(\mathbf{R}_t)}. \tag{4.32}$$

The standard deviation of IC measures the consistency of forecast quality over time. Therefore, the active risk is the product of the standard deviation of IC (consistency), the square root of N (breadth), the risk-model tracking error (risk budget), and the dispersion of actual returns (opportunity).

The ratio of Equation 4.31 to Equation 4.32 produces the IR

$$IR = \frac{\overline{IC_t}}{\text{std}(IC_t)}. \tag{4.33}$$

The IR is the ratio of the average IC to the standard deviation of IC.

4.3.1 Fundamental Law of Active Management

Grinold (1989) proposed the Fundamental Law of Active Management (FLAM) — IR is the product of IC and the square root of breadth. In the case of equity portfolios, the breadth of investment opportunities is understood as the number of stocks available. Grinold derived the result with a

different approach (Problem 4.3). But it is easy to derive it from Equation 4.33. When the standard deviation of IC is

$$\text{std}\left(IC_t\right) = \frac{1}{\sqrt{N}}, \tag{4.34}$$

we have

$$IR = \overline{IC_t}\sqrt{N}\ . \tag{4.35}$$

Thus, the FLAM hinges on the assumption that the standard deviation of IC over time equals $1/\sqrt{N}$. Moreover, under this assumption, the active risk (4.32) reduces to

$$\sigma = \sigma_{\text{model}}\overline{\text{dis}\left(\mathbf{R}_t\right)}. \tag{4.36}$$

Thus, the active risk is close to the target tracking error given in our previous discussion about the dispersion of risk-adjusted returns. Therefore, one can conclude the FLAM depends on the assumption that target tracking error given by the risk model gives an accurate prediction of active risk of alpha factors.

So when is Equation 4.34 true? This assumption is approximately correct if the underlying population correlation coefficient between the risk-adjusted forecasts and the risk-adjusted return is constant over time, and the standard deviation of IC over time is purely because of sampling error. Suppose the underlying population correlation between $\mathbf{F}_t$ and $\mathbf{R}_t$ is ρ, then the standard error of the sample correlation coefficient with a sample of size N is (e.g., see Keeping 1995)

$$\text{stderr}\left(IC_t\right) \approx \frac{\sqrt{1-\rho^2}}{\sqrt{N}}. \tag{4.37}$$

Because the IC is usually small, for example, on a quarterly horizon, most of the quantitative alpha factors have IC less than 0.1, making the numerator of (4.37) close to unity. Therefore, the standard error of IC is indeed close to $1/\sqrt{N}$. However, note the following remark:

- Although the FLAM is theoretically appealing and has wide acceptance by practitioners, the assumption about the standard deviation of IC proves to be too simplistic to be practical. Actually, Grinold did not intend to put forth a descriptive portfolio solution but rather a normative expression to capture the essence of manager skill. For example, it implies the standard deviation of IC is the same for different alpha factors. In the next section, we argue from both theoretical and empirical standpoints that this is hardly true. Past research studies that confirmed the FLAM have done so, using Monte Carlo simulations with normative design rather than descriptive accuracy.

4.3.2 Target Risk, Realized Risk, and *Ex Ante* Risk

The true *ex post* active risk of an active portfolio is not necessarily equal to the targeted risk. This should not be a surprise to anyone, because the targeted risk is only an estimation based on risk models. There are a variety of model errors pertaining to risk models. For instance, Hartmann et al. (2002) studied the measurement error of risk models over a single rebalancing period by analyzing the performance of risk models over a single, relatively short period, during which the examined portfolios are bought and held. The approach is to compare predicted tracking errors of a risk model to the realized tracking errors, using either daily or weekly excess returns, for many simulated portfolios. Hartman et al. (2002) attribute the difference between the estimated risk and the *ex post* tracking error to several reasons: estimation error in covariances in a risk model, time-varying nature of covariances, serial autocorrelations of excess returns, and the drift of portfolio weights over a given period. Depending on how these influences play out in a given period, a risk model can overestimate, as well as underestimate with roughly equal probability, *ex post* tracking errors of simulated portfolios. There is no clear evidence of bias one way or the other.

In contrast, we focus on the active risk of an active portfolio over multiple rebalancing periods, during which the active portfolio is traded periodically, based on the alpha factors. Equation 4.32 reveals a potential bias in the target risk that might be due to an entirely different reason — variability in the IC over time.

It is understandable that the variability of IC plays a role in determining the active risk. For a thought experiment, just imagine two investment strategies, both taking the same risk-model tracking error σ_{model} over time.

The first strategy is blessed with perfect foresight and generates constant excess return every single period. In other words, it has a constant positive IC for all periods such that $\text{std}(IC_t)$ is zero. No sampling error has to be considered. Such a risk-free strategy, admittedly hard to find, has constant excess return, and thus no active risk whatsoever. However, the risk model is not aware of the prowess of the strategy and dutifully predicts tracking error σ_{model} all the time. In this case, the risk model undoubtedly overestimates the active risk. In contrast, the second strategy is extremely volatile with large swings in its excess return, i.e., its IC varies between –1 and +1 with a large $\text{std}(IC_t)$. As a result, its active risk might be much larger than the risk-model estimate. Thus, the two strategies with identical risk-model tracking errors have very different active risks in actuality.

In practice, the difference between active investment strategies is not this extreme. All have some alpha model risk (volatility in IC), but few swing between –1 and +1. However, our experience shows that risk-model tracking error given by various commercially available risk models routinely, and sometimes seriously, underestimates the *ex post* active risk. Other practitioners have also recognized this problem. For example, Freeman (2002) notes that "if a manager is optimizing the long-short portfolio, he or she better assume that the tracking error forecast (of a risk model) will be at least 50% too low." This underestimation could have serious practical consequences.

For this reason, we term $\text{std}(IC_t)$ as strategy risk, because it is tied to an individual investment strategy that employs different alpha factors. It is important to point out the difference between the terminologies used so far. Here is a summary:

- *Risk-model tracking error:* Denoted as σ_{model}, it is the tracking error or the standard deviation of excess returns estimated by a generic risk model, such as BARRA, and it is also referred to as risk-model risk or target tracking error.
- *Strategy risk:* Denoted as $\text{std}(IC_t)$, it is the standard deviation of IC of an investment strategy over time. It is unique to each active investment strategy, conveying strategy-specific risk profile.
- *Active risk:* Denoted as σ, it is the active risk or tracking error of an investment strategy measured by the standard deviation of excess returns over time.

It is possible to segregate the strategy risk into the sample error and true variation in the IC. Assuming the two are independent of each other, we have

$$\left[\mathrm{std}\left(IC_t\right)\right]^2 = \frac{1}{N} + \left[\sigma\left(IC_t\right)\right]^2 \tag{4.38}$$

- Based on the analysis of risk-adjusted IC, the ratio (4.33) serves a good proxy for a factor's efficacy in generating excess returns. This will be used again in Chapter 7 where we use this ratio for multifactor alpha models to derive optimal model weights.

4.3.3 A Better Estimation of IR

In reality, the variability in the dispersion of the risk-adjusted return $\mathrm{dis}\left(\mathbf{R}_t\right)$ is small but nonetheless nonzero. What happens to the IR if we include this variability? The following insight from Equation 4.25 helps us to understand how the interaction between the IC and the dispersion affects the excess return. To produce a high positive excess return for a single period, we need a high and positive IC, as well as a high dispersion. Conversely, when IC is negative, we wish for a low dispersion so that the negative excess return would be small in magnitude. This argument implies that over the long run, the performance will benefit from a positive correlation between the IC (skill) and the dispersion (opportunity). On the other hand, a negative correlation will hurt the average excess return.

The expected excess return including this correlation effect is

$$\overline{\alpha_t} = \sqrt{N}\sigma_{\mathrm{model}}\left\{\overline{IC_t}\,\overline{\mathrm{dis}\left(\mathbf{R}_t\right)} + \rho\left[IC_t, \mathrm{dis}\left(\mathbf{R}_{\mathrm{t}}\right)\right]\mathrm{std}\left(IC_t\right)\mathrm{std}\left[\mathrm{dis}\left(\mathbf{R}_{\mathrm{t}}\right)\right]\right\}. \tag{4.39}$$

The additional term is simply the covariance between the IC and the dispersion, written in terms of the correlation between the IC and the dispersion, and the standard deviations of the IC and the dispersion. This is because for two random variables (x, y) we have $\mathrm{E}\left(xy\right) = \overline{x}\,\overline{y} + \rho\sigma_x\sigma_y$.

The active risk including the variability of the dispersion can also be derived analytically (Problem 4.4). Because the coefficient of variation (the standard deviation over the mean) is much smaller for the dispersion than for the IC, the active risk is approximately unchanged. Combining Equation 4.39 with Equation 4.32 produces the new IR estimate

$$IR = \frac{\overline{IC_t}}{\mathrm{std}\left(IC_t\right)} + \rho\left[IC_t, \mathrm{dis}\left(\mathbf{R}_t\right)\right]\frac{\mathrm{std}\left[\mathrm{dis}\left(\mathbf{R}_t\right)\right]}{\overline{\mathrm{dis}\left(\mathbf{R}_t\right)}}. \tag{4.40}$$

The second term captures the correlation effect on the IR. It has two components. The first is the correlation between the IC and the dispersion over time, and the second term is the coefficient of variation of the dispersion. Note the following remark:

- As we mentioned earlier, the coefficient of variation of the dispersion is usually small. Therefore, the effect of the second term is typically small unless the correlation between the IC and the dispersion gets very high, either positive or negative. For most practical purposes, Equation 4.33, i.e., the first term in Equation 4.40, approximates IR well enough. Nonetheless, Equation 4.40 is an improvement.

4.4 EMPIRICAL EXAMPLES

In the remainder of the chapter, we present some empirical findings concerning active risk and IR of 60 alpha factors, encompassing a wide range of well-known market anomalies. The focus is solely on these statistical measures and not on the detailed description of the factors, which is the subject of the next chapter. The goal of the empirical examination is to demonstrate that Equation 4.32 is a more consistent estimator of *ex ante* active risk, and IR is the ratio of average IC to the standard deviation of IC. These examinations evaluate factors separately rather than jointly. We shall discuss methods of combining multiple alpha factors into a composite, later in Chapter 7.

First, a brief description of the data is in order. We apply the analysis to the universe of stocks in the Russell 3000 index from 1987 to 2003. The data is quarterly, and at the beginning of each quarter, we have available alpha factor values for individual stocks in the universe, constructed from various financial data sources. In addition, we also have available risk factor exposures and specific risk for individual stocks in the universe from the BARRA US E3 equity risk model. Because of data availability and exclusion of outliers, the actual number of stocks is fewer than 3000, and it fluctuates from quarter to quarter. However, the fluctuation is insignificant and does not alter the analysis.

At the beginning of each quarter, we form optimal long-short portfolios for that quarter. Subsequently, cross-sectional analyses of alpha and

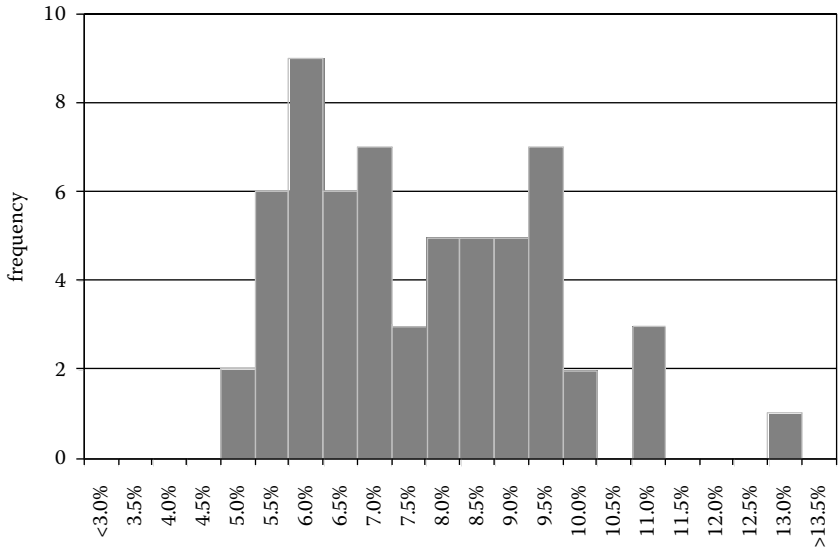


FIGURE 4.1. Histogram of the *ex post* active risk of equity alpha factors. (From Qian, E.E. and Hua, R., *Journal of Investment Management,* Vol. 2, Third Quarter, 2004. With permission.)

IC and dispersion of the risk-adjusted returns are computed on a quarterly basis. We set the constant risk-model tracking error at 2.5% per quarter, or 5% per annum. Additionally, to control risk exposures appropriately, we neutralize active exposures to all BARRA risk factors (13 systematic risk factors and 55 industry risk factors) when rebalancing portfolios each quarter. Hence, the risk-model risk is 100% stock-specific according to the risk model. We collect the results on a quarterly basis and then annualize.

Figure 4.1 shows the histogram of *ex post* active risk of the 60 alpha factors. Although the risk-model tracking error is targeted at 5% for all strategies, the *ex post* active risks differ widely with substantial upward bias, indicating the risk model's propensity to underestimate active risk. The average active risk is 7.7%, and their standard deviation is 1.7%. The highest active risk turns out to be 13.1%, whereas the lowest is just 5.0%. In other words, almost all strategies experienced a higher risk *ex post* than what the risk model predicted. To gauge the risk model's estimation bias in relative terms, we define a scaling constant,

$$\kappa = \mathrm{std}\left(IC\right)\sqrt{N} \approx \frac{\sigma}{\sigma_{\mathrm{model}}}. \tag{4.41}$$

Figure 4.2 shows the histogram of the scaling constant κ for all 60 strategies. Note that, for a majority of strategies, the model underestimates the *ex post* active risk by 50% or more.

Figure 4.3 shows the dispersion of risk-adjusted returns over time. It has an average of 1.01 and a standard deviation of 0.15. By this measure, the BARRA US E3 equity model shows internal consistency.

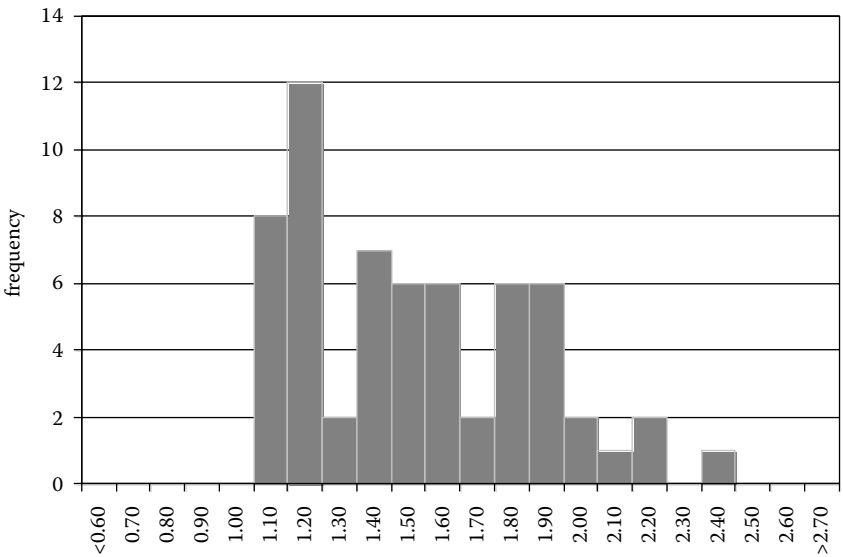


FIGURE 4.2. Histogram of the scaling constant κ . (From Qian, E.E. and Hua, R., *Journal of Investment Management,* Vol. 2, Third Quarter, 2004. With permission.)

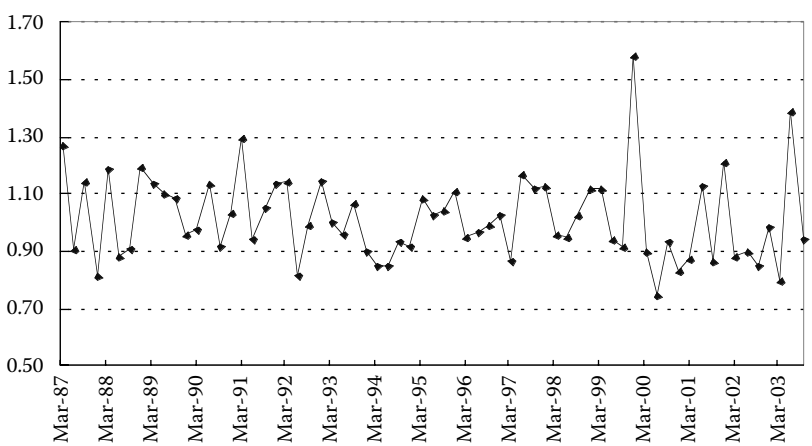


FIGURE 4.3. Dispersion of the risk-adjusted returns.

4.4.1 Two Alpha Factors

The strategy risks of these quantitative factors vary widely. Naturally, one wonders about the statistical significance of their differences. In other words, after appropriately controlling risk exposures specified by BARRA US E3 model in our case, does the standard deviation of ICs provide additional insight regarding the risk profile of a particular alpha factor? The answer to this question is "yes" in many cases. To demonstrate, we select two value factors — gross profit-to-enterprise value (GP2EV) and forward earnings yield based on IBES FY1 consensus forecast (E2P) — for a closer examination.

In Table 4.2, we see that, even though we targeted 5% tracking error for both factors, the realized tracking error is 6.9% for GP2EV and 8.7% for E2P. The average alpha (excess return) for GP2EV is at 6.2% with an IR of 0.90, and the average alpha for E2P is only 3.3% with an IR of 0.38. Next we show the average IC, the standard deviation of IC, and the IR, based on their ratio. As we can see, this approximation is very close to the actual IR based on the excess returns. The average dispersion of risk-adjusted returns is close to 1. Finally, we show the average number of stocks included in the portfolios based on the two factors. The number is lower for E2P because it is based on forward earning forecast, and many firms had no analyst coverage.

We perform two tests on the standard deviation of the ICs. First, we test the statistical significance of the difference between the two strategy risks using the F-test. Assuming both ICs are normally distributed, the ratio of their variance

$$F = \frac{\sigma^2\left(IC_1\right)}{\sigma^2\left(IC_2\right)} \tag{4.42}$$

follows an F-distribution with both degrees of freedom at 66, because both standard deviations are estimated over 67 quarters. Table 4.2 shows

TABLE 4.2 Summary Statistics of Two Value Factors

	Average Alpha	STD of Alpha	IR of Alpha	Average IC	STD of IC	IR of IC	Average Dispersion (R)	Average N
GP2EV	6.2%	6.9%	0.90	2.4%	2.7%	0.91	1.01	2738
E2P	3.3%	8.7%	0.38	1.4%	3.4%	0.41	1.00	2487

Source: From Qian, E.E. and Hua, R., *Journal of Investment Management,* Vol. 2, Third Quarter, 2004. With permission.

that for GP2EV and E2P, the standard deviation of IC is 2.7% and 3.4%, respectively. The variance ratio of the two factors is $3.4^2/2.7^2 = 1.58$, and α equals 0.033. Thus, in this example, there is enough evidence to reject the null hypothesis that these two factors (from the same value category) have the same strategy risk at a 5% confidence level. Our results indicate that the strategy risks of factors selected from different categories, more often than not, are statistically different.

The second test concerns whether the individual factor's strategy risk is significantly higher than the pure sampling error — 1/*N*. We shall use the average of *N* to compute the sampling error, because its variation is negligibly small. For this test, we find the confidence interval of the IC variance, based on the *ex post* value. If we denote the true or population variance by σ^2_{true}, then the ratio

$$\frac{m\sigma^2(IC)}{\sigma^2_{\text{true}}} \tag{4.43}$$

follows a χ^2 distribution with $m = 66$ degrees of freedom. The lower and upper confidence limits for σ^2_{true} are given, respectively, by

$$\sigma_1^2 = \frac{m\sigma^2(IC)}{\chi_1^2}, \qquad \sigma_2^2 = \frac{m\sigma^2(IC)}{\chi_2^2}. \tag{4.44}$$

The values of χ_1^2 and χ_2^2 are given by

$$\mathrm{P}\left(\chi^2 \geq \chi_1^2\right) = \frac{\alpha}{2}, \quad \mathrm{P}\left(\chi^2 \leq \chi_2^2\right) = \frac{\alpha}{2}. \tag{4.45}$$

For a chi-square distribution with 66 degrees of freedom, the values of χ_1^2 and χ_2^2 corresponding to $\alpha = 1\%$ are 99.3 and 40.2, respectively. Given the sample variance of each factor we use (4.44) to derive the limits for the IC variances, and we take their square roots as the confidence limits of the standard deviation of IC. Table 4.3 shows the results for both factors. For the factor GP2EV, the sample IC standard deviation is 2.7%, and the 99% confidence interval is between 2.2% and 3.5%. At the same time, the sampling error based on $N = 2738$ is only 1.9%, which lies outside the confidence interval. Thus, we can conclude that the true IC standard deviation is significantly higher than the sampling error. The same is true

TABLE 4.3 The 99% Confidence Interval for the Standard Deviation of IC and Sampling Error of IC

	STD of IC	Lower Limit	Upper Limit	Sampling Error	Average N
GP2EV	2.7%	2.2%	3.5%	1.9%	2738
E2P	3.4%	2.8%	4.4%	2.0%	2487

for the earning yield. Its 99% confidence interval is $(2.8\%, 4.4\%)$, but the sampling error is only 2.0%. In fact, the significance is much higher than the 99% indicated here (Problem 4.6).

4.4.2 *Ex Ante* Estimate of Active Risk and Information Ratio

The empirical results show that active risk consists of two components: risk-model tracking error and strategy risk, consistent with Equation 4.32. Merely using the sampling error ($1/\sqrt{N}$) could severely underestimate the active risk of an active strategy. Based on this observation, practitioners can use strategy risk in conjunction with a risk model to obtain a more consistent active risk forecast. As an illustration, we divide the sample period into two halves: in-sample period (1986–1994) and out-of-sample period (1995–2003). In the in-sample period, we estimate κ according to Equation 4.41 for each of the 60 equity strategies. Then, in the out-of-sample period, we adjust the risk-model tracking error by $1/\kappa$, using strategy-specific κ to compensate the risk model's bias in estimating active risk. In other words, the adjusted risk-model target tracking error is $\sigma_{\text{model}}/\kappa$. Because κ is greater than one for almost all alpha factors, we have effectively lowered our target tracking error according to the values of κ.

Figure 4.4a shows the distribution of *ex post* active risks in the out-of-sample period, when we set the target tracking error at $5\%/\kappa$ (the adjusted risk-model tracking error), and, for comparison, Figure 4.4b shows active risk of portfolios targeting the same tracking error at 5% (the original risk-model tracking error). We would like to emphasize again that the adjusted risk-model tracking error σ^*_{model} is unique to each equity strategy depending on its κ estimate, whereas the risk-model tracking error σ_{model} is the same for all strategies. From these two histograms, it is obvious that σ^*_{model} is a more consistent estimator of active risk. The average *ex post* active risk is 4.7% when using σ^*_{model} and 7.6% when using σ_{model}. Thus, the expected *ex post* active risk is much closer to our target of 5% with no bias when using the adjusted risk-model tracking error. The standard deviation of *ex post* active risk is 0.76% when using σ^*_{model} and

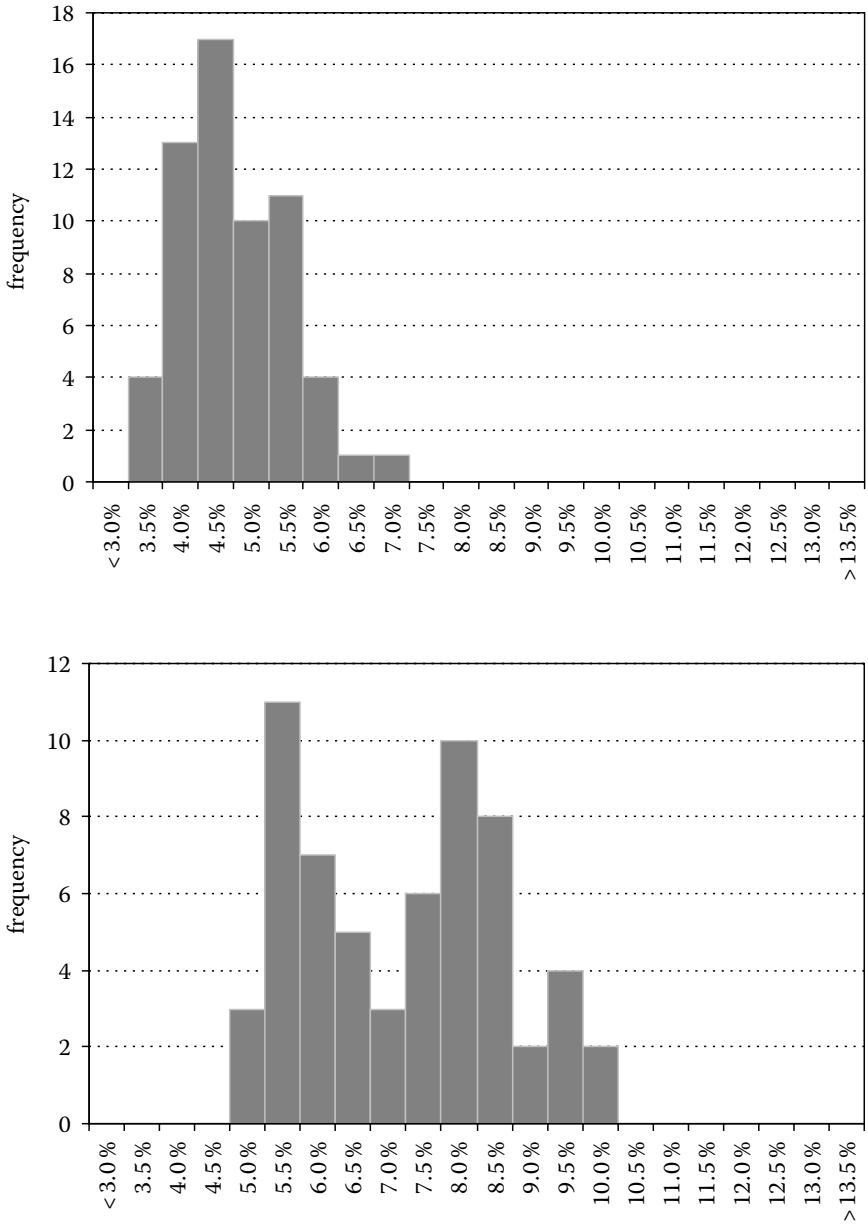


FIGURE 4.4. Histogram of the *ex post* active risks: (top) using adjusted risk-model tracking error (1995–2003) and (bottom) using 5% risk-model tracking error (1995–2003). (From Qian, E.E. and Hua, R., *Journal of Investment Management,* Vol. 2, Third Quarter, 2004. With permission.)

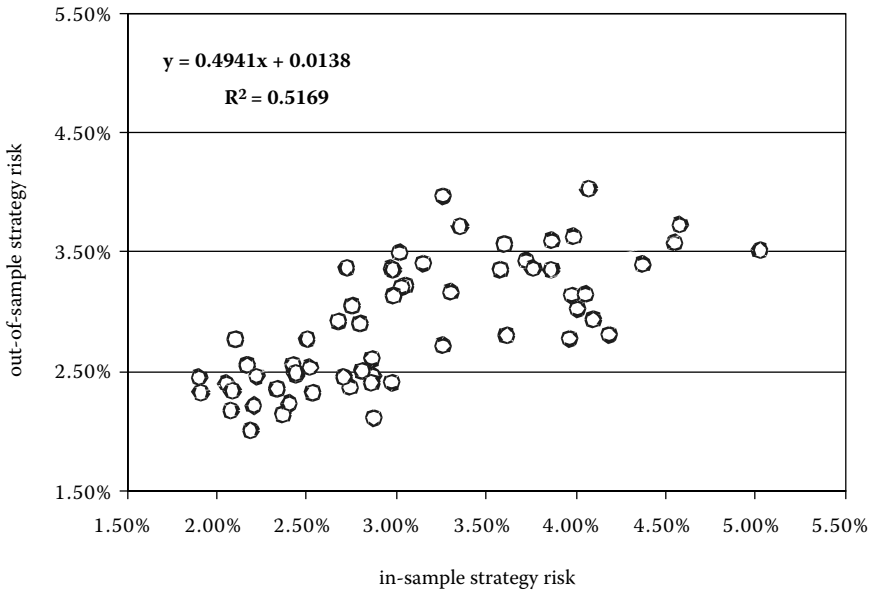


FIGURE 4.5. Scatter plot of in-sample strategy risk vs. out-of-sample strategy risk. (From Qian, E.E. and Hua, R., *Journal of Investment Management,* Vol. 2, Third Quarter, 2004. With permission.)

1.45% when using σ_{model}. It is apparent that in this shorter period, the risk model experienced the similar problem of underestimating the true active risks of many strategies.

The application of the scaling constant κ in the preceding estimation constitutes a simplistic form of forecasting strategy risk — using the strategy risk estimated in the in-sample period as the forecast of the out-of-sample period. Our simplistic forecasting method assumes that strategy risk persists from the in-sample to the out-of-sample period. One implication of this methodology is the relative ranking of strategy risks remains the same in both periods. Figure 4.5 is the scatter plot of strategy risks measured in the in-sample period (x-axis) vs. the out-of-sample period (y-axis). The R-squared of the regression, using in-sample strategy risks to explain the variability of out-of-sample strategy risks, is 52%. Hence, it is plausible that, with this simple forecast method in conjunction with Equation 4.32, active managers can improve their ability to assess portfolio active risk and IR.

PROBLEMS

4.1 Correct Equation 4.6 when the weights are not dollar neutral. This result would be applicable to long-short hedge funds with a long bias.

4.2 We obtain purified alpha by a weighted cross-sectional regression of raw forecast vs. risk factors. It seeks to minimize the following function

$$\text{MSE} = \sum_{i=1}^{N} \frac{\left(f_i - n_0 - n_1 b_{1i} - \cdots - n_K b_{Ki}\right)^2}{\sigma_i^2}. \tag{4.46}$$

Prove that the solution of the regression coefficients is identical to the Lagrangian multipliers of the risk-adjusted forecasts. Is the correlation coefficient between the purified alpha and realized return the same as the risk-adjusted IC?

4.3 Derive the Fundamental Law of Active Management based on expected excess return of individual securities. Assume the risk-adjusted forecasts are normalized such that $\text{dis}\left(\mathbf{F}_t\right) = 1$.

(a) What is the equation for the risk-aversion parameter?

(b) Suppose the expected residual return is the product of volatility, IC, and score (Grinold 1994), prove $\text{E}\left(R_i\right) = IC_t F_i$.

(c) Take the expectation of Equation 4.19 and show that

$$\frac{\bar{\alpha}_t}{\sigma_{\text{model}}} \approx IC_t \sqrt{N} \tag{4.47}$$

(d) Interpret Equation 4.47 as a "one-period IR" — the ratio of expected excess return to the risk-model risk.

4.4 We derive variance of a product of two normal random variables x, y.

(a) Prove: $\text{E}\left(xy\right) = \bar{x}\,\bar{y} + \rho\sigma_x\sigma_y$.

(b) Prove:

$$\text{E}\left(x^2 y^2\right) = \sigma_x^2\sigma_y^2 + 2\rho^2\sigma_x^2\sigma_y^2 + \bar{x}^2\sigma_y^2 + \sigma_x^2\bar{y}^2 + \bar{x}^2\bar{y}^2 + 2\rho\overline{xy}\sigma_x\sigma_y. \tag{4.48}$$

(c) Prove:

$$\mathrm{Var}(xy) = \sigma_x^2\sigma_y^2 + \rho^2\sigma_x^2\sigma_y^2 + \bar{x}^2\sigma_y^2 + \bar{y}^2\sigma_x^2 . \tag{4.49}$$

(d) Show when $\frac{\sigma_y}{\bar{y}} << 1$ and $\frac{\sigma_y}{\bar{y}} << \frac{\sigma_x}{\bar{x}}$, the variance can be approximated by

$$\mathrm{Var}(xy) = \bar{y}^2\sigma_x^2 . \tag{4.50}$$

This approximation justifies using of Equation 4.32 for the active risk even when the dispersion of risk-adjusted returns is not constant.

4.5 Estimate standard deviation of IC by numerical simulation. Suppose for a portfolio of 500 (***N***) stocks, the average IC is 0.05. Simulate forecasts and returns as a bivariate normal distribution with zero means and standard deviation one and calculate the realized IC. Select number of periods as *M*.

(a) Assuming there is no variation in the IC, show that the standard deviation of the realized IC approaches $1/\sqrt{N}$.

(b) Suppose the IC is not constant over time, and its intrinsic variation is 0.05. Then, for each period, the IC is drawn from a normal distribution of mean 0.05 and standard deviation of 0.05. Simulate cross-sectional forecasts and returns based on the drawn IC and calculate the realized IC. Verify Equation 4.38.

4.6 For the factor gross profit to enterprise value (GP2EV), with 99% confidence coefficient, the lower limit of IC standard deviation is 2.2%, higher than the sampling error of 1.9%.

(a) What is the minimal value of χ_1^2 that would make the sampling error fall into the confidence interval?

(b) Find the probability $\mathrm{P}\left(\chi^2 \geq \chi_1^2\right)$.

(c) Repeat question (a) and (b) for the factor E2P.

REFERENCES

BARRA, *United States Equity — Version 3, Risk Model Handbook,* BARRA Inc., 1998.

Clarke, R., de Silva, H., and Thorley, S., Portfolio constraints and the fundamental law of active management, *Financial Analyst Journal,* Vol. 58, No. 5, 48–66, September–October 2002.

Freeman, J.D., Portfolio construction and risk management: Long-short/market-neutral portfolios, in *AIMR Conference Proceeding: Hedge Fund Management,* pp. 41–46, 2002.

Grinold, R.C., The fundamental law of active management, *Journal of Portfolio Management,* Vol. 15, No. 3, 30–37, Spring 1989.

Grinold, R.C., Alpha is volatility times IC times score, *Journal of Portfolio Management,* Vol. 20, No. 4, 9–16, Summer 1994.

Grinold, R.C. and Kahn, R.N., *Active Portfolio Management*, McGraw-Hill, New York, 2000.

Hartmann, S., Wesselius, P., Steel, D., and Aldred, N., Laying the foundations: Exploring the pitfalls of portfolio construction and optimization, working paper, ABN AMRO, 2002.

Keeping, E.S., *Introduction to Statistical Inference*, Dover, New York, 1995.

Qian, E.E. and Hua, R., Active risk and information ratio, *Journal of Investment Management,* Vol. 2, Third Quarter, 2004.

CHAPTER 5

Quantitative Factors

In Chapter 4, we developed an analytic framework to evaluate alpha factors. We now take a closer look at the typical quantitative strategies (alpha factors) comprising three broad categories: value, momentum, and quality. First, value factors seek to identify securities which are trading at bargain prices, which is attributable to investors' excessive pessimism. Second, momentum factors ride winners and expel losers, exploiting investors' inability to incorporate public information in a timely manner. Third, quality factors identify companies that are more likely to create shareholder value by avoiding the agency problem trap. In this chapter, we explore the fundamental underpinnings of these factors, along with the relevant academic literature. We also examine factor construction and historical performance.

5.1 VALUE FACTORS

Value investing is a time-tested cornerstone of active security selection. The prescription is to buy stocks that have relatively low prices translated into ratios deflated by fundamental criteria such as dividends, book value, earnings, cash flows, or other measures of firm value. Benjamin Graham, in the book *The Intelligent Investor*, associated value with a margin of safety, which enables the investment to withstand adverse business developments. Warren Buffet termed Graham's value philosophy as the "cigar butt" approach to investing and said, "A cigar butt found on the street that has only one puff left in it may not offer much of a smoke, but the 'bargain purchase' will make that puff all profit."

A long list of academic literature has focused on documenting the value phenomenon, beginning with Basu (1977) and replicated by Jaffe et al. (1989), Chan et al. (1991), and Fama and French (1992) all showing that stocks with high fundamentals-to-price ratios (say, earnings-to-price) earn higher average returns. Rosenberg et al. (1985) demonstrate that stocks with high book-to-market ratios outperform the market. Additionally, Chan et al. (1991) find that a high ratio of cash-to-price also predicts higher returns. Finally, Cohen and Polk (1998) illustrate that industry adjustment to the book-to-market improves the Sharp ratio of portfolio excess returns.

Although academics agree that value stocks provide above-market returns, they have considerable disagreements about whether this premium is a compensation for risk taking (beta) or a systematic exploitation of irrational behavioral biases (alpha). Fama and French (1993, 1996) suggest that the value premium is simply a compensation for higher systematic risk, namely, financial distress. They assert that companies with high book-to-market ratios are under greater financial distress and more vulnerable to any downturns of the business cycle. In contrast, Lakonishok et al. (1994) suggest that the value premium can be traced to investor's biased cognitive inference that incorrectly extrapolates the past earnings growth rate of firms. They suggest that investors are overly optimistic about firms that have done well in the past and are overly pessimistic about those that have done poorly. As a result, glamorous (low book-to-market) stocks attract naive investors who push up the prices and, hence, lower the expected returns of these securities. Lending more credence to this hypothesis, Rozeff and Zaman (1998) argue that insider buying escalates as stocks change from the low cash-to-price to the high cash-to-price category. Given that insiders know more than the general public about company prospects, this supports the hypothesis that value premium is not solely related to financial distress.

5.1.1 Value Measures

There are a variety of ways to characterize a firm's intrinsic value. We can define cheapness as high cash flow yield, high earnings yield, high dividend yield, or high book-to-market value. Whereas cash flow and earnings yield emphasize the profitability of existing operations, asset value ratio is a measure of liquidation value, and dividend yield relates to dividend payout policy, which typically conveys management's assessment of long-term profitability. Because stakeholders can be defined narrowly as

TABLE 5.1 Commonly Used Value Measure

	Equity	**Enterprise**
Cash Flows	CFO to Market Value FCF to Market Value	CFO to EV FCF to EV EBITDA to EV Gross Profit to EV
Earnings	Net Income to Market Value IBES FY1 Forecast to Market Value IBES Twelve-month Forecast to Market Value	NOPAT to EV
Dividends	Indicated Dividend Yield Dividends plus Net Share Repurchase to Market Value	Dividends minus External Financing to EV
Asset Value	Book to Price	Net Operating Assets to EV Sales to EV

equity holders or broadly as *enterprise holders* (including both equity and bond holders), matching the right intrinsic value with its corresponding market value is an important consideration when computing value ratios. Take earnings yield as an example. For equity holders, earnings yield is a ratio of levered earnings (or net income before extraordinary items on the income statement) divided by the market value of equity. In contrast, for the enterprise version of earnings yield, the numerator is the unlevered earning (or net operating income after tax, aka NOPAT), and the denominator is the enterprise value that equals market value of equity plus market value of debt[1] minus excess cash. Table 5.1 lists commonly used value factors by their intrinsic measure and their stakeholder. (Please refer to the Appendix A5.1 for a detailed description of how we construct these value factors with the Compustat database.)

5.1.2 Value vs. Valuation: A Clarification

We now clarify the philosophical difference between value and valuation investing — two popular approaches that are often mislabeled by practitioners as being interchangeable. As defined above, value investing seeks to buy the lowest priced stocks and sell the highest priced stocks without considering the company's future growth prospect or profitability. As such, value strategies typically purchase securities issued by firms with low return on equity (ROE) and high financial leverage — a reflection of cigar-butt investing. In comparison, valuation investing seeks to purchase

securities whose market values are significantly lower than their fair valuations determined by companies' profitability and growth prospect. (In Chapter 6, we will review valuation investing in detail.)

Let us use the book-to-price (B2P) ratio as an example. Value investors (sometimes referred to as *deep value)* buy the highest B2P stocks, whereas valuation investors examine B2P ratios in conjunction with ROE measures so that the analysis is relative when selecting bargain purchases. OLS regression is a common method to derive fair valuation quantitatively. It establishes the equilibrium pricing of ROE empirically and estimates the extent to which market prices deviate from the equilibrium valuation. Equation 5.1 presents the regression formula incorporating the relationship between B2P and ROE, along with the coefficient estimate over the sample period for stocks in the Russell 3000 universe.[2] In this case, the valuation investor buys securities with the highest regression residuals ε_i, reflecting the portion of cheapness, i.e., B2P not explained by cross-sectional differences in ROE. High ROE should command low B2P. Cheap stocks are those that have high B2P readings — after conditioning on ROE. The mean coefficients and *t*-statistics (in parentheses) are then computed, based on the Fama–MacBeth regression method:

$$\begin{array}{lcll} \mathrm{B2P}_i & \sim & 66 & -\ 0.33\,\mathrm{ROE}_i + \varepsilon_i \\ & & (132) & (-32.6) \end{array} \tag{5.1}$$

The *t*-stat of ROE is –32.6, indicating a persistent negative correlation between ROE and B2P. That is, high ROE companies tend to have low B2P ratio and *vice versa*. Figure 5.1 plots the estimated coefficient of ROE through time. The correlation is quite stable in the sample period with the noticeable exception during the stock market bubble of 1999 and 2000.

To further illustrate the difference, Table 5.2 lists the top 10 stocks in the two strategies at the end of 2004 along with B2P, ROE, and debt-to-asset ratio (D/A). Panel A presents the value strategy that buys stocks with high B2P ratios, low ROEs, and high financial leverage. Panel B presents the strategy of buying higher exposure to ROE and lower exposure to financial leverage.

5.1.3 Important Practical Considerations

To implement a robust value strategy, one has to carefully consider the following practical issues:

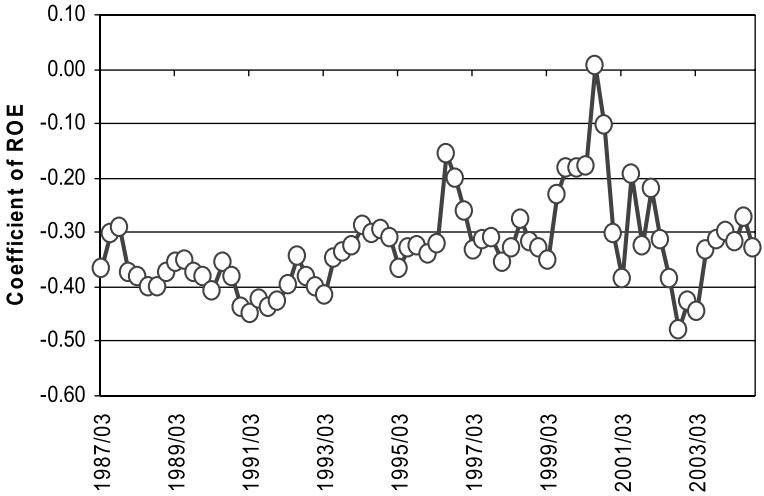


FIGURE 5.1. Time series of regression coefficient in Equation 5.1.

TABLE 5.2 Panel A — Top Ten Names for Value Strategy

Sector	Ticker	B/P	ROE	D/A
Discretionary	BBI	12.09	–28.38	0.03
Materials	PCU	6.94	19.66	0.04
Utilities	CPN	3.57	1.10	0.66
Discretionary	TWRAQ	3.45	–30.87	0.49
Financials	GNW	3.25	8.83	0.05
Industrials	FADV	2.85	1.66	0.22
Staples	PTMK	2.56	3.70	0.42
Financials	NFS	2.44	9.44	0.01
Discretionary	MECA	2.43	–18.95	0.32
Discretionary	XIDE	2.40	186.33	0.20
	Average:	4.20	15.25	0.24

TABLE 5.2 Panel B — Top Ten Names for Valuation Strategy

Sector	Ticker	B/P	ROE	D/A
Discretionary	XIDE	2.40	186.33	0.20
Technology	SOHU	1.32	36.25	0.05
Materials	PCU	6.94	19.66	0.04
Financials	CNO	0.94	98.88	0.04
Industrials	USG	1.05	26.97	0.00
Financials	CSWC	0.98	25.43	0.03
Financials	JNC	0.92	28.25	0.31
Telecom	TALK	0.84	64.03	0.09
Financials	LFG	1.31	15.77	0.15
Financials	GBL	1.38	15.01	0.39
	Average:	1.81	51.66	0.13

Earnings yield vs. PE ratio: To facilitate cross-sectional comparison, earnings yield should be used instead of PE ratio. Between positive and negative earning companies, the earnings yield measure provides a correct rank ordering, whereas the PE ratio mistakenly makes negative earnings companies more attractive as the lower PE is considered to be cheaper.

Peer group selection: Because cheapness is a relative concept determined through peer group comparison, how peer groups are constructed becomes an important consideration. For example, when cheapness is measured relative to the entire investable universe, it may result in a persistent sector bias — buying sectors that are consistently cheaper (such as utilities) and shorting sectors that are more expensive (like technology). In practice, sector classifications are commonly used as the peer group for several reasons. First, it avoids persistent sector bets due to persistent, cheap, or expansive valuation. Second, commonly used sector definitions provide a reasonable number of securities in each sector, thus facilitating a robust cross-sectional comparison. (This might not be true for many industry or other partioning schemes in which the number of firms is limited.) Third, companies, within the same sector, face similar operating challenges, such as economic cyclicality or secular changes induced by technological innovations, and share comparable operation characteristics such as margin, financial leverage, and growth rate. Lastly, many risk models (like BARRA) formally include sectors in the specification of portfolio risk.

Stock- or enterprise-based ratios: Value ratios can reflect either stockholder interests or the larger circle of enterprise holder interests. What are the pros and cons to consider in deciding the preferred choice? The difference between stock- and enterprise-based ratios relates to financial leverage. An unlevered (no debt) company will have the same ratio for both measures, whereas a higher financial leverage firm creates different readings. Stock-based ratios, like E/P ratios, are more sensitive to economic cycles than enterprise-based ratios like NOPAT/EV, especially for those cyclical sectors such as basic material and energy. Because of the artificial influence induced by financial leverage, the PE ratio prefers higher leveraged firms when the economy is at its peak and unlevered firms when it is at its trough, even if these companies are of the same cheapness measured by NOPAT/EV. As such, we recommend enterprise-based measures for companies in cyclical industries, whose growth rate is tightly tied to the overall growth of the economy.

5.1.4 Historical Performance of Value Factors

How do the performances stack up for the typical value factors? We consider eight value measures: cash flows from operations to enterprise value (CFO2EV), EBITDA (Earnings before Interest, Taxes, Depreciation, and Amortization) to enterprise value (EBIDTA2EV), trailing 12-month earnings yield (E2PFY0), earnings yield of IBES's EPS concensus estimate of the next fiscal year (E2PFY1), dividends plus net repurchases to market value (BB2P), net external financing to enterprise value (BB2EV), B2P, and sales-to-enterprise value (S2EV). Factors are evenly selected from all categories to facilitate a cross-category comparison of historical performance and their correlations.

To begin, we disclose the key elements in computing historical factor performance. This same methodology will also apply to other backtest results illustrated in the rest of this chapter.

1. Rank raw factor values by percentile within each sector to provide a more robust estimation and to avoid persistent sector bets.
2. The Russell 3000 Index is used as the sample universe through time to avoid survivorship bias.
3. We exclude the financial sector from this backtest because some ratios lose their meaning for financial companies. For example, one of the components in CFO (Cash Flow from Operating Activities) calculation is the year-over-year change in working capital, a concept that is meaningless for financial firms as they do not have inventory.
4. The backtesting sample period spans from 1986 to 2004.
5. For the risk-adjusted information coefficient (IC) calculation, we set the exposures to beta, size, and size nonlinearity to zero.
6. Three-month forward returns are used to compute historical performance.
7. Portfolios are rebalanced on a quarterly basis to correspond to the forward-return horizon and to avoid an overlapping performance period that typically results in high serial correlation of factor returns and biased standard error estimates.

Table 5.3 shows historical performances of value factors and their required turnover. The first three columns report time series statistics: risk-adjusted IC-average, t-statistics, and information ratio (IR). The next three columns show the same set of statistics for raw IC. The last two columns relate to portfolio turnover. Cross-sectional factor autocorrelation (CFA)

TABLE 5.3 Historical Performance of Value Factors (ICs)

	Performance						Turnover	
	ICa	**t(ICa)**	**IR(ICa)**	**IC**	**t(IC)**	**IR(IC)**	**CFA**	**TO**
CFO2EV	**6.66%	9.57	1.13	**7.20%	6.92	0.82	83.7%	151%
EBITDA2EV	**5.25%	6.72	0.79	**5.76%	5.17	0.61	86.6%	151%
E2PFY0	**3.89%	5.09	0.60	**4.33%	3.81	0.45	86.0%	154%
E2PFY1	**3.31%	3.67	0.43	*3.07%	2.50	0.29	86.0%	158%
BB2P	*2.65%	2.87	0.34	**3.72%	3.41	0.40	88.1%	125%
BB2EV	**4.24%	5.72	0.67	**5.13%	5.46	0.64	79.2%	167%
B2P	1.43%	1.46	0.17	1.54%	1.52	0.18	93.2%	121%
S2EV	**3.67%	3.79	0.45	**3.77%	3.51	0.41	96.0%	98%

Note: * = 90% confidence level; ** = 95% confidence level.

measures cross-sectional correlation of factor scores between two successive periods. TO is the quarterly turnover of long-short portfolios with 5% targeted tracking error. More analysis regarding portfolio turnover is provided in Chapter 8.

Most noticeable in Table 5.3 is the consistent excess returns delivered by these value factors, the very reason most active managers embrace value investing as a cornerstone of their investment principles. In Table 5.3, the achieved positive excess returns are significant at conventional statistical significance levels, with B2P being the only exception. In general, these results are robust across different performance measures: risk-adjusted IC (ICa) and traditional IC. Additionally, IR of ICa is generally higher than that of IC, whereas the average ICa is lower than the average of IC, reflecting the importance of using a refined risk process in assessing factor efficacy. For better visualization, Figure 5.2 presents a box chart of risk-adjusted ICs, including higher moments of the IC distribution. Aside from the positive shift in mean, most distributions also exhibit positive skew, with BB2EV being the most pronounced one. This general tendency of positive skew provides an additional benefit of using value factors that is not captured by IR. Note the following remark:

- Three observations are of interest. First, cash flow yield is the most relevant category in forecasting future returns, whereas asset value is the least. This perhaps reflects the notion that investors are generally more concerned about a firm's ability to generate cash flows as a going concern than a firm's liquidation value. Second, within the earnings yield category, using trailing, reported earnings provides

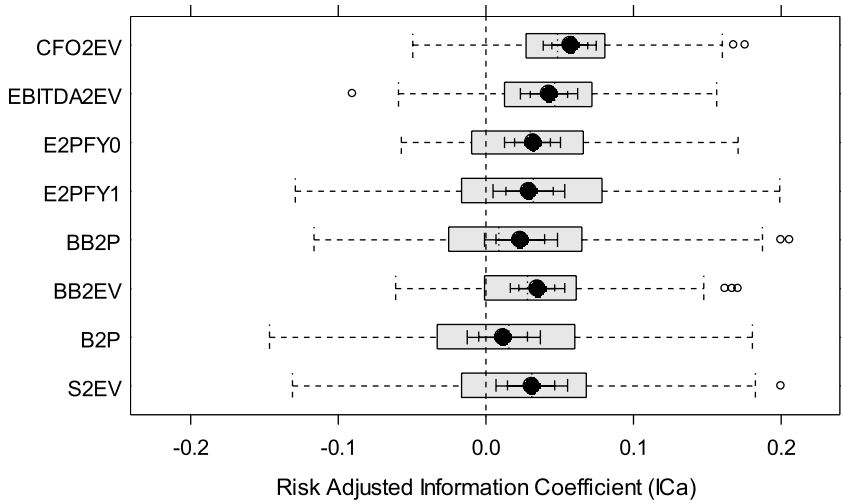


FIGURE 5.2. Box plots of risk-adjusted IC for value factors.

a more effective forecast than using IBES FY1 EPS estimate. Using reported EPS not only provides a higher average IC but also exhibits a lower standard deviation of IC, leading to a significantly better IR. The finding contradicts a popular but misguided belief commonly held by practitioners that forward-looking EPS forecast is a better gauge of value than the reported EPS, since the forward EPS encapsulates information pertaining to future developments. Conversely, empirical evidence supports (1) return predictability mostly arising from investor's under- or overreactions to the reported earnings and (2) sell-side estimates failing to provide forward-looking information, orthogonal to the information contained in the reported earnings. The third observation is that the required turnover of value factors varies between 100 and 150% per annum with CFA between 85 and 95% on a quarterly basis.

Table 5.4 shows the excess returns of decile portfolios for the selected value factors. They are computed in the following manner:

1. In the beginning of each period, ten decile portfolios are formed based on factor values of each security. That is, the top decile portfolio contains the top 10% of the securities possessing the highest factor values, the second decile portfolio contains the second highest 10% of securities, etc.

TABLE 5.4 Decile Performance for Value Factors

	Worst	2	3	4	5	6	7	8	9	Best
CFO2EV	**−3.40	**−1.70	**−1.15	**−0.62	0.02	−0.23	**1.08	**1.48	**2.18	**2.33
	(−5.55)	(−4.05)	(−3.95)	(−2.85)	(0.08)	(−0.71)	(3.87)	(4.49)	(5.21)	(5.60)
EBITDA2EV	**−2.99	*−1.17	*−0.74	*−0.68	−0.31	0.09	*0.78	**1.14	**2.06	**1.83
	(−4.12)	(−2.51)	(−2.11)	(−2.58)	(−1.13)	(0.34)	(2.20)	(3.33)	(4.93)	(4.65)
E2PFY0	**−2.58	**−1.41	−0.66	0.04	0.05	0.10	**0.80	**1.08	**1.26	**1.32
	(−3.80)	(−3.08)	(−1.95)	(0.15)	(0.19)	(0.38)	(2.88)	(2.68)	(3.06)	(3.09)
E2PFY1	*−1.84	−0.74	−0.41	0.22	−0.15	**−0.92	*0.73	**1.08	*0.96	1.07
	(−2.21)	(−1.52)	(−1.12)	(0.65)	(−0.54)	(−3.46)	(2.23)	(2.72)	(2.27)	(1.90)
BB2P	**−2.04	*−0.90	*−0.86	0.06	0.27	0.13	0.43	*0.63	*1.01	**1.27
	(−3.56)	(−2.03)	(−2.50)	(0.19)	(0.67)	(0.86)	(1.93)	(2.23)	(2.39)	(2.97)
BB2EV	**−1.50	**−1.48	**−1.16	**−0.93	−0.23	0.21	0.41	**1.15	**1.64	**1.90
	(−2.66)	(−3.91)	(−4.36)	(−3.91)	(−1.08)	(0.53)	(1.78)	(4.64)	(4.56)	(5.17)
B2P	*−1.18	−0.32	0.03	0.00	−0.18	0.05	**0.59	0.32	0.58	0.11
	(−2.64)	(−0.92)	(0.09)	(−0.01)	(−0.84)	(0.36)	(2.69)	(1.21)	(1.58)	(0.18)
S2EV	**−1.73	−0.82	*−0.66	−0.24	−0.15	−0.15	0.30	**0.91	**1.10	**1.42
	(−2.96)	(−1.95)	(−2.34)	(−1.11)	(−0.85)	(−0.80)	(1.21)	(3.11)	(2.87)	(3.21)

Note: * = 90% confidence level; ** = 95% confidence level.

2. Excess return of each decile portfolio is the difference between the equally weighted average security returns in the decile portfolio and the equally weighted return of the whole universe.
3. Time series average of decile portfolio, excess returns, and their t-statistics (shown in parentheses) are reported in Table 5.4.

Excess returns of decile portfolios facilitate a robust examination of whether buying the cheapest (or the most expensive) set of stocks delivers superior (inferior) investment performance. The decile results also offer an examination of return linearity in the value dimension. Examining the performance of the top two and the bottom two deciles reveals that six of the eight tested factors are capable of delivering both extreme winners and losers with statistical significance. The two exceptions are B2P and earnings yield using IBES estimate (E2PFY1). B2P is a weak differentiator of winners, and only the seventh decile provides statistically significant positive excess returns. In addition, the sixth decile of E2PFY1 has significantly negative returns. CFO2EV delivers the most compelling performance, whose excess returns are not only monotonically increasing from the worst to the best but also statistically significant for the top and bottom four deciles.

5.1.5 Macro Influences on Value Factors

The efficacy of factors to forecast future returns is not constant. It varies across different stocks and through time. We shall provide a more detailed analysis of the cross-sectional and time series variability in Chapter 9 and Chapter 10 and show how to capture these differences to build dynamic models. In this section, we give an overview of how macroeconomic regimes influence the return profile of value strategies. Understanding how strategy returns correlate with macroeconomic variables benefits practitioners in two ways: first, it highlights the potential risk (or deficiency) of employing value strategies during problematic regimes with low or even perverse returns to value. Second, active managers can use their understanding of the economic (risk) cycle to navigate through different market environments by varying factor exposures tactically.

Table 5.5 provides a contemporaneous examination of strategy returns with two market-based variables and one interest rate variable as conditioning factors. They are: (1) *growth-value* markets, defined as the return difference between the Russell 3000 Growth Index and the Russell 3000 Value Index; (2) *up–down* stock markets, defined as the capitalization-weighted return of the Russell 3000 Index; and (3) *up–down* bond markets,

TABLE 5.5 Macro Influences on Value Factors

	Growth vs. Value				Market				Yield Curve Shift			
	Value	**Neutral**	**Growth**	**F**	**Down**	**Flat**	**Up**	**F**	**Down**	**Neutral**	**Up**	**F**
CFO2EV	**9.4%	**4.8%	**2.9%	**14.2	**8.3%	**4.4%	**4.4%	**5.3	**5.3%	**5.7%	**6.1%	0.2
EBITDA2EV	**7.9%	**4.1%	0.9%	**14.0	**6.8%	*2.2%	**3.9%	*4.8	**3.4%	**4.9%	**4.6%	0.5
E2PFY0	**5.3%	**3.1%	1.1%	*4.1	**5.6%	0.6%	**3.4%	**6.3	**3.6%	**3.9%	2.0%	0.9
E2PFY1	**7.4%	*2.5%	−1.0%	**12.1	**5.4%	0.4%	*3.0%	*3.5	*3.0%	*3.9%	2.0%	0.5
BB2P	**6.0%	1.6%	−0.6%	**6.5	**5.6%	0.8%	0.7%	*4.1	−0.1%	2.2%	**4.9%	*3.4
BB2EV	**6.4%	**2.4%	*1.6%	**6.9	**6.3%	**2.4%	1.8%	**6.2	1.8%	**3.7%	**5.0%	2.5
B2P	**6.6%	1.0%	**−4.0%	**22.6	2.1%	−0.5%	2.0%	1.1	−1.1%	1.7%	*3.0%	2.3
S2EV	**8.4%	2.0%	−1.2%	**17.7	**5.0%	*2.5%	1.7%	1.5	0.6%	**3.8%	**4.8%	2.5

Note: * = 90% confidence level; ** = 95% confidence level.

defined by the parallel shift of the U.S. Treasury yield curve — up, neutral, or down. For each regime variable, we first sort the full backtesting sample periods into three equal subsamples. We report average of risk-adjusted ICs and its t-statistics for each subsample along with an F-test showing the significance of IC variance through the introduction of the designated macrovariable. The F-test result answers the question whether market environments significantly influence performance of value investing.

Table 5.5 shows value strategy demonstrated better performance when value index outperforms growth index, when the market drops, and when the interest rate increases. Basically, value investing is a defensive strategy, other things being equal. Among the three macrovariables, value growth is most significant, whereas yield curve shift is the least, as indicated by their F values. Note the following points:

- Cash flow yield (CFO2EV) is the most consistent factor across all market regimes and provides significant positive returns in all market regimes! In contrast, the least consistent is dividend yield (BB2P), because its F values are significant for all three macrovariables. This reflects the dynamic nature of investor's preference toward high dividend paying stocks. Investors seem to only favor high-yielding securities when (1) value outperforms growth, (2) the market goes down, and (3) the interest rates go up.
- B2P and S2EV are the two factors that provide the best opportunity for factor timing, as their F statistics are the highest across different value-growth regimes. When timed correctly, active managers could exploit both factors' perverse performances in growth markets by forming portfolios that are negatively exposed to these factors.

5.1.6 Correlations among Value Factors and Their ICs

At any given time, factors scores have cross-sectional correlations. Over time, factor ICs also have time series correlations. As we discuss in Chapter 7, these types of correlations are interconnected but not the same. The IC correlations provide insight into how the market is pricing the valuation factors overtime. That is, when earnings-based valuation is working to add positive returns, is it also the case for cash flow and asset-based factors? Table 5.6 shows correlations among value factors. As expected, time series correlations of the various value ICs are generally high, ranging from 60 to 90%, thus indicating limited opportunity for diversification. Table 5.6 also shows the average cross-sectional score correlations,

TABLE 5.6 Time Series IC Correlations (Upper Echelon) and Average Cross-Sectional Factor Correlations (Lower Echelon) of Value Factors

	CFO2EV	EBITDA2EV	E2PFY0	E2PFY1	BB2P	BB2EV	B2P	S2EV
CFO2EV	—	80.2%	58.0%	68.1%	74.6%	81.8%	69.9%	76.2%
EBITDA2EV	60.8%	—	81.8%	89.3%	73.6%	72.8%	76.4%	82.4%
E2PFY0	41.6%	66.6%	—	86.8%	54.8%	52.7%	46.4%	55.8%
E2PFY1	39.3%	62.5%	71.9%	—	58.0%	57.3%	69.2%	73.8%
BB2P	33.7%	29.3%	29.4%	24.2%	—	89.9%	61.7%	77.5%
BB2EV	43.0%	32.7%	25.3%	22.5%	57.6%	—	56.6%	75.7%
B2P	32.7%	36.0%	23.9%	25.1%	24.0%	22.4%	—	80.1%
S2EV	39.1%	49.6%	31.3%	36.9%	25.5%	29.5%	51.4%	—

Note: * = 90% confidence level; ** = 95% confidence level.

and it is also interesting to note that they (as shown in the upper echelon) are generally lower than the corresponding time-series IC correlation (as shown in the lower echelon). The rank correlations of factor scores across stocks will typically exhibit lower readings than the correlation across the market pricing of the factors because factor scores contain more noise.

5.2 QUALITY FACTORS

Similar to fundamental research, quality factors assess the health of a firm's business and the competence of its company management, based on information reported in the financial statements. In aggregation, these factors signal a firm's ability to create shareholder value in the future by decomposing a firm's quality into two categories:

1. *Competitiveness of business economics:* Competitive business operation is the engine that creates shareholder wealth. A firm's competitive advantages, typically stemming from efficient operations, intellectual innovation, or market dominance, enable the firm to deliver abnormal profits that are above the cost of capital.
2. *Competency of company management:* Competent and honest company management is the conduit that transfers the maximum amount of wealth created by the firm's business operation to shareholders. As such, competent management translates effective business decisions into profits that accrue primarily to their shareholders instead of more self-serving alternative motivations, often referred to as the agency problem. Factors in this category attempt to measure the extent of any agency problem, wherein the company management acts on its own behalf at the expense of the shareholders.

Measured properly, quality factors identify companies whose operations are sufficiently competitive to generate abnormal business profits, and whose management delivers business profits directly to shareholders without falling prey to agency problems.

For illustrative purpose, we offer examples in four financial ratios to measure the competitiveness of a firm: (1) return on net operating assets (RNOA), (2) cash flow return on investments (CFROI), (3) operating leverage (OL), and (4) increase in operating leverage (OLinc). Intuitively, RNOA and CFROI are proxies for competitiveness because high RNOA or CFROI firms deliver above-average investment returns when compared with their peer groups. Operating leverage adds a bit of complexity, as it measures

how much a firm borrows from its suppliers or customers through the regular course of business operations. Operating leverage is typically a less expensive way of borrowing cash when compared with financial leverage. Thus, in order to minimize borrowing costs (a form of operating expense), firms with strong bargaining power over their suppliers or customers typically increase operating leverages in an effort to decrease financial leverage. OL is selected as a proxy of a firm's bargaining power.

We select several factors to detect the presence of an agency problem. These signals are earnings manipulation (an excessive increase in accounting accruals), excessive capital expenditures, and excessive external financing. The first two are symptomatic of the excessive use of cash by company management at the expense of returning cash to shareholders; the third signal highlights the unwarranted sourcing of cash by management resulting in shareholder dilution. Two specific factors are chosen to illustrate each phenomenon. Working capital increase (WCinc) and net noncurrent asset increase (NCOinc) are earnings manipulation category signals; incremental capital expenditures (icapx) and capital expenditure growth (capxG) rank firm's capital expenditures, and external financing to net operation asset (XF) and share count increase (shareInc) measure the amount of cash raised through external financing. Please refer to the Appendix A5.2 provided at the end of this chapter for a detailed description of how these quality factors are computed from the Compustat database.

5.2.1 Relationship among Quality Factors

Cash is the linkage connecting quality factors. Factors measuring competitiveness also gauge the level of cash flows generated through business activities. RNOA and CFROI both measure cash generated through business transactions, and OL and OLinc measure cash borrowed from suppliers or customers. In other words, competitiveness factors measure the cash raised through the regular course of business operations; the bigger the number is, the more competitive the business economics are. Agency problem-related factors measure the excessive use of cash as well as the amount of cash raised through external financing. WCinc and NCOinc estimate the use of cash in current and noncurrent accruals, and icapx and capxG measures cash used in capital expenditures to facilitate long-tem growth. Lastly, XF measures the amount of cash raised through debt or equity offerings in either private or public placements.

Equation 5.2 depicts the relationship connecting the aforementioned quality factors. (Refer to the Appendix (Equation A5.3) provided at the end of this chapter for a detailed derivation.)

$$\Delta NOA + \Delta CASH = \Delta XF + NI = \Delta WC + \Delta NCO + \Delta CASH\,. \tag{5.2}$$

The terms in the equation are defined as follows:

ΔNOA : Change in net operation assets
ΔXF : Cash flow through external financing activities
NI: Net income in the current period
ΔWC : Change in net current assets (or working capitals)
ΔNCO : Change in net noncurrent assets
$\Delta CASH$: Change in the cash level on the balance sheet from prior period

Dividing Equation 5.2 by prior period's NOA, it becomes

$$XF + RNOA = WCinc + NCOinc + \Delta CASH/NOA \tag{5.3}$$

Equation 5.3 is the decomposition of change in NOA. The left-hand side shows the sources of cash, whereas the right-hand side shows the uses of cash. Cash can be raised either organically through business activities (RNOA) or externally through financing activities (XF). Raised cash can either be invested in working capital (WCinc) or noncurrent asset (NCOinc) through capital expenditure programs, or be left unused in the cash account (ΔCASH/NOA).

5.2.2 Academic Research on Managerial Behavior and Market Inefficiency

Over the last 20 years, researchers have tried to understand the pattern of managerial behavior in reporting corporate earnings. Hayn (1995) contended that firms manage earnings in order to prevent reporting losses. Plotting the distribution of annual earning per share (EPS) for the period 1963–1990, she found a concentration of reported earnings observations just in excess of zero, and a dearth of reported earnings just below zero. She noted, "These results suggest that firms whose earnings are expected to fall just below zero earnings point engage in earnings manipulations to help them across the red line." Burgstahler and Dichev (1997) also concluded that 30 to 40% of firms that would otherwise report small losses manage earnings to report small profits. Degeorge et al. (1999) developed a model to illustrate how companies manipulate their earnings in order to avoid 1) the possibility of red ink, 2) the threat of not being able to sustain recent performance, and 3) concern about not meeting analyst

expectations. Healy and Kaplan (1985) assert that managers manipulate earnings to exceed a benchmark if they can; if they cannot, they take a big shortfall in order to stockpile earnings that can be used in future reporting periods, a phenomenon known as the "big bath."

To further understand managerial behavior, academic researchers examine managers who are unable to report profits and as a result must report losses. Given managers' heightened concern with litigation (Kasznik and Lev 1995) and their vast increase in ownership of stock options, managers are likely to mitigate their tendency to report losses that are below analyst estimates in general and well below analyst estimates in particular.

In contrast, when it comes to managing profit surprise, Levitt (1998) found that managers attempt to report profits that meet or slightly beat analyst estimates. Practitioners maintain that the negative market implication of reporting profits slightly short of analyst estimates is very significant. As a result, if managers are unable to report quarterly earnings that just meet or slightly beat analyst estimates, they may manipulate accruals in order to report small positive surprise earnings and avoid small negative ones (Burgstahler and Eames 2003).

To quantify earnings management, Jones (1991), Dechow et al. (1995), Sloan (1996), and Jeter and Shivakumar (1999) proposed methods to estimate expected accruals after controlling for changes in a firm's economic condition, such as the growth rate. In summary, this body of research separates reported earnings into three components: discretionary accruals, nondiscretionary accruals, and a cash flow component. Discretionary accruals gauge company management's subjectivity in estimating accruals and reporting earnings, and are used to proxy the level of earnings management at each firm. Nondiscretionary accruals represent the expected level of accruals that are needed to accommodate the firm's growth.

Two extensions of accrual measures were introduced recently after its initial discovery by Healy (1985). First, Hribar and Collins (2002) showed that accruals can also be measured directly from the statement of cash flows. They assert that a cash flow statement based measure is superior to a balance sheet based measure, because balance sheet measures are often contaminated by the nonarticulated changes in current accounts, resulting from mergers and acquisitions, discontinued operations, and currency translations. Second, Richardson et al. (2005) expanded Healy's narrow definition of accruals, which focuses on current operating

TABLE 5.7 Historical Performance of Quality Factors

	Performance						Turnover	
	ICa	**t(ICa)**	**IR(ICa)**	**IC**	**t(IC)**	**IR(IC)**	**CFA**	**TO**
RNOA	**3.05%	3.67	0.43	**3.64%	3.45	0.41	89.3%	130%
CFROI	**5.43%	7.74	0.93	**5.68%	5.75	0.69	83.7%	147%
OL	**3.66%	7.73	0.91	**2.95%	7.99	0.94	91.1%	124%
OLinc	**3.61%	9.46	1.12	**3.12%	9.88	1.16	59.8%	253%
WCinc	**−3.98%	−8.00	−0.94	**−3.52%	−7.87	−0.93	65.2%	247%
NCOinc	**−3.15%	−5.83	−0.69	**−3.62%	−6.38	−0.75	79.5%	179%
icapx	**−2.99%	−6.00	−0.71	**−2.34%	−4.81	−0.57	92.4%	111%
capxG	**−1.99%	−4.51	−0.53	**−2.54%	−4.60	−0.54	75.9%	182%
XF	**−4.50%	−8.14	−0.96	**−5.07%	−6.75	−0.80	75.6%	177%
shareInc	**−2.28%	−4.44	−0.52	**−2.50%	−3.36	−0.40	81.9%	142%

accruals (primarily ΔWC), to accommodate long-term operating accruals (ΔNCO) and the change in the net financial assets (ΔFIN *).

Why does accrual predict future returns? There are two schools of thoughts. Sloan (1996) shows that the accrual component of earnings is less persistent than the cash flow component due to managerial subjectivity involved in estimating accruals. He suggests that the investor fails to comprehend the fact that firms manage their earnings by manipulating reported accruals and thus create marketing mispricing. Alternatively, Fairfield et al. (2003) attribute the return predictability to the market mispricing of growth in NOA. They suggest that the lower persistence of accruals is likely to result from the conservative bias in accounting and/or the diminishing economic return to marginal investments due to competition.

5.2.3 Historical Performance of Quality Factors

Table 5.7 displays the historical performance of selected quality factors. To control for the level differences of these ratios across different sectors, factor values are ranked within each sector to facilitate proper peer comparison. All signals generate excess returns, significant at 1% level. Factors measuring competitiveness deliver significant positive returns, pointing to the importance of investing in firms with strong business economics. In contrast, factors gauging the severity of agency problems show significant

* is defined as the change in short-term and long-term investments minus the change in total debt and preferred stocks.

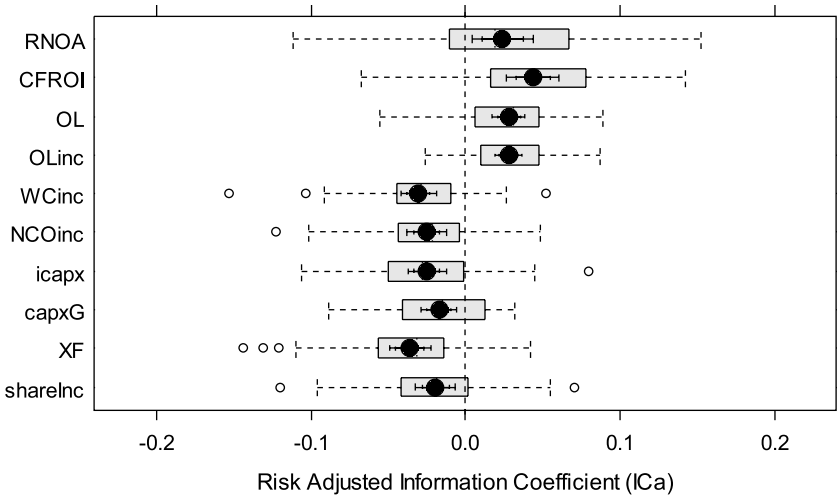


FIGURE 5.3. Box plots of risk-adjusted IC for quality factors.

negative returns, underscoring the importance of avoiding firms that manipulate earnings, pursue excessive capital investment, or engage in excessive equity or debt issuance.

Figure 5.3 shows the box chart of risk-adjusted ICs for quality factors. Comparing the IC distribution of quality factors with that of value factors (as shown in Figure 5.2), it can be seen that the statistical significance is more pronounced for quality factors than value measures evidenced by the higher average IC and lower standard deviation of IC (or strategy risk). For example, 75% of the IC distribution of quality factors falls in the same direction (positive or negative), as predicted, with RNOA and capxG as the only two exceptions. By this measure, quality factors have delivered an astonishing record of consistency — most worked in more than 75% of our sample periods between January 1987 and March 2005! The smaller strategy risk indicates that quality factors are more consistently priced by the market than value factors and are less subject to macroeconomic or behavioral influences in a temporal sense. As such, value factors are better candidates for factor-timing than quality factors, as value factors have higher time series dispersion, which represents the opportunity to apply timing skill.

Table 5.7 also reveals that quality factor requires higher turnover than value factors indicated by both lower CFA and higher TO in the last two columns. This is true for OLinc, WCinc, NCOchg, and capxG because they represent the change of financial ratios measured between two successive financial statements. OLinc, WCinc, and NCOinc are measured

between two successive balance sheet statements, and capxG is computed using two successive cash flow statements.

Table 5.8 shows the excess returns of decile portfolios for quality factors. Interestingly, returns to competitiveness-related factors exhibit linear relationships, whereas returns to agency problem-related factors do not. Firms with high RNOA, CFROI, OL, and OLinc delivered significant excess returns (as shown in the 8th, 9th, and 10th deciles); and firms with inferior business economics destroy shareholder wealth at 1% statistical significance (as shown in the 1st, 2nd, and 3rd deciles). It is also interesting to note that the wealth destruction by inferior firms is more pronounced than the wealth creation by superior firms, both in the level of excess returns and t-statistics. This phenomenon can perhaps be traced to the market structure wherein most active managers are bounded by long-only portfolio mandates, which limit their ability to short stocks issued by inferior firms, or to the disposition effect wherein investors hold onto their losers (inferior firms) for too long.

Agency problem-related factors exhibit nonlinear return relationships, with XF being the only exception. This nonlinear return response makes intuitive sense. For icapx and capxG, the agency problem implies that the act of pursuing excessive capital expenditure programs by management is detrimental to shareholders as it is a symptom of the company management pursuing their own interests at the expense of shareholders. Such reasoning does not apply to the other extreme, and it is misguided to extrapolate that the lowest capital spenders are the most beneficial to shareholders. In fact, firms that underspend capital risk losing their competitive advantage and future growth prospects, both of which destroy shareholder value as well. The best sets of firms are those who embark on conservative capital expenditure programs (2nd and 3rd deciles) instead of the ones not spending at all (1st decile).

For accruals-related factors (WCinc and NCOinc), higher readings signal the possibility of earnings manipulation in which the company management defers costs from the current period to future periods and shifts revenue recognitions from future periods to the current period. As such, high accounting accruals are detrimental to shareholders because the earnings of the current period are artificially inflated to look good at the expense of future periods. Eventually, these inflated earnings will revert, often violently, causing a precipitous drop in stock prices. Does the accrual phenomenon exhibit a linear relationship? In other words, does it enhance shareholder value when firms engage in the opposite extreme — pushing

TABLE 5.8 Decile Performance for Quality Factors

	Lowest	2	3	4	5	6	7	8	9	Highest
RNOA	**−2.20	**−1.20	**−0.78	−0.03	0.23	0.49	0.67	**0.88	**1.02	**0.92
	(−3.15)	(−2.84)	(−2.85)	(−0.14)	(1.03)	(1.47)	(1.82)	(2.69)	(3.03)	(2.86)
CFROI	**−2.74	**−1.97	**−1.40	−0.24	0.26	0.62	**1.21	**1.45	**1.45	**1.35
	(−4.39)	(−4.38)	(−4.57)	(−1.09)	(1.50)	(1.16)	(3.65)	(4.99)	(4.79)	(4.56)
OL	**−1.21	**−0.97	**−0.50	−0.26	0.30	−0.12	**0.83	**0.61	**0.60	**0.72
	(−3.74)	(−4.48)	(−3.12)	(−1.55)	(1.64)	(−0.58)	(4.87)	(4.09)	(3.38)	(2.97)
OLinc	**−1.18	**−0.65	**−0.74	0.00	−0.14	0.09	0.36	*0.41	**1.07	**0.77
	(−3.34)	(−3.03)	(−4.13)	(−0.01)	(−0.64)	(0.35)	(1.38)	(2.25)	(5.61)	(2.91)
WCinc	0.58	**0.88	**0.89	0.37	0.23	0.05	−0.06	−0.10	**−0.77	**−2.08
	(1.96)	(4.21)	(5.62)	(1.79)	(0.98)	(0.19)	(−0.29)	(−0.48)	(−3.33)	(−6.65)
NCOinc	0.60	**0.90	**0.93	**0.65	*0.47	−0.18	0.13	*−0.41	**−0.97	**−2.12
	(1.73)	(3.74)	(3.43)	(3.22)	(2.20)	(−0.58)	(0.71)	(−2.17)	(−4.31)	(−5.69)
icapx	−0.08	**0.77	*0.50	0.38	0.25	0.22	−0.12	0.07	*−0.50	**−1.50
	(−0.22)	(3.50)	(2.33)	(1.75)	(1.41)	(0.57)	(−0.69)	(0.35)	(−2.25)	(−5.25)
capxG	0.13	**0.73	**0.72	**0.68	0.48	−0.06	0.24	*−0.45	**−0.81	**−1.68
	(0.38)	(3.17)	(3.24)	(2.96)	(1.93)	(−0.18)	(1.29)	(−2.60)	(−3.19)	(−3.77)
XF	**1.48	**1.57	**1.07	**0.82	0.44	−0.11	**−0.58	**−1.18	**−1.38	**−2.11
	(7.02)	(5.08)	(4.34)	(3.49)	(0.91)	(−0.67)	(−2.98)	(−5.08)	(−4.38)	(−3.75)
shareInc	0.47	*0.73	0.57	**0.82	**0.52	**−0.82	0.05	−0.14	−0.62	**−1.57
	(1.97)	(2.18)	(1.67)	(3.00)	(3.25)	(−2.95)	(0.26)	(−0.54)	(−1.89)	(−4.05)

Note: * = 90% confidence level; ** = 95% confidence level.

revenue into future periods and pulling costs into the current period? Building up negative accruals, which will dramatically inflate future earnings at the expense of the current one, is still misleading and potentially a sign of dishonest company management. In fact, this extreme negative accrual buildup is called a "big bath" in accounting literature. This phenomenon happens when company management realizes that there is no way to make the current period's earnings look good and pursues an alternative extreme by making current period look even worse in order to inflate future earnings. These firms do not deliver the best excess returns. The best firms are those who exercise truthful, conservative accounting practices in terms of earnings recognition (2nd and 3rd deciles, but not the 1st).

XF exhibits a linear relationship, a stark contrast to both the capital expenditure and accruals factors. Linearity of returns to XF is likely related to information asymmetry, which explains the positive excess return deciles as well as the negative deciles. Information asymmetry posits that (1) company management knows more than the general investment public due to its access to private information, and (2) management is inclined to retain the cash instead of paying it back to the shareholders due to the costs associated with external financing activities. As a result, company management pays cash back to shareholders (through dividend, buyback, or debt repayment) only when their outlook for the firm is rosy. Hence, paying back to shareholders signals a positive assessment of the firm's business environment by management a phenomenon known as *management signaling*. Most interestingly, the statistical significance is more pronounced for companies embarking on buyback programs than firms pursing excessive external financing.

5.2.4 Macro Influences on Quality Factors

Table 5.9 examines the return profile of quality factors under different market environments. When compared with the results of value factors (as shown in Table 5.5), quality factors are generally less sensitive to the changes in macroenvironments than value factors, indicated by smaller F statistics. Two observations are worth noting:

- Agency-problem-related factors deliver higher negative excess returns in value environment than growth. Combined with the fact that the agency problem is more pronounced for growth stocks, it is logical to conclude that growth stocks with symptoms of an agency

TABLE 5.9 Macro Influences on Quality Factors

	Growth vs. Value				Market				Yield Curve Shift			
	Value	**Neutral**	**Growth**	**F**	**Down**	**Flat**	**Up**	**F**	**Down**	**Neutral**	**Up**	**F**
RNOA	−0.3%	**2.9%	**4.6%	**5.4	**3.9%	2.0%	1.3%	1.5	**4.8%	2.0%	0.3%	*4.3
CFROI	1.9%	**4.3%	**7.0%	**8.5	**5.8%	**4.3%	**2.9%	2.4	**5.7%	**4.3%	**2.9%	2.1
OL	**3.0%	**3.4%	**2.0%	1.4	**2.4%	**3.2%	**2.7%	0.4	*1.4%	**3.7%	**3.3%	*4.4
OLinc	**3.1%	**3.3%	**2.0%	2.1	**3.1%	**2.9%	**2.5%	0.4	**3.1%	**2.7%	**2.5%	0.3
WCinc	**−3.4%	**−3.4%	**−2.4%	0.8	**−3.2%	**−3.6%	**−2.4%	0.9	*−1.7%	**−3.1%	**−4.4%	*4.5
NCOinc	**−4.7%	*−1.8%	−1.1%	**7.8	**−3.9%	*−1.5%	**−2.1%	2.9	*−1.6%	**−3.1%	**−2.9%	1.3
icapx	**−4.2%	**−1.8%	−1.5%	*4.9	**−2.9%	**−2.9%	*−1.7%	0.9	**−2.5%	*−1.9%	**−3.1%	0.6
capxG	**−3.4%	−0.8%	−1.1%	**5.1	**−3.1%	−0.9%	−1.1%	*3.6	−0.8%	**−2.0%	**−2.3%	1.5
XF	**−5.2%	**−2.7%	**−2.9%	*3.4	**−5.8%	**−3.1%	**−2.0%	**7.7	**−2.7%	**−3.8%	**−4.3%	1.3
shareInc	**−3.5%	−0.7%	*−1.7%	*3.8	**−3.7%	−0.9%	*−1.3%	*4.3	−0.6%	**−1.9%	**−3.4%	*3.4

Note: * = 90% confidence level; ** = 95% confidence level.

problem are most severely penalized when the market's sentiment shifts from the pursuance of growth to the pursuit of value. It is a time when investors are most worried about the pace of the economic growth and rethink the expected returns on investments, giving rise to a dramatic shrinkage in the duration assumption of discounted cash flow (DCF) valuation.

- RNOA and CFROI work best in a growth environment, and returns to both factors are significantly influenced by value growth regimes as indicated by F-statistics. Change in DCF duration again plays a role in this phenomenon. When DCF duration lengthens during a growth regime, cross-sectional ranking of valuation becomes more correlated with RNOA or CFROI, thus generating higher returns to both factors.

5.2.5 Correlations among Quality Factors and Their ICs

Table 5.10 reports the correlations among quality factors: the upper echelon shows time series correlation of risk-adjusted ICs, and the lower echelon reports the average of cross-sectional correlation of factor scores. The two shaded areas contain correlations between competitiveness-related factors that provide positive excess returns and agency-problem-related factors, which, in contrast, deliver negative excess returns. Boldfaced correlation numbers highlight significant diversification opportunities among quality factors! Because we use the negative of the agency-problem-related factors when combining them with competitiveness-related factors, a positive IC correlation actually translates into a negative IC correlation. For example, RNOA and NCOinc provide an incredible opportunity to diversify risk and to improve the combined IR, as their IC correlation is astonishingly high (48%), whereas their IC averages are of different signs. Table 5.12 simply highlights an important lesson for active managers — maximizing the diversification benefit among quality factors.

5.3 MOMENTUM FACTORS

The momentum phenomenon is typically partitioned into two categories: price momentum and earnings momentum. Price momentum is akin to technical analysis, which uses past price and volume information to predict future security returns. However, unlike the myriad of technical indicators (and their loose interpretations), price momentum was debated

TABLE 5.10 Time Series IC Correlations (Upper Echelon) and Average Cross–Sectional Factor Correlations (Lower Echelon) of Quality Factors

	RNOA	CFROI	OL	OLinc	WCinc	NCOinc	icapx	capxG	XF	shareInc
RNOA	—	78.6%	***(–16.2%)***	***(–14.7%)***	***45.3%***	***48.0%***	***32.7%***	***45.3%***	(–4.5%)	***26.3%***
CFROI	59.4%	—	***(–7.9%)***	***(–5.8%)***	***13.4%***	***31.0%***	***9.0%***	***26.2%***	(–23.5%)	***5.7%***
OL	17.6%	25.5%	—	23.3%	(–40.3%)	(–30.6%)	(–20.0%)	(–12.2%)	(–37.6%)	(–18.2%)
OLinc	(–8.6%)	12.9%	21.7%	—	(–33.8%)	(–43.8%)	(–18.1%)	(–13.3%)	(–25.2%)	(–5.2%)
WCinc	15.7%	(–24.2%)	(–14.4%)	(–39.5%)	—	26.2%	26.9%	32.3%	30.8%	24.1%
NCOinc	15.0%	2.4%	(–18.2%)	(–34.3%)	12.9%	—	50.1%	66.4%	55.2%	65.2%
icapx	18.3%	9.2%	(–7.6%)	(–13.7%)	9.6%	41.3%	—	40.2%	41.7%	42.7%
capxG	13.5%	2.7%	(–3.4%)	(–14.0%)	13.8%	37.4%	41.3%	—	48.3%	59.9%
XF	(–14.0%)	(–28.8%)	(–9.8%)	(–22.3%)	24.7%	41.9%	21.4%	21.5%	—	67.7%
shareInc	(–8.2%)	(–13.3%)	(–8.8%)	(–5.0%)	9.9%	20.3%	5.2%	14.3%	25.4%	—

Note: * = 90% confidence level; ** = 95% confidence level.

and documented by academic researchers who applied modern statistical techniques to assess trends and reversals, and proposed behavorial explanations to justify the existence of these price patterns. Earnings momentum focuses on past earnings changes as well as the movement of forecasted earnings, i.e., earnings revision factors. Traditional earnings revision techniques make use of changes in consensus earnings estimates supplied by sell-side analysts as a proxy for market sentiment (bullish vs. bearish) toward a particular stock. This section provides an academic literature review for price momentum, whereas the next section focuses on earning momentum.

Jegadeesh and Titman (1993) document that when forming portfolios based on past returns, the past-winner portfolios will outperform the past-loser portfolios over the next 2 to 12 months during 1965 to 1989 in the U.S. markets. This phenomenon is referred to as *intermediate-term price momentum continuation*. However, the authors also find that past winners underperformed past losers in the first month after portfolio formation. This anomaly is called *short-term price momentum reversal*.

Price momentum anomalies and research have drawn considerable attention as well as criticism. For many skeptics who have a hard time comprehending how such a simplistic strategy can generate abnormal returns, the price momentum anomaly is considered as a result of data mining from empirical finance researchers. Since price momentum was initially documented in the US market, testing its existence in non-US markets can be considered as an out-of-sample test to assess the robustness of this phenomenon across global equity markets. With this in mind, Rouwenhorst (1998) applies the same price momentum strategy in 12 European countries and finds similar results during 1980 to 1995. The evidence rejects the notion that price momentum is a result of data mining and argues for an alternative explanation.

To understand whether excess return from price momentum is simply a risk premium in disguise, Fama and French (1993) attempted to used their three-factor ICAPM framework (market, price-to-book, and market-cap) to explain intermediate-term price momentum anomaly. To their dismay, they conceded that this anomaly cannot be explained by a premium associated with these previously documented systematic risks. Later on, Fama and French's three-factor model was extended to include momentum as the fourth-priced risk factor and the four factor model becomes the new standard of asset pricing tests.

To explain the price momentum anomaly, Daniel et al. (2001) suggest that investor's overconfidence and biased self-attribution (i.e., cognitive dissonance) causes a biased revision of investor's expectations in response to new information. In response to new information, investors tend to underreact in the beginning and then overreact in the long term. Chan et al. (1996) document that the price momentum anomaly is partially attributable to underreactions to earnings news (aka *earnings momentum*). Hong, Lim, and Stein (2000) suggest that slow diffussion of information into prices (most evident for bad news) causes an initial underreaction to news. More recently, Grinblatt and Han (2005) linked the momentum to the disposition effect — investors' tendency to sell winners and keep losers. Frazzini (2006) develops further analysis based on capital gains (or losses) associated with individual stocks.

To summarize the above findings, the price momentum anomaly is commonly attributed to:

Behavioral bias: Investors are more confident about their own private information concerning a company than about public information; and this causes an initial underreaction to news. Such initial underreaction eventually leads to long-term overreactions. Furthermore, the degree of underreaction is influenced by investors' mental accounting.

Imperfect information. Company-specific information is delayed and uncertain as the management of a company has strong incentives to promote good news and to hide bad news. This leads to delayed and autocorrelated market reactions to bad news. Again, the agency problem is at work here.

Imperfect market structure: Because most institutional money managers are not allowed to short-sell stocks, "informed" money managers are able to fully arbitrage good news by purchasing enough shares of that company, but are unable to fully arbitrage bad news due to the no short sell constraints.

To ascertain whether the efficacy of a price momentum strategy varies across different market segments, Hong and Stein (1999) found the following: First, the profitability of price momentum strategy declines sharply with firm size; in other words, even though price momentum strategy is still profitable for large-cap stocks, it is predominantly a mid- and small-cap phenomenon. Second, with holding size fixed, price momentum

strategy works better among stocks with low analyst coverage. Finally, the effect of analyst coverage is greater for stocks that are past losers than for past winners. This means price momentum strategy is more effective in identifying losers than winners.

5.3.1 Earnings Momentum Anomaly

For more than 20 years, the earnings revision phenomenon has been extensively documented by a large amount of academic literature. Givoly and Lakonishok (1979) conclude that market reaction to analysts' earnings revisions is relatively slow. In addition, Givoly and Lakonishok (1980) show that an investor who acts upon analysts' earnings revisions can consistently outperform a buy-and-hold policy after transaction costs.

Further studies find that large earnings revisions are more indicative of subsequent earnings revisions and price drifts. Hawkins et al. (1984) find that portfolios comprised the 20 stocks with the largest monthly upward revisions in consensus estimates subsequently experienced positive abnormal returns 75% of the time. Kerrigan (1984) shows that, when the EPS forecast for a stock is subject to a large revision, any subsequent revisions within the year tend to be in the same direction. Richards and Martin (1979) find that revisions in the first quarter represent new information but the revisions in subsequent quarters do not. Dowen and Bauman (1991) find that earnings revision anomaly is not explained by the small firm effect (Dowen and Bauman 1986), nor is it explained by the neglect effect (Arbel et al. 1983).

5.3.2 Historical Performance of Momentum Factors

In this section, we sample three price momentum factors and three earnings momentum factors to illustrate the historical performance of momentum strategies. For price momentum, the past 1-month return (ret1) captures the short-term reversal phenomenon, the past 9-month return excluding the first trailing month (ret9) captures the intermediate-term continuation of price momentum, and risk-adjusted 9-month return (adjRet9) captures interactions between past return and residual risks. In the earnings momentum category, the change in the consensus EPS estimate between today and 9 month ago measures the 9-month earnings revisions (earnRev9). Further, the ratio of the number of analysts upgrading EPS estimate minus the number of analysts downgrading divided by total number of analysts during the last 9 months measures earnings diffusion

TABLE 5.11 Historical Performance of Momentum Factors

	Performance						Turnover	
	ICa	**t(ICa)**	**IR(ICa)**	**IC**	**t(IC)**	**IR(IC)**	**CFA**	**TO**
ret1	*−2.88%	−2.68	−0.32	−0.72%	−0.63	−0.07	3.0%	432%
ret9	**7.20%	4.79	0.56	**6.12%	3.97	0.47	62.7%	263%
adjRet9	**6.29%	4.20	0.49	**6.42%	4.49	0.53	61.1%	279%
earnRev9	**3.90%	3.20	0.38	**3.95%	3.77	0.44	63.7%	244%
earnDiff9	**5.10%	3.90	0.46	**4.67%	4.23	0.50	72.1%	220%
ltgRev9	**2.22%	3.99	0.47	**1.80%	3.19	0.38	37.0%	312%

Note: * = 90% confidence level; ** = 95% confidence level.

(earnDiff9). Note that earnRev9 measures the magnitude of change in EPS levels, whereas earnDiff9 is mainly a directional measure ignoring the magnitude of EPS changes. Lastly, the change in long-term growth rate estimate during the trailing 9 months, ltgRev9, reflects a slower moving view of long-term profitability.

Unlike the ranking process applied to value and quality factors, the performance of momentum factor is computed without sector neutralization. As a result, momentum back-testing results as shown in this section capture not only stock-specific momentum but also sector/industry momentum.

Results in Table 5.11 show momentum factors deliver significant positive excess returns (1987–2004); ret1, which captures 1-month reversal, delivers negative excess returns, as expected. Examining the IC stability through time, momentum factors are generally more variable than quality factors, suggesting that momentum factors are more susceptible to shifts in macroeconomic environments, similar to the observation for value strategies. Figure 5.4 shows the box plots of risk-adjusted ICs for momentum factors.

In implementing momentum strategies, it is most striking that considerable portfolio turnover is an onerous requirement to maintain proper exposures. The average turnover for momentum, quality, and value factors are 292, 169, and 141%, respectively. Compared with value strategies, momentum strategies require more than twice the turnover, and quality strategies require about 20% more. Clearly, momentum investing is a *demander of liquidity,* whereas value investing is more a *supplier of liquidity.* This is important for active managers. Implementation costs (more in Chapter 8) induced by maintaining proper factor exposures must be

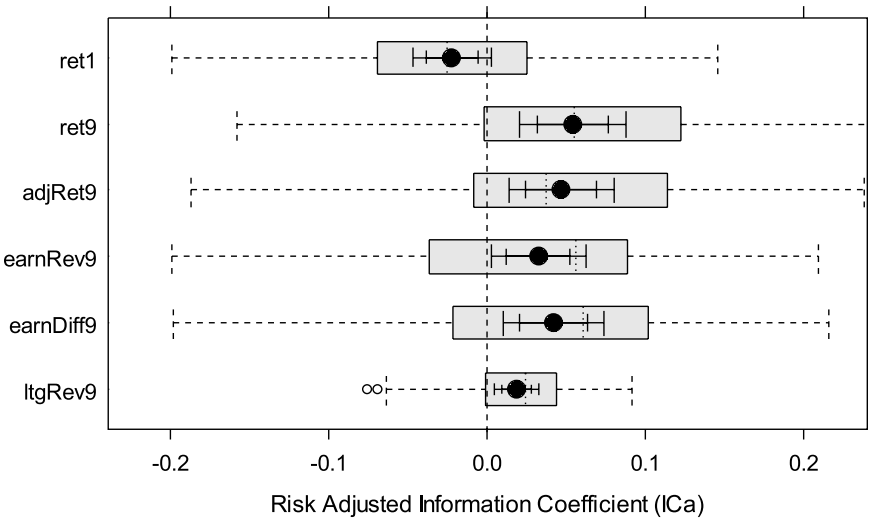


FIGURE 5.4. Box plots of risk-adjusted ICs for momentum factors.

considered in conjunction with the theoretical strategy profit when incorporating value, quality, and momentum strategies into the final model.

Table 5.12 reports decile performance of momentum factors. Three observations stand out: First, in terms of the short-term reversal factor (ret1), stocks with highest trailing 1-month returns deliver the worst performance in the subsequent 3 months (10th decile). However, this phenomenon is nonlinear, as the worst 1-month losers (1st decile) also delivered negative excess returns. Second, adjusting price momentum by its contemporaneous residual risk enhances consistency of performance. When compared with ret9, adjRet9 delivers better t-statistics in the 2nd, 3rd, 4th, 7th, 8th, and 9th deciles. Third, earnings momentum factors generally work for the best and worst ranking stocks. However, the linearity of return response looks distorted by the sixth decile, delivering significant negative excess returns across all three earning momentum factors. Upon a closer examination, this abnormal negative return is an artifact of how missing values are treated in the decile ranking process. Because stocks with missing scores historically delivered significant negative excess returns, excluding them from the analysis eliminates the anomaly pertaining to the 6th decile, thus achieving a better linear result. However, we would caution readers on concluding that a missing earnings momentum score is a signal for underperformance, as survivorship bias (in how IBES populates historical EPS estimates) may play a role in this seemingly anomalous finding.

TABLE 5.12 Decile Performance for Momentum Factors

	Lowest	2	3	4	5	6	7	8	9	Highest
ret1	−1.31	0.46	*0.74	*0.64	0.64	0.46	0.43	−0.07	−0.58	*−1.40
	(−1.33)	(0.85)	(2.21)	(2.42)	(1.85)	(1.34)	(1.24)	(−0.20)	(−1.37)	(−2.13)
ret9	**−3.75	*−1.50	−0.20	0.03	0.52	0.16	0.45	0.80	*1.34	*2.15
	(−3.16)	(−2.63)	(−0.49)	(0.08)	(1.24)	(0.48)	(1.11)	(1.87)	(2.61)	(2.58)
adjRet9	**−2.54	**−1.61	**−1.40	−0.54	0.11	0.06	*0.79	**1.17	**1.66	**2.30
	(−3.55)	(−2.88)	(−3.94)	(−1.58)	(0.41)	(0.21)	(2.58)	(3.34)	(3.59)	(3.10)
earnRev9	**−1.67	−0.66	−0.38	0.06	−0.08	*−0.46	*0.68	**1.22	*1.02	0.26
	(−3.61)	(−1.68)	(−1.26)	(0.24)	(−0.34)	(−2.21)	(2.27)	(3.85)	(2.63)	(0.50)
earnDiff9	*−1.31	−0.62	*−0.62	0.01	−0.35	**−0.63	0.04	0.59	0.77	**2.13
	(−2.53)	(−1.62)	(−2.12)	(0.03)	(−1.83)	(−3.31)	(0.13)	(1.86)	(1.90)	(4.21)
ltgRev9	−0.86	−0.24	0.04	0.33	0.30	**−1.40	*0.74	0.26	0.12	*0.71
	(−1.89)	(−0.87)	(0.19)	(1.32)	(1.03)	(−3.65)	(2.38)	(1.05)	(0.41)	(2.18)

Note: * = 90% confidence level; ** = 95% confidence level.

5.3.3 Macro Influences on Momentum Factors

Table 5.13 examines the return profile of momentum factors under different market regimes. Momentum profits are considerably lower and statistically insignificant when the value index outperforms the growth index. Combining this observation with the fact that momentum is more important for growth stocks (see Chapter 9), we conclude that the major portion of momentum return comes from high-growth stocks in a market environment when the growth index outperforms the value index. Shifts in the yield curve and changes in credit spread also significantly influence momentum profits.

5.3.3.1 Correlations among Momentum Factors and Their ICs

Table 5.14 reports correlations among momentum factors: the upper echelon shows time series correlations of ICs and the lower echelon shows the average of cross-sectional correlations of factor values. Similar to value factors, IC correlations of momentum strategies are generally lower than correlations of factor values. Also, short-term reversal (ret1) provides potential diversification benefit to other momentum strategies as correlations are significantly positive (boldfaced numbers), whereas returns are of different signs. However, one has to be mindful of its high turnover.

TABLE 5.13 Macro Influence on Momentum Factors

	Growth vs. Value				Yield Curve Shift				Credit Spread			
	Value	**Neutral**	**Growth**	**F**	**Down**	**Flat**	**Up**	**F**	**Shrank**	**Neutral**	**Widen**	**F**
ret1	–2.0%	**–3.3%	–1.4%	0.5	0.1%	*–3.0%	**–3.8%	2.1	**–5.1%	0.2%	–2.0%	*3.8
ret9	1.3%	**5.4%	**9.4%	*4.8	**9.1%	*4.6%	2.2%	*3.3	2.5%	**7.9%	*5.6%	2.0
adjRet9	0.9%	**4.8%	**8.2%	*3.9	**8.1%	*4.0%	1.8%	2.9	1.8%	**7.0%	*5.1%	2.0
earnRev9	–0.5%	**4.4%	**5.7%	*3.9	**7.4%	3.2%	–0.9%	**6.5	0.3%	**6.0%	3.2%	2.9
earnDiff9	0.1%	**5.4%	**7.0%	*4.1	**8.6%	*4.2%	–0.3%	**6.8	1.0%	**7.2%	4.2%	3.0
ltgRev9	0.9%	**2.3%	**2.4%	1.1	**2.4%	1.8%	1.3%	0.5	**2.2%	**2.6%	0.7%	1.5

Note: * = 90% confidence level; ** = 95% confidence level.

TABLE 5.14 Time Series IC Correlations (Upper Echelon) and Average Cross-Sectional Factor Correlations (Lower Echelon) of Momentum Factors

	ret1	ret9	adjRet9	earnRev9	earnDiff9	ltgRev9
ret1	—	**49.5%**	**48.2%**	**41.7%**	**43.9%**	**33.7%**
ret9	6.4%	—	98.7%	78.1%	78.3%	54.0%
adjRet9	4.3%	95.6%	—	79.6%	79.1%	55.0%
earnRev9	10.9%	51.7%	52.8%	—	97.7%	48.9%
earnDiff9	11.3%	52.5%	53.6%	80.1%	—	47.8%
ltgRev9	3.2%	20.8%	21.0%	20.2%	19.6%	—

APPENDIX

A5.1 FACTOR DEFINITION

This section illustrates how factors are constructed from the Compustat database. When applicable, we show the *Compustat Quarterly* item number in parentheses as a reference within each formula.

CFO2EV : Cash flow from operations to enterprise value

$$\frac{CFO_{(108)} + intExp_{(022)} \times (1 - tax\_rate)}{market\_cap + debt_{(045\ \&\ 051)} + pfd_{(055)} - cash_{(036)}}$$

EBITDA2EV : Earnings before interest, taxes, and depreciation to enterprise value

$$\frac{sales_{(002)} - COGS_{(030)} - SG\&A_{(001)}}{market\_cap + debt_{(045\ \&\ 051)} + pfd_{(055)} - cash_{(036)}}$$

E2PFY0 : Trailing 12-month earnings to market capitalization

$$\frac{income\_before\_extraordinary_{(025)}}{market\_cap}$$

E2PFY1 : IBES FY1 earnings to market capitalization

$$\frac{IBES\_FY1\_EPS \times shares\_outstanding}{market\_cap}$$

BB2P : Net buyback to market capitalization

$$\frac{dividend_{(089)} + equity\_repurchase_{(093)} - equity\_issuance_{(084)}}{market\_cap}$$

BB2EV : Net external financing to enterprise value

$$\frac{dividend_{(089)} + equity\_repurchase_{(093-084)} - debt\_repurchase_{(092-075-086)}}{market\_cap + debt_{(045+051)} + pfd_{(055)} - cash_{(036)}}$$

B2P : Book-to-market capitalization

$$\frac{common\_equity_{(059)}}{market\_cap}$$

S2EV : Sales to enterprise value

$$\frac{sales_{(002)}}{market\_cap + debt_{(045\ \&\ 051)} + pfd_{(055)} - cash_{(036)}}$$

RNOA : Return on net operating assets

$$\frac{income_{(008)} + intExp_{(022)} \times (1 - tax\_rate)}{equity_{(059)} + debt_{(045\ \&\ 051)} + pfd_{(055)} - cash_{(036)}}$$

CFROI : Cash flow from operations to net operating assets

$$\frac{CFO_{(108)} + intExp_{(022)} \times (1 - tax\_rate)}{equity_{(059)} + debt_{(045\ \&\ 051)} + pfd_{(055)} - cash_{(036)}}$$

OL : Operating liability to net operating assets

$$\frac{total - assets_{(044)} - equity_{(059)} - debt_{(045\ \&\ 051)} - pfd_{(055)}}{equity_{(059)} + debt_{(045\ \&\ 051)} + pfd_{(055)} - cash_{(036)}}$$

OLinc : Change in the ratio of operating liability to net operating assets

$$OL_t - OL_{t-1}$$

WCinc : Change in working capitals to assets

$$\frac{WC_t - WC_{t-1}}{assets_{(044)}},$$

$$where \quad WC = cur\_assets_{(040)} - cash_{(036)} - cur\_liab_{(049)} + st\_debt_{(045)}$$

NCOinc : Change in net noncurrent assets to assets

$$\frac{NCO_t - NCO_{t-1}}{assets_{(044)}},$$

$$where \quad NCO = TA_{(044)} - cur\_assets_{(040)} - TL_{(054)} + cur\_liab_{(049)} + lt\_debt_{(051)}$$

icapx : Capital expenditures minus depreciation expense

$$\frac{capex_{(090)} - depreciation_{(005)}}{assets_{(044)}}$$

capxG : Growth in capital expenditures

$$\frac{capex_t - capex_{t-1}}{assets_{(044)}}$$

XF : Net external financing to net operating assets

$$\frac{dividend_{(089)} + equity\_repurchase_{(093-084)} - debt\_repurchase_{(092-075-086)}}{equity_{(059)} + debt_{(045+051)} + pfd_{(055)} - cash_{(036)}}$$

shareInc : Change in shares outstanding from 1 year ago

$$\frac{shares_t - shares_{t-1}}{shares_{t-1}}$$

ret1 : Trailing 1-month return

ret9 : Trailing 9-month returns skipping the first trailing month

adjRet9 : Risk-adjusted 9-month return

$$\frac{ret9}{residual\_risk}$$

earnRev9 : Change in IBES EPS estimate during the last 9 months

$$\frac{\text{mean}(EPS_t) - \text{mean}(EPS_{t-9})}{\text{std}(EPS_t)}$$

earnDiff9 : IBES EPS diffusion during the last 9 months

$$\frac{\#\_of\_up\_anaysts - \#\_of\_down\_anaysts}{\#\_of\_analysts}$$

ltgRev9 : Change in IBES long-term growth estimate during the last 9 months

$$\frac{\text{mean}(LTG_t) - \text{mean}(LTG_{t-9})}{\text{std}(LTG_t)}$$

A5.2 NET OPERATING ASSETS (NOA)

Most fundamental signals focus on the decomposition and analysis of a firm's NOA, which is the amount of assets deployed to generate business profits. Several quality factors listed above are ratios based on NOA. Now we take a closer look at its derivation.

NOA can be derived from the balance sheet of a firm by rearranging its asset, liability, and owner's equity accounts to reflect: (1) how NOA is financed and (2) where NOA is deployed. Table A5.1 shows the structure of a balance sheet by connecting a firm's assets with its liabilities and shareholders equity. To facilitate a discussion on NOA, each balance sheet account is sorted into four categories (shown in parentheses): operating assets (OA), operating liabilities (OL), financial assets (FA), and financial liabilities (FL). To simplify this discussion, we drop minority interest and preferred stock from this illustration.

TABLE A5.1 Balance Sheet Classification

	Assets		Liab & Owner's Eq
+	cash (FA)	+	st_debt (FL)
+	CA - cash (OA)	+	CL - st_debt (OL)
		+	lt_debt (FL)
+	NCA (OA)	+	NCL - lt_debt (OL)
		+	EQ (FL)
=	TA	=	TA

Note: CA = current assets; NCA = non-current assets; CL = current liabilities; NCL = non-current liabilities; EQ = owner's equity; cash = cash and short-term investments; st_debt = debt in current liabilities; lt_debt = long-term debt; and TA = total assets.

As shown in Equation 5.4, there are two ways to decompose NOAs. In the analysis of the firm's business operations (the operating side), NOA is the net of operating assets (OA) and operating liabilities (OL). OA measures assets deployed to generate business activities (PP&E and inventory) and activities of lending to supplier or customers (accounts receivables). OL reflects borrowing from business partners (suppliers, customers, IRS, or even employees) in the form of accounts payable, tax payable, or pension liabilities. Alternatively, NOA can also be analyzed from the firm's financing perspective, which equals the net of financial liabilities (FL) and financial assets (FA), representing the net investments supplied by enterprise holders (both debt and equity). Assuming that the need for holding cash (or short-term investments) is transitory, NOA calculation deducts cash from FL, pretending as if cash were paid back to enterprise holders.

$$\text{NOA} = \text{OA} - \text{OL} = \text{FL} - \text{FA}\,. \tag{5.4}$$

Table A5.2 shows the rearranged balance sheet. The left-hand side illustrates how investments are deployed for operating activities (the use of cash), whereas the right-hand side demonstrates how investments are raised (the source of cash). Furthermore, operating activities can also be decomposed into working capital (WC) and net noncurrent assets (NCO) by netting current asset with current liabilities and noncurrent assets with noncurrent liabilities, respectively. Combining short-term debt with long-term debt, the financing side becomes debt plus equity minus cash. Equation 5.4 can now be recast as

$$\text{NOA} = \text{WC} + \text{NCO} = \text{debt} + \text{equity} - \text{cash}\,. \tag{5.5}$$

TABLE A5.2 Rearranged Balance According to Net Operating Asset

	(OA - OL)		(FL - FA)
+	CA - cash (OA)	+	st_debt (FL)
–	CL - st_debt (OL)	+	lt_debt (FL)
+	NCA (OA)	+	EQ (FL)
–	NCL - lt_debt (OL)	–	cash (FA)
=	NOA	=	NOA

Equation 5.5 shows the level of NOA at a given time. Equation 5.6 shows the change in NOA from a prior period by taking the first-order difference of (5.5). Decomposition of ΔNOA is readily apparent for the operating side, which includes changes in both working capital and net noncurrent assets. The financing side requires some explanation. The change in debt equals the net of debt issuances and debt repayments during the current period. Change in equity comprises two components: (1) the net of equity issuances and buybacks, and (2) retained earnings that are equal to the net of income and dividend. By aggregating all financing components, the financing side becomes a combination of net external financing (XF) and net income (income). Equation 5.6 illustrates the decomposition of change in NOA.

$$\begin{aligned}\Delta\text{NOA} &= \Delta\text{WC} + \Delta\text{NCO} = \Delta\text{debt} + \Delta\text{equity} - \Delta\text{cash} \\ &= \text{XF} + \text{income} - \Delta\text{cash}\end{aligned} \quad . \tag{5.6}$$

REFERENCES

Arbel, A., Carvell, S., and Strebel, P., Giraffes, institutions and neglected firms, *Financial Analysts Journal*, Vol. 39, No. 3, 57, May–June 1983.

Arnott, R.D., The use and misuse of consensus earnings, *Journal of Portfolio Management*, Vol. 11, No. 3, 18, Spring 1985.

Basu, S., Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis, *Journal of Finance*, 663–682, June 1977.

Benjamin, G., *The Intelligent Investor*, HarperCollins, New York, 1973.

Burgstahler, D. and Dichev, I., Earnings management to avoid earnings decreases and losses, *Journal of Accounting and Economics*, Vol. 24, No. 1, 99, December 1997.

Burgstahler, D.C. and Eames, M.J., Earnings management to avoid losses and earnings decreases: Are analysts fooled?, *Contemporary Accounting Research*, Vol. 20, No. 2, 253, Summer 2003.

Chan, L.K.C., Hamao, Y., and Lakonishok, J., Fundamentals and stock returns in Japan, *Journal of Finance*, Vol. 46, No. 5, 1739, December 1991.

Chan, L.K.C., Jegadeesh, N., and Lakonishok, J., Momentum strategies, *Journal of Finance*, Vol. 51, No. 5, 1681, December 1996.

Cohen, R.B. and Polk, C.K., The Impact of Industry Factors in Asset-Pricing Tests, working paper, Harvard University, 1998.

Daniel, K.D., Hirshleifer, D., and Subrahmanyam, A., Overconfidence, arbitrage, and equilibrium asset pricing, *Journal of Finance*, Vol. 56, No. 3, 921, June 2001.

Dechow, P.M., Sloan, R.G., Sweeney, A.P., Detecting earnings management, *The Accounting Review*, Vol. 70, No. 2, 193–215, April 1995.

Degeorge, F., Patel, J., and Zeckhauser, R., Earnings management to exceed thresholds, *Journal of Business*, Vol. 72, No. 1, 1, January 1999.

Dowen, R.J. and Bauman W.S., The relative importance of size, P/E, and neglect, *Journal of Portfolio Management*, Vol. 12, No. 3, 30, Spring 1986.

Dowen, R.J. and Bauman, W.S., Revisions in corporate earnings forecasts and common stock returns, *Financial Analysts Journal*, Vol. 47, No. 2, 86, March–April 1991.

Fama, E.F. and French, K.R., The cross-section of expected stock returns, *Journal of Finance*, Vol. 47, No. 2, 427, June 1992.

Fama, E.F. and French, K.R., Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, Vol. 33, No. 1, 3, February 1993.

Fama, E.F. and French, K.R., Multifactor explanations of asset pricing anomalies, *Journal of Finance,* Vol. 51, 55–84, 1996.

Fairfield, P.M., Whisenant, J.S., and Yohn T.L., Accrued earnings and growth: implications for future profitability and market mispricing, *The Accounting Review*, Vol. 78, No. 1, 353–372, January 2003.

Frazzini, A., The disposition effect and under-reaction to news, *Journal of Finance*, Vol. 61, No. 4, 2006.

Givoly, D. and Lakonishok, J., The information content of financial analysts' forecasts of earnings: Some evidence on semi-strong inefficiency, *Journal of Accounting and Economics*, 1979.

Givoly, D. and Lakonishok, J., Financial analysts' forecasts of earnings: Their value to investors, *Journal of Banking and Finance*, 1980.

Grinblatt, M. and Han, B., Prospect theory, mental accounting, and momentum, forthcoming, *Journal of Financial Economics*, 2005.

Hawkins, E.H., Chamberlin, S.C., and Daniel, W.E., Earnings expectations and security prices, *Financial Analysts Journal*, Vol. 40, No. 5, 24, September–October 1984.

Hayn, C., The information content of losses, *Journal of Accounting and Economics*, Vol. 20, No. 2, 125, September 1995.

Healy, P.M., The effect of bonus schemes on accounting decisions, *Journal of Accounting and Economics,* Vol. 7, 85–107, April 1985.

Healy, P.M. and Kaplan, R.S., The effect of bonus schemes on accounting decisions/comment, *Journal of Accounting and Economics*, Vol. 7, No. 1–3, 85, April 1985.

Hong, H. and Stein, J.C., A unified theory of underreaction, momentum trading, and overreaction in assets markets, *Journal of Finance*, Vol. 54, No. 6, 2143, December 1999.

Hong, H., Lim, T., and Stein, J.C., Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies, *Journal of Finance*, Vol. 55, No. 1, 265, February 2000.

Hribar, P. and Collins, D.W., Errors in estimating accruals: Implications for empirical research, *Journal of Accounting Research*, Vol. 40, No. 1, 105–134, March 2002.

Jaffe, J., Keim, D.B., and Westerfield, R., Earnings yields, market values, and stock returns, *Journal of Finance*, Vol. 44, No. 1, 135, March 1989.

Jegadeesh, N. and Titman, S., Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance*, Vol. 48, No. 1, 65, March 1993.

Jeter, D.C. and Shivakumar, L., Cross-sectional estimation of abnormal accruals using quarterly and annual data: Effectiveness in detecting event-specific earnings management, *Accounting and Business Research*, Vol. 29, No. 4, 299, Autumn 1999.

Jones, J.J., Earnings management during import relief investigations, *Journal of Accounting Research*, Vol. 29, No. 2, 193–228, Autumn 1991.

Kasznik, R. and Lev, B., To warn or not to warn: Management disclosures in the face of an earnings surprise, *The Accounting Review*, Vol. 70, No. 1, 113, January 1995.

Kerrigan, T.J., When forecasting earnings, it pays to watch forecasts, *Journal of Portfolio Management*, Vol. 10, No. 4, 19, Summer 1984.

Lakonishok, J., Shleifer, A., and Vishny, R.W., Contrarian investment, extrapolation, and risk, *Journal of Finance*, Vol. XLIX, No. 5, 1541–1578, December 1994.

Levitt, A. Jr., The numbers game, *The CPA Journal*, Vol. 68, No. 12, 14, December 1998.

Richards, R.M. and John, D.M., Revisions in earnings forecasts: How much response? *Journal of Portfolio Management*, Vol. 5, No. 4, 47, Summer 1979.

Richardson, S.A., Sloan, R.G., Soliman, M.T., and Tuna, I., Accrual reliability, earnings persistence and stock prices, *Journal of Accounting and Economics*, Vol. 39, No. 3, 437, September 2005.

Rosenberg, B., Reid, K., and Lanstein, R., Persuasive evidence of market inefficiency, *Journal of Portfolio Management*, Vol. 11, No. 3, 9, Spring 1985.

Rouwenhorst, K.G., International momentum strategies, *Journal of Finance*, Vol. 53, No. 1, 267, February 1998.

Rozeff, M.S. and Zaman, M.A., Overreaction and insider trading: Evidence from growth and value portfolios, *Journal of Finance*, Vol. LIII, No. 2, 701–716, April 1998.

Scott, J., Stumpp, M., and Xu, P., Behavioral bias, valuation, and active management, *Financial Analysts Journal*, Vol. 55, No. 4, 49, July–August 1999.

Scott, J., Stumpp, M., and Xu, P., News, not trading volume, builds momentum, *Financial Analysts Journal*, Vol. 59, No. 2, 45, March–April 2003.

ENDNOTES

1. In practice, book value of debt is used to proxy the market value due to data availability issue.
2. To avoid undue influence of outliers and to provide a more robust estimation, we use the market-relative percentile ranking of B2P and ROE in each cross-sectional regression.

CHAPTER 6

Valuation Techniques and Value Creation

Valuation investing seeks to find bargain purchases at prices that are significantly below the intrinsic value. Valuation techniques model the intrinsic value of a firm by forecasting the economics of the firm's business operations and its ability to create shareholder values on a forward-looking basis. For active managers, valuation techniques can complement traditional alpha factors (outlined in Chapter 5) in bottom-up security selection. Valuation is about investing in firms whose economic net worth is likely above its market price; in contrast, quantitative factors seek to arbitrage inefficiencies rooted in behavioral phenomenon. One might think that valuation approach has a lot in common with value factors such as price-to-book, earning yield, etc. But this is not the case, because the former is based on forward-looking economic forecast and requires an explicit forecast of the future, whereas the latter uses a snapshot of the firm's current status as a proxy for its future.

In the investment uses industry, valuation analysis has been used mostly by fundamental equity analysts, both the sell side and buy side, who follow individual companies, estimate their business growth, and calculate the fair value of company stocks. It might seem odd to some that quantitative equity managers would have any use for it. But one must remember that fundamental analysis does contain information, some of which has been used in quantitative models. For example, fundamental analysts issue near-term earning estimates and revisions estimate revision, which have found their way into quantitative factors. It has also been known that aggregate forward-earning forecast for the broad market such as the S&P

500 Index predicts market returns, but not necessarily the actual earnings. It is our view that valuation analysis using multiperiod long-term forecasts by fundamental analysts, when applied appropriately, can add value to quantitative investment processes.

In fact, many aspects of valuation analysis are quantitative in nature. The techniques are built on rational economic forecasts that can be traced to many normative assumptions, such as rationality, perpetuity, mean reversion, or even the validity of CAPM. However, similar to many economic models, valuation techniques place more importance on internal consistency rather than descriptive accuracy.

In this chapter, we will first illustrate a discounted cash flow (DCF) framework. We shall pay particular attention to three subjects: the definition of free cash flow (FCF), drivers of value creation, and the forecasting technique for the fade period. We then extend the one-path, one-life valuation technique into a multipath scenario analysis that provides a distribution of firm valuations. This probabilistic valuation framework is more suitable for forecasting excess returns for active managers given the inherent uncertainty of forecasting the future.

6.1 VALUATION FRAMEWORK

Valuation frameworks take three forms: dividend discount models, discounted cash flow analysis, or economic-value-added approaches. Implemented correctly, all should arrive at the same valuation outcome. In this section, we focus on the discounted cash flow (DCF) framework. As its name implies, DCF defines the intrinsic value of a firm as the sum of the present values of all future cash flows accrued to shareholders in perpetuity. The ultimate goal of the valuation analysis is to compare the resulting intrinsic value to the current equity market value and infer equity return forecast with the relative difference. For instance, if the current stock price is at $10 and the DCF value is $12, the stock is assumed to be undervalued by 20%.

Mathematically, the firm's intrinsic value is given by

$$PV = \sum_{t=1}^{\infty} \frac{f_t}{\left(1+r\right)^t} . \tag{6.1}$$

But how do we estimate cash flows from $t = 1$ to infinity, and what is the appropriate discount rate r? To provide an accurate DCF valuation, we must lay the groundwork for many issues. First, we should understand the components of firm value from both the operating and finance

perspectives. Second, we need to define the notion of free cash flow to shareholders and identify the important drivers and sources that create shareholder value. Third, analysts usually only provide explicit forecasts for one business cycle, generally 5 to 10 years. How do we model business economics and forecast beyond this explicit period? Fourth, we need a framework to estimate the discount rate, consistent with the firm's growth prospect and associated risks.

6.1.1 Firm Value: A Component-Based Approach

A firm's intrinsic enterprise value is not the same as its market value. It is a gauge of a firm's economic net worth in total. It is the sum of operating value, excess cash, and the market value of other nonconsolidated equity investments. For most firms, the majority of the firm value is in the operating value, derived from its future business activities, which is the hardest to estimate. As we shall discuss later, the operating value is the sum of the present value of future free cash flows to the firm (FCFF).

We can also view the firm value from a finance perspective; a firm owes debt to bondholders and preferred stockowners and is owned by minority interest and shareholders of equity. Figure 6.1 shows different operating and finance components of the enterprise value of a firm. Equating the two, we derive a fair equity value by subtracting market value of debt, preferred stocks, and minority interests from the total firm value in Figure 6.2. This is the general framework, and we now discuss each component in detail.

6.1.1.1 Operating Value

Operating value represents the value generated through business activities with the assumption that the company is a going concern and the value will continue in perpetuity. It equals the sum of the present value of all future FCFF that are generated each year through the regular course of business operations. We shall have a detailed definition of FCFF in the next section. Conceptually, FCFF equals the after-tax operating income plus non–cash expenses less the increase in working capitals and capital expenditures (CAPEX).

Figure 6.3 illustrates how operating value is consummated. It is a three-step process according to Equation 6.1: (1) forecasting FCFF on an annual basis in perpetuity, (2) deriving the present value of FCFF discounted by the weighted average cost of capital (WACC) (we shall explain this term shortly), and (3) summing all present values.

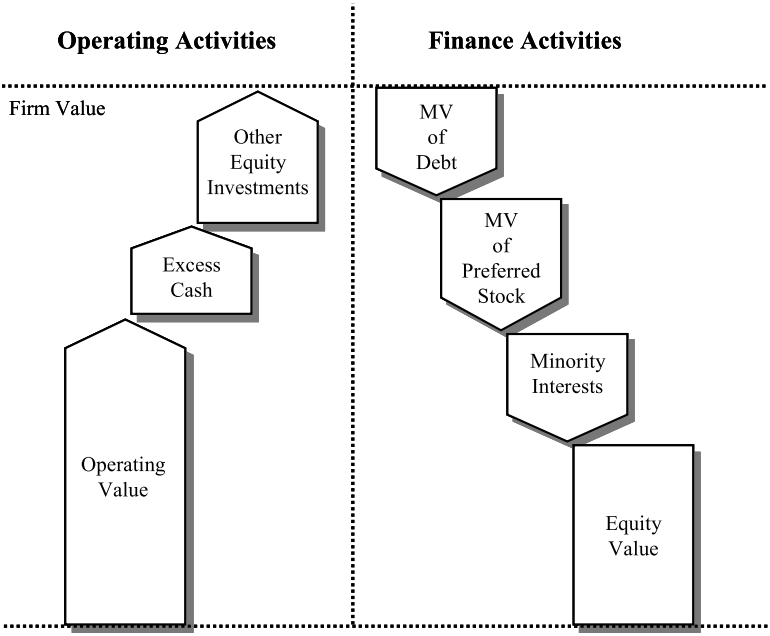


FIGURE 6.1. Components of firm value.

	Operating Value
+	Excess Cash and Marketable Securities
+	Other Equity Investments
=	Firm Value (or Enterprise Value)
−	Market Value of Debt
−	Market Value of Preferred Stocks
−	Minority Interests
=	Equity Value
÷	Shares outstanding
=	Fair Equity Value per Share

FIGURE 6.2. Definition of fair value per share.

As shown in Figure 6.3, it is useful to separate operating value into existing operations and growth opportunities. The former represents the portion of the firm value should there be no firm growth, whereas the latter gauges the portion of the firm value generated from future growth opportunities. Mathematically, we have

$$\begin{aligned} \mathrm{OV} &= \sum_{t=1}^{\infty} \frac{\mathrm{FCF}_t}{(1+\mathrm{WACC})^t} = \sum_{t=1}^{\infty} \frac{\mathrm{FCF}_0}{(1+\mathrm{WACC})^t} + \sum_{t=1}^{\infty} \frac{\mathrm{FCF}_t - \mathrm{FCF}_0}{(1+\mathrm{WACC})^t} \\ &= \frac{\mathrm{FCF}_0}{\mathrm{WACC}} + \sum_{t=1}^{\infty} \frac{\mathrm{FCF}_t - \mathrm{FCF}_0}{(1+\mathrm{WACC})^t} \end{aligned} \tag{6.2}$$

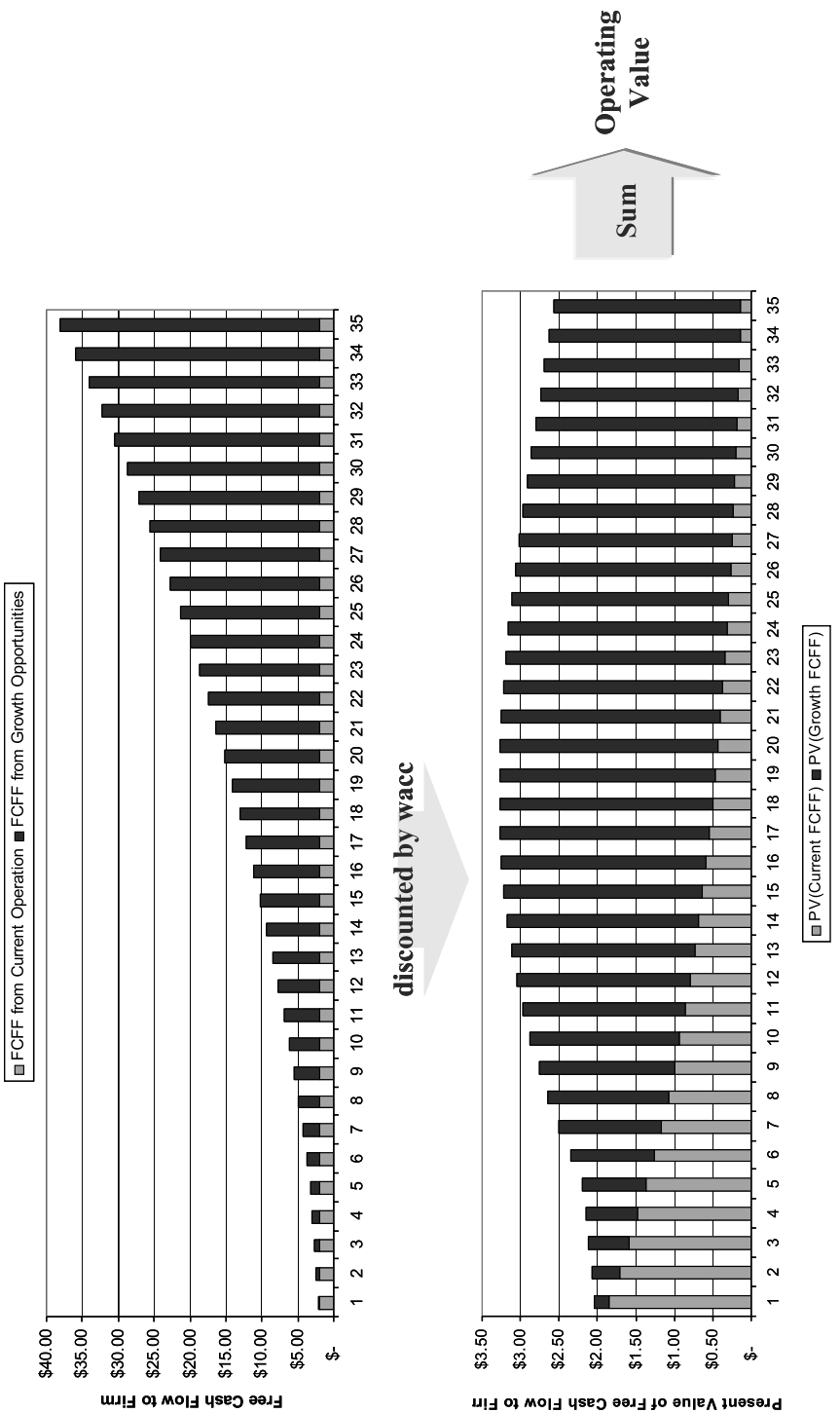


FIGURE 6.3. Operating value from discounted free cash flow.

Naturally, the existing business would account for a bigger portion of the operating value for firms in low-growth industries, whereas the growth opportunity term would account for a bigger portion for those in high-growth industries. Equation 6.3 shows this decomposition under the assumptions that growth rate g is a constant and the discount rate WACC is greater than g.

$$\frac{\text{existing}}{\text{OV}} = \frac{\text{WACC} - g}{\text{WACC} \cdot (1+g)}, \quad \frac{\text{growth}}{\text{OV}} = \frac{g \cdot (1+\text{WACC})}{\text{WACC} \cdot (1+g)}. \tag{6.3}$$

Example 6.1

A hypothetical firm grows its FCF at a 5% annual pace perpetually, and its WACC is 9%. Then

$$\frac{\text{existing}}{\text{OV}} = \frac{9\% - 5\%}{9\% \cdot (1+5\%)} = 42\%, \quad \frac{\text{growth}}{\text{OV}} = \frac{5\% \cdot (1+9\%)}{9\% \cdot (1+5\%)} = 58\%.$$

If the growth rate is 7% instead of 5%, the growth portion of OV increases to 79%.

Focusing on the percentage of value from growth relative to the total operating value, we have

$$\frac{\text{growth}}{\text{OV}} = \frac{g}{1+g}\left(1 + \frac{1}{\text{WACC}}\right) \approx g\left(1 + \frac{1}{\text{WACC}}\right). \tag{6.4}$$

The approximation is valid when the growth rate is not too large. This shows that by holding WACC constant, the percentage of value from growth opportunities is close to a linear function of the growth rate g, whereas when holding the growth rate constant, the percentage is a decreasing function of WACC. This is intuitive because a higher growth rate increases the value of future cash flows but a higher discount rate reduces the present value of future cash flows. By taking its partial derivatives, Equation 6.4 can also be used to derive the relationship between the

change in the ratio and the changes in the growth rate and the discount rate (see Problem 6.2).

6.1.1.2 Excess Cash or Marketable Securities

Excess cash or marketable securities represent the amount of liquid financial instruments that are not required in supporting business operations and can be distributed to enterprise holders. Excess cash is induced by a temporary imbalance of cash flows between operating and finance activities, and this imbalance will eventually be eliminated through cash distributions to either equity or debt holders. It is unnecessary to have a separate DCF analysis of cash instruments because their value is accurately reflected in their market price.

6.1.1.3 Other Nonconsolidated Equity Investments

Equity investments in other business entities that are not consolidated in the FCFF forecast should be included as a separate line item in addition to the operating value. Analysts should avoid double counting the value of a subsidiary by including its valuation impact in both the operating value and other equity investments. In theory, one should try to estimate the fair value of the equity investments through some valuation techniques, which certainly create an additional layer of work. However, when a subsidiary is publicly traded and its value represents a small portion of the firm value, we can simply use the market value of the subsidiary as the product of market value per share and the number of shares held by the firm.

Now that we have covered all the items of the firm value on the operating side, we shall discuss items from a finance perspective.

6.1.1.4 Market Value of Debt and Preferred Stocks

Ideally, the market value, rather than the book value, of debt and preferred stocks should be used in a DCF analysis. However, practitioners rarely use market value for several reasons. First, most equity analysts and managers lack access to pricing databases of fixed-income instruments. Second, most corporate debt today is of the variable rate variety and those by definition should trade close to book, barring some unusual features.

Therefore, book value is typically used in lieu of market value when estimating the fair value of debt and preferred stocks, albeit analysts are encouraged to use market value and to discover fair value whenever possible. Again, when debt and preferred stocks is a small portion of the firm value, this should not be an issue.

6.1.1.5 Minority Interests

Minority interests arise when a third party owns some percentage of one of the firm's consolidated subsidiaries. Typically, minority interest represents a small portion of firm value and only in rare instances does it become significant. Similar to debt, market value of minority interest is the preferred choice. However, there is no market pricing for minority interest; thus, the estimated fair value is used instead. There are two commonly adopted approaches. The first is to use the book value of minority interest reported on the balance sheet. The second approach is to estimate minority interest as a portion of the gross equity value. Gross equity value is the residual of the firm value after subtracting the market value of debt and preferred stocks. The appropriate portion is determined by the ratio of minority interest expense (reported in the income statement) divided by recurring earning, i.e.,

$$\text{MinorityInterest} = \frac{\text{MinorityIntExpense}}{\text{RecurringEarning}} \times (\text{FirmValue} - \text{debt} - \text{preferredStk}).$$

Recurring earning excludes extraordinary items; it is earning before tax (EBT) minus tax expense and plus equity earnings.

6.1.1.6 Other Considerations

Figure 6.1 shows the major components of the intrinsic value of equity. Other adjustments are often made by fundamental analysts in order to achieve a more accurate estimation. For example, on the operating side, other risk provisions are typically deducted from the firm value. On the finance side, the dilution effect of option grants is captured by either scaling up shares outstanding or adjusting the gross equity value downward.

6.2 FREE CASH FLOW

Being the center of DCF analysis, FCF is the portion of a company's operating cash flows that is available for distribution to enterprise holders without any adverse impact on the firm's current or future business economics, such as growth, competitive advantage, profitability, or return on investments. To facilitate the discussion, it is helpful to have a basic understanding of how the business operates. Figure 6.4 shows a conceptual diagram of the flow of a typical business operation and the ownership structure between enterprise holders (creditors and shareholders) and the

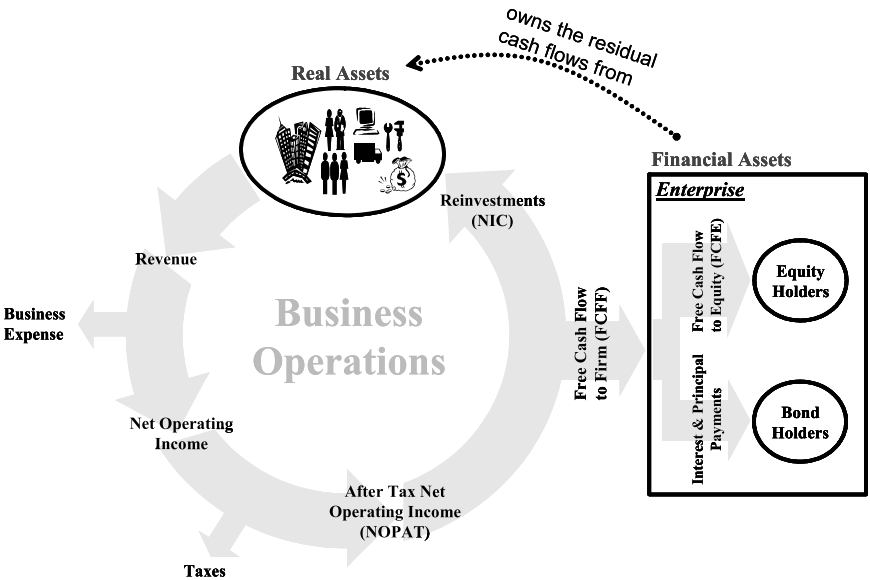


FIGURE 6.4. Business operations and free cash flow.

physical entity of a firm. Several interesting points are discussed in the following text.

- Enterprise holders own financial assets, and this ownership grants them the right to claim the residual cash flow generated through business activities. In this ownership structure, enterprise holders are the principals who provide capital, whereas the company management is the agent who acts on the enterprise holders' behalf in running daily business operations. In addition, creditors have a higher seniority in exercising their claim on the residual cash flow than shareholders. For example, interest payments must be made before dividends can be distributed.
- In terms of the business flow, a firm employs both physical and intellectual assets to conduct its business activities to produce goods. Physical assets include property, plant, and equipment (PP&E) and working capital; intellectual assets are the company management team and the employees. The economics of a business starts with revenue — the gross proceeds received from customers who buy company goods. Business profit is the residual portion of the revenue

after deducting business expenses and taxes — net operating income after tax (NOPAT). A portion of the NOPAT is plowed back as reinvestment in order to sustain the firm's growth and competitive advantage. Should NOPAT be larger than the reinvestment, the firm generates a positive FCF that can be distributed to enterprise holders. On the other hand, if the reinvestment is larger than NOPAT, FCF is negative, and the firm would need to engage in external financing to solicit additional capital from enterprise holders to fund the reinvestment.

- There are two types of FCF: free cash flow to firm (FCFF) and free cash flow to equity (FCFE). The former is the residual cash flow available to enterprise holders, whereas the latter is the residual cash flow available to equity holders only, after principal and interest payment have been made to debt holders.

6.2.1 Definition of FCF

In Figure 6.5, we define FCF from items in income and cash flow statements. Starting with the revenue, FCFE is the residual portion after subtracting four major components: operating expenses, taxes, incremental investments, and payments to creditors. We provide some detail for each component as follows:

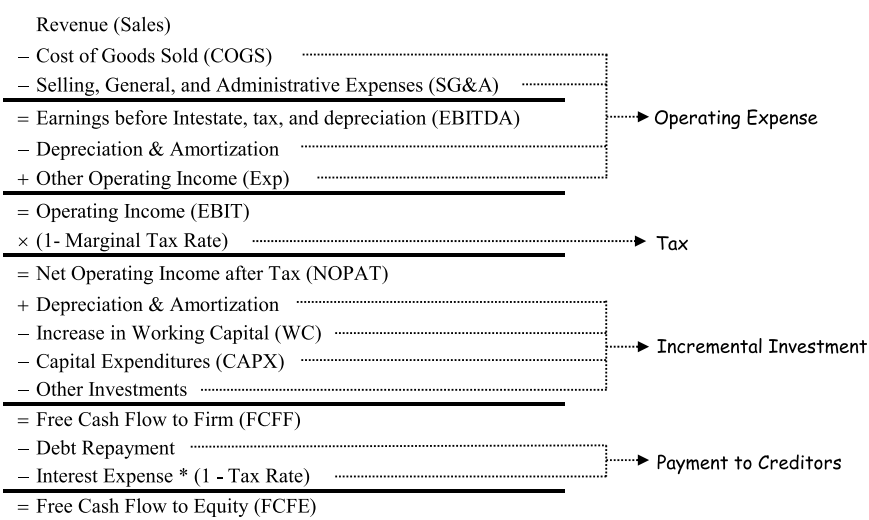


FIGURE 6.5. Definition of free cash flow.

- *Operating expenses*: These can be divided into three categories. They are cost of goods sold (COGS), selling, general, and administrative costs (SGA), and depreciation expense. COGS arises from costs associated with raw material and labor in manufacturing goods for customers or in delivering services to them; SGA is the necessary overhead incurred on the corporate level to support sales/marketing activities, legal, or human resource functions; and depreciation expense comes from the aging of fixed assets such as PP&E. Operating income is revenue less the operating expense. A related concept is the operating margin, which is the operating income divided by revenue; it measures the profitability of a firm's business operations. Holding revenue constant, the lower the operating expenses, the more profitable is the business.
- *Taxes*: Tax includes levies from all levels of government: federal, state, city, or local. In general, statutory marginal tax rate should be used and short-term fluctuations in tax rate, due to prior losses or tax incentive programs, should be adjusted on a one-time basis. As mentioned before, operating income after tax is NOPAT.
- *Incremental investments*: A firm regularly reinvests a portion of NOPAT in itself in order to expand its business operations and to sustain its competitive advantage. Incremental investments consist of three parts: an increase in working capital (ΔWC), the incremental capital expenditure (ICAPEX), and other investments. Although an increase in working capital reflects the additional resources needed for fueling short-term growth, capital expenditure expands the capacity of business operation in order to achieve long-term firm growth. As shown in Figure 6.6, working capital is the net of current assets and current liabilities. ICAPEX is the portion of capital expenditure that is above depreciation and amortization (DA) expense; in essence, it represents the economic addition to PP&E. Lastly, other investment includes outlays for acquisitions, which generate nonorganic firm growth. NOPAT after incremental investments is FCFF.

	Inventory
+	Accounts Receivable
+	Other Current Assets
–	Accounts Payable
–	Other Current Liabilities
=	Working Capital

FIGURE 6.6. Definition of working capital.

- *Payment to creditors*: Finally, there is payment to creditors, including interest expense and debt repayment. FCFF after payment is FCFE.

To summarize, FCFF for a given period is NOPAT less the incremental investments, which is the change in a firm's capital

$$\text{FCFF} = \text{NOPAT} - \Delta\text{Capital}\,. \tag{6.5}$$

6.2.2 Linkage between Operating and Finance Cash Flows

By its definition, in the long run, FCFF must equal payments to (or contributions from) enterprise holders. But in the short run, this balance does not necessarily hold, and the temporal differences are reflected in the change in the cash account on the balance sheet and the change in external financing from the enterprise holders, i.e.,

$$\Delta\text{CASH} = \text{FCFF} + \Delta\text{XF}\,. \tag{6.6}$$

Thus, if there is no change in the cash account, a negative FCFF means that an additional capital infusion is required from either shareholders or creditors, whereas a positive FCFF implies that a portion of NOPAT will be distributed to enterprise holders. In general, a temporary difference between FCFF and cash flow from finance activities results in a change in the cash account.

6.2.3 Agency Problem and Economic Forecast

An economic forecast typically focuses on a firm's business and ignores the behavioral idiosyncrasies of company management — the agent — and it further assumes that all agents behave rationally. In the case of a DCF model, analysts often assume that the company management will act in the best interest of its shareholders and, conversely, shareholders will trust their company management when asked to contribute additional capital. Such tacit assumptions are necessary to derive an internally consistent firm fair value.

However, as illustrated by a long list of empirical research outlined in the previous chapter, the reality is quite different because of the agency problem where the management does not always act in the best interests of their shareholders. For example, an abnormal increase in inventory could be interpreted rationally as a reflection of a short-term spike of

demand. However, the agency problem might describe such an increase as a symptom of earnings management (or even worse, earnings manipulation) wherein costs are shifted from the current period to future periods for the purpose of boosting reported earnings. The inconsistency between the two interpretations is exacerbated by the fact that most fundamental analysts seek answers/guidance directly from the company management, potentially resulting in a rosier forecast than what reality would otherwise suggest. This underscores the importance of using a quantitative alpha model in conjunction with valuation techniques to perform bottom-up security selection. Quantitative models can help navigate around behavioral idiosyncrasies, whereas valuation techniques provide economic forecasts based on the assumption of rationality.

6.3 MODELING THE BUSINESS ECONOMICS OF A FIRM

An integrated analysis of a firm's business economics — a firm's ability to create shareholder value — starts with the ratio of return on incremental capital (RIC), followed by the decomposition of the RIC ratio, and ends with a detailed analysis and forecast of various components that build up the FCFF forecast. As we shall see later, modeling business economics focuses solely on a firm's operating activities and ignores finance decisions.

6.3.1 Return on Incremental Capital

RIC measures the expected incremental earnings generated by a dollar of additional investment into a firm's business operations, defined as the ratio $\text{RIC} = \Delta\text{Income}/\Delta\text{Capital}$. Finance decisions are ignored because this ratio is indifferent to the source of the additional capital, whether it is debt financing, equity financing, or NOPAT. It focuses on the question of how much profit can be generated through incremental operating activities. Because RIC measures the productivity of a firm in total, ΔIncome equals the change in ΔNOPAT, and $\Delta\text{Capital}$ equals the change in net operating assets (ΔNOA). So we can write

$$\text{RIC} = \frac{\Delta\text{NOPAT}}{\Delta\text{NOA}}. \tag{6.7}$$

The difference between the RIC and the cost of capital is the economic value creation (EVC) of a firm, i.e., $\text{EVC} = \text{RIC} - \text{WACC}$.

6.3.2 Decomposition of RIC

Incremental capital investments, which equals the change in net operating asset, generate additional sales or revenues, which in turn translates to additional income. By introducing ΔSales into Equation 6.7, we can decompose RIC into two major value drivers — profitability and scalability, measured by $\Delta\text{NOPAT}/\Delta\text{Sales}$ and $\Delta\text{Sales}/\Delta\text{NOA}$ respectively. Hence,

$$\text{RIC} = \frac{\Delta\text{NOPAT}}{\Delta\text{NOA}} = \frac{\Delta\text{NOPAT}}{\Delta\text{Sales}} \times \frac{\Delta\text{Sales}}{\Delta\text{NOA}} = \text{profitability} \times \text{scalability}. \quad (6.8)$$

Profitability gauges the expected profit margin per one dollar of incremental sales, whereas scalability reflects the additional capital investments that are required to generate one more dollar of incremental sales. The two measures vary widely across industries and across firms within the same industry. The determinants of these two measures depend on the nature of the business.

- *Profitability*: A firm's profitability depends on the *competitive structure* of the industry as well as the part of the *value system* in which a firm's business model resides. The business model determines how much economic value the firm creates, between its upstream suppliers and its downstream customers. The competitive structure governs the portion of economic value that can be retained by the firm. Michael E. Porter (1985) provides structured analyses of both.
- *Scalability*: Scalability depends on the nature of the business. For example, capital-intensive industries are often less scalable, and consequently it is typically harder for firms in these industries to create shareholder value through growth. In contrast, industries with low fixed cost are the prime candidates for business expansions.

6.3.3 Further Decompositions of RIC

Equation 6.8 can be further decomposed into its underlying drivers by

$$\begin{aligned} \text{NOPAT} &= (\text{Sales} - \text{COGS} - \text{SGA} - \text{DA}) \cdot (1 - \text{taxRate}) \\ \Delta\text{NOA} &= \Delta\text{WC} + (\text{CAPEX} - \text{DA}) + \Delta\text{otherAssets} \end{aligned}. \quad (6.9)$$

Assuming the tax rate does not change, substituting Equation 6.9 into Equation 6.8 yields

$$\text{profitability} = \left(1 - \frac{\Delta\text{COGS}}{\Delta\text{Sales}} - \frac{\Delta\text{SGA}}{\Delta\text{Sales}} - \frac{\Delta\text{DA}}{\Delta\text{Sales}}\right) \times (1 - \text{taxRate})$$
$$\frac{1}{\text{scalability}} = \frac{\Delta\text{WC}}{\Delta\text{Sales}} + \left(\frac{\text{CAPEX}}{\Delta\text{Sales}} - \frac{\text{DA}}{\Delta\text{Sales}}\right) + \frac{\Delta\text{otherAssets}}{\Delta\text{Sales}} \tag{6.10}$$

Profitability of a firm depends on the following four subcomponents:

- Cost of goods sold (COGS): It contains both labor and raw material costs. It measures direct costs in producing final products. In order to be successful, firms subject to price competition must have a lower-than-industry COGS structure.
- Selling, general, and administrative expense (SGA): It contains costs associated with marketing expenses and corporate overhead, such as human resource, legal, or administrative functions. For firms relying on product differentiation, a higher-than-industry SGA is typically required to maintain their competitive advantage.
- Depreciation and amortization (DA): Depreciation is associated with the use of tangible, long-term assets — PP&E. Amortization is the charge against acquired, nontangible assets, such as patents.
- Tax rate: Tax rate is a percentage of the net operating income paid for all governmental levies.

Salability has the following three subcomponents:

- Change in working capital (ΔWC): This is associated with the additional resources that are needed to accommodate short-term growth needs, such as proper level of inventory, increase in accounts receivable, etc.
- Incremental capital expenditures (ICAPEX): It represents the net of CAPEX and DA. It is the additional capital investments in non-current assets to expand operating capacity in order to achieve higher long-term growth.
- Change in other assets: This item captures other forms of investments that are not part of the prior two categories.

Figure 6.7 summarizes the structure of RIC and all the relevant components discussed so far.

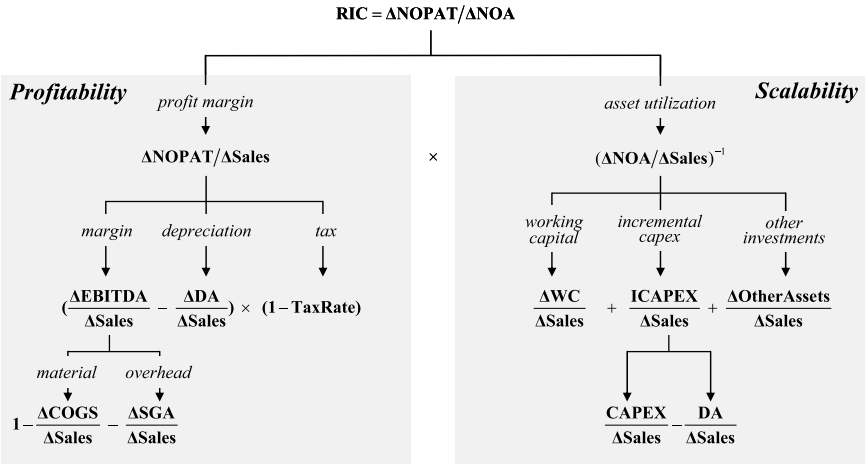


FIGURE 6.7. Modeling business economics.

6.3.4 RIC Decomposition and FCFF Forecast

We shall use the decompositions of RIC to forecast FCFF. Starting with Equation 6.5, we have

$$\begin{aligned}\text{FCFF}_t &= \text{NOPAT}_t - \Delta\text{Capital}_t = \text{NOPAT}_t - \Delta\text{NOA}_t \\ &= \text{Sales}_t \frac{\text{NOPAT}_t}{\text{Sales}_t} - \Delta\text{Sales}_t \frac{\Delta\text{NOA}_t}{\Delta\text{Sales}_t} \\ &= \text{Sales}_t \left[\frac{\text{NOPAT}_t}{\text{Sales}_t} - \frac{\Delta\text{Sales}_t}{\text{Sales}_t} \left(\frac{\Delta\text{Sales}_t}{\Delta\text{NOA}_t} \right)^{-1} \right]\end{aligned} \tag{6.11}$$

The first ratio $\text{NOPAT}_t / \text{Sales}_t$ is the profit margin. For simplicity, we assume it is constant and estimated based on historical measures. The second ratio $\Delta\text{Sales}_t / \text{Sales}_t$ is the revenue growth rate g_{t+1}. The third ratio is the scalability measure defined earlier. Equation 6.11 becomes

$$\text{FCFF}_t = \text{Sales}_t \left[\text{profitability}_t - g_{t+1} \left(\text{scalability}_t \right)^{-1} \right]. \tag{6.12}$$

- The FCFF margin $\text{FCFF}_t / \text{Sales}_t$ is $\text{profitability}_t - g_{t+1} \left(\text{scalability}_t \right)^{-1}$. Intuitively, FCFF margin, at time t, is positively correlated with a firm's profitability and scalability, and negatively correlated with the growth rate due to the required reinvestment.

6.3.5 Firm Value

As a first approximation, we derive the firm operating value using the DCF model by assuming the firm will grow perpetually at a constant growth rate *g*. Profitability and scalability are also assumed to be constants denoted as $\bar{p}$ and $\bar{s}$ respectively to represent their expected values. In addition, the appropriate WACC is *w*, which is greater than *g*. Then the firm value is given by

$$\mathrm{OV} = \sum_{t=1}^{\infty} \frac{S_0\left(1+g\right)^t (\bar{p} - g/\bar{s})}{\left(1+w\right)^t} = S_0(\bar{p} - g/\bar{s})\frac{1+g}{w-g}\,. \tag{6.13}$$

The barred variables denote expected value and S_0 is the initial sales at time O.

Example 6.2

A hypothetical firm currently generates one dollar of sales S_0. Its profitability and scalability are 10% and 2, respectively. Its sales will grow at a 5% annual pace perpetually, and its WACC is 9%. The fair value for this firm is

$$\frac{\$1 \times (10\% - 5\%/2)(1+5\%)}{(9\% - 5\%)} = \$1.97\,.$$

The FCFF margin is $7.5\% = 10\% - 2^{-1} \cdot 5\%$, and RIC is equal to 20% = $10\% \cdot 2$. The EVC of this firm is 11%.

6.3.5.1 Sensitivities

Based on (6.13), we can derive the sensitivities of the firm value to the various inputs. We have

$$\begin{aligned}
&\frac{\Delta \mathrm{OV}}{\mathrm{OV}} = \frac{1}{\bar{p} - g/\bar{s}}\Delta\bar{p}, \quad \frac{\Delta \mathrm{OV}}{\mathrm{OV}} = \frac{g}{s^2\left(\bar{p} - g/\bar{s}\right)}\Delta\bar{s} \\
&\frac{\Delta \mathrm{OV}}{\mathrm{OV}} = \frac{\Delta S_0}{S_0}, \quad \frac{\Delta \mathrm{OV}}{\mathrm{OV}} = \frac{-1}{(w-g)} \cdot \Delta w \qquad . \\
&\frac{\Delta \mathrm{OV}}{\mathrm{OV}} = \left[-\frac{1}{s\left(\bar{p} - g/\bar{s}\right)} + \frac{1}{\left(1+g\right)} + \frac{1}{(w-g)}\right]\Delta g
\end{aligned} \tag{6.14}$$

TABLE 6.1 Sensitivity of DCF Inputs

Input	Sensitivity
Profitability	13.33
Scalability	0.17
Growth	19.29
Sales	1
Weighted average cost of capital	−25

Table 6.1 shows the sensitivity of the fair value for each DCF input in our example. For instance, 1% increase in profitability would results in 13% increase in fair value. In terms of the absolute magnitude of sensitivities, the fair value is most sensitive to the WACC estimate, followed by growth rate and profitability. The scalability is the least sensitive input.

6.4 COST OF CAPITAL

So far we have denoted the discount rate as WACC. We provide this explicitly in this section. The cost of capital represents the opportunity costs of all the capital providers — creditors and shareholders — whose funds can be invested in other opportunities. The WACC is simply the sum of cost of capital for each of the capital provider times their proportion of the capital structure. Most valuation and corporate finance books discuss the estimation of WACC extensively. We shall skip a detailed discussion of its construction and instead highlight several important, practical considerations for equity managers.

$$\text{WACC} = \frac{k_s \cdot S + k_b \cdot (1 - \text{taxRate}) \cdot B + k_p \cdot P}{V}. \tag{6.15}$$

In the definition, k_s, k_b, and k_p are the cost of equity, debt, and preferred stocks, respectively; and S, B, and P are the market values of equity, debt, and preferred stocks, respectively. The total market value of the firm is $V = S + B + P$.

The cost of equity k_s is determined by the risk of equity investment, and it is common for practitioners to use a required return from a risk model as the cost of equity. The cost of debt is determined primarily by the corporate bond yield and the same is true for preferred stock. Note the following:

- The discount rate must be consistent with the type of cash flow estimation. Mismatching these estimations would invariably result in

erroneous estimation of operating value. For example, if FCFE is the estimated cash flow, its discount rate should be the cost of equity. On the other hand, WACC is the appropriate rate to discount FCFF.

- The discount rate estimation should be kept as simple as possible. Complex methodology not only diverts valuable resources that could otherwise be devoted to forecast FCFF but also typically yields inferior *ex post* performance. Often, complex and questionable WACC estimation is fudged in order to achieve a "valuation target," simply because the fair value is most sensitive to a unit change in WACC estimate as illustrated in the previous section.

A check on the WACC can be done by looking at the yields on the company's debt or the yields implicit to its credit rating. Generally, equity holders would want around 2% more than the cost of a company's long-term (10 years) debt.

6.5 EXPLICIT PERIOD, FADE PERIOD, AND TERMINAL VALUE

To forecast FCFF into perpetuity, the DCF valuation framework breaks the forecasting horizon into three periods — the explicit period, fade period, and constant growth period. Our discussion thus far has focused on modeling the business economics in the explicit period, which typically spans over 5 to 10 years. The fade period is the forecasting horizon beyond the explicit period during which the firm matures and gradually loses its competitive advantage. Two economic principles must be upheld when forecasting FCFF in the fade period.

RIC fades to WACC: Economic theory suggests that in the long run competition will eventually eliminate all economic value creation (EVC=RIC–WACC), which reflects a firm's ability to deliver higher return on investments than the opportunity cost (WACC).

Growth rate fades to long-run GDP growth: It is unrealistic to assume that a company can grow faster than the economy for an extended period of time, because the sales of such a company will eventually be bigger than the total output of the economy. Economic theory also suggests that the long-term risk-free rate provides an unbiased proxy for the economic growth rate. Thus, sales growth should fade to long-term risk-free rate in perpetuity.

Mathematically, we have

$$\begin{aligned} \mathrm{RIC}_{t+1} &= (\mathrm{RIC}_t - \mathrm{WACC}) \times F_t + \mathrm{WACC} \\ g_{t+1} &= (g_t - r_f) \times F_t + r_f \end{aligned} \tag{6.16}$$

In the formula, the long-term risk-free rate is r_f, and F_t is the fade function that declines from 1 to 0 during the fade period. Given RIC and growth forecasts, FCFF in the fade period can be derived as

$$\begin{aligned} \mathrm{FCFF}_t &= \mathrm{NOPAT}_t - \Delta\mathrm{NOA}_t \\ &= \mathrm{NOPAT}_t - \Delta\mathrm{NOPAT}_{t+1} / \mathrm{RIC}_{t+1} \\ &= \mathrm{NOPAT}_t - \mathrm{NOPAT}_t \times g_{t+1} / \mathrm{RIC}_{t+1} \\ &= \mathrm{NOPAT}_t \times (1 - g_{t+1} / \mathrm{RIC}_{t+1}) \end{aligned} \tag{6.17}$$

Figure 6.8 shows an example of an exponential decay (fade function) applied to the RIC and growth rate forecasts in the fade period. Exponential decay is characterized by the half-life — the amount of time it takes the value of the function to drop by one half. In the example, the half-life is 6 years.

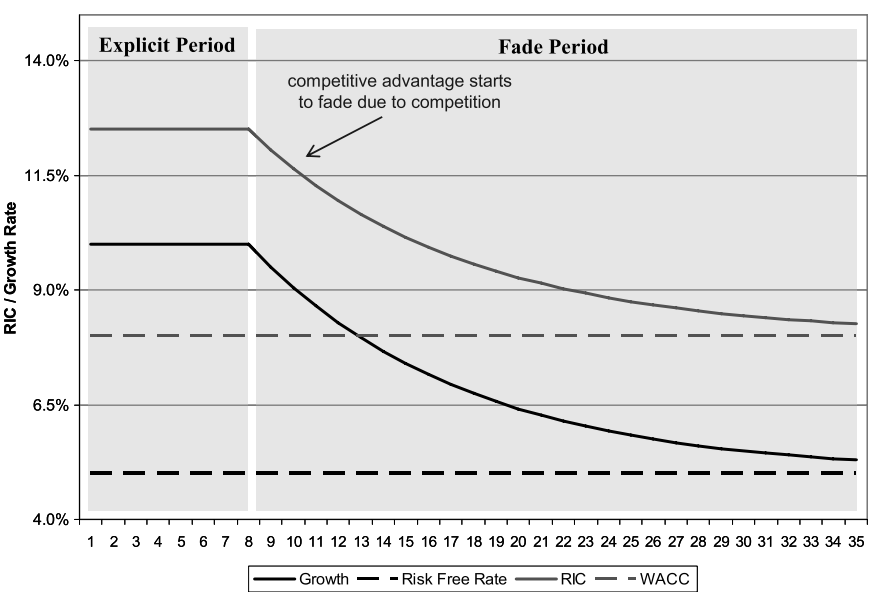


FIGURE 6.8. The fade period.

		2005/12/25
COE	Local Risk Free Rate [A]	3.94%
	+ Global Equity Risk Premium [B]	3.50%
	+ Company Premium/Discount [C]	1.00%
	= **Cost Of Equity [D=A+B+C]**	**8.4%**
COD	Local Risk Free Rate [A]	3.94%
	+ Global Debt Risk Premium [E]	2.00%
	+ Company Premium/Discount [C]	1.00%
	= **Cost Of Debt [F=A+E+C]**	**6.9%**
weight	Price per Share [G]	37.95
	´ Shares Outstanding [H]	78.5
	= **Market Value of Equity [I=G*H]**	2979.1
	Book Value of Debt [J]	21
	MVE% [K=I/(I+J)]	**99.3%**
	MVD% [L=J/(I+J)]	**0.7%**
	WACC [M=(D*K)+(L*F)]	**8.4%**

FIGURE 6.9. Weighted average cost of capital for CAKE.

Lastly, in the final stage after the fade period, the firm grows at the constant risk-free rate with the RIC the same as the WACC. A terminal value can be obtained for the remaining FCFF.

6.6 AN EXAMPLE: CHEESECAKE FACTORY, INC. (CAKE)

We have established the entire DCF process for firm valuation. To illustrate how it applies in practice, we devote this section to evaluate the intrinsic value of Cheesecake Factory, Inc. (ticker: CAKE), a popular restaurant chain specializing in upscale casual dining. We will start with the estimation of the discount rate using a straightforward approach. The RIC and its subcomponents are then modeled for the Cheesecake Factory, Inc., to pave the way for FCFF forecasts. In addition, the operating value is estimated as the summation of three time periods discussed above: explicit period, fade period, and terminal value. Finally, equity value is consummated and compared with the current market value.

6.6.1 Weighted Average Cost of Capital (WACC)

Figure 6.9 shows WACC as a weighted average of (1) cost of equity (COE) and (2) cost of debt (COD). Their weighting is proportional to the market value of equity and the book value of debt. Our methods of estimating COE and COD are simple but practical. For example, COE consists of

three parts: a local risk-free rate, the global equity risk premium, and a company-specific premium (or discount). Although the local risk-free rate changes from country to country, the global equity risk premium is the same for all companies. We also note that our risk-free rate is nominal (instead of real); therefore in order to be consistent, our FCFF forecasts are also estimated on a nominal basis. Company-specific premium is a catch-all term, based on different beliefs of how assets are priced. If one subscribes to the notion of CAPM, the company-specific premium reflects each company's beta to the market. Should one use the Fama–French three-factor model, the catch-all term would reflect the company's exposures to market capitalization, book-to-price, and beta. COD has a similar structure, and CAKE has no preferred stock.

6.6.2 Return on Incremental Capital (RIC) and FCFF

Figure 6.10 shows the RIC forecast in the explicit period and the FCFF forecast for FY1. To ensure that the RIC forecast is realistic and possibly errs on the conservative side, it is useful to prepare a side-by-side comparison with the 5-year historical average and IBES consensus estimates. The RIC forecast for CAKE is 13.8%, with the profit margin being 7.3% and asset utilization being 1.89. That is, CAKE is expected to retain 7.3¢ as profit for every dollar of sales and it is expected to generate $1.89 of incremental sales per one dollar of reinvestment. With the WACC estimated at 8.4%, CAKE is expected to deliver abnormal return of 5.4% (= 13.8% – 8.4%) to its shareholders — a positive value company.

For the fiscal year 1 (FY1), assuming CAKE's sales is $1399 million with the NOPAT margin being 7.3%, CAKE will earn $102 million (= 1399 * 7.3%). The expected reinvestment (or ΔNOA) is $154 million, which equals the product of FY1 sale ($1399 million), sale growth (20.8%), and the inverse of scalability ($0.53 = 1.89^{-1}$). Because the expected reinvestment ($154 million) is greater than the expected NOPAT ($102 million), CAKE has a negative FCFF of $52 million. In other words, CAKE is expected to raise $52 million of cash through external financing in FY1, largely due to its extraordinary pace of growth at 20.8% per annum which cannot be funded through internal cash generation.

6.6.3 Operating Value

As shown in Figure 6.11, the operating value is estimated as the sum of three parts: (1) the present value of FCFF in the explicit period, (2) the present value of FCFF in the fade period, and (3) the present value of the

		Estimate	5 Yr. Hist Avg	IBES Est
sales	**Sales Growth Rate [A]**	**20.8%**	**22.8%**	**19.96%**
	FY1 Sales [B]	**1399**	**-**	**1399**
profitability	EBITDA Margin [C]	14.7%	13.9%	14.9%
	- Depr / Sales [D]	3.5%	3.4%	3.5%
	= Operating Margin [E=C-D]	11.2%	10.5%	11.4%
	- Tax Rate [F]	34.8%	35.6%	34.7%
	= **NOPAT Margin [G=E*(1-F)]**	**7.3%**	**6.7%**	**7.4%**
scalability	CAPEX / ΔSales [H]	72.2%	57.7%	-
	- Depr / ΔSales [I]	16.8%	14.3%	-
	= ICAPEX / ΔSales [J=H-I]	55.4%	43.4%	-
	+ ΔWorking Capital / ΔSales [K]	-2.5%	-2.9%	-
	+ ΔNet Other Assets / ΔSales [L]	0.0%	-1.9%	-
	= **ΔSales / ΔNOA [M=1/(J+K+L)]**	**1.89**	**2.59**	-
EVA	**RIC [N=G*M]**	**13.8%**	**17.5%**	-
	- WACC [O]	8.4%	8.4%	-
	= **value creation [=N-O]**	**5.3%**	**9.1%**	-

		12/2006 (E)
NOPAT	Current year's forecasted sales [=B]	1399
	EBITDA [P=A*C]	205
	- Depr & Amort [Q=A*D]	49
	= Operating Income [R]	156
	- Taxes [S=R*F]	54
	= **NOPAT [T=R-S]**	**102**
ΔNOA	CAPEX [U=H*A*B]	210
	- Depr & Amort [Q=A*D or A*B*I]	49
	= ICAPEX [R=U-Q]	161
	+ Δ Working Capital [S=K*A*B]	-7
	+ ΔNet Other Assets [T=L*A*B]	0
	= **ΔNOA [U=R+S+T]**	**154**
	= **FCFF [=T-U]**	**-52**

FIGURE 6.10. Business economics and FCFF forecasts of CAKE.

terminal value. In the explicit period (2006–2010), RIC and growth stay constant resulting in the same FCFF margin in these years. This means NOPAT, ΔNOA, and FCFF all grow at the same rate as sales.

In this example, we choose a fade period of almost 40 years. Choosing different fade horizons does not change the valuation result materially, as long as its duration is greater than 30 years. The following steps are worth noting in the computation.

Explicit Period	**2006/12**	**2007/12**	**2008/12**	**2009/12**	**2010/12**
Growth	20.8%	20.8%	20.8%	20.8%	20.8%
RIC	13.8%	13.8%	13.8%	13.8%	13.8%
Sales	1399	1689	2040	2464	2976
NOPAT	102	123	148	179	217
ΔNOA	154	186	224	271	327
FCFF	**-52**	**-63**	**-76**	**-91**	**-110**
t	1.01	2.01	3.01	4.01	5.01
PV(FCFF)	**-48**	**-53**	**-59**	**-66**	**-74**

Fade Period	**2011/12**	**2012/12**	**2013/12**	**2014/12**	**2015/12**	•••	**2046/12**	**2047/12**	**2048/12**	**2049/12**
Growth	20.0%	18.4%	16.9%	15.6%	14.5%		4.3%	4.3%	4.3%	4.2%
RIC	13.2%	12.7%	12.3%	11.9%	11.6%		8.5%	8.5%	8.5%	8.5%
NOPAT	260	308	360	416	476		3707	3866	4031	4201
ΔNOA	374	423	471	519	567		1870	1937	2006	2079
FCFF	**-114**	**-115**	**-112**	**-103**	**-91**		**1836**	**1929**	**2025**	**2123**
t	6.01	7.01	8.01	9.01	10.01	•••	41.01	42.01	43.01	44.01
PV(FCFF)	**-70**	**-65**	**-58**	**-50**	**-40**		**67**	**65**	**63**	**61**

Terminal Value	**2050/12**
NOPAT	4378
Terminal Value	**52021**
t	45.01
PV(Terminal Value)	**1370**

FIGURE 6.11. Explicit period, fade period, and terminal value for CAKE.

- *RIC fade*: RIC is exponentially faded at 10% each year from 13.8% to WACC 8.4%. This results in a RIC of 13.2% for 2011.
- *NOPAT in 2011*: NOPAT for the year 2011 is based on 2010 NOPAT and 2010 ΔNOA and the 2011 RIC from the preceding step. We have

$$\begin{aligned}\text{NOPAT}_{2011} &= \text{NOPAT}_{2010} + \Delta\text{NOPAT}_{2011} \\ &= \text{NOPAT}_{2010} + (\Delta\text{NOA}_{2010} \times \text{RIC}_{2011}) \\ &= 217 + (327 \cdot 13.2\%) = 260\end{aligned}$$

- *Growth fade*: The growth rate in 2011 is calculated as (NOPAT_{2011}/ NOPAT_{2010} –1), which equals 20%. It is then exponentially faded at 10% each year to the long-term risk-free rate of 4.2%.
- *ΔNOA estimation*: Because ΔNOA is defined as the required reinvestment in order to achieve next year's NOPAT growth target, it is estimated by $\text{NOPAT}_t \times (g_{t+1} / \text{RIC}_{t+1})$.
- *Terminal value*: Lastly, the terminal value is a perpetual valuation of a firm with no growth.[1] Specifically, CAKE is expected to generate $4378 million of NOPAT in 2050, and its NOPAT will stay at that level in years beyond 2050, as well. Because CAKE is not expected to achieve any NOPAT growth after year 2050, it is also not expected

		% of EV	Value
operating activities	Operating Value from Existing Business	35%	1,001
	+ Operating Value from Growth	62%	1,744
	= Operating Value	**97%**	**2,745**
	+ Excess cash and marketable securities	3%	75
	+ MV of equity and other investments	-	-
	– MV of provision for risks and charges	-	-
finance decisions	**= Firm Value/Enterprise Value**	**100%**	**2,820**
	– MV of debt, pref & other obligations	1%	21
	– MV of minority interests	-	-
valuatio	= Equity value	99%	2,799
	÷ Shares Outstanding		78.5
	= Equity value / share		**35.64**
	Current price / share		37.95
	Under / (over) valued %		**-6%**

FIGURE 6.12. Valuation summary for CAKE.

to reinvest in its business operations. Thus, ΔNOA is expected to be 0 for years beyond 2050, and NOPAT is equal to FCFF. Terminal value is $52,021 million ($4,378 million divided by 8.4%). Finally, the terminal value of $52,021 million is discounted back to today and is worth $1,370 million.

6.6.4 Valuation Summary

Based on the DCF calculation of operating value, Figure 6.12 shows the detailed valuation components for CAKE. Setting the enterprise value (or total firm value) to 100%, we can break down the contributions from each valuation component in percentage terms. According to Figure 6.12, the operating value is the biggest slice, accounting for 97% of the enterprise value; within the operating value, CAKE's future growth prospect is the biggest contributor, delivering 62% of the enterprise value. In all, as of the date we conducted this valuation analysis, CAKE is fairly priced by the market at a small premium of 6%. Based on this analysis of valuation components, it is clear that the intrinsic value of CAKE is mostly dependent on its future growth rate. As seen from the table, CAKE uses NOPAT plus additional capital infusion to expand its operating assets in order to sustain its growth. As a result, FCFF is negative for the initial years and only

turns positive after more than 10 years. Should it deviate from the current forecast of 20.8%, CAKE's relative premium/discount from its current stock price will change as well, perhaps significantly so.

This example therefore also highlights the sensitivity of valuation analysis to the underlying growth assumptions. We shall now introduce multipath sensitivity analysis to firm valuation and devise various ways to obtain the standard error of fair value.

6.7 MULTIPATH DISCOUNTED CASH FLOW ANALYSIS

So far, our discussion has focused on how to model the set of value drivers, such as RIC or growth rate, as DCF inputs to forecast a company's cash flows and to determine its enterprise value (EV). In reality, *ex post* realizations of these drivers are subject to many exogenous influences. For example, different economic environments, boom or bust, would influence the expected growth rate of a particular company and subsequently result in a different EV estimation. The same argument is true for the forecasts of a firm's profitability and scalability, which jointly determine the RIC forecast. This highlights the stochastic nature of DCF analysis, in which FCFF is never certain. Using one single set of DCF inputs to determine EV is inadequate at least and erroneous at worst. This is similar to the dilemma of valuing mortgage-backed-securities (MBS), whose cash flow is uncertain due to the prepayment option of homeowners and its sensitivity to changes in the interest rate. In the DCF analysis, FCFF depends more on management's execution of the business plan, and the outcome can be probabilistic. Therefore, a probabilistic approach to the firm valuations is warranted. Indeed, competent analysts model the future as a set of possible outcomes and use probability distribution to quantify the likelihood of each scenario.

Similar to MBS valuation, we shall use Monte Carlo simulation to determine a distribution of EVs in a two-step process.

- Model inputs as random variables: Similar to a scenario analysis, parametric or nonparametric statistical techniques can be applied to determine the joint probability distribution of DCF inputs. In this section, we use a multivariate normal distribution.

- Monte Carlo simulation: We simulate DCF inputs based on their distribution and then derive an array of EV for all possible scenarios. Expected EV then becomes a probability weighted average. It is important to note that the expected EV no longer represents a particular scenario; instead it is an unbiased forecast incorporating all possible outcomes.

We will start with the sensitivity analysis that helps to identify important DCF inputs. Inputs with high sensitivity ought to be forecasted with more care. We then show how to conduct a multipath discounted cash flow (MDCF) analysis through Monte Carlo simulation. Finally, we construct a set of new valuation analytics incorporating statistical measures (to be viewed in conjunction with the valuation upside) and discuss their relevance to investment decision making. We shall continue to use CAKE as an example.

6.7.1 Sensitivity Analysis

The aim of sensitivity is to determine how much fair value changes given changes in the underlying inputs. For instance, for the Cheesecake Factory, Inc., an investment manager would ask, "Is CAKE an attractive investment if it were to deliver an 8% NOPAT margin instead of 7.3% (from the original forecast)? How sensitive is CAKE's valuation upside to different NOPAT margin inputs?"

Mathematically, if the valuation is a linear function of the input, we need to consider the first derivative (or slope) of valuation with respect to the input. On the other hand, if the function is nonlinear, we also need to at least consider the second derivative (or curvature). This is entirely analogous to the concept of duration/convexity in bond analysis and delta/gamma in option analysis. We shall in fact use delta/gamma for the first and second derivatives.

Use x to represent a particular DCF input and U to represent the corresponding valuation upside. Suppose x_0 is the base case for the DCF input, and $U(x_0)$ is the valuation upside. We can then vary the DCF input by $\pm\ \Delta x$ and compute the resulting valuation upside $U(x_0 \pm \Delta x)$. Then, the two sensitivity measures are

$$\begin{aligned} \text{delta} &= \frac{U(x_0+\Delta x)-U(x_0-\Delta x)}{2\Delta x} \\ \text{gamma} &= \frac{U(x_0+\Delta x)+U(x_0-\Delta x)-2U(x_0)}{(\Delta x)^2} \end{aligned} \tag{6.17}$$

In term of graphical interpretations, delta measures the slope of the tangency line passing through the base case, and gamma depicts the curvature. A positive delta indicates that the tangency line is upward sloping; alternatively, it means that valuation upside goes up as the DCF input

x increases. A positive gamma indicates a convex curve, and a negative gamma indicates a concave curve. A convex curve is more beneficial to investors when compared to a concave curve. When a curve is concave, the magnitude of the change in upside is greater when the input value goes up than when it goes down.

6.7.2 CAKE as an Example

In the preceding section, we discussed the base case of CAKE's DCF analysis. Figure 6.13 shows a graphical illustration of CAKE's sensitivity analysis. Panel A contains inputs variables related to profitability; panel B and panel C relate to scalability and WACC, respectively. Among all inputs, valuation upside is most sensitive to changes in WACC, followed by EBITDA, depreciation, and growth rate. CAKE's valuation outcome is least sensitive to changes in the tax rate, working capital, and ICAPEX. In terms of the curvature, WACC is again the most pronounced one.

Figure 6.14 shows delta, gamma, and valuation upsides of CAKE under different scenarios. As expected, the WACC's delta is the largest followed by EBITDA, depreciation, and growth — confirming previous graphical observation. Deltas of incremental capital expenditures, working capital change, and tax rate are relatively small and inconsequential. For example,

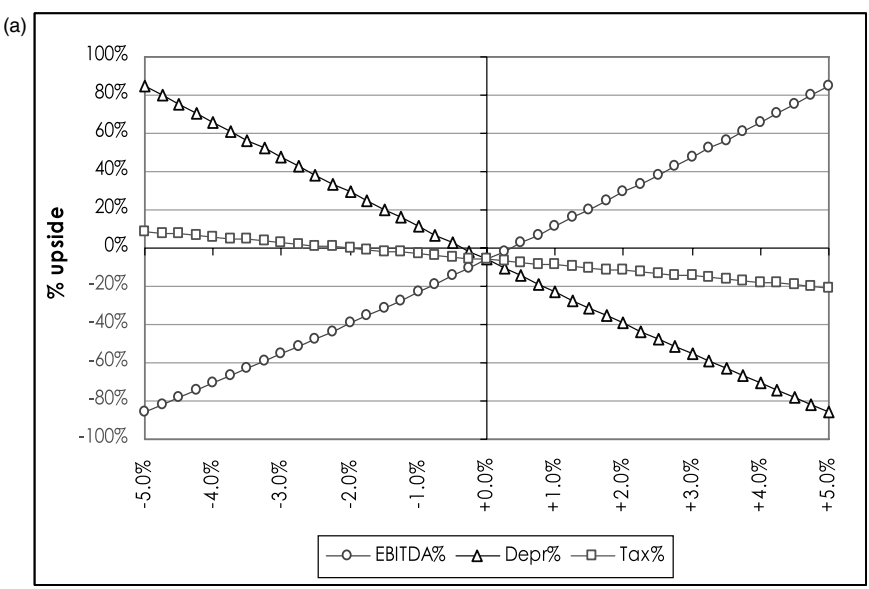
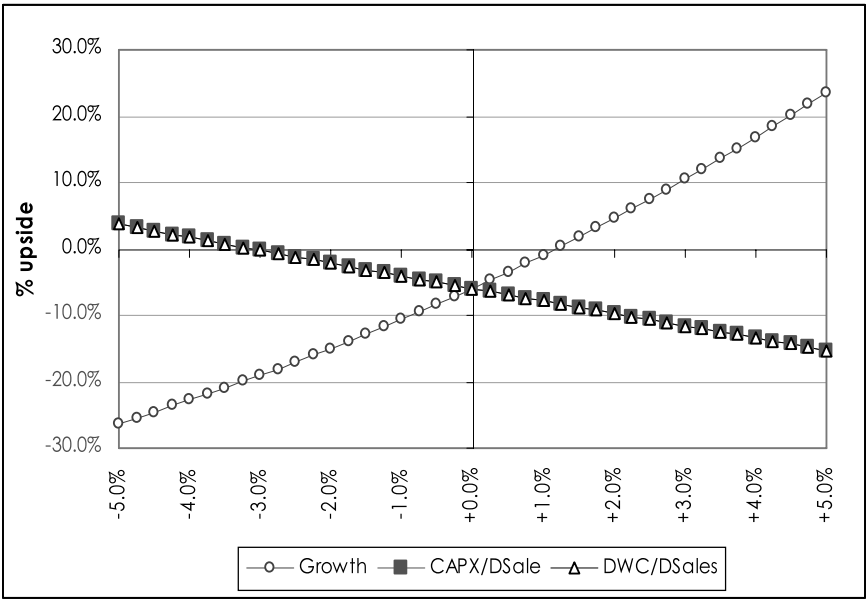


FIGURE 6.13. Sensitivity of DCF inputs: (a) profitability ratios, (b) scalability ratios, and (c) WACC.

(b)



(c)

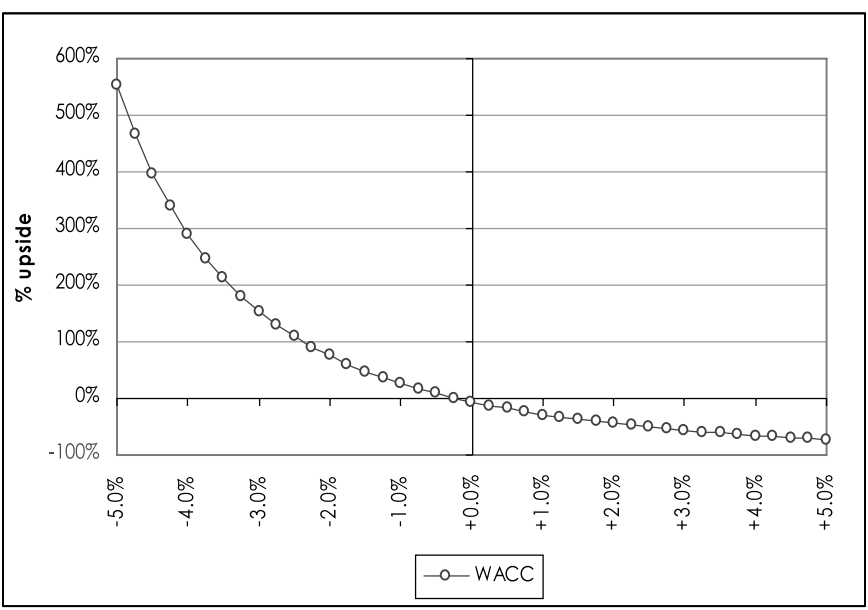


FIGURE 6.13. (continued)

	Sensitivity		% upside				
	Delta	Gamma	+2.0%	+1.0%	0.0%	-1.0%	-2.0%
EBITDA%	17.2	42.4	29.2%	11.4%	-6.0%	-23.0%	-39.5%
Depr%	-17.2	42.4	-39.5%	-23.0%	-6.0%	11.4%	29.2%
Tax%	-2.9	1.2	-11.8%	-8.9%	-6.0%	-3.0%	-0.1%
Growth	4.9	37.3	4.5%	-0.9%	-6.0%	-10.7%	-15.0%
ICAPX/ΔSales	-1.9	1.9	-9.7%	-7.9%	-6.0%	-4.1%	-2.2%
ΔWC/ΔSales	-1.9	1.9	-9.7%	-7.9%	-6.0%	-4.1%	-2.2%
WACC	-27.4	1073.9	-44.3%	-28.3%	-6.0%	26.4%	75.3%

FIGURE 6.14. Delta and gamma of DCF inputs.

a 1% change in WACC (i.e., from 8.4 to 9.4%) results in a 27% change in valuation upside (i.e., from −6% to −33%), whereas a 1% change in the tax rate induces only about a 3% change in upside. In other words, a change in the WACC is ten times more influential than a change in the tax rate of the same amount. Gammas for most inputs are inconsequential — meaning curves are fairly linear — except for WACC. It is also interesting to note that all gammas are positive.

Delta and gamma can be used to approximate the new valuation upside given a change in the input from the base case. This is a useful tool to gauge the upside of a new scenario without going through a full DCF analysis. Based on a Taylor expansion, we have

$$U\left(x_0+\Delta x\right)\approx U\left(x_0\right)+\text{delta}\cdot\Delta x+\frac{1}{2}\text{gamma}\cdot\left(\Delta x\right)^2. \tag{6.18}$$

6.8 MULTIPATH DCF ANALYSIS (MDCF)

The sensitivity analysis can test the robustness of the firm value evaluation. But it does not provide a distribution of possible outcomes. The MDCF approach provides that distribution by simulating DCF inputs according to an appropriate distribution and then computing corresponding firm values. As a result, MDCF not only properly gauges the expected firm valuation, or valuation upside when compared to the market value, but also provides a standard error estimate that can be used to ascertain the confidence of a particular DCF valuation.

Naturally, companies in high-growth, competitive industries, such as technology, would exhibit larger standard errors reflecting the uncertainly of these firms' future cash flows, when compared with firms in low-growth, stable industries, such as utilities. This difference can also

be said about firms that are more transparent in their reporting practice vs. those that are more opaque. For investment managers, quantitative and fundamental alike, an accurate standard error estimate is crucial to investment success, because portfolios should be formed on basis of both return and risk. This risk/return trade-off might be apparent to quantitative managers; it is not so for fundamental analysts, many of whom still use a single-path DCF approach and recommend the buy highest upside stocks, an action that subjects their portfolios to higher volatility due to greater forecast errors. For example, high valuation upside may be an artifact of high forecast error. In contrast, we advocate using standard error in conjunction with expected valuation upside to derive an error-adjusted upside that is better suited for active valuation investing.

6.8.1 Modeling DCF Inputs as Random Variables

We first model DCF inputs as random variables that are normally distributed, parameterized by both the mean and the covariance matrix. We continue to use CAKE as an example and model the EBITDA margin and growth rate as the only two random variables by holding all other inputs as constants. We select these two inputs because valuation upside is most sensitive to these two company-specific inputs, as shown in the previous section.

Panel A of Figure 6.15 shows CAKE's EBITDA margin and growth rate through time, including forward-looking IBES forecasts. A covariance matrix is modeled using an exponential weighting scheme, which puts more emphases on IBES forward information and less weight on the portion of history that are more distant from today. We choose a decay ratio of 15% to construct the covariance estimate as shown in Equation 6.18. To accommodate a reasonable starting point, we set u_0 and σ_0 to the equally weighted mean and standard deviation of the whole sample. Panel B of Figure 6.15 shows the covariance matrix estimate and the calculation follows Equation 6.20. α is the decay ratio, μ and σ are the mean and the standard deviation estimate for each time period, and $f_{i,t}$ is the observation of either growth rate or EBITDA margin at time t.

$$\begin{aligned}
\mu_{i,t} &= \alpha \cdot f_{i,t} + (1-\alpha) \cdot \mu_{i,t-1} \\
\sigma_{i,t}^2 &= \alpha \cdot (f_{i,t} - \mu_{i,t})^2 + (1-\alpha) \cdot \sigma_{i,t-1}^2 \\
\sigma_{ij,t} &= \alpha \cdot (f_{i,t} - \mu_{i,t})(f_{j,t} - \mu_{j,t}) + (1-\alpha) \cdot \sigma_{ij,t-1}
\end{aligned} \tag{6.19}$$

(a)

Date	Sales	EBITDA%	Growth	std(EBITDA)	std(Growth)	corr(EBITDA,Growth)
12/2007 (E)	1677.82	14.67%	19.95%	1.28%	4.72%	-66.75%
12/2006 (E)	1398.74	15.10%	18.37%	1.35%	4.90%	-64.33%
12/2005 (E)	1181.63	14.85%	21.91%	1.34%	4.75%	-56.75%
12/2004	969.23	13.70%	25.25%	1.34%	4.98%	-52.52%
12/2003	773.84	14.30%	18.69%	1.44%	5.40%	-52.65%
12/2002	651.97	14.28%	20.93%	1.48%	5.07%	-45.11%
12/2001	539.13	13.52%	23.01%	1.51%	4.90%	-35.04%
12/2000	438.28	13.65%	26.13%	1.60%	4.89%	-29.90%
12/1999	347.48	12.02%	31.02%	1.67%	5.18%	-25.79%
12/1998	265.22	9.90%	27.15%	1.81%	5.57%	-25.02%
12/1997	208.59	10.03%	30.12%	1.65%	5.99%	-38.51%
12/1996	160.31	12.16%	36.82%	1.32%	6.49%	-47.45%
12/1995	117.17	13.81%	36.89%	1.34%	6.27%	-37.36%
12/1994	85.59	14.66%	27.69%	1.45%	5.64%	-48.94%
12/1993	67.03	13.60%	-	1.50%	6.09%	-54.40%

(b)

	Growth	EBITDA%
Growth	0.002224	-0.000404
EBITDA%	-0.000404	0.000164

FIGURE 6.15. Stochastic modeling of DCF inputs: (a) time-series data and (b) covariance estimate.

Panel A reveals three interesting operating characteristics of CAKE's business.

- Negative correlation between margin and growth: CAKE's profitability is significantly negatively correlated with its growth rate. As CAKE's business started to mature, it delivered higher EBITDA margin with lower revenue growth. For example, between 1994 and 1997, CAKE's sales expanded at an annualized rate of 32% and delivered 12.7% EBITDA margin on average. In contrast, between 2005 and 2007, CAKE's sales growth is expected to slow down to 20.1% per annum with its EBITDA margin increasing to 14.9%.
- Growth rate is more volatile than EBITDA margin: This phenomenon is generally true for most firms. Company management has more control over the EBITDA margin, through the use of corporate budgeting process and internal expense control, than its sales growth, which has many exogenous influences such as consumer preference or the economy.
- Operating risk decreases as CAKE's business matures: The volatility of CAKE's EBITDA margin and growth rate has decreased significantly over its history. This phenomenon is also typically true for most successful firms.

(a)

EBITDA%	Growth Rate										
	8.98%	**11.34%**	**13.70%**	**16.06%**	**18.42%**	**20.77%**	**23.13%**	**25.49%**	**27.85%**	**30.21%**	**32.56%**
11.45%	-66%	-65%	-64%	-62%	-61%	-59%	-56%	-53%	-50%	-47%	-42%
12.09%	-62%	-60%	-58%	-55%	-52%	-48%	-44%	-39%	-33%	-26%	-18%
12.73%	-57%	-55%	-52%	-48%	-43%	-38%	-32%	-24%	-15%	-5%	7%
13.38%	-53%	-49%	-45%	-40%	-34%	-28%	-19%	-9%	3%	17%	34%
14.02%	-48%	-44%	-39%	-33%	-25%	-17%	-6%	6%	21%	39%	60%
14.66%	-44%	-39%	-32%	-25%	-16%	-6%	7%	22%	40%	62%	88%
15.30%	-39%	-33%	-26%	-17%	-7%	5%	20%	38%	59%	85%	115%
15.94%	-34%	-27%	-19%	-9%	2%	17%	34%	54%	79%	109%	144%
16.58%	-30%	-22%	-13%	-1%	12%	28%	48%	71%	99%	133%	173%
17.22%	-25%	-16%	-6%	7%	22%	40%	62%	88%	119%	157%	203%
17.86%	-20%	-11%	1%	15%	31%	52%	76%	105%	140%	182%	233%

(b)

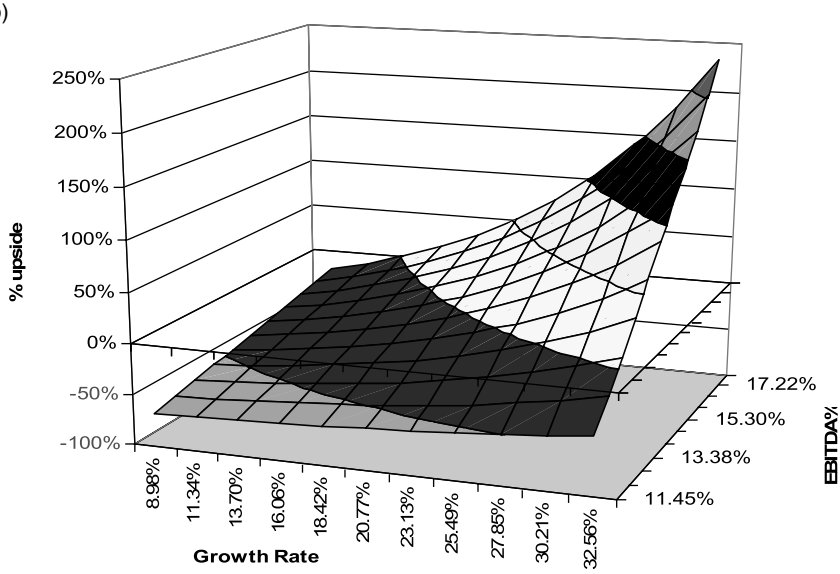


FIGURE 6.16. Monte Carlo simulation of valuation upside: (a) table and (b) graph.

6.8.2 Monte Carlo Simulation

In this illustration, Monte Carlo simulation is conducted by simultaneously varying both EBITDA margin and growth rate, creating 121 plausible scenarios. Figure 6.16 shows CAKE's valuation upsides under each scenario; Figure 6.17 shows the probability of each scenario according to the bivariate normal distribution. Starting from the base case highlighted by a gray-shaded background in Panel A of both exhibits, 11 possible values are selected for each DCF input by symmetrically increasing and decreasing the base case input by one half of a standard deviation each time. Panel A of Figure 6.16 tabulates valuation upsides derived from the 91 different combinations of EBITDA margin and growth rate, and Panel B of Figure 6.16 uses a surface graph to visually illustrate the changes in

(a)

EBITDA%	Growth Rate 8.70%	11.11%	13.53%	15.94%	18.36%	20.77%	23.19%	25.60%	28.02%	30.43%	32.85%
11.28%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.07%	0.18%	0.29%	0.30%	0.32%
11.96%	0.00%	0.00%	0.00%	0.01%	0.04%	0.15%	0.39%	0.67%	0.72%	0.50%	0.30%
12.63%	0.00%	0.00%	0.01%	0.05%	0.23%	0.71%	1.39%	1.75%	1.41%	0.72%	0.29%
13.31%	0.00%	0.01%	0.05%	0.27%	0.95%	2.16%	3.14%	2.93%	1.75%	0.67%	0.18%
13.98%	0.00%	0.04%	0.23%	0.95%	2.50%	4.21%	4.54%	3.14%	1.39%	0.39%	0.07%
14.66%	0.02%	0.15%	0.71%	2.16%	4.21%	5.26%	4.21%	2.16%	0.71%	0.15%	0.02%
15.33%	0.07%	0.39%	1.39%	3.14%	4.54%	4.21%	2.50%	0.95%	0.23%	0.04%	0.00%
16.01%	0.18%	0.67%	1.75%	2.93%	3.14%	2.16%	0.95%	0.27%	0.05%	0.01%	0.00%
16.68%	0.29%	0.72%	1.41%	1.75%	1.39%	0.71%	0.23%	0.05%	0.01%	0.00%	0.00%
17.36%	0.30%	0.50%	0.72%	0.67%	0.39%	0.15%	0.04%	0.01%	0.00%	0.00%	0.00%
18.03%	0.32%	0.30%	0.29%	0.18%	0.07%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%

(b)

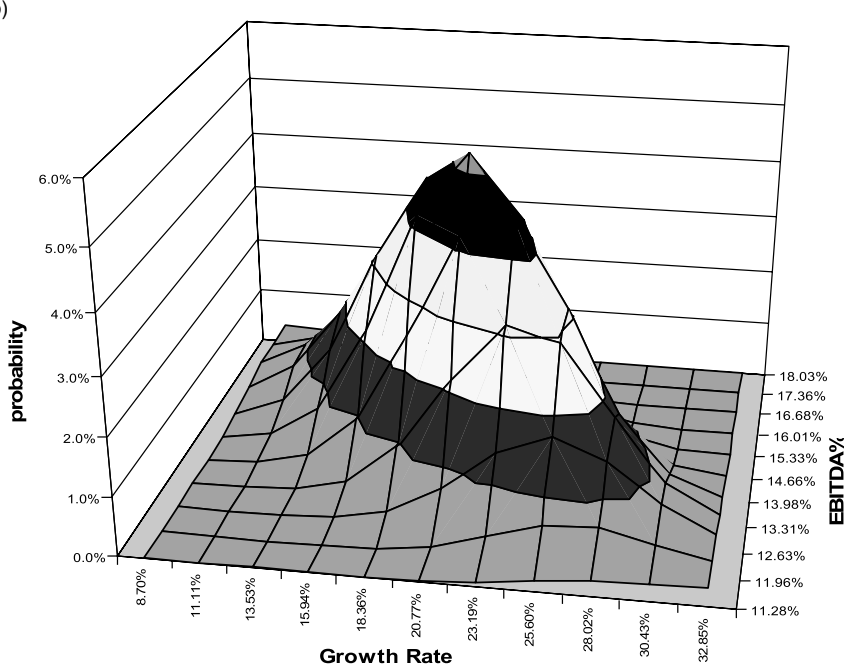


FIGURE 6.17. Probability distribution (bivariate normal): (a) table and (b) graph.

upside. Panel A of Figure 6.17 presents a discrete form of the bivariate normal probability distribution, and Panel B illustrates it graphically. Note the following:

- The base case scenario produces a negative 6% upside, which is the same as shown in Figure 6.12; the probability of the base case scenario is 5.3%, given the covariance estimate shown in Figure 6.15.
- The best scenario is when both the EBITDA margin and growth rate are the highest, delivering a 247% upside. Similarly, the worst case is

when both inputs are the lowest, producing a 68% downside. However, both scenarios are extremely unlikely to happen, and their probabilities are close to zero. The near-zero probability is due to not only extreme values of both inputs, but also to the negative correlation between the growth and margin. If the correlation were significantly positive, probabilities of these extreme cases would have been more likely. This highlights the importance of the correlation matrix in MDCF analysis, which further captures each firm's unique competitive environment by incorporating the dynamics among DCF inputs.

Figure 6.18 graphically displays other interesting DCF analytics across all likely scenarios. As shown in Panel A, CAKE needs to borrow cash to finance its growth and its FCF margin would turn positive when it were to slow down revenue expansion and maintain higher EBITDA margin. Panel B reveals that CAKE's economic value creation is directly linked to the level of EBITDA margin. This is somewhat artificial by construction, as we hold scalability a constant in this set of Monte Carlo simulations. Interested readers can include scalability as an additional random variable in the construction of simulated scenarios. Lastly, the amount of operating value, coming from growth opportunities, is jointly determined by both the EBITDA margin and growth rate. It is the highest when both inputs are at their peaks.

6.8.3 Analytical Results of MDCF

MDCF provides a new set of analytics that are better suited for active security selection by incorporating forecast errors. For example, instead of investing in stocks with positive expected valuation upside, active managers should select underpriced stocks with small standard deviations of upside. Similarly, active managers should underweight overvalued stocks with small forecast errors. This suggests a ratio of expected upside to the standard deviation as an alternative value measure.

The following formulas show the construction of MDCF analytics associated with valuation upside, denoted by U.

$$\bar{U} = \sum p_i \times U_i, \qquad \text{std}(U) = \sqrt{\sum p_i \times (U_i - \bar{U})^2}$$
$$t(U) = \frac{\bar{U}}{\text{std}(U)}, \qquad \text{prob}(U > 0) = \sum_{U_i > 0} p_i \tag{6.20}$$

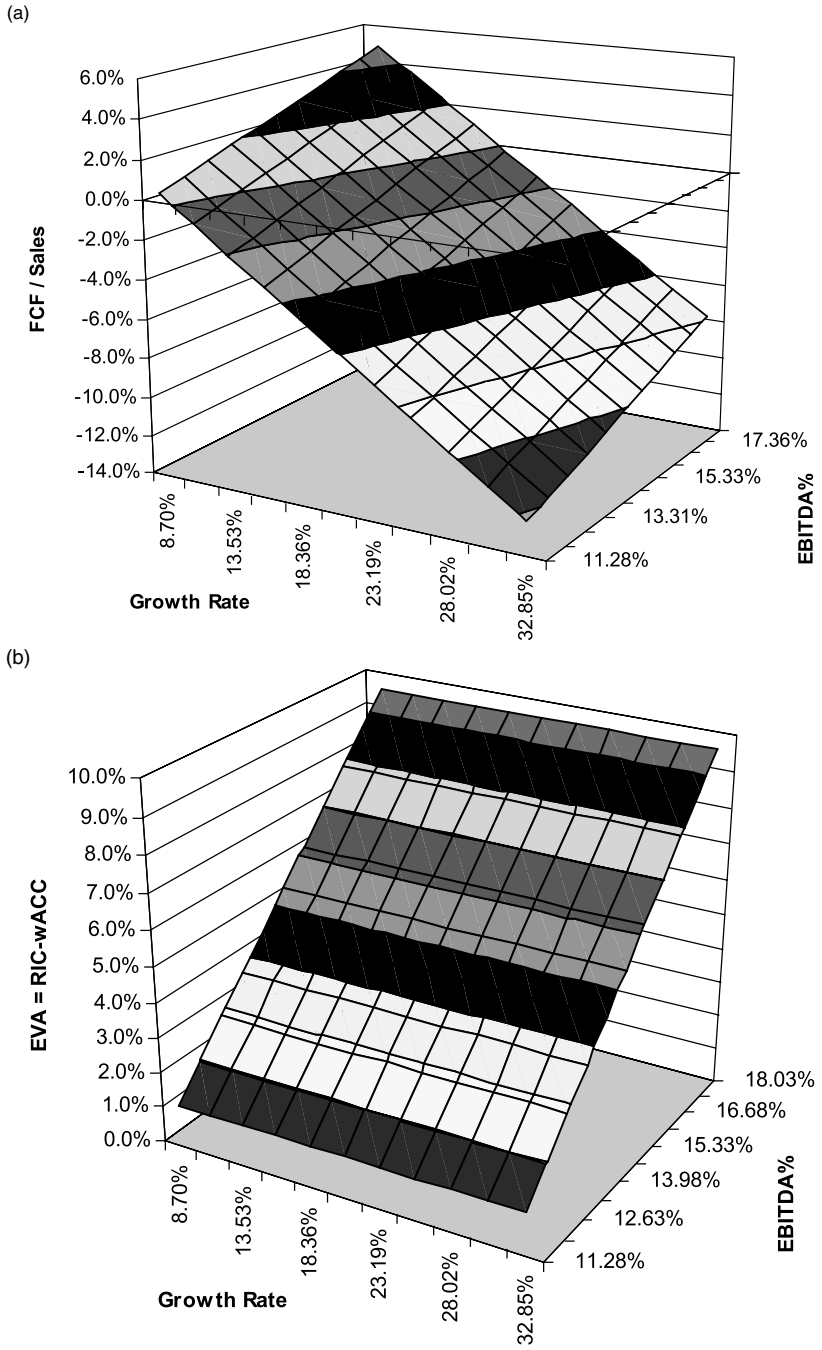


FIGURE 6.18. Other DCF analytics: (a) FY1 free cash flow margin (FCFF/Sales), (b) FY1 economic value added (EVA = RIC – WACC), and (c) percentage of operating value from growth.

(c)

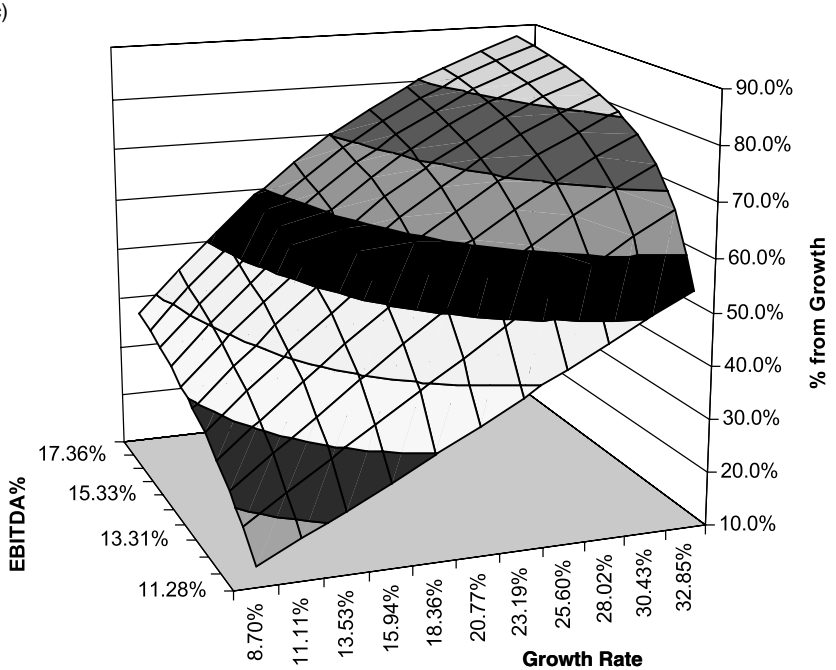


FIGURE 6.18. (continued)

Figure 6.19 shows these measures for CAKE, along with other statistics. Comparing the expected valuation downside (–7%) with the forecast error of 19.6%, CAKE's overpricing is not significant with a t-stat of –0.34. That is, CAKE's valuation could easily become an upside, should its business fundamental improve from the current forecast.

More observations can be obtained from the MDCF analysis. First, CAKE is likely to engage in external financing in FY1 in order to sustain its sales expansion. The probability of having enough internally generated cash in FY1 is only 12.8%. Second, without a doubt, CAKE creates positive shareholder value (where RIC > WACC) — a quality company that is expected to generate excess returns for its shareholders. The probability of having a positive value creation is 99.9% — near certainty! Lastly, these statistics reconfirms that future growth opportunity plays an important role in determining CAKE's operating value. The expected percentage of operating value coming from growth is 59.6%; about 88.4% of the time growth will account for more than half of CAKE's operating value.

Based on the aforementioned analysis, CAKE is a high-quality, growing firm, which derives much of its firm value from growth opportunities.

Percent Upside	
E(% Upside):	-7%
STD(% Upside)	19.6%
t(% Upside)	-0.34
Upside Probability	33.8%
Downside Probability	66.0%
FCF Margin (FCFF/Sales)	
E(FCFF/Sales)	-3.7%
Prob(FCFF/Sales > 0)	12.8%
Economic Value Added	
E(EVA)	5.3%
Prob(EVA > 0)	99.9%
Operating Value	
E(% from Existing Bus):	37.4%
E(% from Growth):	59.6%
Prob(% from Existing Bus > 50%):	88.4%

FIGURE 6.19. Multipath DCF analytics.

CAKE would require external financing in FY1 and beyond, in order to sustain its business expansion. It is currently slightly overpriced. However, if its business economics remain strong, this overpricing could quickly turn into underpricing. As such, CAKE's investment appeal should not be rejected simply based on the current overpricing alone.

6.9 SUMMARY

Discovering attractive investment opportunities takes two different forms — one stemming from arbitraging behavioral inefficiencies and the other built on rational economic forecast. Valuation techniques belong to the latter and model a firm's intrinsic value based on many normative assumptions: rationality, perpetuity, going concern, mean reversion, or the validity of CAPM. Valuation analysis is a technique that helps active managers to better understand the business economics of a firm from the following perspectives.

- What is the business model and what are the competitive advantages?
- What are the set of value drivers and how does competition affect them?
- How sensitive is each DCF input and how does a change in each input affects valuation outcome?
- What is the standard error of valuation upside and what is the statistical confidence of having a positive upside?

Although a one-path, one-life DCF analysis provides an estimation of the firm value, it is inadequate, often reflecting overconfident and possible erroneous belief of a single analyst. Instead, the multipath discounted cash flow (MDCF) analysis should be used to properly account for other plausible scenarios and their probabilities. The distribution of upside estimation from such analysis should provide more robust information for active managers.

PROBLEMS

6.1 Derive formula in Equation 6.2 with the following assumptions: (1) WACC is the discount rate, (2) g is the perpetual growth rate of FCF, and (3) $FCFF_0$ is the free cash flow to the firm at year 0.

6.2 Given Equation 6.4, show that the change in the ratio of value from growth opportunities to the total operating value is given by

$$\Delta\left(\frac{\text{growth}}{\text{OV}}\right)=\left(1+\frac{1}{\text{WACC}}\right)\frac{\Delta g}{(1+g)^2}$$
$$\Delta\left(\frac{\text{growth}}{\text{OV}}\right)=-\frac{g}{1+g}\,\frac{\Delta(\text{WACC})}{(\text{WACC})^2} \tag{6.21}$$

6.3 Prove that the book value equals the present value of future cash flows when discount rate equals expected rate of return on investment.

6.4 Derive the firm operating value of (6.13).

6.5 One way of estimating required capital expenditure is to correlate historical capital expenditures (CAPEX_t) with next year's sales increase (ΔSales_{t+1}) directly. However, the stability of such direct estimation of CAPEX/ΔSales is poor, because ΔSales is typically volatile through time. Alternatively, it can be estimated as follows. Derive the formula below:

$$E\left(\frac{\text{CAPEX}}{\Delta\text{Sales}}\right)=E\left(\frac{\text{DA}}{\text{Sales}}\right)\times g^{-1}+E\left(\frac{\text{nPPE}}{\text{Sales}}\right),$$

where DA is depreciation and amortization, g is the growth rate of sales, and nPPE is net property, plant, and equipment.

6.6 Repeat MDCF analysis of the Cheesecake Factory, Inc., and include scalability ratio as an additional random variable.

REFERENCES

Porter, M.E., *Competitive Strategy: Creating and Sustaining Superior Performance*, The Free Press, New York, 1985.

ENDNOTES

1. The assumption of no growth simplifies the computation of terminal values. Should one assume that a firm grows at the risk-free rate perpetually at the terminal period, one also needs to estimate the scalability ratio in the terminal period to compute the expected reinvestment rate each year.

CHAPTER 7

Multifactor Alpha Models

In Chapter 4 (see also Qian & Hua 2004), we presented an analytic framework to evaluate individual alpha factors based on the risk-adjusted information coefficient (IC). The ratio of average IC to the standard deviation of IC serves as a proxy for the information ratio (IR) of active strategies that employ the alpha factors. We then devoted the next two chapters to the examination of several alpha factors on an individual basis. In practice, alpha models almost always employ multiple factors instead of a single one. So then, the question naturally arises: how to blend these factors optimally into a composite alpha model? The combination of these factors is not restricted to quantitative factors. For instance, some investment firms conduct both fundamental and quantitative researches. How to combine them into a single forecasting process, in terms of ranking or scores, presents a similar challenge.

In this chapter, we extend the analytic framework to derive factor weights in a multifactor alpha model. Our objective is to maximize the IR of the multifactor model. The approach is similar to a mean–variance optimization. The difference is that we now replace a portfolio of stocks with a portfolio of factors. Thus, average IC and standard deviation of IC resemble the expected return and risk of dollar neutral, risk-neutral factor portfolios. In addition, correlations between ICs of different factor portfolios also play an essential role in delivering the diversification benefits. It is important to note that the correlation between ICs is not the same as the correlation between factor scores. The former is the correlation of returns

to factor portfolios across time, whereas the latter is the cross-sectional correlation of factor scores at a given time. We will show that the correlations among ICs play a crucial role in determining the optimal alpha model weights, whereas correlations among factor scores play a secondary role. Theoretically, it is tempting to assume that the two are identical, but empirical evidence seems to prove the contrary.

This chapter consists of four sections. In the first section, we derive the analytical expression of the composite IC of a multifactor alpha model for a single period. We define a *multifactor model* as one that linearly combines scores of individual alpha factors to create a composite forecast (i.e., a composite score), and a composite IC is the IC of the composite score. The efficacy (or the expected performance) of a multifactor alpha model becomes the IR of its single-period ICs through time. A similar approach is illustrated in Chapter 4. In the second section, the analytical expression of a composite IR is derived with the assumption that cross-sectional factor-score correlations do not change over time. This time invariant assumption makes analytical derivations tractable, so we can solve for the optimal model weighting that achieves the highest IR of the composite forecast. In the third section, we discuss the important difference between cross-sectional factor score correlation and time-series IC correlation in the context of multifactor model building. We also suggest a practical procedure to deal with the time variability of factor-score correlations. In the last section, we examine the statistical linkage between our model optimization framework and the Fama–MacBeth regression procedure. Specifically, we provide cautionary notes to practitioners who would like to apply a Fama–MacBeth-like regression framework to derive optimal model weights.

7.1 SINGLE-PERIOD COMPOSITE IC OF A MULTIFACTOR MODEL

As in Chapter 4, we will first consider a single-period excess return of a multifactor model, which is a linear combination of M factors $\left(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_M\right)$ with the weight vector $\mathbf{v} = \left(v_1, v_2, \cdots, v_M\right)'$. The weight vector, once selected, shall remain constant over time. To put it differently, we are solving for the optimal weighting of a constant linear multifactor model. There are more complex alpha models that could be nonlinear and/or dynamic. We shall cover them in later chapters.

To link model performance to realistic portfolio implementation, we assume all factors are risk-adjusted according to the analytical framework

illustrated in Chapter 4. Therefore, the composite risk-adjusted factor is a linear combination:

$$\mathbf{F}_c = \sum_{i=1}^{M} v_i \mathbf{F}_i \,. \tag{7.1}$$

The composite will also be risk-adjusted in the sense that the associated active portfolio will be neutral to all risk factors and is mean–variance optimal. Now we treat the composite factor $\mathbf{F}_c$ as a single factor and use the analytic framework presented in Chapter 4.

Recall from Chapter 4 that the single-period excess return of an alpha factor is expressed as a function of the covariance between the factor and the risk-adjusted return. To clarify the notation, $\mathbf{F}_{c,t}$ represents the risk-adjusted composite factor available at the beginning of period t, whereas $\mathbf{R}_t$ is the risk-adjusted return during period t.

$$\begin{aligned} \alpha_t &= \frac{(N-1)}{\lambda_t} \operatorname{cov}\left(\mathbf{F}_{c,t}, \mathbf{R}_t\right) \\ &= \frac{(N-1)}{\lambda_t} \operatorname{corr}\left(\mathbf{F}_{c,t}, \mathbf{R}_t\right) \operatorname{dis}\left(\mathbf{F}_{c,t}\right) \operatorname{dis}(\mathbf{R}_t) \end{aligned} \tag{7.2}$$

The covariance between the composite factor and the risk-adjusted return is a linear combination of covariances between individual factors and the risk-adjusted return:

$$\begin{aligned} \operatorname{cov}\left(\mathbf{F}_{c,t}, \mathbf{R}_t\right) &= \operatorname{cov}\left(\sum_{\mathrm{i}=1}^{M} v_i \mathbf{F}_{i.t}, \mathbf{R}_t\right) = \sum_{\mathrm{i}=1}^{M} v_i \operatorname{cov}\left(\mathbf{F}_{i.t}, \mathbf{R}_t\right) \\ &= \left[\sum_{\mathrm{i}=1}^{M} v_i IC_{i,t} \operatorname{dis}\left(\mathbf{F}_{i,t}\right)\right] \operatorname{dis}(\mathbf{R}_t) \end{aligned} \tag{7.3}$$

In the second line of the preceding equation, we have expressed the covariances in terms of ICs and dispersions. Also recall from Chapter 4 that the risk-aversion parameter is calibrated such that the active portfolio would have a targeted tracking error. The relationship in the case of a composite alpha factor is

$$\lambda_t = \frac{\sqrt{N-1}\,\mathrm{dis}\left(\mathbf{F}_{c,t}\right)}{\sigma_{\mathrm{model}}}. \tag{7.4}$$

The dispersion of the composite factor depends on the model weights and cross-sectional covariances among different factor scores. Denoting the cross-sectional covariance between two factors by $\phi_{ij,t} = \mathrm{cov}\left(\mathbf{F}_{i,t}, \mathbf{F}_{j,t}\right)$ and the factor covariance matrix by $\mathbf{\Phi}_t = \left(\phi_{ij,t}\right)_{i,j=1}^{M}$, the dispersion of the composite is given by

$$\mathrm{dis}\left(\mathbf{F}_{c,t}\right) = \sqrt{\mathbf{v}'\mathbf{\Phi}_t\mathbf{v}}\,. \tag{7.5}$$

Substituting Equation 7.5, Equation 7.4, and Equation 7.3 into Equation 7.2 yields

$$\alpha_t = IC_{c,t}\sqrt{N-1}\,\sigma_{\mathrm{model}}\mathrm{dis}(\mathbf{R}_t)\,. \tag{7.6}$$

Further,

$$IC_{c,t} = \mathrm{corr}\left(\mathbf{F}_{c,t}, \mathbf{R}_t\right) = \frac{\sum_{\mathrm{i}=1}^{M} v_i IC_{i,t}\mathrm{dis}\left(\mathbf{F}_{i,t}\right)}{\sqrt{\mathbf{v}'\mathbf{\Phi}_t\mathbf{v}}}. \tag{7.7}$$

Equation 7.6 provides the excess return of a multifactor alpha model. It is essentially of the same form as in the single-factor case, except that the IC is that of a composite factor given in (7.7) instead of a single one. The composite IC is a linear combination of individual factor ICs, and the weights are factor weight v_i times the ratio of individual factor dispersion to composite factor dispersion. Among the four terms in (7.6), the number of stocks, the target tracking error, and the dispersion of risk-adjusted returns have either little or no time-series variation, so we shall assume that they are constant throughout the remainder of the chapter. The composite IC, on the other hand, has many time-varying components, including the ICs of the underlying alpha factors $IC_{i,t}$, their cross-sectional dispersions $\mathrm{dis}\left(\mathbf{F}_{i,t}\right)$, and their covariance matrix $\mathbf{\Phi}_t$.

Example 7.1

Suppose we have two factors $\mathbf{F}_1$ and $\mathbf{F}_2$. In a given period, we have $\text{dis}(\mathbf{F}_1)=1$ and $\text{dis}(\mathbf{F}_2)=0.5$, and the factor correlation is 0.5. Then the factor covariance matrix is

$$\mathbf{\Phi}=\begin{pmatrix} 1 & 0.5\cdot 1\cdot 0.5 \\ 0.5\cdot 1\cdot 0.5 & 0.5^2 \end{pmatrix}=\begin{pmatrix} 1 & 0.25 \\ 0.25 & 0.25 \end{pmatrix}.$$

Suppose we equally weight these two factors; the dispersion of the composite factor is

$$\text{dis}(\mathbf{F}_c)=\sqrt{\mathbf{v}'\mathbf{\Phi}\mathbf{v}}=\left[\begin{pmatrix} 0.5 & 0.5 \end{pmatrix}\begin{pmatrix} 1 & 0.25 \\ 0.25 & 0.25 \end{pmatrix}\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}\right]^{1/2}$$

$$=\sqrt{0.5^2+0.25\cdot 0.5^2+2\cdot 0.25\cdot 0.5^2}=0.66$$

Example 7.2

Suppose that, in the given period, the ICs of factor 1 and factor 2 are 0.15 and 0.20, respectively. Then the IC of the composite factor is

$$IC_c=\frac{\sum_{i=1}^{M} v_i IC_{i,t}\text{dis}(\mathbf{F}_{i,t})}{\sqrt{\mathbf{v}'\mathbf{\Phi}_t\mathbf{v}}}=\frac{0.5\cdot 0.15\cdot 1+0.5\cdot 0.20\cdot 0.5}{0.66}=0.11+0.08=0.19.$$

In this case, the composite IC is greater than the IC of factor 1 but less than that of factor 2.

The previous examples illustrate the relationship between the composite IC and individual ICs for a single period. The major purpose of optimal alpha modeling is to maximize the IR over multiple periods, which depends not only on the average IC but also on the standard deviation of IC. It seems highly unlikely that there exists a full analytic solution for the weight vector $\mathbf{v}$ that maximizes the IR based on (7.6) because $\mathbf{v}$ appears in a quadratic form in the denominator. There are several possible approaches to solving this problem. One involves analytical approximation, and another involves transformation of alpha factors into orthogonal factors. We shall start with analytical approximation by assuming the factor correlation to be constant through time. Factor orthogonalization and factor-score correlations are also discussed in the second half of this chapter.

7.2 OPTIMAL ALPHA MODEL: AN ANALYTICAL DERIVATION

In this section, we derive an analytical expression of the optimal model weighting that achieves the highest information ratio, under the assumption that the factor covariance matrix stays unchanged over time. We first explore how factor standardization affects the IC of a composite factor. Then, the analytical expression of IR is derived for a composite multifactor alpha model, linking the composite IR to the time-series of ICs of each individual alpha factor. Based on this expression of composite IR, we solve analytically for the optimal model weighting that achieves the highest composite IR. In this derivation, we assume that model weighting is also time invariant. Lastly, we provide a brief discussion of why maximizing the single-period IC of a composite model does not achieve optimality.

7.2.1 Factor Standardization

If we assume that the factor covariance matrix is time invariant, the composite IC becomes a constant linear combination of model weights and individual ICs. To simplify things further, we standardize all individual factors such that their dispersion is always unity over time, i.e., $\text{dis}\left(\mathbf{F}_{i,t}\right)=1$, for all i,t. It is common to standardize all factors in practice, and there are several potential benefits for doing so. First, it "equalizes" the contribution of individual factors to the overall model for a given set of model weights. Second, it immunizes the composite model from changes in the dispersions of the factors, thus reducing portfolio turnovers associated with such changes. More importantly, there is little direct empirical evidence indicating that such turnover adds value. Note the following:

- Standardizing individual factors before combining them into an alpha model amounts to rescaling the model weights putting factors in the same units for comparison. Moreover, as the dispersions of factors change over time, the rescaling weights are also time varying. In other words, standardizing factors actually leads to implicit time-varying alpha models.

Example 7.3

We will standardize factor 2 in Example 7.1, whose original dispersion for the given period is 0.5, by multiplying it by 2. The first factor is already standardized. Suppose we still equally weight the two standardized factors; the effective weights on the original factors are 1/3 and 2/3. Suppose

also that during the next period, the dispersion of factor 1 changes to 0.5, whereas the dispersion of factor 2 changes to 1. We would standardize the factor 1 by doubling it while leaving factor 2 untouched. In this period, an equally weighted model of the standardized factor would imply an effective weight of 2/3 and 1/3 on the original factors.

With factor standardization, the composite IC for time t is

$$IC_{c,t} = \frac{1}{\sqrt{\mathbf{v}'\mathbf{\Phi}\mathbf{v}}} \sum_{i=1}^{M} v_i IC_{i,t} = \frac{1}{\tau} \sum_{i=1}^{M} v_i IC_{i,t} \,. \tag{7.8}$$

The covariance matrix $\mathbf{\Phi}$ reduces to the correlation matrix of factors because all factors are standardized. The composite the IC can be seen as a linear combination of the ICs of the underlying factors scaled by a constant τ, which is the dispersion of the composite factor (7.5). Another important feature of Equation 7.8 is that the composite IC remains unchanged if the factor weights are all scaled by the same constant.

7.2.2 IR of the Composite IC

We now calculate the expected IC and the standard deviation of IC to obtain the IR. We start with a two-factor example.

Example 7.4

If there are two factors, then we have

$$IC_{c,t} = \frac{1}{\sqrt{v_1^2 + v_2^2 + 2v_1 v_2 \rho_{12}}} \left(v_1 IC_{1,t} + v_2 IC_{2,t} \right) = \frac{1}{\tau} \left(v_1 IC_{1,t} + v_2 IC_{2,t} \right). \tag{7.9}$$

The correlation between the two factors is ρ_{12}, which, for the moment, is assumed to be constant over time. The expected composite IC is a linear combination of individual ICs is

$$\overline{IC_c} = \frac{1}{\tau} \left(v_1 \overline{IC_1} + v_2 \overline{IC_2} \right), \tag{7.10}$$

and the standard deviation of the IC is

$$\begin{aligned} \mathrm{std}\left(IC_c\right) &= \frac{1}{\tau} \mathrm{std}\left(v_1 IC_{1,t} + v_2 IC_{2,t} \right) \\ &= \frac{1}{\tau} \sqrt{v_1^2 \sigma_{IC_1}^2 + v_2^2 \sigma_{IC_2}^2 + 2 v_1 v_2 \rho_{12,IC} \sigma_{IC_1} \sigma_{IC_2}} \end{aligned} \tag{7.11}$$

The IC correlation between the two factors is denoted by $\rho_{12,IC}$, and the standard deviations of ICs are σ_{IC_1} and σ_{IC_2}. The IR, in this case the ratio of average IC to the standard deviation of IC, is

$$IR_c = \frac{\left(v_1 \overline{IC}_1 + v_2 \overline{IC}_2\right)}{\sqrt{v_1^2 \sigma_{IC_1}^2 + v_2^2 \sigma_{IC_2}^2 + 2 v_1 v_2 \rho_{12,IC} \sigma_{IC_1} \sigma_{IC_2}}}. \tag{7.12}$$

For a general model with M factors, we can denote the average IC by a vector $\overline{\mathbf{IC}} = \left(\overline{IC}_1, \overline{IC}_2, \cdots, \overline{IC}_M\right)'$, and the IC covariances by matrix $\mathbf{\Sigma}_{IC} = \left(\rho_{ij,IC}\right)_{i,j=1}^{M}$. Then the average and standard deviation of a composite IC are

$$\begin{aligned} \overline{IC}_c &= \frac{1}{\tau} \sum_{i=1}^{M} v_i \overline{IC}_i = \frac{1}{\tau} \mathbf{v}' \cdot \overline{\mathbf{IC}} \\ \operatorname{std}\left(IC_c\right) &= \frac{1}{\tau} \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} v_i v_j \rho_{ij,IC} \sigma_{IC_i} \sigma_{IC_i}} = \frac{1}{\tau} \sqrt{\mathbf{v}' \cdot \mathbf{\Sigma}_{IC} \cdot \mathbf{v}} \end{aligned} \tag{7.13}$$

and the IR is

$$IR_c = \frac{\sum_{i=1}^{M} v_i \overline{IC}_i}{\sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} v_i v_j \rho_{ij,IC} \sigma_{IC_i} \sigma_{IC_i}}} = \frac{\mathbf{v}' \cdot \overline{\mathbf{IC}}}{\sqrt{\mathbf{v}' \cdot \mathbf{\Sigma}_{IC} \cdot \mathbf{v}}}. \tag{7.14}$$

- The scale constant τ — the dispersion of the composite factor, which depends on cross-sectional factor-score correlations — has completely dropped out of the IR equation. However, the time-series IC correlations remain, and the IC correlation matrix determines the standard deviation of composite IC over time, and thus its active risk.

7.2.3 Optimal Model Weights

We can now find the optimal model weights that maximize the IR (7.14) of the composite alpha factor. We note that IR in (7.14) assumes that the

cross-sectional factor-score correlation matrix is a constant through time. As we can see, although the IR optimization problem is similar to mean–variance optimization, there are important differences. The objective function is the mean/standard deviation ratio, and there is no risk-aversion parameter. As a result, any constant multiple of optimal weights will also be optimal because they give rise to the same IR. In theory, there is no need for the weight to sum up to 100%. However, in practice, we often do so customarily.

This is an unconstrained optimization. Taking the partial derivative of (7.14) with respect to the weights yields

$$\frac{\partial\left(IR_c\right)}{\partial \mathbf{v}} = \frac{\overline{\mathbf{IC}}}{\sqrt{\mathbf{v}' \cdot \mathbf{\Sigma}_{IC} \cdot \mathbf{v}}} - \frac{\left(\mathbf{v}' \cdot \overline{\mathbf{IC}}\right)\mathbf{\Sigma}_{IC} \cdot \mathbf{v}}{\left(\mathbf{v}' \cdot \mathbf{\Sigma}_{IC} \cdot \mathbf{v}\right)^{3/2}} . \tag{7.15}$$

Equating the partial derivatives to zero, we have

$$\left(\mathbf{v}' \cdot \mathbf{\Sigma}_{IC} \cdot \mathbf{v}\right)\overline{\mathbf{IC}} = \left(\mathbf{v}' \cdot \overline{\mathbf{IC}}\right)\mathbf{\Sigma}_{IC} \cdot \mathbf{v} . \tag{7.16}$$

The solution for the optimal weights is

$$\mathbf{v}^* = s\mathbf{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}} , \tag{7.17}$$

where s is an arbitrary, generally positive constant. We can select s such that the sum of its optimal weights is 1. Substituting the optimal weights into (7.14) gives the optimal IR:

$$IR^* = \sqrt{\overline{\mathbf{IC}}' \cdot \mathbf{\Sigma}_{IC}^{-1} \cdot \overline{\mathbf{IC}}} . \tag{7.18}$$

- The optimal weight (7.17) is akin to the mean–variance solution for the optimal portfolio of securities including cash. It is identical to the solution of optimal manager selections for investment consultants, where the "managers" in this case are alpha factors. This indicates that the weight of an alpha factor in the composite depends not only on its own risk/return trade-off but also on its IC correlation with other factors' ICs.
- The optimal weight $\mathbf{v}^*$ can also be derived from an OLS regression without an intercept term. Britten-Jones (1998) shows that mean–variance (MV) optimal weights in general can be obtained this way.

One of the benefits of this alternative approach is that we can obtain standard errors for the optimal weights. We leave the proof as an exercise (see Problem 7.4).

Example 7.5

We illustrate the optimal model weights in a two-factor case in which

$$\begin{aligned} v_1 &= \frac{s}{1-\rho_{12,IC}^2}\left(\frac{\overline{IC}_1}{\sigma_{IC_1}^2} - \frac{\rho_{12,IC}\overline{IC}_2}{\sigma_{IC_1}\sigma_{IC_2}}\right) \\ v_2 &= \frac{s}{1-\rho_{12,IC}^2}\left(\frac{\overline{IC}_2}{\sigma_{IC_2}^2} - \frac{\rho_{12,IC}\overline{IC}_1}{\sigma_{IC_1}\sigma_{IC_2}}\right) \end{aligned} . \tag{7.19}$$

Equation 7.19 states that the optimal weight of a factor is determined by two terms. The first term is the ratio of the average IC to the variance of IC. The second term, carrying a negative sign, is proportional to the IC correlation and the average IC of the other factor. Therefore, if a factor has high IC correlations with other factors, then its model weight will be negatively affected. On the other hand, if a factor has low and/or negative IC correlations with other factors, its model weight will be positively affected.

For a model with two factors, the optimal IR can also be explicitly written as

$$IR^* = \frac{\sqrt{IR_1^2 + IR_2^2 - 2\rho_{12,IC}IR_1IR_2}}{\sqrt{1-\rho_{12,IC}^2}} . \tag{7.20}$$

For two factors with given IRs, the optimal IR will be higher if their IC correlation is lower. Figure 7.1 plots the optimal IR as a function of IC correlation for given values of two individual IRs. The two IRs are 1.0 and 0.5, respectively. As the IC correlation changes from −0.5 to 0.5, the optimal IR declines from 1.5 to 1.0. When the IC correlation is at −0.5, there are strong diversification benefits between the two factors, and the combined optimal IR is much higher than both individual IRs. However, as the IC correlation increases, the diversification benefit shrinks. When it reaches 0.5 and above, the benefit disappears entirely unless one is willing to bet against one of the factors (see Problem 7.6), i.e., when the optimal weight becomes negative.

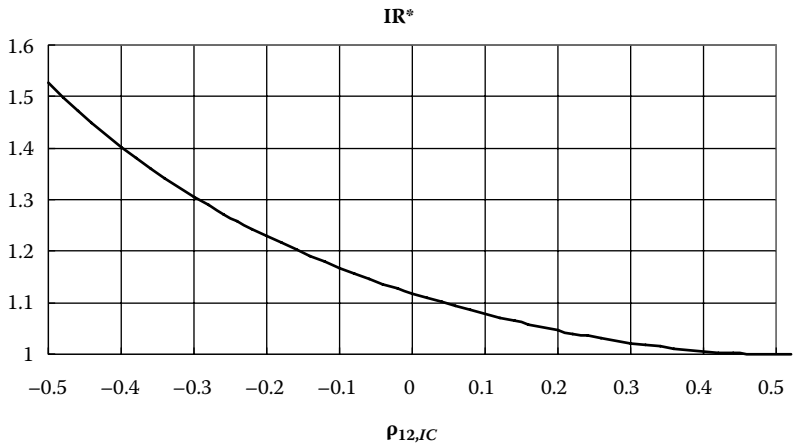


FIGURE 7.1. The optimal IR as a function of IC correlation between the two factors whose IRs are 1.0 and 0.5, respectively.

Although such a factor model is theoretically correct, in practice it is highly improbable to implement such a solution. This is so because, when the IC correlation is high and positive, the optimal model will try to arbitrage one factor against another, i.e., place positive weight on the factor with higher IR, and negative weight on the factor with lower IR. Thus, the outcome of such a model is extremely sensitive to the estimation accuracy of the IR difference. If the model happens to be wrong in this regard, it would put the wrong weights on the wrong factors.

7.2.4 An Empirical Example

To illustrate an empirical application of Equation 7.17, we select one factor from each factor category discussed in Chapter 5: cash flow from operation to enterprise value (CFO2EV) from the value category, external financing (XF) from the quality category, and the 9-month price momentum (Ret9) from the momentum category. For each factor, we calculated the risk-adjusted IC on a quarterly basis using the Russell 3000 as the stock universe. The time span of our data is from 1987 to 2004 — 72 quarters in total. We also compute the average IC and the standard deviation of IC for the three factors so that we can derive the optimal alpha model weights based on the three factors.

The average ICs and the standard deviation of ICs are listed in Table 7.1 together with the annualized IR. Because we use quarterly data, the annualized IR is simply twice the ratio of average IC to the standard deviation

TABLE 7.1 Average IC and Standard Deviation of IC for the Three Factors

	CFO2EV	XF	Ret9
Average IC	0.06	0.04	0.05
Standard deviation	0.05	0.04	0.09
Annualized IR	2.09	1.91	1.10

TABLE 7.2 Weights of Alpha Models and Corresponding IR

	IR	CFO2EV	XF	Ret9
$\mathbf{w}_1$	2.68	38%	50%	12%
$\mathbf{w}^*$	3.23	69%	–1%	32%

of IC. As we can see from this table, both the value factor CFO2EV and the quality factor XF have high IR mainly due to a low standard deviation of IC, i.e., the excess returns associated with these two factors tend to exhibit low volatility. On the other hand, the momentum factor has the same level of average IC as the other two, but its standard deviation is almost twice as high, resulting in lower IR for the factor.

With standard deviations of IC and the IC correlation matrix (in Table 7.4), we construct the IC covariance matrix and then derive the optimal alpha model that maximizes IR, using (7.17). The weights of the optimal model are shown as $\mathbf{w}^*$ in Table 7.2. In this case, we have 69% in CFO2EV and 32% in Ret7, but –1% in XF. The XF factor itself has an IR of 1.91, but because it is highly correlated with the factor CFO2EV, which has a higher IR and lower correlation with Ret9, the XF factor gets no weight in the optimal alpha model. To see the importance of IC correlation more directly, we also derive another set of weights with a diagonal IC covariance matrix by letting IC correlations be zero. This is shown as $\mathbf{w}_1$ in Table 7.2 and has 50, 38, and 12% in XF, CFO2EV, and Ret9, respectively. However, the IR of this model is only 2.68, whereas the maximum IR with $\mathbf{w}^*$ is 3.23.

7.2.5 Maximum Single-Period IC

We have found the optimal model weights $\mathbf{v}$ that maximize the multiperiod IR. One could also focus on model weights that maximize the single-period IC. The optimal weights for a single-period IC depend on the average ICs and the factor correlation matrix $\mathbf{\Phi}$.

From (7.8), we take the partial derivative with respect to $\mathbf{v}$ to obtain the optimality condition. Following steps similar to (7.16) and (7.17), we obtain

$$\mathbf{v} = s\mathbf{\Phi}_t^{-1}\overline{\mathbf{IC}}. \tag{7.21}$$

The solution is proportional to the inverse of the factor covariance (or correlation) matrix times the IC.

If the factor correlation matrix remains constant over time, (7.21) is also the solution that achieves *the maximum average IC* over multiple periods. However, the efficacy of an alpha model is not in the average IC but in the ratio of the average IC to the standard deviation of IC. The weights in (7.21) totally ignore the standard deviation of IC. Therefore, there is no guarantee that its IR would be high. A prime example of factors with high average IC but high standard deviation of IC is the 1-month price reversal factor. In addition, the 1-month reversal factor tends to have low factor correlation with other low-frequency factors. Hence, a model that maximizes the average IC would have significant weight in the 1-month price reversal factor. However, such a model is likely to have a low IR and, to make matters worse, extremely high turnover. We shall discuss the subject of portfolio turnover in detail in later chapters.

7.3 FACTOR CORRELATION VS. IC CORRELATION

The optimal model weights depend strongly on IC correlations but not on factor correlations. We have shown that, when we assume that the factor correlations stay constant over time, it completely drops out of the analysis as far as IR is concerned. Although it is important to distinguish between them, the two are in fact interrelated. In this section we analyze their relationship.

7.3.1 Relationship in a Single Period

We continue to use the two-factor case as an example. Suppose that, for a single period, the two standardized factors have a factor correlation $\phi_{12,t} = \text{corr}\left(\mathbf{F}_{1,t}, \mathbf{F}_{2,t}\right)$. The ICs of the two factors for the period will be constrained by the factor correlation. Imagine the case where the factor correlation is unity; then we know that the two factors are essentially identical and the two ICs must be the same. On the other hand, if the factor correlation is –1, then the two ICs must be the opposite of each other. However, when the factor correlation falls somewhere between these two extreme cases, it leads to a much looser constraint on the two ICs.

For general cases, the two ICs — $IC_{1,t}$ and $IC_{2,t}$ — together with $\phi_{12,t}$ forms a 3×3 correlation matrix:

$$C = \begin{pmatrix} 1 & IC_{1,t} & IC_{2,t} \\ IC_{1,t} & 1 & \phi_{12,t} \\ IC_{2,t} & \phi_{12,t} & 1 \end{pmatrix}. \tag{7.22}$$

Because C has to be positive definite, its determinant must be nonnegative. We have

$$\begin{aligned} \det C &= \begin{vmatrix} 1 & \phi_{12,t} \\ \phi_{12,t} & 1 \end{vmatrix} - IC_{1,t} \begin{vmatrix} IC_{1,t} & \phi_{12,t} \\ IC_{2,t} & 1 \end{vmatrix} + IC_{2,t} \begin{vmatrix} IC_{1,t} & 1 \\ IC_{2,t} & \phi_{12,t} \end{vmatrix} \\ &= 1 - \phi_{12,t}^2 - IC_{1,t}^2 - IC_{2,t}^2 + 2\phi_{12,t} IC_{1,t} IC_{2,t} \geq 0 \end{aligned} \tag{7.23}$$

or

$$IC_{1,t}^2 + IC_{2,t}^2 - 2\phi_{12,t} IC_{1,t} IC_{2,t} + \phi_{12,t}^2 - 1 \leq 0. \tag{7.24}$$

For a given factor correlation, the expression on the left side describes an ellipse on the $\left(IC_{1,t}, IC_{2,t}\right)$-plane, and the two ICs must lie inside the ellipse. Figure 7.2 plots the ellipse and the region within for a factor correlation of 0.5. The major axis of the ellipse lies on the line $IC_{1,t} = IC_{2,t}$, and the minor axis on the line $IC_{1,t} = -IC_{2,t}$. This is true as long as $\phi_{12,t} \geq 0$. When the factor correlation is negative, the two axes switch places. Statistically, the two ICs can be anywhere inside the ellipse. As seen from the graph, the possibilities are numerous: they can be both positive, both negative, or have opposite signs.

Another way to look at the influence of the factor correlation on the two ICs is to express IC_2 in terms of IC_1, ϕ_{12}, and a residual IC, $IC_{\varepsilon 2,1}$, as

$$IC_2 = \phi_{12} \cdot IC_1 + \sqrt{1 - \phi_{12}^2} \cdot IC_{\varepsilon 2,1}. \tag{7.25}$$

Here, we suppress the subscript t for clarity. The residual IC, $IC_{\varepsilon 2,1}$, is the correlation between security returns and the residual factor score of $\mathbf{F}_2$ after netting out $\mathbf{F}_1$. Because the correlation between the two factors is ϕ_{12} and the two factors are standardized, the residual factor, $\boldsymbol{\varepsilon}_{2,1}$, is simply $\boldsymbol{\varepsilon}_{2,1} = \mathbf{F}_2 - \phi_{12}\mathbf{F}_1$ and it is orthogonal to $\mathbf{F}_1$. It is easy to prove that the correlation $IC_{\varepsilon 2,1}$ between the residual factor $\boldsymbol{\varepsilon}_{2,1}$, and the return is related to other terms by (7.25). Furthermore, as $\boldsymbol{\varepsilon}_{2,1}$ is orthogonal to $\mathbf{F}_1$, the residual

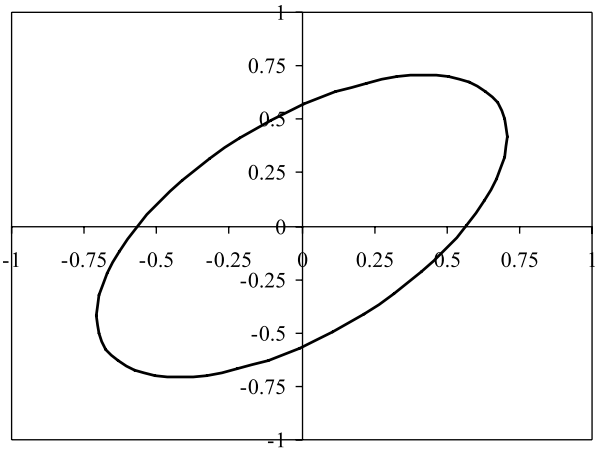


FIGURE 7.2. Feasible region of IC for two factors with correlation of 0.5.

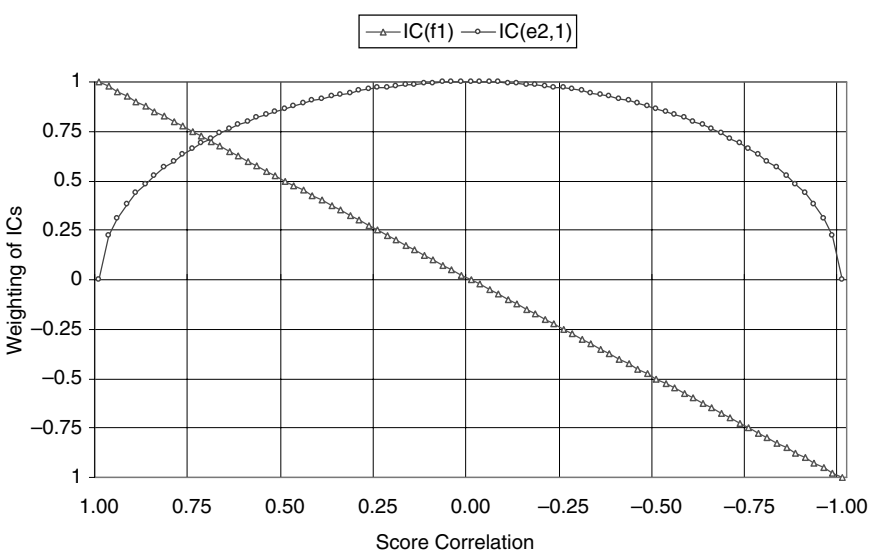


FIGURE 7.3. Weighting of ICs with score correlations.

correlation $IC_{\varepsilon 2,1}$ is completely free, i.e., it can be any number between –1 and 1. Based on (7.25), IC_2 can be as high as $IC_2 = \phi_{12} \cdot IC_1 + \sqrt{1-\phi_{12}^2}$ and as low as $IC_2 = \phi_{12} \cdot IC_1 - \sqrt{1-\phi_{12}^2}$.

We can also interpret IC_2 as a weighted, linear combination of IC_1 and $IC_{\varepsilon 2,1}$ whose weighting is a function of the score correlation, ϕ_{12} . Figure 7.3 shows how the weighting of IC_1 and $IC_{\varepsilon 2,1}$ varies with ϕ_{12} . The influence

of IC_1 is linearly proportional to ϕ_{12}, ranging from 1 to –1, whereas the influence of $IC_{\varepsilon 2,1}$ is not only always positive but also a concave function. As such, $IC_{\varepsilon 2,1}$ generally exhibits more influence in determining IC_2 than IC_1. For example, when ϕ_{12} is equal to 0.975 — extremely close to a perfect-score correlation — the weights for IC_1 and $IC_{\varepsilon 2,1}$ are 0.975 and 0.222, respectively, implying that $IC_{\varepsilon 2,1}$ still commands a material influence. In contrast, when factor scores are close to being uncorrelated, such as ϕ_{12} being equal to 0.025, the weights for IC_1 and $IC_{\varepsilon 2,1}$ are 0.025 and 0.9997, respectively. In this instance, the influence of IC_1 is no longer material.

7.3.2 Multiperiod IC Correlations

The discussion so far has focused on the ICs and factor correlation of a single period, and they are calculated based on a cross section of two risk-adjusted forecast vectors and risk-adjusted returns of N stocks. As we extend from a single period to multiple periods, all three correlation coefficients in matrix (7.22) fluctuate, forming time-series or distributions. For instance, $IC_{1,t}$ and $IC_{2,t}$ each has sample (theoretical) and empirical distributions. Our interest is on the statistical properties of their distribution.

One of the major findings from Chapter 4 is that, even though the naive estimation for the standard deviation of IC is $1/\sqrt{N}$ or the sampling error, with N being the number of stocks, empirically the IC standard deviation for the majority of alpha factors we considered, is much higher than the naive estimation. With two or more factors, we are interested in the correlation between their ICs over time because they play a crucial role in determining the IR of multifactor alpha models. In this section, we first present a naive estimation of the IC correlation and then examine IC correlations empirically.

One naive estimate of IC correlation follows the general theory of sample covariance matrix based on a multivariate normal distribution. Under certain assumptions, the sample covariance matrix follows a Wishart distribution (see Muirhead 1982), and the covariance between the ICs is given by the following equation:

$$\operatorname{cov}\left(IC_{1,t}, IC_{2,t}\right) = \frac{1}{N}\left(\bar{\phi}_{12} + \overline{IC}_1 \cdot \overline{IC}_2\right). \tag{7.26}$$

The left-hand side is the covariance between the two ICs. On the right-hand side, N is the number of stocks; the barred variables are the averages

of factor correlations and the averages of ICs. In practice, the average IC of the alpha factors is usually small. We approximate Equation 7.26 by

$$\operatorname{cov}\left(IC_{1,t}, IC_{2,t}\right) = \operatorname{std}\left(IC_1\right)\operatorname{std}\left(IC_2\right)\operatorname{corr}\left(IC_1, IC_2\right) \approx \frac{1}{N}\bar{\phi}_{12}. \quad (7.27)$$

Therefore, we have

$$\operatorname{corr}\left(IC_1, IC_2\right) \approx \frac{\bar{\phi}_{12}}{N\operatorname{std}\left(IC_1\right)\operatorname{std}\left(IC_2\right)}. \quad (7.28)$$

Equation 7.28 is the naïve estimation of the IC correlation. Furthermore, when the standard deviations of ICs are solely due to sampling error, they are equal to $1/\sqrt{N}$, i.e., $\operatorname{std}\left(IC_1\right) = \operatorname{std}\left(IC_2\right) = 1/\sqrt{N}$. If that were the case, then the IC correlation would be approximately the same as the average factor correlation, i.e., $\operatorname{corr}\left(IC_1, IC_2\right) \approx \bar{\phi}_{12}$.

When the standard deviations of ICs are greater than the sampling error, the IC correlation, as demonstrated in Chapter 4 and according to (7.28), should be *in theory* of the same sign as the factor correlation but less than the factor correlation. For models with more than two factors, Equation 7.28 applies to every pairwise IC correlation.

- Previous researchers seem to have focused solely on factor correlation, ignoring IC correlation. For analysis of multiperiod IR, we have established a theoretical link between the IC correlation and the factor correlation, which is only valid under the most ideal assumptions. Although the link provides some theoretical justification for previous research using factor correlation, it also highlights their limitation.

Example 7.6

If the average factor correlation is 0.5, $N = 1000$, and if the standard deviations of both ICs are $1/\sqrt{N}$, i.e., 0.032, then the IC correlation should also be 0.5. However, if the standard deviations of IC are 0.04 and 0.05, respectively, the IC correlation should be $0.5/(1000 \times 0.04 \times 0.05) = 0.25$, half the factor correlation.

TABLE 7.3 Average and Standard Deviation of Factor Correlations

Average (Stdev)	CFO2EV	XF	Ret9
CFO2EV	1.00 (0.00)	0.31 (0.09)	–0.04 (0.10)
XF		1.00 (0.00)	0.06 (0.06)
Ret9			1.00 (0.00)

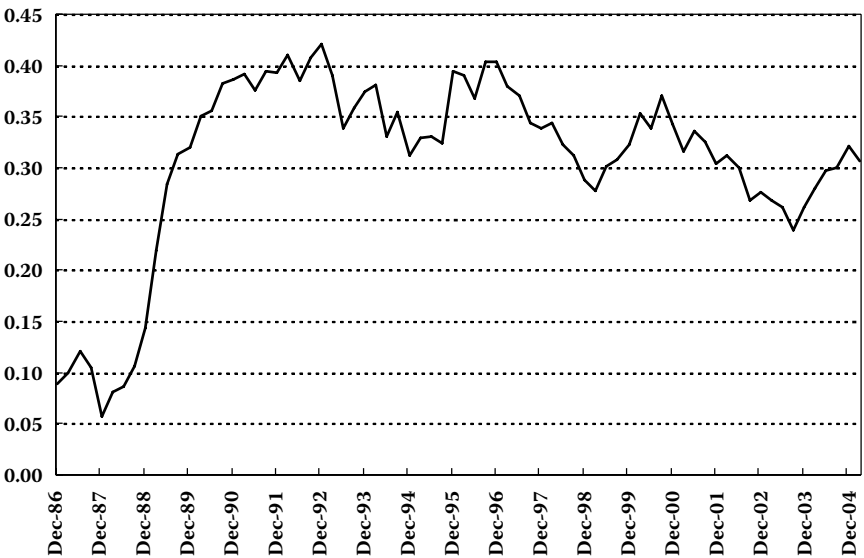


FIGURE 7.4. Quarterly factor correlations between CFO2EV and XF.

7.3.3 Empirical Examination of Factor Correlation and IC Correlation

It is probably safe to say that, in reality, many simplifying assumptions underlying theoretical models of the stock market break down. For instance, stock returns are generally not normally distributed. We also saw another example in Chapter 4 in the standard deviation of IC. We will now examine another case concerning the IC correlation.

Continuing the empirical example in the last section, Table 7.3 shows the average and standard deviation of factor correlations over the entire period. It is interesting to note that the correlation between CFO2EV and XF has an average of 0.31 and a standard deviation of 0.09, so it is significantly positive. The correlation between CFO2EV and Ret9 is slightly negative, whereas the correlation between XF and Ret9 is slightly positive. Figure 7.4 plots the time series of the factor correlations between CFO2EV and XF. It is initially low in 1987 and then increases to around 0.4 in 1990. Since then it has been fluctuating between 0.3 and 0.4.

TABLE 7.4 The IC Correlations of Three Factors

	CFO2EV	XF	Ret9
CFO2EV	1.00	0.73	−0.50
XF		1.00	−0.22
Ret9			1.00

TABLE 7.5 Sampling Errors of Time-Series IC Correlations

	ρ	std(ρ)	2-std Interval
ρ(IC_XF, IC_CFO2EV)	0.73	0.08	(0.56, 0.89)
ρ(IC_RET9, IC_CFO2EV)	−0.50	0.10	(−0.71, −0.29)
ρ(IC_RET9, IC_XF)	−0.22	0.12	(−0.45, 0.02)

The correlations of risk-adjusted ICs for the three factors are presented in Table 7.4. We note that they are significantly different from the factor correlations seen in Table 7.3. For example, the IC correlation between CFO2EV and XF is 0.73, which is significantly higher than the average factor correlation of 0.31, indicating that the diversification benefit between these two factors is not as strong as it would seem. On the other hand, the IC correlation between CFO2EV and Ret9 is −0.5, which is significantly lower than the factor correlation between the two. This seems to be a general phenomenon for value factors and price momentum factors as the IC diversification between them is significantly better than what the factor correlation would otherwise indicate. Lastly, the IC correlation between the quality factor XF and the price momentum factor Ret9 is slightly negative.

In our example, two out of the three IC correlations are significantly different from the factor correlations even if we take into account the variability of factor correlations over the entire period. We can calculate the confidence interval of IC correlations to provide another perspective. The standard deviation of IC correlation is approximately given by in the sample IC and the number of quarters Q (Keeping, 1995)

$$\mathrm{std}(\rho_{IC}) = \frac{(1-\rho_{IC}{}^2)}{\sqrt{Q-1}}\sqrt{1+\frac{11\rho_{IC}{}^2}{2(Q-1)}}\,. \tag{7.29}$$

Table 7.5 shows the sampling error of the time-series IC correlations as well as their two standard deviation confidence intervals. All three cross-sectional score correlations fall out of their corresponding confidence

interval. In fact, for the first two pairs, their average factor correlations lie outside the three standard deviations confidence interval.

7.4 COMPOSITE ALPHA MODEL WITH ORTHOGONALIZED FACTORS

Our analysis so far has focused on building composite models with the risk-adjusted factors. We have shown that the optimal weights of factors depend on average ICs and the covariance matrix of ICs. This provides important insights into factor diversification: factors with low IC correlations are more desirable than factors with high IC correlation, as the previous example illustrates.

We have made several simplifying assumptions, though. First, we standardized all risk-adjusted factors so that their cross-sectional dispersions remain unity. Second, we assumed that correlations among factors are constant over time. These assumptions made the problem of optimizing IR analytically tractable and led to our solution for the optimal weights and insight about factor diversification.

However, factor correlations are time varying, as we have shown in the last section in Figure 7.4. The fact that the variation in factor correlations is relatively small compared to the IC volatility justifies our approximation approach. Nevertheless, it would be desirable to derive a solution without this simplification. We can do so with orthogonalized factors. Factor orthogonalization can be viewed as another step in preprocessing factors along with factor standardization. When the procedure is carried out in every time period, the factor correlations will always be zero and thus constant.

When the factors are both orthogonal and standardized, the single-period IC of a composite (7.8) reduces to

$$IC_{c,t} = \frac{1}{\sqrt{\mathbf{v}' \cdot \mathbf{v}}} \sum_{i=1}^{M} v_i IC_{i,t} \ . \tag{7.30}$$

Because the ICs are now the only terms that vary in time, the IR of the model will be exactly that of (7.14), and the previous solution of optimal weights applies without any approximation.

7.4.1 Gram–Schmidt Procedure

A common mathematical technique, the Gram–Schmidt procedure sequentially makes each factor orthogonal to previously orthogonalized

factors. Suppose we have M factors $\left(\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_M\right)$ that have been standardized. With no particular order, the first factor $\mathbf{F}_1$ will be the first orthogonal factor, i.e., $\mathbf{F}_1^o = \mathbf{F}_1$, with the superscript denoting orthogonalized factors. Then the second orthogonal factor is defined as

$$\mathbf{F}_2^o = \frac{1}{\sqrt{1-\hat{\rho}_{21}^2}}\left(\mathbf{F}_2 - \hat{\rho}_{21}^2 \mathbf{F}_1^o\right), \tag{7.31}$$

where $\hat{\rho}_{21} = \rho_{21}$ is the cross-sectional correlation between $\mathbf{F}_2$ and $\mathbf{F}_1^o$, which is the same as the correlation between $\mathbf{F}_2$ and $\mathbf{F}_1$. The orthogonlized factor $\mathbf{F}_2^o$ is the factor $\mathbf{F}_2$ with the effect of $\mathbf{F}_1^o$ taken out. The ratio $1\big/\sqrt{1-\hat{\rho}_{21}^2}$ makes $\mathbf{F}_2^o$ standardized. Moving on to the third factor, let $\hat{\rho}_{31}$ and $\hat{\rho}_{32}$ be the correlation between $\mathbf{F}_3$ and $\mathbf{F}_1^o$ and $\mathbf{F}_3$ and $\mathbf{F}_2^o$, respectively, which are calculated after we have derived the orthogonalized factor. Then,

$$\mathbf{F}_3^o = \frac{1}{\sqrt{1-\hat{\rho}_{32}^2 - \hat{\rho}_{31}^2}}\left(\mathbf{F}_3 - \hat{\rho}_{32}\mathbf{F}_2^o - \hat{\rho}_{31}\mathbf{F}_1^o\right) \tag{7.32}$$

is a standardized factor orthogonal to both $\mathbf{F}_2^o$ and $\mathbf{F}_1^o$. In general, suppose $\left(\mathbf{F}_1^o, \cdots, \mathbf{F}_{p-1}^o\right)$ are orthogonalized factors; then, for the factor $\mathbf{F}_p$, we first calculate its correlations with $\left(\mathbf{F}_1^o, \cdots, \mathbf{F}_{p-1}^o\right)$ and denote them by $\left(\hat{\rho}_{p1}, \cdots, \hat{\rho}_{p,p-1}\right)$. The orthogonalized factor is given by

$$\mathbf{F}_p^o = \frac{1}{\sqrt{1-\hat{\rho}_{p1}^2 - \hat{\rho}_{p2}^2 - \cdots - \hat{\rho}_{p,p-1}^2}}\left(\mathbf{F}_p - \hat{\rho}_{p1}\mathbf{F}_1^o - \hat{\rho}_{p2}\mathbf{F}_2^o - \cdots - \hat{\rho}_{p,p-1}\mathbf{F}_{p-1}^o\right). \tag{7.33}$$

The factor $\mathbf{F}_p^o$ is proportional to the component of $\mathbf{F}_p$, which is uncorrelated with the previous orthogonlized factors.

Orthogonal factors produced by the Gram–Schmidt procedure can attest whether or not the original factors have independent information about forward returns. This is true if the IC of an orthogonalized factor is still positive and significant. However, if the IC of an orthogonalized factor becomes insignificant or even changes sign, its weight in the optimal model will likely change dramatically.

7.4.2 Optimal Model with the Gram–Schmidt Procedure

How do we combine the orthogonalized factors into an optimal alpha model? Recall the solution for weights of the optimal alpha model that is

TABLE 7.6 Average IC and Standard Deviation of IC for the Three Orthogonalized Factors

	CFO2EV.o	XF.o	Ret9.o
Average IC	0.06	0.02	0.05
Standard deviation	0.05	0.03	0.09
Annualized IR	2.09	1.36	1.15

given by $\mathbf{v}^* = s\boldsymbol{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}}$ in (7.17), where $\boldsymbol{\Sigma}_{IC}^{-1}$ is the inverse of the IC covariance matrix, $\overline{\mathbf{IC}}$ is the average IC of the factors, and s is a scalar. The optimal model of orthogonalized factors follows the same form. We illustrate it with the three factors used in the previous example: cash flow from operating to enterprise value (CFO2EV), external financing (XF), and 9-month return (Ret9). In the Gram–Schmidt procedure, we have picked CFO2EV as the first factor, XF as the second, and Ret9 as the third.

Table 7.6 lists the average IC, the standard deviation of the orthogonalized factors, and the IR. As CFO2EV is the first factor, the orthogonalized version CFO2EV.o is the same as the original factor. The second factor XF.o differs significantly from the original factor. Compared to Table 7.3, both the average IC and the standard deviation of IC decrease, and the IR is less than that of the original factor. The reason is that the factor correlation between XF and CFO2EV is reasonably high, and hence the orthogonalization procedure greatly affects XF. On the other hand, the last factor Ret9 has little correlation with the other two factors, so Ret.o is almost the same as Ret9.

- As the example shows, the Gram–Schmidt procedure affects factors that have high correlations with other factors. This is especially true for factors in the same factor category: for example, earning yield and dividend yield in the value category.

Table 7.7 shows the IC correlations of the orthogonalized factors. In general, we should expect ICs of the orthogonalized factors to be less correlated than the original factors because their factor correlations are constructed to be zero. This seems to be true for two pairs of factors. Factors CFO2EV.o and XF.o have IC correlation of 0.34 compared to the IC correlation of 0.73 for CFO2EV and XF. Factors XF.o and Ret9.o have IC correlation of –0.03 compared to the IC correlation of –0.22 for the original factors (Table 7.2). However, the other IC correlation between CFO2EV.o and Ret9.o shows no change.

TABLE 7.7 The IC Correlations of Three Orthogonalized Factors

	CFO2EV.o	XF.o	Ret9.o
CFO2EV.o	1.00	0.34	−0.50
XF.o		1.00	−0.03
Ret9.o			1.00

TABLE 7.8 Weights of Alpha Models and Corresponding IR Based on the Three Orthogonalized Factors

	IR	CFO2EV.o	XF.o	Ret9.o
$\mathbf{w}_1$	2.85	40%	47%	13%
$\mathbf{w}^\star$	3.30	61%	9%	30%

Table 7.8 shows the sets of weights of optimal alpha models based on the orthogonalized factors — one with the full IC covariance matrix and the other with diagonal IC covariance matrix. Compared to Table 7.4, the optimal weight $\mathbf{w}^\star$ has a positive 9% in XF.o, and the IR increases slightly. The IR of $\mathbf{w}_1$ shows greater improvement from that of Table 7.4 because the IC correlations of the orthogonalized factors play a lesser role in determining the optimal IR. Note the following:

- Another method of factor orthogonalization is principal component analysis, or PCA. The principal components (PC) of $\left(\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_M\right)$ are their linear combinations. The first PC is the linear combination of $\left(\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_M\right)$ that has the largest cross-sectional dispersion, and the second PC is the combination of $\left(\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_M\right)$ uncorrelated to the first PC that has the largest cross-sectional dispersion, and so on. The PCA technique is theoretically appealing, but it has one practical difficulty. Because principal components are unique up to a change in signs, one has to ensure that "same" PCs are selected over time. This could be a challenge if the correlation structure of factors changes drastically over time.

7.5 FAMA–MACBETH REGRESSION AND OPTIMAL ALPHA MODEL

Although most practitioners recognize the benefit of combining multiple alpha sources in terms of IR improvement, their approaches to construct a multifactor alpha model vary widely. The analytical framework developed so far in this book relies on the risk-adjusted ICs of individual factors and

their correlations. One of the key facts for a multifactor alpha model is that the excess returns from individual factors are essentially additive; the overall excess return is a linear combination of individual excess returns, whereas the factor correlations enter the linear combination through a scaling factor.

There are practitioners who employ other statistical framework and derive forecasts based on empirical asset pricing back-test procedure, such as the Fama–MacBeth (1973) regression, which consists of a series of cross-sectional OLS regressions. Even though the Fama–MacBeth regression is simple to implement and intuitively appealing, it is used in most asset pricing studies to ascertain whether a factor is priced. The question is whether it provides an analytical foundation for combining multiple alpha sources.

To answer this question, we should first give an economic interpretation of the regression coefficients in a cross-sectional OLS regression. The key question is whether the regression coefficients represent the excess returns of certain active portfolios, and, if they do, what are the alpha factors behind these active portfolios?

7.5.1 Univariate OLS Regression

When there is just one independent factor in the cross-sectional regression, the interpretation is straightforward. Suppose the regression takes the form

$$\mathbf{r}_t = \alpha_t + \beta_t \mathbf{f}_t . \tag{7.34}$$

Then the coefficient is

$$\beta_t = \frac{\operatorname{cov}\left(\mathbf{r}_t, \mathbf{f}_t\right)}{\operatorname{var}\left(\mathbf{f}_t\right)} = \frac{\operatorname{corr}\left(\mathbf{r}_t, \mathbf{f}_t\right)\operatorname{dis}\left(\mathbf{r}_t\right)}{\operatorname{dis}\left(\mathbf{f}_t\right)} . \tag{7.35}$$

When the factor is standardized, the regression coefficient is IC times the dispersion of realized returns, i.e.,

$$\beta_t = \operatorname{corr}\left(\mathbf{r}_t, \mathbf{f}_t\right)\operatorname{dis}\left(\mathbf{r}_t\right) . \tag{7.36}$$

Comparing Equation 7.36 with Equation 7.6, we see that, in this case, the regression coefficient is proportional to the excess return of an active portfolio based on the factor.

7.5.2 OLS Regression with Multiple Factors

When there are multiple factors, the OLS regression coefficients are no longer the ICs of individual factors, unless the factors are uncorrelated. However, what are their economic interpretations in the context of excess returns? To develop insight into this question, we consider the case with two factors and derive the coefficients explicitly. The regression equation is

$$\mathbf{r}_t = \alpha_t + \beta_{1,t}\mathbf{f}_{1,t} + \beta_{2,t}\mathbf{f}_{2,t} . \tag{7.37}$$

The coefficients in terms of variances and covariances are given by

$$\begin{pmatrix} \beta_{1,t} \\ \beta_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & \rho_t \\ \rho_t & 1 \end{pmatrix}^{-1} \begin{pmatrix} IC_1 \\ IC_2 \end{pmatrix} \mathrm{dis}\left(\mathbf{r}_t\right). \tag{7.38}$$

Again, we have assumed that the factors are standardized, with variance being 1, and ρ_t denotes the factor or score correlation. Inverting the matrix and multiplying the ICs gives

$$\begin{aligned} \beta_1 &= \frac{1}{1-\rho^2}\left(IC_1 - \rho IC_2\right)\mathrm{dis}\left(\mathbf{r}\right) \\ \beta_2 &= \frac{1}{1-\rho^2}\left(IC_2 - \rho IC_1\right)\mathrm{dis}\left(\mathbf{r}\right) \end{aligned} \tag{7.39}$$

We have suppressed subscript t for clarity. The coefficients are combinations of ICs, with the factor correlation entering as one of the weights. When the two factors are uncorrelated, the coefficients are identical to the univariate regression coefficients.

The economic interpretation of β_1 is the *marginal* return contribution of $\mathbf{f}_1$ after netting out the influence of $\mathbf{f}_2$. Similarly, β_2 represents the *marginal* return contribution of $\mathbf{f}_2$ after controlling the influence of $\mathbf{f}_1$. To see this, we note that both β_1 and β_2 can be derived from two separate univariate OLS regressions with cross-sectional return as the dependent variable. For instance, to derive β_1, we first regress $\mathbf{f}_1$ against $\mathbf{f}_2$:

$$\mathbf{f}_1 = \rho\mathbf{f}_2 + \boldsymbol{\varepsilon}_{1,2} . \tag{7.40}$$

The residual is then $\boldsymbol{\varepsilon}_{1,2} = \mathbf{f}_1 - \rho\mathbf{f}_2$. To be consistent with factor standardization, we standardize the residual so that its cross-sectional dispersion is unity:

$$\tilde{\boldsymbol{\varepsilon}}_{1,2} = \frac{\mathbf{f}_1 - \rho\mathbf{f}_2}{\sqrt{1-\rho^2}}. \tag{7.41}$$

ICs of both $\tilde{\boldsymbol{\varepsilon}}_{1,2}$ (standardized residual) and $\boldsymbol{\varepsilon}_{1,2}$ (raw residual) are the same:

$$\widetilde{IC_1} = \frac{\operatorname{cov}\left(\tilde{\boldsymbol{\varepsilon}}_{1,2}, \mathbf{r}\right)}{\operatorname{dis}\left(\mathbf{r}\right)} = \frac{IC_1 - \rho IC_2}{\sqrt{1-\rho^2}}. \tag{7.42}$$

In the second univariate regression, let $\beta_{\mathbf{r},\boldsymbol{\varepsilon}_{1,2}}$ be the coefficient estimate of a cross-sectional regression, wherein the cross-sectional return, $\mathbf{r}_t$, is the dependent variable, and raw residual of $\boldsymbol{\varepsilon}_{1,2}$ is the independent variable. As the following equation shows, $\beta_{\mathbf{r},\boldsymbol{\varepsilon}_{1,2}}$ is exactly the same as β_1

$$\beta_{\mathbf{r},\boldsymbol{\varepsilon}_{1,2}} = \frac{\operatorname{cov}(\boldsymbol{\varepsilon}_{1,2}, \mathbf{r})}{\operatorname{var}(\boldsymbol{\varepsilon}_{1,2})} = \frac{\operatorname{cov}(\mathbf{f}_1 - \rho\mathbf{f}_2, \mathbf{r})}{1-\rho^2} = \frac{IC_1 - \rho \cdot IC_2}{1-\rho^2} \cdot \operatorname{dis}(\mathbf{r}) = \beta_1. \tag{7.43}$$

Similarly, the IC of factor 2 with factor 1 regressed out is

$$\widetilde{IC_2} = \frac{IC_2 - \rho IC_1}{\sqrt{1-\rho^2}}. \tag{7.44}$$

Comparing Equation 7.39, Equation 7.42, and Equation 7.44 shows that multivariate regression coefficients are related to *residual ICs* as

$$\begin{aligned} \beta_1 &= \frac{1}{\sqrt{1-\rho^2}} \widetilde{IC_1} \operatorname{dis}\left(\mathbf{r}\right) = \beta_{\mathbf{r},\boldsymbol{\varepsilon}_{1,2}} \\ \beta_2 &= \frac{1}{\sqrt{1-\rho^2}} \widetilde{IC_2} \operatorname{dis}\left(\mathbf{r}\right) = \beta_{\mathbf{r},\boldsymbol{\varepsilon}_{2,1}} \end{aligned} \tag{7.45}$$

- The residual IC is, in essence, the information coefficient of a composite factor whose weights are related to the factor correlation. For example, $\widetilde{IC_1}$ is the IC of factor $\tilde{\boldsymbol{\varepsilon}}_{1,2} = \left(\mathbf{f}_1 - \rho\mathbf{f}_2\right) \Big/ \sqrt{1-\rho^2}$. Depending on the factor correlation, the residual IC could be very different from the IC of the individual factor.

Example 7.7

Suppose $IC_1 = 0.2$, $IC_2 = 0.1$, and $\rho = 0.8$. Then the residual ICs are $\widetilde{IC_1} = \left(0.2 - 0.8 \cdot 0.1\right) \Big/ \sqrt{1-0.8^2} = 0.2$ and $\widetilde{IC_2} = \left(0.1 - 0.8 \cdot 0.2\right) \Big/ \sqrt{1-0.8^2}$ = –0.1. Even though both factors have positive ICs, one residual IC is positive and the other is negative! This is due to the high correlation between the two factors. If the correlation is reduced to 0.5 from 0.8, the residual ICs are $\widetilde{IC_1} = 0.17$ and $\widetilde{IC_2} = 0.0$, respectively. The second factor is rendered as having no information.

When the factor correlation is negative, the residual ICs are going to be higher than the original ICs. The lesson is that one should not interpret multivariate regression coefficients as returns to alpha factors; instead, they are *marginal* returns to alpha factors after netting out influences from other factors. Especially, they should not be used in performance attribution of alpha factors. This is particularly problematic or simply wrong when the factors from the same category have high correlations, as we have seen in Chapter 5. For instance, earnings yield and cash flow yield tend to have high factor-score correlation, as both are constructed with the price as the denominator. Just because one worked better than the other in terms of higher IC, we cannot conclude that the lesser one had a negative contribution to the portfolio return.

7.5.3 Fama–MacBeth Regression and Asset Pricing Tests

Fama–Macbeth regression is commonly used by academic researchers to ascertain whether a factor is priced by the market through time after controlling for other known, priced factors such as beta, book-to-price, size, or price momentum. The procedure consists of a series of multiple OLS regressions for each cross section of securities. In each regression, cross-sectional returns form the dependent variable; and independent variables consist of two parts: control variables and a set of tested factors. Control variables are deployed to ensure that the tested pricing phenomenon was not subsumed by other known pricing phenomena. In other words, it is a test of whether the factor in question provides incremental pricing

information. For illustrative purpose, let us assume that $\mathbf{f}_1$ is a control variable and $\mathbf{f}_2$ is the factor in question. Each cross-sectional regression at time t is formulated as $\mathbf{r}_t = \alpha_t + \beta_{1,t}\mathbf{f}_{1,t} + \beta_{2,t}\mathbf{f}_{2,t}$. Factor $\mathbf{f}_2$ is considered as a priced factor if its time series t-stat $t = \overline{\beta}_2 / \text{std}(\beta_{2,t})$ is significantly different from zero. In other words, should $t(\beta_{2,t})$ be significantly different from zero, then $\mathbf{f}_2$ is said to be priced by the market after controlling for the known asset pricing phenomenon of $\mathbf{f}_1$.

Equation 7.45 shows this residual effect directly because it connects the OLS regression coefficients to the ICs of residual factors. When factor correlation ρ is stable and the return dispersion is constant, it is easily seen that the Fama–MacBeth t-stat is proportional to the IR of residual factors.

The interpretation of multivariate regression coefficients as coefficients of univariate regressions of return vs. residual factors provides critical insight into the results of the Fama–MacBeth regression. It turns out that this interpretation remains true as we add control variables (or risk factors) and more alpha factors into the OLS regression. Suppose we have

$$\mathbf{r} = \alpha + b_1\mathbf{I}_1 + \cdots + b_K\mathbf{I}_K + \beta_1\mathbf{f}_1 + \cdots + \beta_L\mathbf{f}_L, \tag{7.46}$$

where $(\mathbf{I}_1, \cdots, \mathbf{I}_K)$ are control variables and $(\mathbf{f}_1, \cdots, \mathbf{f}_L)$ are alpha factors, then the coefficient β_j can be obtained in the following steps for each cross section at a given time t, and these steps are repeated through time to derive a time series of estimates of β_j (see appendix for proof).

- Step 1: We regress factor $\mathbf{f}_j$ against all control variables and remaining alpha factors simultaneously.
- Step 2: We take the residual of the regression in Step 1 and run a univariate regression of returns against the residual to obtain β_j.

Similar to Equation 7.45, the coefficient β_j is related to the IC of the residual, the dispersion of the actual return, and the dispersion of the residual.

- There is a connection between the residual IC and the IC of the purified alpha in Chapter 4. The purified alpha is an alpha signal with the risk factors regressed out. The residual IC that is contained in the multivariate regression (7.46) is the IC of an alpha signal with not only the risk factors but also all other alpha factors regressed out. It

is an alpha signal so "pure" that it is orthogonal to both risk factors and other alpha factors.

7.5.4 Multifactor Model through Fama–MacBeth Regression

Although multivariate regression coefficients should be interpreted as return sensitivities to residual factor scores, a naive application of the Fama–MacBeth regression in deriving factor returns and optimal model weighting would result in erroneous model estimation due to factor-score correlations. There are two methods to alleviate the problem. First, recall if the factors are uncorrelated, and then the coefficients become sensitive to the factors and proportional to the factors' ICs. Thus, one simple way to avoid the collinear problem is to sequentially orthogonalize factor scores through the Gram–Schmidt procedure before each cross-sectional OLS regression. Then, using the coefficients, we can estimate the average ICs and covariances of IC to derive the optimal alpha model. This is the same model derived under the Gram–Schmidt procedure.

In the second method, one may choose not to orthogonalize the factors. Given the interpretation of regression coefficients in the Fama–MacBeth regression, one can still construct a multifactor model using the regression coefficients based on residual ICs. As we have shown, the residual IC can be easily derived from the Fama–MacBeth regression coefficients. We can find optimal weights that maximize the IR of the residual ICs, i.e., the average of residual IC to its standard deviation. This is similar to our approach of finding optimal weights based on the ICs of individual factors. However, there is one crucial difference. Models constructed through the Fama–MacBeth regression coefficients are no longer models for the original factors. Rather, they should be used as models of the residual factors. To apply the weights of the model, one must first find the residual factors by performing multivariate regression on each factor against all other factors and compute a weighted sum of the residual factors as the composite model.

The procedure to find the optimal weights of residual factors is analogous to the previous procedure for the original factors. We shall not repeat it here. We focus instead on the connection between the two sets of models: the model that maximizes the IR of the original factors and the model that maximizes the IR of the residual factors. First, it should be noted that the optimal model of the residual factors could be transformed into a model of the original factors because the residual factors themselves are linear combination of the original factors. For instance, for two-factor cases, the

residual factors are $\tilde{\boldsymbol{\varepsilon}}_{1,2} = \left(\mathbf{f}_1 - \rho\mathbf{f}_2\right) \Big/ \sqrt{1-\rho^2}$ and $\tilde{\boldsymbol{\varepsilon}}_{2,1} = \left(\mathbf{f}_2 - \rho\mathbf{f}_1\right) \Big/ \sqrt{1-\rho^2}$. If the model weights for the residual factors are $\tilde{v}_1$ and $\tilde{v}_2$, we have

$$\tilde{v}_1\tilde{\boldsymbol{\varepsilon}}_{1,2} + \tilde{v}_2\tilde{\boldsymbol{\varepsilon}}_{2,1} = \frac{\left(\tilde{v}_1 - \rho\tilde{v}_2\right)}{\sqrt{1-\rho^2}}\mathbf{f}_1 + \frac{\left(\tilde{v}_2 - \rho\tilde{v}_1\right)}{\sqrt{1-\rho^2}}\mathbf{f}_2 = v_1\mathbf{f}_1 + v_2\mathbf{f}_2 . \tag{7.47}$$

Conversely, a model of original factors can be transformed to a model of residual factors:

$$\tilde{v}_1 = \frac{v_1 + \rho v_2}{\sqrt{1-\rho^2}}, \tilde{v}_2 = \frac{v_2 + \rho v_1}{\sqrt{1-\rho^2}} . \tag{7.48}$$

Because of this linear transformation between the two sets of models, optimal models that maximize the information ratio utilizing either original factors or standardized residual factors are identical, provided that the factor correlations are constant over time. This is because the relationship between the residual IC and the original IC, and the relationship between the standardized residual factor and the original factors are identical (see, for example, Equations 7.41 and 7.42).

For the general case, denoting this constant linear relationship by matrix $\mathbf{P}$, we have

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{P}\cdot\mathbf{f} \text{ and } \mathbf{IC}_{\tilde{\boldsymbol{\varepsilon}}} = \mathbf{P}\cdot\mathbf{IC} . \tag{7.49}$$

The average residual IC and its covariance matrix are related to the average of the original IC and its covariance matrix by $\overline{\mathbf{IC}_{\tilde{\boldsymbol{\varepsilon}}}} = \mathbf{P}\cdot\overline{\mathbf{IC}}$ and $\boldsymbol{\Sigma}_{IC_\varepsilon} = \mathbf{P}'\boldsymbol{\Sigma}_{IC}\mathbf{P}$. The optimal weights (see Problem 7.9) for the residual factors are simply

$$\mathbf{v}_\varepsilon = \mathbf{P}^{-1}\boldsymbol{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}} = \mathbf{P}^{-1}\mathbf{v} , \tag{7.50}$$

where $\mathbf{v} = \boldsymbol{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}}$ is the optimal weights for the original factors. Therefore, the two composites with respective optimal weights are equal:

$$\mathbf{v}_{\boldsymbol{\varepsilon}}'\cdot\tilde{\boldsymbol{\varepsilon}} = \mathbf{v}'\mathbf{P}^{-1}\mathbf{P}\mathbf{f} = \mathbf{v}'\mathbf{f} . \tag{7.51}$$

- Another alternative for constructing a multifactor alpha model using Fama–MacBeth regression is to apply it directly to a predetermined combination of alpha factors plus risk factors from the outset (Yang, 2005). Unlike the multivariate setting, we now have just one composite alpha factor whose regression coefficient is directly linked to its IC after the effects of the risk factors are netted out. There is no residual effect involving other alpha factors. This is a version of purified alpha for a composite factor, and the regression coefficient is simply the multifactor IC times the dispersion of actual returns. When we carry out Fama–MacBeth regression over multiple time periods, the t-stat of the regression coefficient is a proxy of the IR for the predetermined combination of the alpha factors. This serves as a good indicator of portfolio performance for the given model. To find the optimal alpha model, however, we have to search for the optimal weights that maximize the t-stats of the regression coefficients by numerical means.

PROBLEMS

7.1 Calculate the dispersion and IC of the composite factor in Example 7.1 and 7.2 if the factor weights are 1/3 and 2/3, respectively.

7.2 Prove that the model weights that maximize single-period IC of (7.8) is (7.21).

7.3 Verify (7.17) to satisfy Equation 7.16. Find the value of s so that the sum of the model weights equals 1.

7.4 Assume that there are M alpha factors whose ICs are measured over T periods. We derive the optimal model weight $\mathbf{v}$ that maximizes IR by the following OLS regression:

$$\underset{(T\times 1)}{\mathbf{i}} = \underset{(T\times M)}{\mathbf{IC}} \times \underset{(M\times 1)}{\mathbf{v}} + \underset{(T\times 1)}{\mathbf{u}},$$

where $\mathbf{i}$ is a vector of ones — a constant dependent variable — $\mathbf{IC}$ is the observed IC matrix from the independent variables , $\mathbf{v}$ is the regression coefficients, and $\mathbf{u}$ is the error vector.

Prove that

(a) $\mathbf{v} = \left(\mathbf{IC}'\mathbf{IC}\right)^{-1}\left(\mathbf{IC}' \cdot \mathbf{i}\right)$;

(b) $\mathbf{IC}'\mathbf{IC} = \mathbf{\Sigma}_{\mathbf{IC}} + \overline{\mathbf{IC}} \cdot \overline{\mathbf{IC}}'$;

(c) $(\mathbf{IC}'\mathbf{IC})^{-1} = \mathbf{\Sigma}_{IC}^{-1} + \dfrac{\left(\mathbf{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}} \cdot \overline{\mathbf{IC}}'\mathbf{\Sigma}_{IC}^{-1}\right)}{1 + \overline{\mathbf{IC}}'\mathbf{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}}}$;

(d) $\mathbf{v} = \dfrac{\mathbf{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}}}{\mathbf{1} + \overline{\mathbf{IC}}'\mathbf{\Sigma}_{IC}^{-1}\overline{\mathbf{IC}}}$.

7.5 Derive the optimal IR (7.20) for two-factor models.

7.6 Extend Figure 7.1 to the full range of IC correlation from –1 to 1. Show that, when the IC correlation is greater than 0.5, the optimal model weight of factor 2 is negative.

7.7 Prove that factor $\mathbf{F}_p^o$ in (7.33) is orthogonal to $\left(\mathbf{F}_1^o, \cdots, \mathbf{F}_{p-1}^o\right)$.

7.8 Given two residual terms $\boldsymbol{\varepsilon}_{1,2} = \mathbf{f}_1 - \rho_t \mathbf{f}_2$ and $\boldsymbol{\varepsilon}_{2,1} = \mathbf{f}_2 - \rho_t \mathbf{f}_1$, calculate their correlation coefficient.

7.9 Derive Equation 7.48.

7.10 (a) Suppose the standardized residual factors are related to the original factor through $\tilde{\boldsymbol{\varepsilon}} = \mathbf{P} \cdot \mathbf{f}$. Prove that $\mathbf{IC}_{\tilde{\varepsilon}} = \mathbf{P} \cdot \mathbf{IC}$. (b) With averages and covariance matrix of residual ICs given by $\overline{\mathbf{IC}}_{\tilde{\varepsilon}} = \mathbf{P} \cdot \overline{\mathbf{IC}}$ and $\mathbf{\Sigma}_{IC_\varepsilon} = \mathbf{P}'\mathbf{\Sigma}_{IC}\mathbf{P}$, show that the optimal weights for the standardized residual factors are related to the optimal weights for the original factors by $\mathbf{v}_\varepsilon = \mathbf{P}^{-1}\mathbf{v}$.

APPENDIX

In this appendix, we prove that a multivariate linear regression can be decomposed into two separate regressions: one between independent variables and the other between a dependent variable and the residual of the first regression. This property is inherent to the multivariate regression.

A7.1 INVERSE OF A PARTITIONED MATRIX

We first present the following result for the inverse of a nonsingular matrix. Given a square matrix $\mathbf{\Sigma}$, we partition it as block matrix in which the diagonal blocks $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ are nonsingular square matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}. \tag{7.52}$$

Define

$$\begin{aligned} \mathbf{\Sigma}_{11,2} &= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \\ \mathbf{\Sigma}_{22,1} &= \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} \end{aligned}. \tag{7.53}$$

Then the inverse is given by

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{\Sigma}_{11,2}^{-1} & -\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22,1}^{-1} \\ -\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11,2}^{-1} & \mathbf{\Sigma}_{22,1}^{-1} \end{pmatrix}. \tag{7.54}$$

A7.2 DECOMPOSITION OF MULTIVARIATE REGRESSION

For a multivariate regression $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, the coefficient vector is given by $\boldsymbol{\beta} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}$. Suppose all variables have zero mean. The covariance matrix of independent variables $\mathbf{x}$ is $\Sigma = \left(\sigma_{ij}\right)_{i,j=1}^{K}$, the standard deviation of the dependent variable y is σ_y, and the correlations between the independent variables and the dependent variable are $\left(s_1, \cdots, s_K\right)$. Then the regression coefficient can be written as

$$\boldsymbol{\beta} = \Sigma^{-1}\mathbf{s}. \tag{7.55}$$

The vector $\mathbf{s}$ consists of covariances between the independent variables and the dependent variable, i.e.,

$$\mathbf{s} = \left(s_1\sigma_1\sigma_y, \cdots, s_K\sigma_K\sigma_y\right)'. \tag{7.56}$$

We partition the independent variables into

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix},$$

where $\mathbf{x}_1$ consists of k_1 factors and $\mathbf{x}_2$ consists of k_2 factors, and $k_1 + k_2 = k$. The coefficient vector $\boldsymbol{\beta}$ and the vector $\mathbf{s}$ can also be partitioned into

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \quad \mathbf{s} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix}.$$

The covariance matrix $\boldsymbol{\Sigma}$ can also be written as in (7.52), in which case $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are the covariance matrices for $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}'$ is the covariance matrix between $\mathbf{x}_1$ and $\mathbf{x}_2$. According to (7.55), we have

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{\Sigma}^{-1}\mathbf{s} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix}.$$

Using the inverse matrix (7.54) gives

$$\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11,2}^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22,1}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11,2}^{-1} & \boldsymbol{\Sigma}_{22,1}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix}.$$

We now focus our attention on the coefficient $\boldsymbol{\beta}_1$ and obtain

$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{11,2}^{-1}\mathbf{s}_1 - \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22,1}^{-1}\mathbf{s}_2 . \tag{7.57}$$

Next, we carry out the two-stage regression. First, we regress $\mathbf{x}_1$ against $\mathbf{x}_2$. As both dependent and independent variables are vectors in general, the regression coefficient is in fact a matrix in a form similar to (7.55) and it equals $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Hence, the residual of this regression is

$$\boldsymbol{\varepsilon}_{1,2} = \mathbf{x}_1 - \mathbf{x}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} . \tag{7.58}$$

The second regression is to regress y vs. the residual $\boldsymbol{\varepsilon}_{1,2}$. Denoting the regression coefficient by $\tilde{\boldsymbol{\beta}}_1$, we can write its solution in the same form as (7.55), with the covariance matrix being that of the residuals and the vector $\mathbf{s}$ being the covariances between y and the residuals; i.e.,

$$\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_{1,2}}^{-1} \operatorname{cov}\left(y, \boldsymbol{\varepsilon}_{1,2}\right) . \tag{7.59}$$

The covariance matrix of $\boldsymbol{\varepsilon}_{1,2}$ is

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_{1,2}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{11,2}. \tag{7.60}$$

The covariances between y and the residuals are

$$\operatorname{cov}\left(y, \boldsymbol{\varepsilon}_{1,2}\right) = \mathbf{s}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{s}_2. \tag{7.61}$$

Combining these, we have

$$\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\Sigma}_{11,2}^{-1}\mathbf{s}_1 - \boldsymbol{\Sigma}_{11,2}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{s}_2. \tag{7.62}$$

To prove $\boldsymbol{\beta}_1 = \tilde{\boldsymbol{\beta}}_1$ from Equation 7.57 and Equation 7.62, we need to prove that

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22,1}^{-1} = \boldsymbol{\Sigma}_{11,2}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1},$$

or

$$\boldsymbol{\Sigma}_{11,2}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22,1}.$$

Substituting (7.53) into the preceding matrices gives

$$\left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right).$$

Multiplying the matrices leads to an identity

$$\boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{7.63}$$

Equation 7.63 furnishes our proof for $\boldsymbol{\beta}_1 = \tilde{\boldsymbol{\beta}}_1$.

REFERENCES

Britten-Jones, M., Portfolio optimization and Bayesian regression, Q Group Conference, 1998.

Fama, E.F. and MacBeth, J.D., Risk, return and equilibrium: Empirical tests, *Journal of Political Economy,* Vol. 81, 607–636, 1973.

Goodwin, T.H., The information ratio, *Financial Analysts Journal*, Vol. 54, No. 4, 34–43, July–August 1998.

Grinold, R.C., The fundamental law of active management, *Journal of Portfolio Management*, Vol. 15, No. 3, 30–37, Spring 1989.

Grinold, R.C., Alpha is volatility time IC score or real alphas don't get eaten, *Journal of Portfolio Management*, Vol. 20, No. 4, 9–16, Summer 1994.

Keeping, E.S., *Introduction to Statistical Inference,* Dover Publications, 1995, New York.

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

Qian, E. and Hua, R., Active risk and information ratio, *Journal of Investment Management*, Vol. 2, No. 3, 20–34, 2004.

Sorensen, E.H., Qian, E., Hua, R., and Schoen, R., Multiple alpha sources and active management, *Journal of Portfolio Management,* Vol. 31, No. 2, 39–45, Winter 2004.

Yang, J., private communication, 2005.

Part III

CHAPTER 8

Portfolio Turnover and Optimal Alpha Model

The delivered value of an investment process relies on two parts: the theoretical value of the alpha skill (the gross paper profit) and the cost of implementation (the unrealized paper profit). The larger the former and the smaller the latter, the happier is the investor. Clearly, the total assets under management influence the latter. A strategy might be profitable with small assets under management and unprofitable with larger assets under management; as assets grow, transaction costs grow. Recently, Kahn and Shaffer (2005) pointed out that one remedy to the "size" problem is to reduce portfolio turnover. This is a sensible suggestion. However, their work is based on a hypothetical relationship between turnover and expected alpha that might be too general to be applicable.

In Chapter 7, we developed a framework to construct an optimal alpha model in the absence of transaction costs (Sorensen, Qian, Schoen, Hua 2004). In this chapter, we present an analytical extension to integrate alpha models with portfolio turnover. In practice, many alpha models are not constructed in such an integrated framework. Typically, managers adopt an alpha model first (with little consideration given to turnover) and then throw the list into an optimizer, setting turnover constraints to handle the transactions costs. There are two drawbacks to this two-step process: (1) it creates difficulty in knowing the true effectiveness of the alpha model, and (2) it does not allow managers to adjust the alpha model along the way as the assets under management grow.

The majority of implementation costs are related to trading. These costs could be exchange fees, broker commissions, bid/ask spread, and market

impact on prices when buying or selling stocks. We shall discuss these in detail in a later chapter. In general, the trading cost varies from stock to stock; for a given trade size, it is lower for large liquid stocks and higher for small illiquid stocks. On an aggregated portfolio level, the total cost should be proportional to the amount of trading or portfolio turnover. Therefore, as a first step to estimate transaction costs, we shall estimate portfolio turnover of different quantitative factors and their associated investment strategies. We then integrate both "paper" alpha as well as transaction costs into model construction by optimizing IR under various turnover constraints.

The issue of portfolio turnover is closely related to the information horizons of forecasts. If the information horizon of a factor is short, it only predicts returns within a short period after information about factor becomes known; then we need to update the information frequently and rebalance the portfolio, causing high portfolio turnover. On the other hand, if the information horizon of a factor is long, it has predictive power long after the factor became known; we only need to update the factor and rebalance the portfolio infrequently. The portfolio turnover associated with such factors will be low. Depending on the predictive power of different factors, the optimal alpha model may favor one kind of factors over another kind.

In this chapter, we first examine portfolio turnover of fixed-weight portfolios due only to rebalance. We then present a general discussion about the information horizon and derive an analytical formula for portfolio turnover conditioned on changes in forecasts.[1] This solution allows us to estimate portfolio turnover for different quantitative alpha factors and related investment strategies. We find that portfolio turnover can be endogenous in a complete system, and factor autocorrelation is a key exogenous ingredient. We then present an analytic framework for building an optimal alpha model with turnover constraints. In the final section of the chapter, we analyze the effect of bypassing small trades — a common practice by portfolio managers, on portfolio turnover and portfolio returns.

8.1 PASSIVE PORTFOLIO DRIFT

Weights of a passive or buy-and-hold portfolio would drift, purely due to price changes of the securities. Suppose the portfolio weights at the beginning of a period are $\mathbf{w} = \left(w_1, \cdots, w_N\right)'$ and they sum to one, i.e.,

$$\mathbf{i}' \cdot \mathbf{w} = \sum_{i=1}^{N} w_i = 1 .$$

Also, assume the returns for the period are $\mathbf{r} = \left(r_1, \cdots, r_N\right)'$. Then, the portfolio return for the period is

$$r_p = \mathbf{w}' \cdot \mathbf{r} = \sum_{i=1}^{N} w_i r_i \ .$$

The new portfolio weight is given by

$$w_i^d = \frac{w_i\left(1+r_i\right)}{1+r_p}, \quad i = 1, \cdots, N \ . \tag{8.1}$$

Compared to the old weights, the difference for a given stock is

$$\Delta w_i = w_i^d - w_i = \frac{w_i\left(1+r_i\right)}{1+r_p} - w_i = \frac{w_i\left(r_i - r_p\right)}{1+r_p} \ . \tag{8.2}$$

- When the weight of a stock is positive (a long position), it is easy to see that $\Delta w_i > 0$ if $r_i > r_p$ and $\Delta w_i < 0$ if $r_i < r_p$. In other words, the weight would drift higher (lower) if its return is higher (lower) than the portfolio return. On the other hand, if the weight of a stock is negative (a short position), the opposite is true: the weight would drift lower (higher) if its return is higher (lower) than the portfolio return. In essence, the winning long positions get longer, whereas the losing short positions get shorter.

Example 8.1

For a two-stock portfolio with equal weight of 50% each, suppose the returns are 10% and 20%, respectively. The portfolio return is then 15%. The new portfolio weights are

$$w_1^d = \frac{0.5\left(1+0.1\right)}{1.15} = 47.8\%, \ w_2^d = \frac{0.5\left(1+0.2\right)}{1.15} = 52.2\% \ .$$

Example 8.2

We have a long-short portfolio of two stocks, whose weights are 100% and –100%, respectively, relative to capital held in cash. Suppose the stocks' returns are 10% and 20%, respectively, and cash returns 2%. The portfolio return is

$$r_p = 100\% \cdot 10\% + 100\% \cdot \left(-20\%\right) + 100\% \cdot 2\% = -8\% .$$

The new weights are

$$w_1^d = \frac{100\%\left(1+10\%\right)}{1-8\%} = 119\%,$$

$$w_2^d = \frac{-100\%\left(1+20\%\right)}{1-8\%} = -130\%,$$

$$w_{cash}^d = \frac{100\%\left(1+2\%\right)}{1-8\%} = 111\%.$$

Note that when the portfolio return r_p is small, the change in weights is approximately

$$\Delta w_i \approx w_i\left(r_i - r_p\right). \tag{8.3}$$

8.2 TURNOVER OF FIXED-WEIGHT PORTFOLIOS

For fixed-weight portfolios, we try to maintain constant portfolio weights over time to correct the portfolio drift. The examples are equally weighted stock portfolios or fixed-weight stock/bond asset allocation portfolios. As we have shown, the weights of a portfolio would change due to the relative returns of the underlying components. Therefore, to maintain the fixed weights, the portfolio needs to be rebalanced periodically.

8.2.1 Turnover Definition

Let us first define portfolio turnover in terms of changes in portfolios weights. If the targeted weights are $\mathbf{w}^{new} = \left(w_1^{new}, \cdots, w_N^{new}\right)'$, and the current portfolio weights are $\mathbf{w}^{old} = \left(w_1^{old}, \cdots, w_N^{old}\right)'$, then the amount of turnover required to move the portfolio to the targeted weights is[2]

$$T = \frac{1}{2}\sum_{i=1}^{N}\left|w_i^{new} - w_i^{old}\right|. \tag{8.4}$$

If the new weight is greater than the current weight, i.e., $w_i^{new} > w_i^{old}$ we need to buy the difference $w_i^{new} - w_i^{old}$. On the other hand, if the new weight is less than the current weight, i.e., $w_i^{new} < w_i^{old}$ we need to sell by $w_i^{old} - w_i^{new}$. Because the amount of buying normally offsets the amount of selling, we divide the total sum of two to obtain the one-way turnover. In practice, some use the two-way turnover, which is double the one-way turnover.

Example 8.3

If we replace a long-only portfolio entirely by another portfolio of new securities, the turnover is 100% because

$$T = \frac{1}{2}\left(\sum w_i^{new} + \sum w_i^{old}\right) = 1, \text{ or } 100\%.$$

- In practice, the portfolio turnover, like other measures, is quoted on an annual basis. Intuitively, a portfolio with 100% turnover turns itself over in 1 year. In other words its average holding period for a stock is 1 year. A turnover of 200% implies the average holding period is 6 months, and a turnover of 50% implies the average holding period is 2 years.

Example 8.4

In Example 8.1, to get back to an equally weighted portfolio, we buy 2.2% of stock 1 and simultaneously sell 2.2% of stock 2. Thus, the one way turnover is 2.2%.

Example 8.5

In Example 8.2, to get back to the original leverage ratio of 100% long, 100% short, and 100% cash, we sell 19% of stock 1 and buy back or cover 30% of stock 2. The turnover is

$$T = \frac{1}{2}\left(19\% + 30\% + 11\%\right) = 30\%.$$

In this example, the amounts of buying and selling are not the same, because one of the portfolio holdings is cash. In fact, we should view the turnover as selling 19% of stock 1 and buying back 19% of stock 2 and buying back additional 11% of stock 2, i.e.,

$$T = \frac{1}{2}\left(19\% + 19\%\right) + 11\% = 30\%.$$

This yields the same answer. The general proof of this statement is left as an exercise.

8.2.2 Turnover due to Drift

For a fixed-weight portfolio, the turnover is solely due to portfolio rebalancing to correct the portfolio drift due to price movement. Therefore, combining Equation 8.4 and Equation 8.2, we have

$$T = \frac{1}{2}\sum_{i=1}^{N}\left|\Delta w_i\right| = \frac{1}{2\left(1+r_p\right)}\sum_{i=1}^{N}\left|w_i\left(r_i - r_p\right)\right|. \tag{8.5}$$

We first gain some insight by considering an equally weighted long-only portfolio, i.e., $w_i = 1/N$. Then,

$$T = \frac{1}{2\left(1+r_p\right)}\sum_{i=1}^{N}\left|\Delta w_i\right| = \frac{1}{2\left(1+r_p\right)N}\sum_{i=1}^{N}\left|r_i - r_p\right|. \tag{8.6}$$

The turnover is thus related to the average of absolute return differences between individual stocks and the portfolio. This is intuitive. When the returns are the same for all stocks, there is no drift of portfolio weights, and therefore there is no need to rebalance. When the return difference or dispersion is large, the drift of portfolio weights is large and leads to a high rebalancing turnover.

We further improve our results and understanding of portfolio turnover by obtaining an analytical approximation for (8.6). We assume stock return r forms a continuous distribution, for simplicity, a normal distribution, $r \sim N\left(\bar{r}, d^2\right)$, where $\bar{r}$ is the average return of stocks, and d is their dispersion. The individual stock returns r_i's are samples from this distribution. Then, the sample average of (8.6) is an approximation of the expectation

$$T = \frac{1}{2\left(1+r_p\right)N}\sum_{i=1}^{N}\left|r_i - r_p\right| \approx \frac{1}{2\left(1+r_p\right)}\mathrm{E}\left(\left|r - r_p\right|\right). \tag{8.7}$$

Note the average return and the portfolio return usually are not the same. However, for an equally weighted portfolio, we have $\bar{r} = r_p$, and therefore $\mathrm{E}\left(\left|r - r_p\right|\right) = \mathrm{E}\left(\left|r - \bar{r}\right|\right)$. Now, $r - \bar{r}$ is normally distributed with zero mean, the expectation of its absolute value can be evaluated analytically. We have (Problem 8.2)

$$\mathrm{E}\left(\left|r - \bar{r}\right|\right) = \sqrt{\frac{2}{\pi}} d \,. \tag{8.8}$$

The expected absolute return difference is the return dispersion times a constant. Combining (8.8) and (8.7) yields the turnover of equally weighted portfolio

$$T \approx \frac{d}{\left(1 + \bar{r}\right)\sqrt{2\pi}} \,. \tag{8.9}$$

The turnover for rebalancing the drift is directly proportional to the cross-sectional dispersion of stock returns during the rebalancing period. Furthermore, the turnover is inversely related to the average return of stocks: higher (lower) returns lead to lower (higher) turnover. However, the effect tends to be small unless the average return is significantly positive or negative.

Example 8.6

Suppose the average stock return is 2% and the dispersion is 15% for a 3-month period, then the turnover for a quarterly rebalanced of an equally weighted portfolio is about 5.9%. The annual turnover would be 23.5%.

8.2.3 More Results on Rebalance Turnover

Most portfolios encountered in practice are not equally weighted. Their weights are not only uneven, but can be both long and short. Furthermore, returns of most portfolios are not necessarily the same as average stock returns. We shall generalize (8.9) to derive rebalance turnover of more general portfolios.

To do so, we shall assume portfolio weights and subsequent returns are independent of each other. This assumption might be incorrect for active portfolios that consistently outperform their benchmark because outperforming implies a consistent positive correlation between the active

weights and the subsequent returns. However, this positive correlation is typically small and the effect on turnover is negligible. For portfolios with fixed weights, this is a reasonable assumption. When the weights and the returns are independent, we recast Equation 8.5 as an expectation of a product of two terms, which can be written as a product of two independent expectations, i.e.,

$$\begin{aligned} T &= \frac{N}{2\left(1+r_p\right)}\frac{1}{N}\sum_{i=1}^{N}\left|w_i\right|\left|\left(r_i-r_p\right)\right| \approx \frac{N}{2\left(1+r_p\right)}\mathrm{E}\left(\left|w\right|\left|\left(r-r_p\right)\right|\right) \\ &= \frac{N}{2\left(1+r_p\right)}\mathrm{E}\left(\left|w\right|\right)\mathrm{E}\left(\left|r-r_p\right|\right) = \frac{\sum_{i=1}^{N}\left|w_i\right|}{2\left(1+r_p\right)}\mathrm{E}\left(\left|r-r_p\right|\right) \end{aligned} \tag{8.10}$$

The expectation of the absolute value of weight is just the average of the absolute weights. For long-only portfolios, the weights are all positive, and the sum is 1. For long-short portfolios, the sum of absolute weights equates to portfolio leverage L. Hence,

$$T \approx \frac{L}{2\left(1+r_p\right)}\mathrm{E}\left(\left|r-r_p\right|\right). \tag{8.11}$$

- With $L = 1$ for long-only portfolios, Equation 8.11 is applicable to both long-only and long-short portfolios. The turnover is, therefore, directly proportional to the portfolio leverage. If a portfolio is 125% long and 25% short, the leverage is 150%. Therefore, the rebalance turnover would be 50% higher than a long-only portfolio with similar characteristics.

When the average stock return $\bar{r}$ differs from the portfolio return r_p, the expectation in (8.11) can still be derived using special functions. The derivation is given as an exercise (Problem 8.3). Using the result, we have

$$T \approx \frac{Ld}{\sqrt{2\pi}\left(1+r_p\right)}\left[1+\frac{\left(\Delta r\right)^2}{2d^2}\right]. \tag{8.12}$$

In (8.12), $\Delta r = r_p - \bar{r}$ is the difference between the portfolio return and the average stock return, L is the leverage of the portfolio, and d is the cross-sectional dispersion of the stock returns.

A notable difference between (8.12) and (8.9) is that any difference between the portfolio return and the average return contributes to higher turnover. The magnitude of the turnover increase depends on the ratio of the return difference to the stock-return dispersion. When the ratio is small, the increase in turnover is small. However, when the ratio is high, the increase in turnover could be significant. Thus, portfolios that either underperform or outperform the market average require higher turnover to rebalance to the original weights than a portfolio with average return.

8.3 TURNOVER DUE TO FORECAST CHANGE

So far, our results on rebalance turnover are derived for portfolios with fixed weights. Although these portfolios are not indexed portfolio, they are not actively managed either, and they tend to have low turnover compared to actively managed portfolios. For active portfolios that are actively managed with an alpha model, it is reasonable to assume that most of the portfolio turnover is caused by changes in the model forecasts, whereas portfolio drift plays a secondary role. Trading a portfolio according to the new model forecasts raises the expected return of the portfolio but also incurs transaction costs associated with portfolio turnover. It is important for managers to balance this trade-off. To do that, we need to know how much turnover is induced by forecast changes.

Consider turnover over a single trading period, in which the active weights change from w_i^t to w_i^{t+1}. We assume the new active weights for each security result from an unconstrained mean–variance optimization based on residual return and residual risk, respectively, at time t and $t + 1$ (from Chapter 4):

$$w_i^t = \frac{1}{\lambda_t}\frac{F_i^t}{\sigma_i}, \quad w_i^{t+1} = \frac{1}{\lambda_{t+1}}\frac{F_i^{t+1}}{\sigma_i}. \tag{8.13}$$

F_i^t and F_i^{t+1} are risk-adjusted forecasts at t and $t + 1$. For simplicity, we have assumed all stock-specific risks remain unchanged, and the number of stocks remains unchanged. If we hold constant the targeted tracking error σ_{model} for the portfolio, then the risk-aversion parameter is given by

$$\lambda_t = \frac{\sqrt{N-1}\,\mathrm{dis}\left(\mathbf{F}^t\right)}{\sigma_{\mathrm{model}}}, \text{ and } \lambda_{t+1} = \frac{\sqrt{N-1}\,\mathrm{dis}\left(\mathbf{F}^{t+1}\right)}{\sigma_{\mathrm{model}}}. \tag{8.14}$$

Substituting (8.14) into (8.13) gives

$$w_i^t = \frac{\sigma_{\mathrm{model}}}{\sqrt{N-1}} \frac{\tilde{F}_i^t}{\sigma_i}, \quad w_i^{t+1} = \frac{\sigma_{\mathrm{model}}}{\sqrt{N-1}} \frac{\tilde{F}_i^{t+1}}{\sigma_i}, \tag{8.15}$$

in which $\tilde{F}_i^t$ and $\tilde{F}_i^{t+1}$ are now standardized with $\mathrm{dis}\left(\tilde{\mathbf{F}}^t\right)=1$, $\mathrm{dis}\left(\tilde{\mathbf{F}}^{t+1}\right)=1$. In other words, they are merely *z*-scores. Note the following:

- During the period from t to $t+1$, the active weight would change to $\tilde{w}_i^t$ due to price movement, and turnover arises when we rebalance portfolio weights from $\tilde{w}_i^t$ to w_i^{t+1}. For the following calculation, we ignore the weight drift and calculate turnover solely due to forecast changes. In most cases, this is an excellent approximation of portfolio turnover, because the majority of the turnover is created by changes in the forecasts.

The portfolio turnover caused by forecast changes, according to definition (8.4), is

$$T = \frac{1}{2}\sum_{i=1}^{N}\left|w_i^{t+1} - w_i^t\right| = \frac{\sigma_{\mathrm{model}}}{2\sqrt{N-1}}\sum_{i=1}^{N}\frac{\left|\tilde{F}_i^{t+1} - \tilde{F}_i^t\right|}{\sigma_i}. \tag{8.16}$$

It is apparent that the turnover is linearly proportional to the target tracking error.

The most difficult aspect of analyzing turnover is dealing with the absolute value function. Our way to solve this problem is to approximate the turnover in Equation 8.16 as the expectation of the absolute difference of two continuous variables that underlie two sets of forecasts. We then rely on standard statistical theory to evaluate various expectations. To this end, we rewrite (8.16) as

$$T = \frac{\sigma_{\mathrm{model}}\sqrt{N}}{2}\frac{1}{N}\sum_{i=1}^{N}\frac{\left|\tilde{F}_i^{t+1} - \tilde{F}_i^t\right|}{\sigma_i} = \frac{\sigma_{\mathrm{model}}\sqrt{N}}{2}\mathrm{E}\left(\frac{\left|\tilde{F}^{t+1} - \tilde{F}^t\right|}{\sigma}\right). \tag{8.17}$$

In order to evaluate the expectation, we assume that the changes in the risk-adjusted forecast and the stock-specific risk are independent. Therefore, (8.17) can be written as

$$T = \frac{\sigma_{\text{model}}\sqrt{N}}{2}\mathrm{E}\left(\left|\tilde{F}^{t+1} - \tilde{F}^{t}\right|\right)\mathrm{E}\left(\frac{1}{\sigma}\right). \tag{8.18}$$

The second expectation can be evaluated as the average of the reciprocals of specific risks. It is immediately clear that the higher the specific risks, the lower the turnover. To evaluate the first expectation, we note that both sets of forecasts have a standard deviation of 1. We further assume they form a bivariate normal distribution with mean 0, and the cross-sectional correlation between the two sets of consecutive forecasts is ρ_f. This is simply the lag 1 autocorrelation of the risk-adjusted forecasts. When the forecast autocorrelation is high, then the change in forecasts is minimal, and the turnover should be low. Conversely, if the forecast autocorrelation is low, then the forecast change is significant, and the turnover will be high.

Because both forecasts are normally distributed, the change $\tilde{F}^{t+1} - \tilde{F}^{t}$ is still a normal distribution with 0 mean and standard deviation $\sqrt{2\left(1-\rho_f\right)}$. We have (Problem 8.2)

$$\mathrm{E}\left(\left|\tilde{F}^{t+1} - \tilde{F}^{t}\right|\right) = \frac{2\sqrt{1-\rho_f}}{\sqrt{\pi}}. \tag{8.19}$$

Substituting (8.19) into (8.18) yields

$$T = \sqrt{\frac{\mathrm{N}}{\pi}}\sigma_{\text{model}}\sqrt{1-\rho_f}\,\mathrm{E}\left(\frac{1}{\sigma}\right). \tag{8.20}$$

Equation 8.20 represents our solution for the forecast-induced turnover of an unconstrained long-short portfolio.[3] It depends on four elements. The turnover is higher:

- The higher the tracking error
- The larger the number of stocks (proportional to the square root of N)

- The lower the forecast autocorrelation (cross-sectional correlation between the consecutive forecasts), $\rho_f = \text{corr}\left(\tilde{F}^{t+1}, \tilde{F}^t\right)$
- The lower the average stock-specific risk

It confirms our intuitions regarding the impact of target tracking error and cross-sectional correlation between forecasts on the turnover. In addition, Equation 8.20 indicates that turnover is proportional to both the square root of N and the targeted tracking error. According to the results of Chapter 4, the paper excess return of a long-short portfolio is similarly proportional to the square root of breadth or N and the target tracking error. This would imply the net expected return also behaves as such.

Example 8.6

When stock-specific risks are the same for all stocks and equals σ_0, the turnover is reduced to

$$T = \sqrt{\frac{N}{\pi}} \frac{\sigma_{\text{model}}}{\sigma_0} \sqrt{1-\rho_f} \,. \tag{8.21}$$

For a long-short portfolio with $N = 500$, $\sigma_{\text{model}} = 5\%$, $\sigma_0 = 30\%$, and $\rho_f = 0.9$, the one-time turnover would be

$$T = \sqrt{\frac{500}{3.1415}} \frac{5\%}{30\%} \sqrt{1-0.9} = 66\% \,.$$

The forecast autocorrelation $\rho_f = \text{corr}\left(\tilde{F}^{t+1}, \tilde{F}^t\right)$ is most relevant for our analysis of turnover. There is considerable intuition behind this. If there were perfect correlation between the forecasts, then the weights are identical, and there is no turnover. When the correlation is not perfect there will be turnover, and at the other extreme: Turnover will be at the maximum if the correlation is –1. In this case, all weights flip signs, and the portfolio reverses itself. The dependence of turnover on the forecast autocorrelation is through function $\sqrt{1-\rho_f}$, which is plotted in Figure 8.1. We can see that the turnover is a decreasing function of forecast autocorrelation. The function behaves close to a linear function for most of the range, and it drops more precipitously when ρ_f is greater than 0.8.

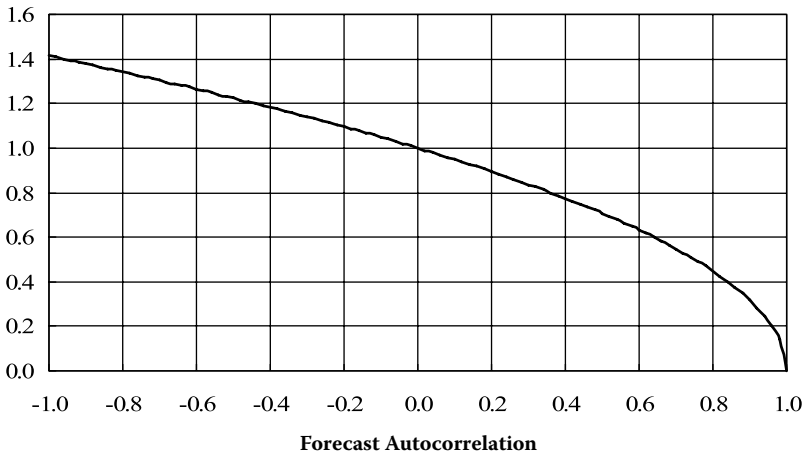


FIGURE 8.1. The dependence of turnover on the forecast autocorrelation.

8.3.1 Leverage and Turnover

Portfolio turnover is also a function of leverage: the higher the leverage, the higher the turnover. To derive the relationship between the two, we first obtain an analytic expression for the leverage. We have

$$\begin{aligned} L &= \sum_{i=1}^{N} \left| w_i \right| = \frac{\sigma_{\text{model}}}{\sqrt{N-1}} \sum_{i=1}^{N} \left| \frac{\tilde{F}_i^t}{\sigma_i} \right| \\ &= \sigma_{\text{model}} \sqrt{N} \mathrm{E}\left(\frac{\left| \tilde{F}^t \right|}{\sigma} \right) = \sigma_{\text{model}} \sqrt{N} \mathrm{E}\left(\left| \tilde{F}^t \right| \right) \mathrm{E}\left(\frac{1}{\sigma} \right) \end{aligned} . \tag{8.22}$$

Because $\tilde{F}^t$ is a standard normal variable, we have $\mathrm{E}\left(\left| \tilde{F}^t \right| \right) = \sqrt{2/\pi}$. Therefore,

$$L = \sqrt{\frac{2N}{\pi}} \sigma_{\text{model}} \mathrm{E}\left(\frac{1}{\sigma} \right). \tag{8.23}$$

Portfolio leverage is proportional to the target tracking error, the square root of N, and the average of the reciprocal of specific risks. Combining (8.23) and (8.20) yields

$$T = \frac{L\sqrt{1-\rho_f}}{\sqrt{2}}. \tag{8.24}$$

The turnover is directly proportional to the leverage. However, note the following:

- Because the turnover is proportional to leverage, it is certain that the transaction costs will increase linearly with leverage. For example, a market-neutral long-short portfolio with 4:1 leverage (200% long and 200% short) would have twice as much turnover as a portfolio with 2:1 leverage (100% long and 100% short).

8.3.2 Forecast Autocorrelations of Quantitative Factors

Table 8.1 shows the serial autocorrelation of a select group of quantitative factors. These factors are risk-adjusted, and we have neutralized all their exposures to the BARRA risk factors in the USE3 risk model. The details are given in Chapter 5. We report the average forecast autocorrelations between quarterly data. These factors fall into three broad categories: momentum, value, and quality. We observe that value factors, in general, have the highest forecast autocorrelation and thus the lowest turnover. Among the three value factors listed, the cash flow factor has the lowest autocorrelation, whereas the book-to-price and earning-to-price have very high autocorrelations.

The momentum factors have the lowest forecast autocorrelation, thus the highest turnover. Interestingly, the long-term growth revision has a very low autocorrelation, implying a short-term investment horizon for the factor. The 9-month price momentum factor and the 9-month earning momentum factor have the same level of autocorrelation, around 0.6. We also note that for price momentum factors, the autocorrelation increases as the time window used for return calculation lengthens up to 12 months.

TABLE 8.1 Summary Statistics of Forecast Autocorrelation of Quantitative Factors

Category	Factors	Avg(ρ_f)
Momentum	EarnRev9	0.64
	Ret9Monx1	0.60
	LtgRev9	0.37
Value	E2PFY0	0.96
	B2P	0.93
	CFO2EV	0.84
Quality	RNOA	0.89
	XF	0.76
	NCOinc	0.80

Therefore, one should use a longer time window to measure price momentum if the objective is to reduce turnover.

The quality factors have autocorrelations between that of value and momentum factors. Return on net operating assets (RNOA) has an autocorrelation of 0.89, whereas external financing (XF) has an autocorrelation of 0.76. The accrual factor or increase in net noncurrent assets NCOinc, has an autocorrelation of 0.80.

8.4 TURNOVER OF COMPOSITE FORECASTS

The preceding sections provide the relationship between the forecast-induced turnover and the forecast autocorrelation. Most alpha models consist of multiple factors. Therefore, to analyze turnover of a composite model, we start from the autocorrelation of composite forecasts, which depends on the autocorrelations of individual factors, as well as cross-correlation of different factors. By changing the model weights of the composite forecast, we not only change the information ratio (IR) of the composite forecast but also its autocorrelation and turnover. We shall study the autocorrelation here and later integrate it into the analysis of optimal information ratio.

8.4.1 Two-Factor Composite

In a two-factor case, the composite forecasts are linear combinations $\mathbf{F}_c = v_1\mathbf{F}_1 + v_2\mathbf{F}_2$, in which both $\mathbf{F}_1$ and $\mathbf{F}_2$ are standardized and v_1 and v_2 are weights. The autocorrelation of the composite factor is

$$\rho_{f_c} = \frac{\operatorname{cov}\left(\mathbf{F}_c^t, \mathbf{F}_c^{t+1}\right)}{\operatorname{std}\left(\mathbf{F}_c^t\right)\operatorname{std}\left(\mathbf{F}_c^{t+1}\right)}. \tag{8.25}$$

The standard deviation of the composite factors is

$$\operatorname{std}\left(\mathbf{F}_c^t\right) = \operatorname{std}\left(\mathbf{F}_c^{t+1}\right) = \sqrt{v_1^2 + v_2^2 + 2v_1v_2\rho_{12}^{t,t}}, \tag{8.26}$$

where $\rho_{1,2}^{t,t}$ is the contemporaneous correlation between the two factors. The covariance is

$$\operatorname{cov}\left(\mathbf{F}_c^t, \mathbf{F}_c^{t+1}\right) = v_1^2\rho_{11}^{t,t+1} + v_2^2\rho_{22}^{t,t+1} + v_1v_2\left(\rho_{12}^{t,t+1} + \rho_{21}^{t,t+1}\right), \tag{8.27}$$

where $\rho_{ij}^{t,t+1}$ is the correlation between $\mathbf{F}_i^t$ and $\mathbf{F}_j^{t+1}$. If we have $i = j$, $\rho_{ij}^{t,t+1}$ is serial autocorrelation of the same factor. If we have $i \neq j$, $\rho_{ij}^{t,t+1}$ is serial cross-correlation between two different factors. Hence, the autocorrelation of the composite factor is

$$\rho_{f_c} = \frac{v_1^2 \rho_{11}^{t,t+1} + v_2^2 \rho_{22}^{t,t+1} + v_1 v_2 \left(\rho_{12}^{t,t+1} + \rho_{21}^{t,t+1}\right)}{v_1^2 + v_2^2 + 2 v_1 v_2 \rho_{12}^{t,t}}. \tag{8.28}$$

- The autocorrelation of the composite factor depends on weights, as well as serial auto- and cross-correlation of factors. It can be seen that the autocorrelation of the composite factor will be high if the two factors have high serial auto- and cross-correlation, but low contemporaneous correlation. This would imply lower portfolio turnover for the composite forecast.

Example 8.7

Suppose the serial autocorrelations of two factors are $\rho_{11}^{t,t+1} = 0.8$ and $\rho_{22}^{t,t+1} = 0.9$, the serial cross-correlations are $\rho_{12}^{t,t+1} = 0.6$ and $\rho_{21}^{t,t+1} = 0.6$, and the contemporaneous correlation $\rho_{12}^{t,t} = 0.5$, then,

$$\rho_{f_c} = \frac{0.8 v_1^2 + 0.9 v_2^2 + 1.2 v_1 v_2}{v_1^2 + v_2^2 + v_1 v_2}.$$

For an equally weighted composite factor $v_1 = v_2 = 0.5$, the serial autocorrelation is 0.97, which is higher than both individual autocorrelations.

All the correlation coefficients can be put into a single correlation matrix — the correlation matrix for the stacked vector $\left(\mathbf{F}_1^{t+1}, \mathbf{F}_2^{t+1}, \mathbf{F}_1^t, \mathbf{F}_2^t\right)$:

$$\mathbf{C} = \begin{array}{c} \mathbf{F}_1^{t+1} \\ \mathbf{F}_2^{t+1} \\ \mathbf{F}_1^t \\ \mathbf{F}_2^t \end{array} \left| \begin{pmatrix} 1 & \rho_{12}^{t,t} & \rho_{11}^{t+1,t} & \rho_{12}^{t+1,t} \\ \rho_{12}^{t,t} & 1 & \rho_{21}^{t+1,t} & \rho_{22}^{t+1,t} \\ \rho_{11}^{t+1,t} & \rho_{21}^{t+1,t} & 1 & \rho_{12}^{t,t} \\ \rho_{12}^{t+1,t} & \rho_{22}^{t+1,t} & \rho_{12}^{t,t} & 1 \end{pmatrix} \right. . \tag{8.29}$$

We shall make use of this correlation matrix later in the chapter when we formulate the problem of optimizing IR under constraint of portfolio turnover constraint. The correlation matrix must be positive definite in general. Therefore, all correlations are not independent.

We shall assume the forecasts have stationary correlation structure, such that $\rho_{ij}^{t_1+s,t_2+s} = \rho_{ij}^{t_1,t_2}$.

8.4.2 Serial Autocorrelation of Moving Averages

When a time series signal is volatile, it can be smoothed using some types of moving averages. In our framework, moving averages can also be thought of as composite factors — a linear combination of new and past information. A natural question is, "why would we use outdated information in the forecasts?" One tends to think that a forecast based on the most recent information is better than the lagged forecast, in terms of more predictive power for subsequent returns, i.e., better IC or better IR. This may be true. However, if the market is not efficient, then there is no reason to believe that the inefficiency could only be exploited with the most recent information.

A second and more pertinent reason to use lagged forecast is that moving averages lead to higher serial autocorrelation and thus lower turnover. Despite possible information decay of lagged forecasts, the trade-off between lost paper profit and saving in transaction cost can lead us to include the lagged forecasts in the composite model.

We analyze the moving averages of forecasts in the same way as we analyzed composite forecasts. Given forecast series $\left(\mathbf{F}^t, \mathbf{F}^{t-1}, \mathbf{F}^{t-2}, \cdots\right)$, we form a moving average of order L as

$$\mathbf{F}_{ma}^t = \sum_{l=0}^{L-1} v_l \mathbf{F}^{t-l} . \tag{8.30}$$

For instance, if $L=2$ then $\mathbf{F}_{ma}^t = v_0\mathbf{F}^t + v_1\mathbf{F}^{t-1}$. The serial autocorrelation of $\mathbf{F}_{ma}^t$ is given by

$$\begin{aligned} \rho_{f_{ma}} &= \frac{\operatorname{cov}\left(v_0\mathbf{F}^t + v_1\mathbf{F}^{t-1}, v_0\mathbf{F}^{t+1} + v_1\mathbf{F}^t\right)}{\operatorname{var}\left(v_0\mathbf{F}^t + v_1\mathbf{F}^{t-1}\right)} \\ &= \frac{v_0v_1 + \left(v_0^2 + v_1^2\right)\rho_f\left(1\right) + v_0v_1\rho_f\left(2\right)}{v_0^2 + v_1^2 + 2v_0v_1\rho_f\left(1\right)} \end{aligned} . \tag{8.31}$$

We use $\rho_f\left(h\right)$ to denote the serial autocorrelation function of $\mathbf{F}^t$ with lag h and $\rho_f\left(0\right)=1$.

For given the serial autocorrelations $\rho_f\left(h\right), h=1,2$, the correlation of (8.31) is a function of the weights, v_0 and v_1. Because the correlation

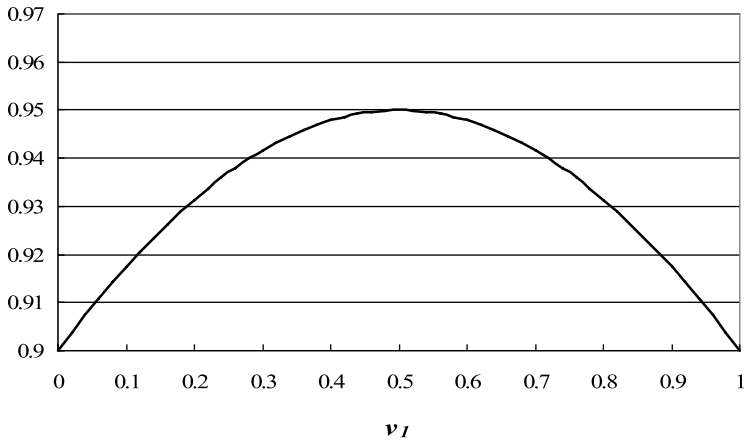


FIGURE 8.2. Serial autocorrelation of forecast moving average with $L = 2$, and $\rho_f(1) = 0.90$, and $\rho_f(2) = 0.81$.

is invariant to a scalar, we assume $v_0 + v_1 = 1$. Figure 8.2 shows a case in which the serial autocorrelation of the moving average is higher than the serial autocorrelation of the forecast itself. Therefore, using moving averages within an alpha model would reduce portfolio turnover. Figure 8.2 plots the correlation of (8.31) as a function v_1 — the weight of the lagged forecast for $\rho_f(1) = 0.90$, $\rho_f(2) = 0.81$. When $v_1 = 0$, the moving average is identical to the original forecast, so the serial autocorrelation is 0.9. As v_1 increases, the lagged forecast is added to the moving average, the serial autocorrelation of $\mathbf{F}_{ma}^t$ increases; it reaches a maximum of 0.95 at $v_1 = 0.5$, when the terms are equally weighted. As v_1 changes from 0.5 to 1, the autocorrelation declines from the maximum to 0.9.

Inclusion of lagged forecast would increase the serial autocorrelation as long as $\rho_f(2)$ is above a certain threshold. When $\rho_f(2)$ is below the threshold, the moving average would actually have a lower serial autocorrelation and thus higher turnover. The value of the threshold is (Problem 8.7)

$$\rho_f(2) = 2\left[\rho_f(1)\right]^2 - 1. \tag{8.32}$$

For example, when $\rho_f(1) = 0.90$, the threshold for $\rho_f(2)$ is 0.62. When $\rho_f(1) = 0.8$, the threshold for $\rho_f(2)$ is only 0.28. These values are easily exceeded for most factors encountered in practice. Thus, it can be concluded in general that using moving averages of forecasts should raise the serial autocorrelation and reduce portfolio turnover.

8.4.3 Composites of Moving Averages

The most general composite model would include moving averages of multiple factors. Putting the previous two sections together, we analyze the autocorrelation of composites of moving averages. The new composites have two dimensions of inclusion: factor dimension and time dimension. Assuming there are M factors, each of which has a moving average of order L, we write the composite as

$$\mathbf{F}_{c,ma}^{t} = \sum_{j=1}^{M} \sum_{l=0}^{L-1} v_{lj} \mathbf{F}_{j}^{t-l} . \tag{8.33}$$

An intuitive way to construct (8.33) is through a two-step process: The first step is to form a moving average for each factor, and the second step is to combine all moving averages together. For expository clarity, we consider the case of two factors and one lag, i.e.,

$$\mathbf{F}_{c,ma}^{t} = v_{01}\mathbf{F}_{1}^{t} + v_{02}\mathbf{F}_{2}^{t} + v_{11}\mathbf{F}_{1}^{t-1} + v_{12}\mathbf{F}_{2}^{t-1} . \tag{8.34}$$

It is still possible to calculate the serial autocorrelation of (8.34) algebraically as in the previous two cases, but the expression is more cumbersome. The autocorrelation can be written succinctly in terms of matrix multiplication. To this end, we denote the weights in (8.34) as a vector, $\mathbf{v} = \begin{pmatrix} v_{01} & v_{02} & v_{11} & v_{12} \end{pmatrix}'$. We consider the stacked vector $\left(\mathbf{F}_1^{t+1}, \mathbf{F}_2^{t+1}, \mathbf{F}_1^{t}, \mathbf{F}_2^{t}, \mathbf{F}_1^{t-1}, \mathbf{F}_2^{t-1}\right)$ and denote its correlation matrix as

$$\mathbf{C} = \begin{array}{c} \mathbf{F}_1^{t+1} \\ \mathbf{F}_2^{t+1} \\ \mathbf{F}_1^{t} \\ \mathbf{F}_2^{t} \\ \mathbf{F}_1^{t-1} \\ \mathbf{F}_2^{t-1} \end{array} \left| \begin{pmatrix} 1 & \rho_{12}^{0,0} & \rho_{11}^{1,0} & \rho_{12}^{1,0} & \rho_{11}^{2,0} & \rho_{12}^{2,0} \\ \rho_{12}^{0,0} & 1 & \rho_{21}^{1,0} & \rho_{22}^{1,0} & \rho_{21}^{2,0} & \rho_{22}^{2,0} \\ \rho_{11}^{1,0} & \rho_{21}^{1,0} & 1 & \rho_{12}^{0,0} & \rho_{11}^{1,0} & \rho_{12}^{1,0} \\ \rho_{12}^{1,0} & \rho_{22}^{1,0} & \rho_{12}^{0,0} & 1 & \rho_{21}^{1,0} & \rho_{22}^{1,0} \\ \rho_{11}^{2,0} & \rho_{21}^{2,0} & \rho_{11}^{1,0} & \rho_{21}^{1,0} & 1 & \rho_{12}^{0,0} \\ \rho_{12}^{2,0} & \rho_{22}^{2,0} & \rho_{12}^{1,0} & \rho_{22}^{1,0} & \rho_{12}^{0,0} & 1 \end{pmatrix} \right. . \tag{8.35}$$

In the matrix, the element is $\rho_{ij}^{l,k} = \mathrm{corr}\left(\mathbf{F}_i^{t+l}, \mathbf{F}_j^{t+k}\right)$. We next denote the 4×4 matrix in the upper-left corner of $\mathbf{C}$ as $\mathbf{C}_4$ and the 4×4 matrix in the upper-right corner of $\mathbf{C}$ as $\mathbf{D}_4$, i.e.,

$$\mathbf{C}_4 = \begin{pmatrix} 1 & \rho_{12}^{0,0} & \rho_{11}^{1,0} & \rho_{12}^{1,0} \\ \rho_{12}^{0,0} & 1 & \rho_{21}^{1,0} & \rho_{22}^{1,0} \\ \rho_{11}^{1,0} & \rho_{21}^{1,0} & 1 & \rho_{12}^{0,0} \\ \rho_{12}^{1,0} & \rho_{22}^{1,0} & \rho_{12}^{0,0} & 1 \end{pmatrix},$$

$$\mathbf{D}_4 = \begin{pmatrix} \rho_{11}^{1,0} & \rho_{12}^{1,0} & \rho_{11}^{2,0} & \rho_{12}^{2,0} \\ \rho_{21}^{1,0} & \rho_{22}^{1,0} & \rho_{21}^{2,0} & \rho_{22}^{2,0} \\ 1 & \rho_{12}^{0,0} & \rho_{11}^{1,0} & \rho_{12}^{1,0} \\ \rho_{12}^{0,0} & 1 & \rho_{21}^{1,0} & \rho_{22}^{1,0} \end{pmatrix}. \tag{8.36}$$

Then, the variance of $\mathbf{F}_{c,ma}^{t}$ is

$$\mathrm{var}\left(\mathbf{F}_{c,ma}^{t}\right) = \mathbf{v}' \cdot \mathbf{C}_4 \cdot \mathbf{v} \tag{8.37}$$

and the covariance

$$\mathrm{cov}\left(\mathbf{F}_{c,ma}^{t}, \mathbf{F}_{c,ma}^{t-1}\right) = \mathbf{v}' \cdot \mathbf{D}_4 \cdot \mathbf{v}\,. \tag{8.38}$$

Therefore, the serial autocorrelation of $\mathbf{F}_{c,ma}^{t}$ is

$$\rho_{f_{c,ma}} = \frac{\mathbf{v}' \cdot \mathbf{D}_4 \cdot \mathbf{v}}{\mathbf{v}' \cdot \mathbf{C}_4 \cdot \mathbf{v}}\,. \tag{8.39}$$

Equation 8.39 is the most general expression of the autocorrelation of a composite model with multiple factors and multiple lags, from which we can derive its corresponding portfolio turnover.

8.5 INFORMATION HORIZON AND LAGGED FORECASTS

The previous sections show that using moving averages of forecasts has the potential benefit of reducing portfolio turnover due to the increase in the serial autocorrelation of the forecasts. However, turnover reduction alone would not achieve the goal of delivering high risk-adjusted excess returns. We must also study their information content in terms of the information coefficient of lagged forecasts, i.e., lagged IC.

Another way of studying the information content of lagged forecasts is to look at the information horizon of a given forecast in terms of its IC

for different return horizons such as one month, three months, or longer, hereafter called the horizon IC. These two ICs are interrelated, as the following analysis shows.

8.5.1 Lagged IC

We denote the IC as the cross-sectional correlation coefficient between the factor value at the start of time t and the security return over time period t: $IC_{t,t} = \text{corr}\left(\mathbf{F}_t, \mathbf{R}_t\right)$. Consider this the standard IC measure. An example is the first quarter return IC. The factor values are observed December 31, and the return period is January to March.

The lagged IC is the correlation coefficient between time t factor values and a later period (lagged 1, 2, or more quarters) return vector, $IC_{t,t+l} = \text{corr}\left(\mathbf{F}_t, \mathbf{R}_{t+l}\right)$, with lag l. For example, using factor readings on December 31, we can correlate lagged returns for later periods (second quarter [$l = 1$]), third quarter [$l = 2$]), and so on. The IC will typically decay in power as the lag increases. The decay rate differs across different types of factors such as momentum and value. Typically, the ICs of momentum factors decay much faster than ICs of value factors.

8.5.2 Horizon IC

Another variant of the standard IC is the horizon IC. We define *horizon IC* as the IC of a factor at a given time, t, for subsequent returns over multiperiod horizons. For example, if we have factor values available at December 31, we are interested in its correlations with cumulative returns of next quarter, next two quarters, next three quarters, etc. We denote $\mathbf{R}_{t,t+h}$ as the risk-adjusted cumulative returns from period t to period $t+h$, horizon IC and denote $IC_t^h = \text{corr}\left(\mathbf{F}_t, \mathbf{R}_{t,t+h}\right), \quad h = 0,1,\cdots,H$ as the horizon IC. For example, IC_t^1 is the standard IC for the return in period t, and IC_t^2 is the correlation between the factor and the return vectors over the next six months (periods 1 and 2).

8.5.3 The Relationship between Lagged IC and Horizon IC

Although the lagged IC typically decays with the lag, the horizon IC often increases with the horizon, at least initially. We assume the cumulative multiperiod return in the horizon IC is related to the single-period return by $\mathbf{R}_{t,t+l} = \left(1+\mathbf{R}_t\right)\left(1+\mathbf{R}_{t+1}\right)\cdots\left(1+\mathbf{R}_{t+l}\right)-1$. When the periods returned are small, it can be approximated by $\mathbf{R}_{t,t+l} \approx \mathbf{R}_t + \mathbf{R}_{t+1} + \cdots + \mathbf{R}_{t+l}$. Using it in the horizon IC yields

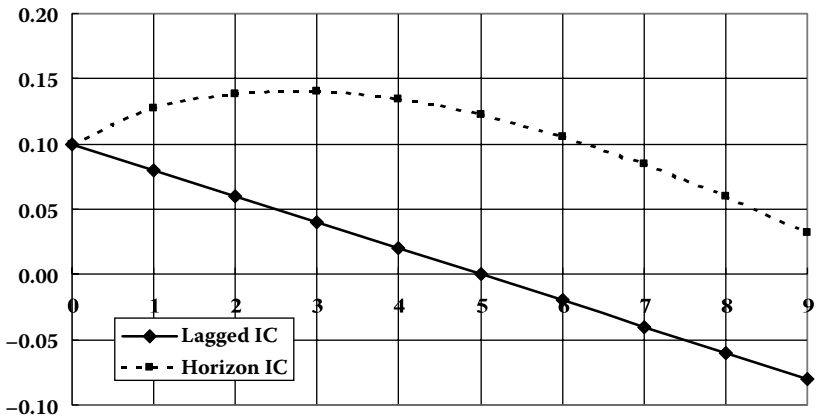


FIGURE 8.3. Lagged IC and horizon IC of a signal.

$$IC_t^l \approx \frac{\operatorname{cov}\left(\mathbf{F}_t, \mathbf{R}_t + \mathbf{R}_{t+1} + \cdots + \mathbf{R}_{t+l}\right)}{\operatorname{dis}\left(\mathbf{F}_t\right)\operatorname{dis}\left(\mathbf{R}_t + \mathbf{R}_{t+1} + \cdots + \mathbf{R}_{t+l}\right)}. \tag{8.40}$$

If we further assume that the risk-adjusted returns from different periods are uncorrelated,[4] then

$$IC_t^l \approx \frac{IC_{t,t} + IC_{t,t+1} + \cdots + IC_{t,t+l}}{\sqrt{l+1}} = \operatorname{avg}\left(IC\right)\sqrt{l+1}\,. \tag{8.41}$$

The horizon IC is an average of lagged ICs times the square root of the horizon length. Note that the horizon IC covers returns of multiple periods, and the lagged ICs cover forecasts of single intervals for future periods. Suppose there is no information decay in the lagged forecasts, i.e., the lagged ICs were the same as the IC with no lag, i.e., $IC_{t,t} = IC_{t,t+1} = \cdots = IC_{t,t+l}$. Then from Equation 8.41 we have $IC_t^l = IC_{t,t}\sqrt{l+1}$. In this case, the horizon IC is IC times the square root of the horizon length, and it therefore increases as the horizon lengthens.

Even when there is information decay, the horizon IC can still initially increase with the horizon length. It would then decline as the horizon lengthens further and the lagged IC declines more rapidly. Figure 8.3 plots one such case, in which the initial period IC is 0.10. The lagged IC is 0.08 with lag 1, 0.06 with lag 2, and so on. It reaches 0 with lag 5 and turns negative thereafter. The horizon IC increases at first. For example, the IC is 0.128 for returns over the next 2 periods and 0.139 for returns over the

next 3 periods. However, the horizon IC is eventually dragged down by the declining lagged ICs.

8.5.4 Horizon IC and the Trading Horizon

The propensity for the horizon IC to increase initially with the horizon does not necessarily mean that we can increase the total IC for a longer trading horizon. Longer trading horizons allow fewer opportunities to rebalance or fewer chances along the time dimension. Therefore, the horizon IC suffers from reduced breadth.

Example 8.9

Suppose both forecasts and returns are of quarterly frequency. The quarterly IC has a mean of 0.1 and a standard deviation of 0.2. Then, the quarterly IR is 0.5, and the annualized IR is $0.5\sqrt{4}=1$. Let us assume the lagged ICs with lag 1, 2, and 3 quarters all behave the same way as the regular IC, and they are all uncorrelated. Then, according to Equation 8.41, the horizon IC of 1 year, or 4 quarters, will have a mean of $4\cdot 0.1/\sqrt{4}=0.2$ and a standard deviation of $\sqrt{4}\cdot 0.2/\sqrt{4}=0.2$. Hence, the annual IR is also 1 — the same as the annualized IR of quarterly trading. There is no difference in terms of the performance. Note the following:

- This example highlights the importance of comparing horizon ICs with different horizons on the same-horizon basis. This can be achieved by simply comparing the horizon IC divided by $\sqrt{l+1}$ the square root of the horizon length. We call this the effective IC for the given horizon. In Example 8.9, the effective IC of the quarterly and annual horizon are the same.

Even though the annualized IR of the quarterly and annual rebalance is identical in this case, the amount of portfolio turnover can be different. In the former case, we trade four times per year so the total portfolio turnover is four times the quarterly turnover. In the latter, we only trade once a year. The question is, "which has less total turnover?"

It is easy to compare the turnover of the two cases using the results derived earlier. According to (8.20), the turnover is proportional to $\sqrt{1-\rho_f}$, in which ρ_f is the serial autocorrelation of the forecasts between trades. Denote the autocorrelation function of the forecast by $\rho_f(h)$. Then, the total turnover for quarterly trading is proportional to $4\sqrt{1-\rho_f(1)}$, whereas the total turnover for annual trading is proportional to $\sqrt{1-\rho_f(4)}$. For instance, if $\rho_f(1)=0.9$ and $\rho_f(4)=0.9^4=0.66$, then

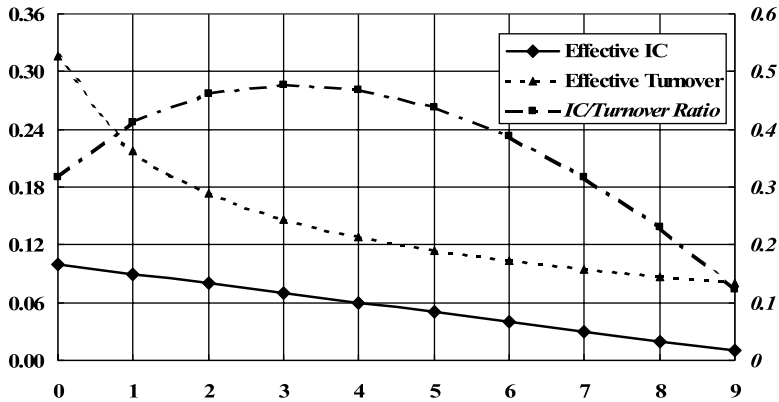


FIGURE 8.4. Effective IC, effective turnover, and their ratio.

$$4\sqrt{1-\rho_f(1)} = 1.26 \text{ and } \sqrt{1-\rho_f(4)} = 0.59 .$$

Under these assumptions, the turnover of annual trading is less than half the turnover of quarterly trading.

We define a ratio of effective IC to effective turnover for a given horizon as

$$\begin{aligned} Q_{IC,T} &= \frac{\text{Effective } IC}{\text{Effective Turnover}} \\ &= \frac{IC_t^l \big/ \sqrt{l+1}}{\sqrt{1-\rho(l+1)} \big/ (l+1)} = \frac{IC_t^l \sqrt{l+1}}{\sqrt{1-\rho(l+1)}} \end{aligned} \tag{8.42}$$

The effective IC is adjusted for trading opportunity, and the effective turnover is the turnover per unit period.

Figure 8.4 plots the effective IC based on the data in Figure 8.3. It declines linearly as the horizon extends. We also plot the effective turnover, assuming the autocorrelation function of the forecast is $\rho(h) = [\rho(1)]^h$ and $\rho(1) = 0.9$. The effective turnover drops rather rapidly at first and then declines steadily as the horizon extends further. As a result, the IC/turnover ratio (scale on the right axis) first increases as the trading horizon extends from one quarter to the second and third quarters. Then, it starts to decrease as the horizon extends beyond four quarters. Note the following:

- The effective IC/turnover ratio provides one convenient way to estimate the trade-off between paper alpha and trading cost for different trading horizons, once the horizon IC and the autocorrelations of the forecast are calculated. We caution that in practice, one should not use it to obtain the optimal rebalance horizon. The ratio itself doesn't reflect the true economic benefit or cost. In practice, the rebalance horizon is often determined by the flow of market and company information (e.g., see Chapter 10).

8.6 OPTIMAL ALPHA MODEL UNDER TURNOVER CONSTRAINTS

The prior analyses on the portfolio turnover due to forecast change and on lagged and horizon ICs provide the foundation for building optimal alpha models under a turnover constraint. The key insight is that one should use lagged forecasts as part of an alpha model, even if the lagged ICs might be weaker than the current ICs, because including lagged forecasts increases forecast autocorrelation and thus lowers the portfolio turnover.

The trade-off between the lagged IC and the forecast autocorrelation determines how much weight an alpha model has in the lagged forecasts. For instance, value factors often have little information decay — the past information is as good as new. In this case, we can assign substantial weights to the lagged value factors. On the other hand, momentum factors tend to lose their luster after a couple of periods. We would need to update them more frequently, and hence assign less weight to the lagged momentum factors.

The constrained optimization, however, lacks an analytical solution. Therefore, we use a numerical solution to derive optimal weights for the factor model.

8.6.1 Constrained Optimization

For expository clarity, we again consider the case of two factors and one lag. The following equation (same as Equation 8.34) would describe an alpha model based on the two factors and their lagged values $\mathbf{F}_{c,ma}^{t} = v_{01}\mathbf{F}_1^t + v_{02}\mathbf{F}_2^t + v_{11}\mathbf{F}_1^{t-1} + v_{12}\mathbf{F}_2^{t-1}$. The autocorrelation of the composite is given by Equation 8.39

$$\rho_{f_{c,ma}} = \frac{\mathbf{v}' \cdot \mathbf{D}_4 \cdot \mathbf{v}}{\mathbf{v}' \cdot \mathbf{C}_4 \cdot \mathbf{v}},$$

where the matrices $\mathbf{C}_4$ and $\mathbf{D}_4$ are defined in (8.36). We shall express the turnover constraint as an equality constraint on the forecast autocorrelation, because we have proven forecast-induced turnover is a function of $\rho_{f_{c,ma}}$, provided the target tracking error, the number of stocks, and the stock-specific risks are given.

The objective is to maximize the IR of the alpha model, which is approximated by the ratio of average IC to the standard deviation of IC. Denote the average IC of $\left(\mathbf{F}_1^t, \mathbf{F}_2^t, \mathbf{F}_1^{t-1}, \mathbf{F}_2^{t-1}\right)$ by $\overline{\mathbf{IC}}$ and the IC covariance matrix by Σ_{IC}, the optimization problem is

$$\begin{aligned} &\text{Maximize: } \mathrm{IR} = \frac{\mathbf{v}' \cdot \overline{\mathbf{IC}}}{\sqrt{\mathbf{v}' \cdot \mathbf{\Sigma}_{IC} \cdot \mathbf{v}}} \\ &\text{subject to: } \rho_{f_{c,ma}} = \frac{\mathbf{v}' \cdot \mathbf{D}_4 \cdot \mathbf{v}}{\mathbf{v}' \cdot \mathbf{C}_4 \cdot \mathbf{v}} = \rho_t \end{aligned} \tag{8.43}$$

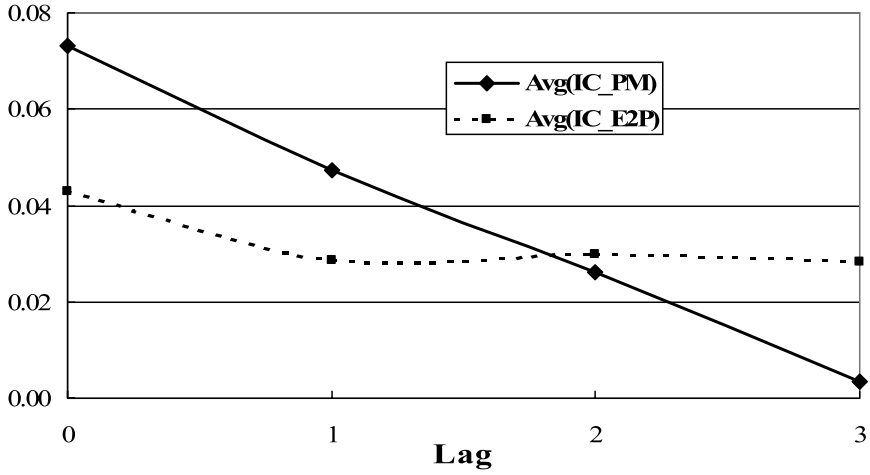
The target autocorrelation is denoted by ρ_t, which we shall vary in different optimization runs. The autocorrelation constraint is quadratic in nature. Thus, (8.43) is a nonlinear optimization with a quadratic constraint, which does not seem to have an analytic solution. However, it is easy to solve with numerical means, and we shall do so in the following example. We note that the problem can be extended to include more factors and multiple lags.

8.6.2 A Numerical Example: The Inputs

We present a numerical example of an optimal alpha model with turnover constraint, using two factors. The first factor mimics a momentum factor in that the IR is high with no lag but decays quickly over time and is based on the 9-month price momentum excluding the last month (Ret9Monx1). The second factor mimics a value factor in that the IR starts out low but decreases very slowly as the lag increases and is based on the earning-to-price ratio of the current fiscal year (E2PFY0) on a sector-relative basis.

Figure 8.5 depicts their behavior in terms of average IC, standard deviation of IC, and IR. We use PM to denote the price momentum factor and E2P to denote the earning yield factor. These sample ICs are derived from the universe of Russell 3000 stocks from 1987 to 2004. Figure 8.5 extends to 3 lags, which, with quarterly data, corresponds to factor values 3 quarters or 9 months ago. From Figure 8.5a, we observe that the average IC of the momentum factor is high when there is no lag, but it decreases

(A) Average IC



(B) Standard Deviation of IC

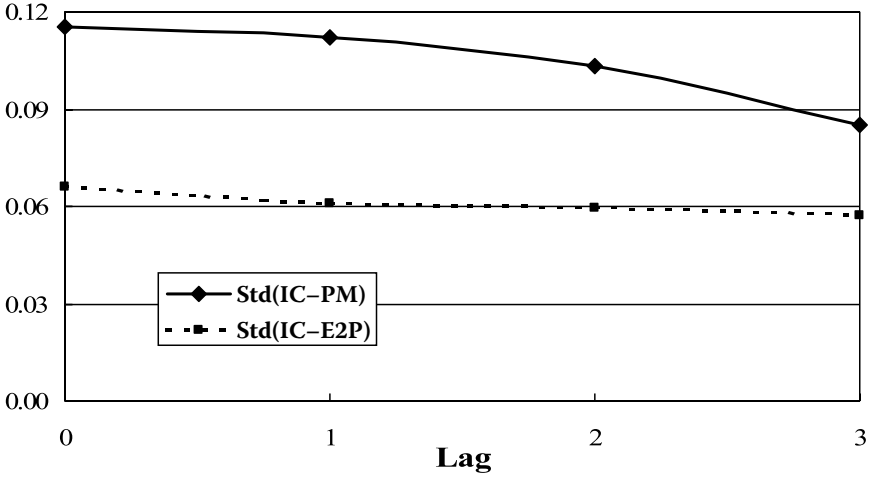


FIGURE 8.5. Average IC, standard deviation of IC, and IR for the price momentum and earning yield factor and their lagged factors: (a) average IC, (b) standard deviation of IC, and (c) IR of IC.

(C) IR of IC

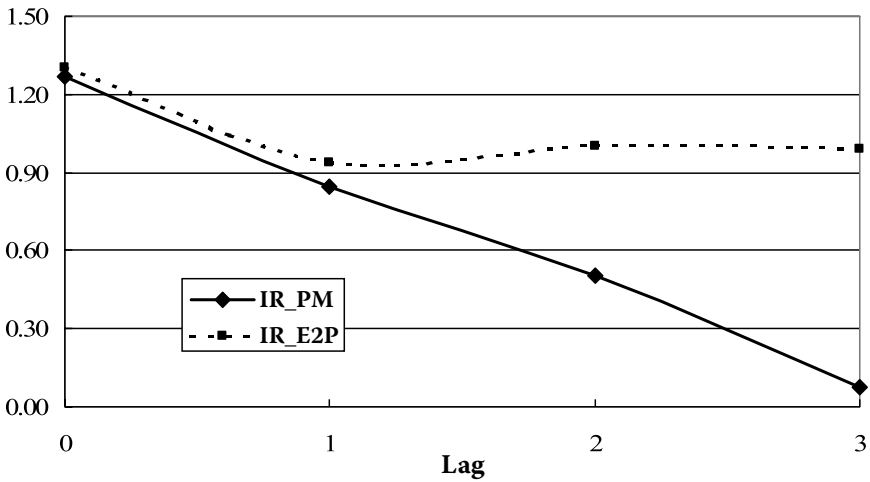


FIGURE 8.5 (continued).

linearly with a rapid rate. When the lag reaches three, the lagged IC is essentially zero, i.e., the momentum factor 9 months ago has no information for next quarter's returns. In contrast, the average IC of the value factor is lower when there is no lag, but it only drops slightly when the lag is one and remains at the same level as the lag increases further. There is little information decay for this value factor, and this remains true when the lag goes beyond three. Our example illustrates the drastically different behavior of the information content of these two factors. Figure 8.5b shows the standard deviations of ICs are relatively stable with respect to the lag for both factors. However, the standard deviation of IC is higher for the momentum factor. Figure 8.5c plots the annualized IR in terms of the ratio of average IC to the standard deviation of IC. As expected, it follows the pattern of average IC.

Figure 8.5 shows that both factors with current value, values from 3, 6, and 9 months ago, all have predictability for returns over the next 3 months. Thus, with two factors and three lags, we have eight different sources of alpha. To compute the IR of a composite model, in addition to the ICs of individual components, we also need IC correlations between them. Table 8.2 provides the correlation matrix of the eight alpha sources. The subscripted numbers denote lags. As we noted in the last chapter, the

TABLE 8.2 The IC Correlation Matrix of Current and Lagged Values for the Price Momentum and Earning Yield Factor

	PM_0	E2P_0	PM_1	E2P_1	PM_2	E2P_2	PM_3	E2P_3
PM_0	1.00	−0.42	0.86	−0.37	0.78	−0.26	0.61	−0.19
E2P_0	−0.42	1.00	−0.44	0.92	−0.31	0.84	−0.29	0.78
PM_1	0.86	−0.44	1.00	−0.45	0.88	−0.36	0.71	−0.30
E2P_1	−0.37	0.92	−0.45	1.00	−0.33	0.94	−0.30	0.86
PM_2	0.78	−0.31	0.88	−0.33	1.00	−0.28	0.83	−0.22
E2P_2	−0.26	0.84	−0.36	0.94	−0.28	1.00	−0.28	0.94
PM_3	0.61	−0.29	0.71	−0.30	0.83	−0.28	1.00	−0.30
E2P_3	−0.19	0.78	−0.30	0.86	−0.22	0.94	−0.30	1.00

momentum factor and value factors tend to have a negative IC correlation, a fact again reflected in the table. For instance, the ICs of PM_0 and E2P_0 have a correlation of -0.42, indicating significant diversification benefit. The diversification extends to the ICs of the lagged forecasts. For example, the ICs of PM_1 and E2P_1 have a correlation of -0.45, and the ICs of PM_0 and E2P_1 have a correlation of -0.37. The IC correlations among the same factors but of different lags are high, indicating less diversification of information. However, note that the correlation drops as the time span increases between the forecasts. For instance, for the PM factor, the correlation is 0.86 between PM_0 and PM_1, 0.78 between PM_0 and PM_2, and 0.61 between PM_0 and PM_3. For the value factor, the correlations are even higher, 0.92 between E2P_0 and E2P_1, 0.84 between E2P_0 and E2P_2, and 0.78 between E2P_0 and E2P_3.

To compute the autocorrelation of a composite factor, we need to specify the factor correlation matrix between factors of different lags, i.e., the matrix **C**. It is displayed in Table 8.3. Notice there are four lags in Table 8.3. This is because we need to consider autocorrelation (with one lag) of forecasts that are made of factors of three lags. We note that correlations among the same factor having different lags are high, with E2P in particular. This is not surprising because high serial autocorrelation of value factors is consistent with their minimal information decay. These values are much smaller for the PM factor: the lag 1 correlation is 0.68, and the lag 2 correlation is 0.40. However, the lag 3 and lag 4 correlation drop nearly to zero. These values indicate that the PM factor can bring more turnover than the value factor, even though its IR is higher. Lastly, we note the correlations between PM and E2P of different lags are small and significantly different from their IC correlations.

TABLE 8.3 The Factor Correlation Matrix of Current and Lagged Values for the Price Momentum and Earning Yield Factor

	PM_0	E2P_0	PM_1	E2P_1	PM_2	E2P_2	PM_3	E2P_3	PM_4	E2P_4
PM_0	1.00	−0.08	0.68	0.00	0.40	0.05	0.09	0.08	0.07	0.09
E2P_0	−0.08	1.00	−0.09	0.94	−0.06	0.84	0.01	0.73	0.03	0.61
PM_1	0.68	−0.09	1.00	−0.08	0.68	0.00	0.40	0.05	0.09	0.08
E2P_1	0.00	0.94	−0.08	1.00	−0.09	0.94	−0.06	0.84	0.01	0.73
PM_2	0.40	−0.06	0.68	−0.09	1.00	−0.08	0.68	0.00	0.40	0.05
E2P_2	0.05	0.84	0.00	0.94	−0.08	1.00	−0.09	0.94	−0.06	0.84
PM_3	0.09	0.01	0.40	−0.06	0.68	−0.09	1.00	−0.08	0.68	0.00
E2P_3	0.08	0.73	0.05	0.84	0.00	0.94	−0.08	1.00	−0.09	0.94
PM_4	0.07	0.03	0.09	0.01	0.40	−0.06	0.68	−0.09	1.00	−0.08
E2P_4	0.09	0.61	0.08	0.73	0.05	0.84	0.00	0.94	−0.08	1.00

TABLE 8.4 The Optimal Weights of the Composite Model for Different Levels of Autocorrelation and Their Optimal IR

ρ_f	IR	PM_0	E2P_0	PM_1	E2P_1	PM_2	E2P_2	PM_3	E2P_3
0.85	2.30	45%	55%	0%	0%	0%	0%	0%	0%
0.86	2.33	43%	57%	0%	0%	0%	0%	0%	0%
0.87	2.36	41%	59%	0%	0%	0%	0%	0%	0%
0.88	2.38	39%	61%	0%	0%	0%	0%	0%	0%
0.89	**2.39**	36%	64%	0%	0%	0%	0%	0%	0%
0.90	2.38	34%	65%	2%	0%	0%	0%	0%	0%
0.91	2.37	31%	65%	4%	0%	0%	0%	0%	0%
0.92	2.36	28%	65%	7%	0%	0%	0%	0%	0%
0.93	2.33	24%	65%	10%	0%	0%	0%	0%	1%
0.94	2.28	21%	58%	12%	4%	0%	1%	0%	4%
0.95	2.21	18%	50%	12%	8%	0%	4%	0%	8%
0.96	2.09	15%	42%	11%	10%	2%	7%	2%	10%
0.97	1.88	11%	32%	8%	14%	5%	12%	5%	14%

8.6.3 A Numerical Example: The Results

Given the inputs, we solve the optimization problem (8.43) for a series of forecast autocorrelations, ranging from 0.85 to 0.97. Note that the autocorrelation of PM is 0.68, and the autocorrelation of E2P is 0.94. The optimal weights for each autocorrelation target ρ_f together with the corresponding IR are presented in Table 8.4.

Note that as ρ_f goes from 0.85 to 0.97, the optimal IR first increases from 2.30 to 2.39 and then decreases to 1.88 when ρ_f reaches 0.97. The highest IR is when the autocorrelation is at 0.89 and the optimal weights

TABLE 8.5 The Aggregated Optimal Weights of the Composite Model with Autocorrelation Targets and Associated IRs

ρ_f	IR	PM	E2P	w_0	w_1	w_2	w_3
0.85	2.30	45%	55%	100%	0%	0%	0%
0.86	2.33	43%	57%	100%	0%	0%	0%
0.87	2.36	41%	59%	100%	0%	0%	0%
0.88	2.38	39%	61%	100%	0%	0%	0%
0.89	2.39	36%	64%	100%	0%	0%	0%
0.90	2.38	35%	65%	98%	2%	0%	0%
0.91	2.37	35%	65%	96%	4%	0%	0%
0.92	2.36	35%	65%	93%	7%	0%	0%
0.93	2.33	34%	66%	88%	10%	0%	1%
0.94	2.28	33%	67%	79%	15%	1%	4%
0.95	2.21	30%	70%	68%	20%	4%	8%
0.96	2.09	30%	70%	57%	21%	9%	13%
0.97	1.88	28%	72%	42%	23%	16%	19%

are 36% PM_0 and 64% E2P_0 with no lagged factors. We remark that this is the unconstrained model because it has the maximum IR. When the autocorrelation target is below 0.9, optimal weights do not contain any lagged factors. When the autocorrelation target is at 0.9 and above, the lagged factors join the optimal model, whereas the weights of PM_0 and of E2P_0 decline. PM_1 is the first lagged forecast to get into the model, and it is followed by E2P_1, E2P_2, and E2P_3. The other two lagged-momentum factors, PM_2 and PM_3, never obtain any significant weight in the model. This is consistent with the information input, because PM_2 and PM_3 have both low IC and low autocorrelation with PM_0. In contrast, all E2P factors have consistent IC and high autocorrelation.

We also assess the aggregated effect of forecast autocorrelation constraints on the factor level and on individual lags. We aggregate Table 8.4 into PM and E2P and into lags of 0, 1, 2, and 3, and show the results in Table 8.5. We see that as ρ_f increases from 0.85 to 0.97, the PM weight decreases from 45 to 28%, whereas the E2P weight increases from 55 to 72%. Meanwhile, the weight with no lag decreases from 100 to 42%, offset by increases in the weights of the lagged factors, first, factors with one lag and, then, factors with two and three lags. However, note the following:

- Although the maximum IR occurs when ρ_f is at 0.89 and the associated optimal model weights include no lagged factors, the model IR declines very little as ρ_f increases. For example, when ρ_f is at

0.93, the model IR is 2.33 vs. the maximum of 2.39. The small fall in the IR implies only a slight drop of the expected alpha, whereas the increase of autocorrelation could lead to much less turnover and thus less transaction cost.

To see the effect of autocorrelation on both the IR and turnover, we calculate the latter, on an annual basis, for a long-short portfolio with $N = 3000$, target risk $\sigma_{\text{model}} = 4\%$, and stock-specific risk $\sigma_0 = 30\%$ according to (8.21). The results are graphed in Figure 8.6. First, note the extremely high turnover when the autocorrelation is low; it is nearly 550% when ρ_f is 0.89. However, the most important feature of the graph is in the different rates of decrease for the IR and turnover as ρ_f increases. Although the turnover drops consistently, the IR changes rather slowly except when the autocorrelation reaches a very high level. Note the following:

- Because the turnover drops more rapidly than the IR, it is easy to see that the maximum net expected return might be achieved with an alpha model at a higher autocorrelation, not at $\rho_f = 0.89$. At higher autocorrelations, we would be likely to include lagged factors in the model.

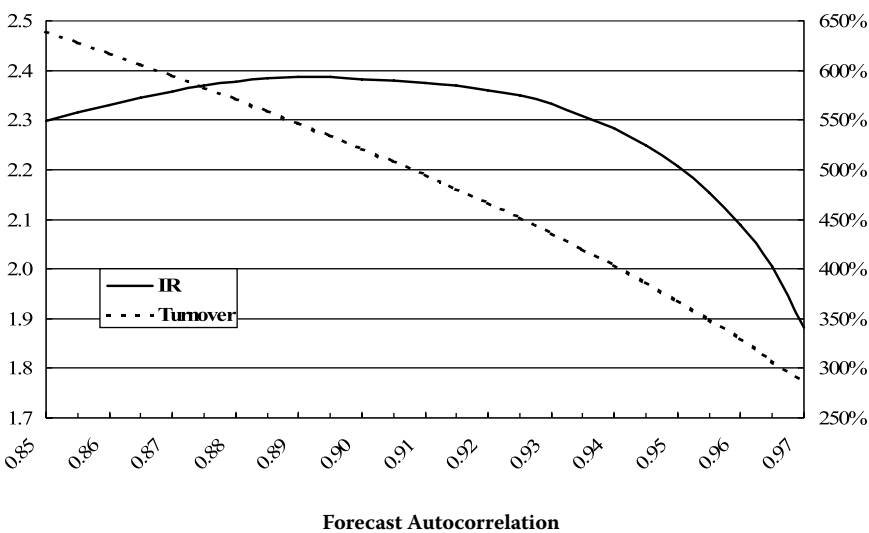


FIGURE 8.6. The IR and portfolio turnover of optimal alpha models with given forecast autocorrelation. The IR scale is on the left axis, and the turnover scale is on the right axis.

TABLE 8.6 The Gross Excess Return and Net Excess Returns under Different Transaction Cost Assumptions for Portfolios

ρ_f	IR	Gross Return	Turnover	Net Return (0.5%)	Net Return (1.0%)	Net Return (1.5%)
0.85	2.30	9.19%	638%	6.00%	2.81%	–0.38%
0.86	2.33	9.32%	617%	6.24%	3.15%	0.07%
0.87	2.36	9.43%	594%	6.46%	3.49%	0.52%
0.88	2.38	9.51%	571%	6.66%	3.80%	0.95%
0.89	2.39	9.55%	547%	6.81%	4.08%	1.35%
0.90	2.38	9.53%	521%	6.93%	4.32%	1.71%
0.91	2.37	9.50%	494%	7.03%	4.56%	2.08%
0.92	2.36	9.44%	466%	7.11%	4.78%	2.45%
0.93	2.33	9.33%	436%	7.15%	4.97%	2.79%
0.94	2.28	9.13%	404%	7.11%	5.09%	3.07%
0.95	2.21	8.83%	369%	6.98%	5.14%	3.30%
0.96	2.09	8.35%	330%	6.70%	5.06%	3.41%
0.97	1.88	7.53%	285%	6.10%	4.68%	3.25%

Note: $N = 3000$, target risk $\sigma_{\text{model}} = 4\%$, and stock-specific risk $\sigma_0 = 30\%$.

To examine explicitly the trade-off between a lower IR and a lower portfolio turnover at higher forecast autocorrelations, we compute the net expected return by imposing different levels of transaction costs. We assume the transaction is a linear proportion of the portfolio turnover. For example, at 50 basis points (bps) or 0.5%, a turnover of 100% would cost us 0.5% of excess return, and a turnover of 200% would cost us 1% of excess return. Table 8.6 lists the gross returns given by the IR times the target tracking error, turnover, and net returns with different transaction cost assumptions.

As expected, the gross return is maximized at $\rho_f = 0.89$, where the IR is at the maximum. However, the net return attains its maximum at higher ρ_f. When the cost is 0.5%, the maximum net return of 7.15% is at $\rho_f = 0.93$, where the gross IR is 2.33 but the turnover drops to 436% from 547%. This model outperforms the model with $\rho_f = 0.89$ by 34 bps per year. When the transaction cost is higher at 1.0%, the maximum net return of 5.14% is at $\rho_f = 0.95$, where the paper IR is 2.21 but the turnover further reduces to 369%. This model outperforms the model with $\rho_f = 0.89$ by 106 bps per year. At 1.5% cost for 100% turnover, the optimal model for net return of 3.41% would be at $\rho_f = 0.96$. This model outperforms the model with $\rho_f = 0.89$ by over 200 bps per year. Alpha models with these autocorrelations would include significant weights of lagged factors (see Table 8.4). Note the following:

- The net return and the optimal model is sensitive to the IR assumption. If the IR is lower than those in the example, then for a given level of cost, the maximum net return is achieved with models with even higher ρ_f. In other words, when the information content of the factors is lower, we need to pay even more attention to reduce portfolio turnover to reduce transaction costs.[5] This inevitably leads to more weight in the lagged factors, especially lagged value factors.

We plot in Figure 8.7 the return data: the gross return, and the net return with three transaction cost assumptions from Table 8.6. The square on each curve denotes the maximum return. As the transaction cost increases, the net return gets lower and lower. This is especially true for the left side of the return curves because of higher turnover. The right side of the curves drops to a lesser extent because the turnover is lower. As a result, the point of maximum net return shifts to the right. Another feature of the graph is that, when the transaction cost is high enough, optimal models with low autocorrelations or high turnover can have negative net returns. In contrast, optimal models with high autocorrelation have a better chance to yield positive net returns.

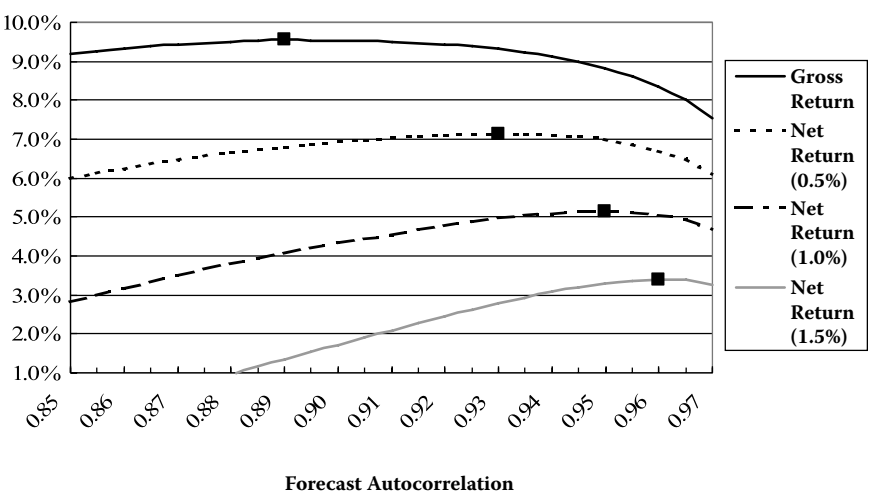


FIGURE 8.7. The gross excess return and net excess returns under different transaction cost assumption for portfolios with N = 3000, target risk $\sigma_{\text{model}} = 4\%$, and stock-specific risk $\sigma_0 = 30\%$.

8.7 SMALL TRADES AND TURNOVER

The discussion so far in this chapter assumes that all trades suggested by optimal portfolios are executed. In practice, portfolio managers often instill their own judgment when implementing portfolio trades recommended by optimization. They might alter the size of certain trades, for example, based on information about the companies not captured by the model or they might elect to ignore small trades based on the belief that these small trades would not have a meaningful impact on the portfolio and its performance.

How do small trades affect portfolio turnover and portfolio performance? In this section, we analyze the trade-off between turnover reduction and performance impact when small trades are neglected.

8.7.1 Alpha Exposure

Leaving small trades out reduces the alpha exposure of an optimal portfolio. We first calculate the alpha exposure or the expected return of a full implementation of optimal weights. It is the sum of active weights times the forecasts. At time t, with optimal weights of Equation 8.13, the alpha exposure is the sum of weight times factor value

$$\alpha^t = \sum_{i=1}^{N} w_i^t f_i^t = \frac{1}{\lambda_t} \sum_{i=1}^{N} \left(F_i^t\right)^2 \approx \frac{N}{\lambda_t} \left[\operatorname{std}\left(\mathbf{F}^t\right)\right]^2. \tag{8.44}$$

Note that the forecasts are not yet standardized. Substituting the risk aversion parameter in (8.14) gives

$$\alpha^t \approx \sqrt{N}\sigma_{\text{model}} \operatorname{std}\left(\mathbf{F}^t\right). \tag{8.45}$$

If we assume $f_i^t = \overline{IC} z_i \sigma_i$, then $\operatorname{std}\left(\mathbf{F}^t\right) = \overline{IC}$ and the alpha exposure is

$$\alpha^t \approx \sqrt{N}\sigma_{\text{model}} \overline{IC}. \tag{8.46}$$

Note that this is the original form of the fundamental law of active management (Grinold 1989).

By the time t+1, the forecasts or alpha factors have changed from f_i^t to f_i^{t+1}. Therefore, the alpha exposure of the portfolio is also changed.

Assuming no drift from t to t+1, the new alpha exposure is the sum product of optimal weights at t and factor value at t+1:

$$\begin{aligned}\alpha^{t,t+1} &= \sum_{i=1}^{N} w_i^t f_i^{t+1} = \frac{1}{\lambda_t}\sum_{i=1}^{N} F_i^t F_i^{t+1} \approx \frac{N}{\lambda_t}\rho_f \operatorname{std}\left(\mathbf{F}^t\right)\operatorname{std}\left(\mathbf{F}^{t+1}\right) \\ &= \rho_f\sqrt{N}\sigma_{\text{model}}\operatorname{std}\left(\mathbf{F}^{t+1}\right) = \rho_f\sqrt{N}\sigma_{\text{model}}\overline{IC}\end{aligned}. \quad (8.47)$$

ρ_f is simply the autocorrelation of the risk-adjusted forecast. Note that the alpha decay or the ratio of $\alpha^{t,t+1}$ to α^t is ρ_f, which is always less than one. So, the alpha exposure declines in proportion to the forecast autocorrelation. Relating to the previous results, we note that the alpha exposure declines slowly with value factors but rapidly with momentum factors.

We opt to analyze the alpha exposure instead of the information coefficient to simplify the analysis. Equation 8.47 can also be expressed in terms of lagged IC. The two are equivalent only if the lagged IC declines according to the forecast autocorrelation. We note that this might be the case in practice.

When we reoptimize at t+1 and rebalance to form a new optimal portfolio, we regain the original exposure. In other words, after all trades, the alpha exposure α^{t+1} reverts back to α^t, with an increase of

$$\Delta\alpha = \alpha^{t+1} - \alpha^{t,t+1} = \left(1-\rho_f\right)\sqrt{N}\sigma_{\text{model}}\overline{IC}. \quad (8.48)$$

The turnover required in the rebalance, to regain the prior alpha exposure, is the turnover caused by the change in forecasts and is given in (8.20),

$$T = \sqrt{\frac{\mathrm{N}}{\pi}}\sigma_{\text{model}}\sqrt{1-\rho_f}\,\mathrm{E}\left(\frac{1}{\sigma}\right).$$

8.7.2 Turnover Reduction of Small Trades

If we elect to ignore small trades, it is obvious that there will be a reduction in turnover. However, it is also likely the alpha exposure will decrease. We are interested in their respective rates of decrease.

Consider a trade-size threshold, below which trades will not be executed. In other words, at time t+1, if the difference between the new optimal weight and the old one is above the threshold, we adopt the new weight. Otherwise, we ignore the trade, and the active weight stays the same. In order to gain some insight regarding the trade-off between alpha exposure and turnover, we consider the case in which all stock-specific risks are

the same. Under this assumption, a trade-size threshold is equivalent to a threshold in forecast difference by the following relationship

$$w_i^{t+1} - w_i^t = \frac{\sigma_{\text{model}}}{\sqrt{N}} \frac{\tilde{F}_i^{t+1} - \tilde{F}_i^t}{\sigma_0}. \tag{8.49}$$

Suppose the threshold is the weight difference ε_w, then the threshold in the standardized forecast difference would be

$$\varepsilon_F = \frac{\varepsilon_w \sqrt{N} \sigma_0}{\sigma_{\text{model}}}. \tag{8.50}$$

The remaining portfolio turnover, excluding trade size below ε_w, is

$$T(\varepsilon_w) = \frac{1}{2} \sum_{|\Delta w_i| > \varepsilon_w} \left| w_i^{t+1} - w_i^t \right| = \frac{1}{2} \frac{\sigma_{\text{model}}}{\sigma_0 \sqrt{N}} \sum_{|\Delta \tilde{F}_i| > \varepsilon_F} \left| \tilde{F}_i^{t+1} - \tilde{F}_i^t \right|. \tag{8.51}$$

By assumption, $\Delta \tilde{F}_i = \tilde{F}_i^{t+1} - \tilde{F}_i^t$ is normally distributed with zero mean and standard deviation $s = \sqrt{2(1-\rho_f)}$, the resulting turnover is related to a conditional expectation of the normal variable

$$\begin{aligned} \sum_{|\Delta \tilde{F}_i| > \varepsilon_F} \left| \Delta \tilde{F}_i \right| &\approx N \cdot \mathrm{E}\left(\left| \Delta \tilde{F}_i \right| \middle| \left| \Delta \tilde{F}_i \right| > \varepsilon_F \right) \\ &= \frac{2N}{\sqrt{2\pi} s} \int_{\varepsilon_F}^{\infty} x \exp\left(-\frac{x^2}{2s^2} \right) dx = N \sqrt{\frac{2}{\pi}} s \exp\left(-\frac{\varepsilon_F^2}{2s^2} \right) \end{aligned} \tag{8.52}$$

Substituting Equation 8.52 into Equation 8.51 yields

$$T(\varepsilon_w) = \frac{\sqrt{N} \sigma_{\text{model}}}{\sqrt{\pi} \sigma_0} \sqrt{1-\rho_f} \exp\left(-\frac{\varepsilon_F^2}{2s^2} \right) = T(0) \exp\left(-\frac{\varepsilon_F^2}{2s^2} \right). \tag{8.53}$$

The reduced turnover with a threshold in the trading size equals the product of the original turnover and an exponential function of the threshold in the forecast difference, which represents the reduction in turnover when small trades are not executed.

Example 8.10

According to Example 8.6, for a long–short portfolio with $N = 500$, $\sigma_{\text{model}} = 5\%$, $\sigma_0 = 30\%$, and $\rho_f = 0.9$, the one-time turnover would be 66%. Suppose we do not execute any trade below 0.3% or 30 bps. The threshold for difference in the risk-adjusted forecast would be

$$\varepsilon_F = \frac{\varepsilon_w \sqrt{N}\sigma_0}{\sigma_{\text{model}}} = \frac{0.3\% \cdot \sqrt{500} \cdot 30\%}{5\%} = 0.40 .$$

The turnover reduction ratio is then

$$\exp\left[-\frac{\varepsilon_F^2}{4\left(1-\rho_f\right)}\right] = \exp\left[-\frac{\left(0.4\right)^2}{4\left(1-0.1\right)}\right] = 0.67 .$$

Therefore, the turnover after eliminating small trades of less than 30 bps would be 67% of the original turnover. Figure 8.8 plots this ratio vs. the threshold in trading size. As the threshold gets larger turnover decreases rather rapidly.

8.7.3 Decrease in Alpha Exposure

To calculate the alpha exposure for a given threshold, we note that the active weights are now a mixture of the optimal weights at t and the optimal weights at t+1: when the forecast difference is below the threshold,

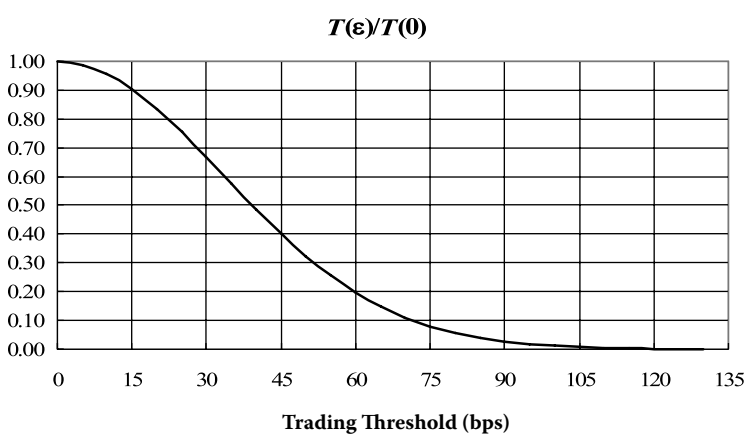


FIGURE 8.8. Portfolio turnover with trading threshold as a ratio of the original turnover ($N = 500$, $\sigma_{\text{model}} = 5\%$, $\sigma_0 = 30\%$, and $\rho_f = 0.9$).

the active weight is unchanged, whereas when the difference is above the threshold, the active weight is rebalanced according to the new forecast. We define a mixed forecast by

$$\tilde{F}_i^* = \begin{cases} \tilde{F}_i^t, \text{ if } \left|\tilde{F}_i^t - \tilde{F}_i^{t+1}\right| \le \varepsilon_F \\ \tilde{F}_i^{t+1}, \text{otherwise} \end{cases}. \tag{8.54}$$

The alpha exposure with a threshold is then the sum of the product of the mixed forecast and the factor value at t+1, i.e.,

$$\begin{aligned} \alpha^{t+1}(\varepsilon_w) &= \frac{1}{\lambda_{t+1}} \sum_{i=1}^{N} \frac{F_i^*}{\sigma_i} f_i^{t+1} = \frac{1}{\lambda_{t+1}} \sum_{i=1}^{N} F_i^* F_i^{t+1} \\ &= \sigma_{\text{model}} \sqrt{N} \mathrm{E}\left(\tilde{F}_i^* \tilde{F}_i^{t+1}\right) \cdot \mathrm{std}\left(\mathbf{F}^{t+1}\right) \\ &= \sigma_{\text{model}} \sqrt{N}\, \overline{IC} \cdot \mathrm{E}\left(\tilde{F}_i^* \tilde{F}_i^{t+1}\right) \end{aligned} . \tag{8.55}$$

We have used $\mathrm{std}\left(\mathbf{F}^{t+1}\right) = \overline{IC}$ in (8.55). Note that when the trading threshold is 0, all trades are executed. We have $\mathrm{E}\left(\tilde{F}_i^* \tilde{F}_i^{t+1}\right) = \mathrm{E}\left(\tilde{F}_i^{t+1} \tilde{F}_i^{t+1}\right) = 1$, and the alpha exposure is fully restored. When the trading threshold is infinity, no trades are executed. We have $\mathrm{E}\left(\tilde{F}_i^* \tilde{F}_i^{t+1}\right) = \mathrm{E}\left(\tilde{F}_i^t \tilde{F}_i^{t+1}\right) = \rho_f$.

For general cases, we evaluate the expectation $\mathrm{E}\left(\tilde{F}_i^* \tilde{F}_i^{t+1}\right)$ analytically in an appendix. We have

$$\mathrm{E}\left(\tilde{F}^* \tilde{F}^{t+1}\right) = 1 + \frac{s\varepsilon_F}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon_F^2}{2s^2}\right) - \frac{s^2}{2} \Phi\left(\frac{\varepsilon_F}{\sqrt{2}s}\right). \tag{8.56}$$

We have used $s = \sqrt{2\left(1-\rho_f\right)}$, and $\Phi(\cdot)$ is the error function. Substituting (8.56) into (8.55) yields

$$\alpha^{t+1}(\varepsilon_w) = \alpha^{t+1}(0)\left[1 + \frac{s\varepsilon_F}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon_F^2}{2s^2}\right) - \frac{s^2}{2} \Phi\left(\frac{\varepsilon_F}{\sqrt{2}s}\right)\right]. \tag{8.57}$$

Figure 8.9 plots the ratio $\alpha^{t+1}(\varepsilon_w)/\alpha^{t+1}(0)$ as a function of the trade threshold using the same parameters as in Figure 8.8. As we can see from the graph, when the threshold is 0, all trades are carried out, and the ratio

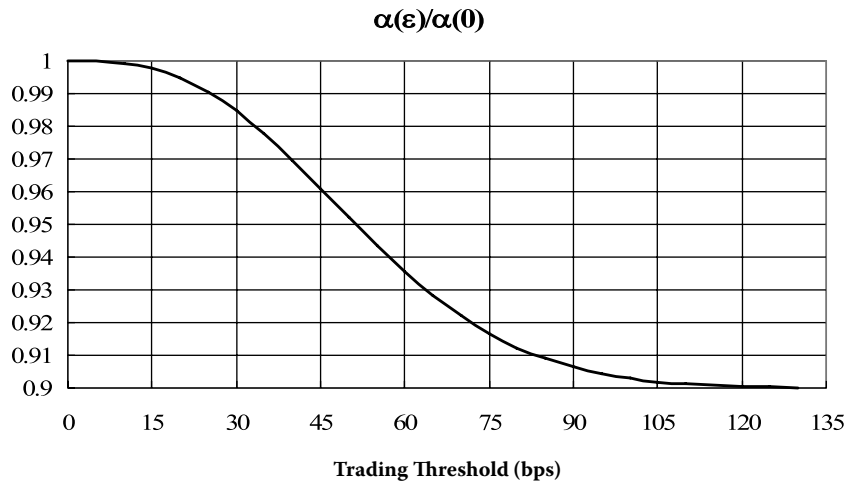


FIGURE 8.9. Ratio of alpha exposure with trading threshold to full exposure ($N = 500$, $\sigma_{\text{model}} = 5\%$, $\sigma_0 = 30\%$, and $\rho_f = 0.9$).

is unity. As the threshold increases, the alpha exposure declines rather slowly at first. For instance, if the size threshold is 30 bps, the alpha exposure is 0.985 of the full exposure. Recall that at 30 bps, the portfolio turnover is 67% of the full turnover. This reveals a favorable trade-off between turnover reduction and loss in alpha exposure. As the trade size further increases, the alpha exposure drops more rapidly. When the size threshold is large enough, very few trades are carried out (see Figure 8.5), the alpha exposure converges to the pretrade level given by (8.47) and, in our example, it is 0.9 of the full exposure.

We can also view the alpha-turnover trade-off directly. The question is how much incremental alpha exposure can be obtained with the remaining trades. Figure 8.10 plots this relationship. The horizontal axis denotes the remaining turnover, as a percentage of the total turnover, and the vertical axis is the alpha increase, also as a percentage of full increase. Obviously, one end point of the curve corresponds to no trades without any alpha pickup, and the other end point of the curve corresponds to all trades and full alpha pickup. The concave shape of the curve indicates that the trade-off is certainly not linear. With 50% turnover, we can get 70% of the alpha increase, and with 60% of turnover the alpha increase would be 80%. Note the following:

- Our analysis does lend some support to the practice of ignoring small trades in portfolio implementation. However, there are a couple of caveats. First, the trade-off between turnover reduction and alpha

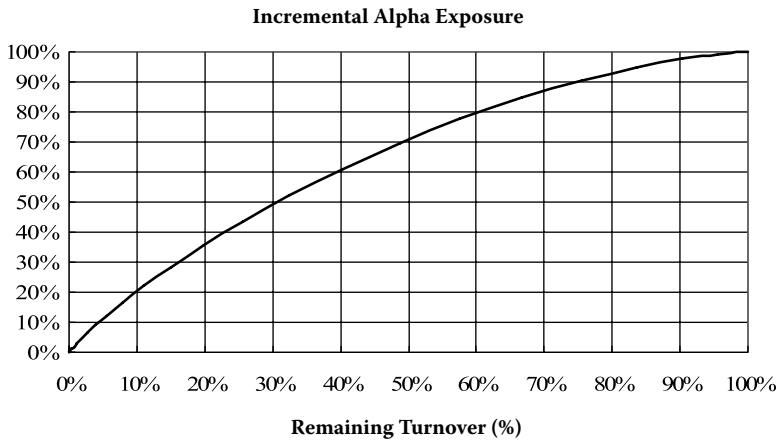


FIGURE 8.10. Percentage of alpha exposure increase as a function of remaining portfolio turnover ($N = 500$, $\sigma_{\text{model}} = 5\%$, $\sigma_0 = 30\%$, and $\rho_f = 0.9$).

exposure reduction has to be carefully weighed in each case, where the target tracking error and number of stocks in the portfolio are important inputs. Second, our analysis considers only a single rebalance. Additional analysis is needed to provide insights to the trade-off between turnover reduction and alpha exposure reduction for multiple-period rebalances. Finally, we note that the analysis needs to be generalized to the impact of small trades on ICs and lagged ICs.

8.7.4 Effect on Tracking Error

Optimal portfolios are often constructed with a targeted tracking error. Does the practice of ignoring small trades have any effect on the tracking error of the portfolio? There are reasons to suspect that any effect, should they exit, is small. Both sets of active weights are derived with the same target tracking error. If all trades are carried out, then the target tracking error should be σ_{model}. At the other extreme, if none of the trades are executed, the tracking error remains at σ_{model}, ignoring portfolio drift.

When small trades are ignored, the active weights are a mixture of old and new, and they are related to the forecast $\tilde{F}_i^*$ defined in (8.54). Therefore, the tracking error of the mixed weights is given by the second moment, or the variance of $\tilde{F}^*$, because it is easy to see $\mathrm{E}\left(\tilde{F}^*\right) = 0$. We have

$$\sigma^* = \sigma_{\text{model}} \mathrm{E}\left[\left(\tilde{F}^*\right)^2\right]. \tag{8.58}$$

In the appendix, we prove that $\mathrm{E}\left[\left(\tilde{F}^{*}\right)^{2}\right]=1$ for all ε_F. Therefore, regardless of the cutoff for the small trades, the tracking error of the portfolio is not affected at all.

PROBLEMS

8.1 (a) Suppose our initial holding is 100% cash, and we invest it fully in a portfolio of stocks. Calculate the turnover using formula (8.4).

(b) Prove that the definition (8.4) is valid when one of the portfolio holdings is cash.

8.2 Suppose the return is normally distributed with zero mean $x \sim N\left(0, d^2\right)$. Prove that

$$\mathrm{E}\left(|x|\right)=\sqrt{\frac{2}{\pi}}d.$$

8.3 Suppose $r \sim N\left(\bar{r}, d^2\right)$. Let $x = r - \bar{r}$, then $x \sim N\left(0, d^2\right)$.

(a) Show that

$$\mathrm{E}\left(\left|r-r_p\right|\right)=\mathrm{E}\left(|x-\Delta r|\right), \text{ with } \Delta r = r_p - \bar{r}.$$

(b) Show that

$$\mathrm{E}\left(|x-\Delta r|\right)=\frac{2d}{\sqrt{2\pi}}\exp\left[-\frac{(\Delta r)^2}{2d^2}\right]+\Delta r\cdot\mathrm{erf}\left(\frac{\Delta r}{\sqrt{2}d}\right),$$

where $\mathrm{erf}(y)=\frac{2}{\sqrt{\pi}}\int_0^y \exp\left(-t^2\right)dt$ is the error function.

(c) Use approximations for the exponential and error functions to show that

$$\mathrm{E}\left(|x-\Delta r|\right)\approx\frac{2d}{\sqrt{2\pi}}\left[1+\frac{(\Delta r)^2}{2d^2}\right].$$

8.4 Our portfolio has 125% long and 25% short so the total weight is still 100%. Suppose it returned 7%, whereas the average stock return is 2%, and the return dispersion is 15%. Calculate the average portfolio turnover required for rebalancing.

8.5 Prove that the forecast change $F^{t+1} - F^t$ has a standard deviation of $\sqrt{2(1-\rho_f)}$.

8.6 Suppose we have three different forecasts, with different levels of autocorrelations at 0.7, 0.8, and 0.9, respectively. Calculate the relative levels of turnover for the three forecasts.

8.7 (a) Prove that the serial autocorrelation of moving average of (8.31) has an extreme value when $v_1 = v_0$.

(b) When is the extreme value a maximum and when is it a minimum?

8.8 Suppose the forecast follow an AR(1) process, i.e., $\mathbf{F}^{t+1} = a\mathbf{F}^t + \boldsymbol{\varepsilon}^t$, where $a < 1$ and the forecast vector $\mathbf{F}^t$ and the error vector $\boldsymbol{\varepsilon}^t$ are independent. Suppose all forecast vectors are standardized with $\text{dis}(\mathbf{F}^t) = 1$.

(a) Show that $\rho_f(1) = a$, $\rho_f(2) = a^2$, and, in general, $\rho_f(L) = a^L$.

(b) Show that for AR(1) process, $\rho_f(2) = a^2$ is always above the threshold of (8.32); hence, moving averages of the forecasts have higher series autocorrelation and lower portfolio turnover.

8.9 Prove the relationship between the lagged IC and the horizon IC (8.41).

8.10 [Grinold and Stuckelman 1993] We optimize a quadratic utility function $\text{U}(w) = fw - 0.5\lambda\sigma^2 w^2$, in which f is the alpha forecast and w is trading amount.

(a) Find the optimal w^* and show that the optimal utility is $\text{U}^*(w^*) = f^2 / (2\lambda\sigma^2)$.

(b) Suppose we wish to cut the trade in half, i.e., $w_{1/2} = w^*/2$; prove that $\text{U}(w_{1/2}) = 0.75\text{U}^*(w^*)$. Therefore, we achieve 75% of value-added by half of portfolio turnover. However, the value-added in this case is not the expected alpha but the utility.

(c) How much is the expected alpha being reduced if the trade is cut in half?

(d) Let $w_k = kw^*$, $0 \le k \le 1$. Plot the utility ratio $\mathrm{U}(w_k)/\mathrm{U}^*(w^*)$ as a function of *k*.

APPENDIX
A8.1 REDUCTION IN ALPHA EXPOSURE

We evaluate the expectation in alpha exposure when small trades are neglected. As defined in the main text, $\tilde{F}^t$ and $\tilde{F}^{t+1}$ are normal random variables with 0 mean, standard deviation 1, and correlation ρ_f. The random variable $\tilde{F}^*$ is defined as

$$\tilde{F}^* = \begin{cases} \tilde{F}^t, \text{ if } \left|\tilde{F}^t - \tilde{F}^{t+1}\right| \le \varepsilon_F \\ \tilde{F}^{t+1}, \text{otherwise} \end{cases}.$$

The alpha exposure of the modified weight is related to the expectation $\mathrm{E}\left(\tilde{F}^* \tilde{F}^{t+1}\right)$.

Because the new variable is contingent on the difference between $\tilde{F}^t$ and $\tilde{F}^{t+1}$, we define a new random variable $x = \tilde{F}^t - \tilde{F}^{t+1}$. We also define $y = \tilde{F}^{t+1}$. Then,

$\tilde{F}^t = x + y$ and $\tilde{F}^{t+1} = y$. It is easy to verify that x and y satisfy the following: $\mathrm{E}(x) = 0$, $\mathrm{E}(y) = 0$, $\mathrm{var}(x) = 2 - 2\rho_f$, $\mathrm{var}(y) = 1$, and $\mathrm{cov}(x, y) = \rho_f - 1$.

Using conditional expectation, we have

$$\begin{aligned}
\mathrm{E}\left(\tilde{F}^* \tilde{F}^{t+1}\right) &= \mathrm{E}\left[\mathrm{E}\left(\tilde{F}^* \tilde{F}^{t+1} \mid x\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(\tilde{F}^t \tilde{F}^{t+1} \mid x,\ |x| \le \varepsilon_F\right)\right] + \mathrm{E}\left[\mathrm{E}\left(\tilde{F}^{t+1} \tilde{F}^{t+1} \mid x,\ |x| \ge \varepsilon_F\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(y(x+y) \mid x,\ |x| \le \varepsilon_F\right)\right] + \mathrm{E}\left[\mathrm{E}\left(y^2 \mid x,\ |x| \ge \varepsilon_F\right)\right] \qquad (8.59) \\
&= \mathrm{E}\left[\mathrm{E}\left(xy \mid x,\ |x| \le \varepsilon_F\right)\right] + \mathrm{E}\left[\mathrm{E}\left(y^2 \mid x\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(xy \mid x,\ |x| \le \varepsilon_F\right)\right] + 1
\end{aligned}$$

Because the conditional distribution of *y* given *x* is

$$y|_x \sim N\left[-\frac{x}{2}, 1-\frac{1-\rho_f}{2}\right] = N\left[-\frac{x}{2}, \frac{1+\rho_f}{2}\right], \tag{8.60}$$

we have $\mathrm{E}\left(y \mid x\right) = -\frac{x}{2}$.

Hence, the remaining expectation in (8.59) is given by integration

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left(xy \mid x,\ |x| \le \varepsilon_F\right)\right] &= \mathrm{E}\left[-\frac{x^2}{2} \mid\ |x| \le \varepsilon_F\right] \\
&= -\frac{1}{2}\frac{1}{\sqrt{2\pi}s}\int_{-\varepsilon_F}^{\varepsilon_F} x^2 \exp\left(-\frac{x^2}{2s^2}\right)dx \\
&= -\frac{1}{\sqrt{2\pi}s}\int_{0}^{\varepsilon_F} x^2 \exp\left(-\frac{x^2}{2s^2}\right)dx .
\end{aligned} \tag{8.61}$$

We have denoted the standard deviation of x as $s = \sqrt{2-2\rho_f}$. Integration by parts and changing integration variable leads to

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left(xy \mid x,\ |x| \le \varepsilon_F\right)\right] &= -\frac{1}{\sqrt{2\pi}s}\int_{0}^{\varepsilon_F} x^2 \exp\left(-\frac{x^2}{2s^2}\right)dx \\
&= \frac{s}{\sqrt{2\pi}} x \exp\left(-\frac{x^2}{2s^2}\right)\Bigg|_0^{\varepsilon_F} - \frac{s}{\sqrt{2\pi}}\int_0^{\varepsilon_F} \exp\left(-\frac{x^2}{2s^2}\right)dx \\
&= \frac{s\varepsilon_F}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon_F^2}{2s^2}\right) - \frac{s^2}{\sqrt{\pi}}\int_0^{\varepsilon_F/\sqrt{2}s} \exp\left(-t^2\right)dt \\
&= \frac{s\varepsilon_F}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon_F^2}{2s^2}\right) - \frac{s^2}{2}\Phi\left(\frac{\varepsilon_F}{\sqrt{2}s}\right).
\end{aligned} \tag{8.62}$$

Therefore,

$$\mathrm{E}\left(\tilde{F}^* \tilde{F}^{t+1}\right) = 1 + \frac{s\varepsilon_F}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon_F^2}{2s^2}\right) - \frac{s^2}{2}\Phi\left(\frac{\varepsilon_F}{\sqrt{2}s}\right). \tag{8.63}$$

A8.1.1 Constancy of Tracking Error

To calculate the tracking error of a portfolio with a trading threshold, we evaluate the expectation $\mathrm{E}\left[\left(\tilde{\mathbf{F}}^{*}\right)^{2}\right]$ in a similar way. Using the same variables x and y, we have

$$\begin{aligned}
\mathrm{E}\left[\left(\tilde{F}^{*}\right)^{2}\right] &= \mathrm{E}\left\{\mathrm{E}\left[\left(\tilde{F}^{*}\right)^{2}\right]\middle| x\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[\left(\tilde{F}^{*}\right)^{2}\right]\middle| x,\ |x| \le \varepsilon_F\right\} + \mathrm{E}\left\{\mathrm{E}\left[\left(\tilde{F}^{*}\right)^{2}\right]\middle| x,\ |x| \ge \varepsilon_F\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[\left(x+y\right)^{2}\right]\middle| x,\ |x| \le \varepsilon_F\right\} + \mathrm{E}\left\{\mathrm{E}\left[y^{2}\right]\middle| x,\ |x| \ge \varepsilon_F\right\} \qquad . \quad (8.64) \\
&= \mathrm{E}\left\{\mathrm{E}\left(x^{2}+2xy+y^{2}\right)\middle| x,\ |x| \le \varepsilon_F\right\} + \mathrm{E}\left\{\mathrm{E}\left(y^{2}\right)\middle| x,\ |x| \ge \varepsilon_F\right\} \\
&= \mathrm{E}\left(y^{2}\right) + \mathrm{E}\left(x^{2}\middle| x,\ |x| \le \varepsilon_F\right) + \mathrm{E}\left[2x\mathrm{E}\left(y\middle| x,\ |x| \le \varepsilon_F\right)\right]
\end{aligned}$$

Previously, we have shown $\mathrm{var}\left(y\right) = \mathrm{E}\left(y^{2}\right) = 1$, and $\mathrm{E}\left(y \mid x\right) = -\frac{x}{2}$. Substituting them into (8.64), we observe the last two terms cancel each other while the first term is unity. Hence,

$$\mathrm{E}\left[\left(\tilde{F}^{*}\right)^{2}\right] = 1 .$$

REFERENCES

Grinold, R.C., The fundamental law of active management, *Journal of Portfolio Management,* Vol. 15, No. 3, 30–37, Spring 1989.

Grinold, R.C. and Stuckelman, M., The value-added/turnover frontier, *Journal of Portfolio Management,* Vol. 19, No. 4, 8–17, Summer 1993.

Kahn, R.N. and Shaffer, J.S., The surprising small impact of asset growth on expected alpha, *Journal of Portfolio Management,* Vol. 32, No. 1, 49–60, Fall 2005.

Qian, E., Hua, R., and Tilney, J., Portfolio turnover of quantitatively managed portfolios, *Proceeding of the 2nd IASTED International Conference, Financial Engineering and Applications*, Cambridge, MA, 2004.

Sorensen, E.H., Qian, E., Hua, R., and Schoen, R., Multiple alpha sources and active management, *Journal of Portfolio Management,* Vol. 31, No. 2, 39–45, Winter 2004.

ENDNOTES

1. Turnover can also be caused by flows in and out of a portfolio. These forced turnovers are not due to portfolio rebalance, and they are easy to analyze. We shall exclude them from our analysis.
2. Our definition of turnover measures the percentage change of the portfolio vs. portfolio capital, which is most relevant in terms of amount of trading. There are other variations that use total portfolio leverage or notational exposures as denominators.
3. For constrained portfolios such as long-only portfolios, the turnover can be substantially less, since constraints work to suppress changes in portfolio weights (Qian et al. 2004).
4. If there is short-term reversion between consecutive period returns, then the horizon IC will be higher.
5. It is not hard to imagine this situation might apply to market segments that are relatively less inefficient, such as U.S. large cap stocks.

CHAPTER 9

Advanced Alpha Modeling Techniques

Quantitative equity portfolio management relies on both the alpha model and the risk model to construct a mean–variance efficient portfolio. The alpha model forecasts the excess return of each security by identifying pricing inefficiencies, whereas the risk model forecasts the covariance structure of the security return. The former delivers value added of active management in the form of portfolio returns in excess of its benchmarks; the latter provides portfolio risk control and diversification benefit. Although each plays a different role, both depend on the assumption of a return generating equation in constructing their forecasts.

In this chapter, we shall take a closer look at the return-generating equation behind most traditional quantitative models and present modeling techniques that provide a structured framework in relaxing many stringent assumptions behind the traditional approach. Specifically, we will first discuss three assumptions behind the commonly used return-generating equation: "one size fits all," "bigger is always better," and "time independence." We will then discuss various advanced modeling techniques that can achieve better alpha forecasts by relaxing the first two assumptions. Both assumptions are cross-sectional in nature. The techniques include contextual alpha modeling, sector modeling, and nonlinear effect modeling. We will address the third assumption in Chapter 10 by highlighting several time-varying modeling techniques.

9.1 THE RETURN-GENERATING EQUATION

Equation 9.1 postulates a generic return-generating equation, which expresses security returns in terms of exposures to factors. Security return is a linear combination of attributed returns to factors that possess cross-sectional explanatory power.

$$r_i = b_{i0} + b_{i1} I_1 + \cdots + b_{iK} I_K + \varepsilon_i \,. \tag{9.1}$$

In the equation, r_i is the return of stock i, $b_{i1}, \cdots, b_{iK}$ are factor exposures of the stock, and $I_1, \cdots, I_K$ are factor returns. The residual portion of security return that is not attributed, is called *security specific return* and is expressed as ε_i. Note that in Equation 9.1 we dropped the subscript of time to simplify the notation. This equation serves as the core of risk models in Chapter 3. The covariance matrix of returns is given by

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_I \mathbf{B}' + \mathbf{S} \,, \tag{9.2}$$

where $\mathbf{\Sigma}_I$ is the factor return covariance matrix, $\mathbf{B}$ is the exposure matrix, and $\mathbf{S}$ is the diagonal specific variance matrix. Equation 9.2 forms the foundation of many commercially available risk models, such as BARRA, Northfield, or Citigroup GRAM. The only difference among them is the set of factors selected. For example, BARRA uses fundamental factors, whereas Northfield employs mostly macro economic factors.

Perhaps, due to its academic origin and popularity in commercial risk models, many active managers also adopt framework similar to (9.1) in constructing their proprietary alpha models. Specifically, they forecast expected return as

$$E(r_i) \propto f_{i1} v_1 + \cdots + f_{iM} v_M \,, \tag{9.3}$$

where $(f_{i1}, \cdots, f_{iM})$ are cross-sectional alpha factors and $(v_1, \cdots, v_M)$ are the factor weights that are related to expected factor returns. Although methods of selecting the factor weights vary greatly among active managers (see Chapter 7 for the discussion), most methods conform to (9.3), which makes the following three unrealistic assumptions.

One size fits all: In Equation 9.3, the factor weights are the same for every security, thus making it a one-size-fits-all approach. However, most

practitioners recognize the conditional nature of factor returns, and their intuitions find significant support from empirical research. For example, Daniel et al. (1999) find that momentum effects are stronger for growth stocks, and Asness (1997) finds that value strategies work, in general, but less so for stock with high momentum.

Bigger is always better: Because (9.3) is linear, it implies that the expected security return is linearly proportional to the factor exposure. For example, if buying cheap stocks is a good thing, then purchasing deep value securities must produce the best investment results. In reality, practitioners are often aware of the fact that deep value securities are often cheap for a reason. For example, Bruce and Morillo (2003) find that expected returns of securities with extreme factor values tend to break away from their linear expectations, sometimes in a fairly dramatic way.

Time independence: The last assumption deals with the constancy of factor weights over time, making it an unconditional model. In reality, factor returns change through time, depending on various macroeconomic regimes or even different calendar events. This time-varying behavior is ignored in (9.3).

In all, the linear one-size-fits-all return-generating equation provides a resilient foundation for risk models. However, the same equation is an inadequate foundation for forecasting the expected security return, mostly due to the linearity assumption. Such inadequacy is born out of the fact that security markets are quasi-efficient wherein many sophisticated managers try to arbitrage the same set of behavioral phenomenon. Simplistic alpha models such as (9.3) deliver inferior portfolio excess returns. In the rest of this chapter, we shall present several advanced modeling techniques.

9.2 CONTEXTUAL MODELING

In practice, linking a stock's ranking signal or factor to expected return and assigning it an appropriate weight is a matter of context. The application of a timely security selection criterion is conditional. Simply — it depends. For example, many researchers demonstrate that value, as a selection variable, is often conditional on the type of firm, other nonvalue factors, the investment horizon, or some other dimension. Sloan (2001) and Beneish et al. (2001) call this interdependency of security factors *contextual*.

Seasoned active managers know that value investing focuses on discovering cheap stocks with a balance of quality; at the same time, growth investing often seeks to balance positive momentum with quality and cheapness. This anecdotal assertion finds substantiation in prior academic studies. For example, Daniel and Titman (1999) find that momentum effects are stronger for growth stocks. Asness (1997) finds that value strategies work, in general, but less so for stocks with high momentum. In a particularly relevant study, Scott et al. (1999) focuses on prospect theory and investor overconfidence. They provide empirical evidence that rational value investors should emphasize cheapness (as in dogs), whereas growth investors should let winners run — with the prospect of future good news. Piotroski (2000) and Mohanram (2004) also demonstrate that one should focus on different sets of financial statement information when analyzing stocks with different book-to-price ratios. Taken together, these studies (and others) point to the importance of analyzing the efficacy of alpha factors within carefully selected security universes — the contextual analysis of active strategies.

9.2.1 Factor Categories

To illustrate contextual dynamics, we introduce five composite factors representing the set of investing philosophies discussed in Chapter 5. Table 9.1 describes the description of these composites. To capture the essence of the value investing that buys cheap stocks, we create the relative value (RV) factor, a composite encompassing two types of cheapness measures: the earnings yield and the asset value. We title this factor relative value because cheapness is gauged in the context of a peer group; and, in this study, we use sector as the peer group for comparison. Additionally, to represent the premise of the fundamental investing, we trace the analysis of the enterprise profitability, accrued to shareholders, into three composite factors: (1) the operating efficiency (OE) factor measuring management's ability to generate shareholder value, (2) the accounting accrual (AA) factor measuring the accuracy and the honesty of a company's financial reporting practice, and (3) the external financing (EF) factor measuring the hazard of self-serving management pursuing corporate expansions at the expense of shareholder wealth. Finally, the philosophy of riding market sentiment in momentum investing is captured in the momentum factor (MO), which consists of the measures of the intermediate-term price momentum, the earnings revision, and the earnings surprise.

TABLE 9.1 Definition of Factor Composites

Composite	Factors
Valuation (RV)	Book-to-price ratio
	Sales to enterprise value
	Earnings yield (historical)
	Earnings yield (IBES FY1)
	EBIT to enterprise value
Operating Efficiency (OE)	Increase in asset turnover ratio
	Level of operating leverage
	Cashflow-from-operation to sales
Accounting Accrual (AA)	Accounting accruals (balance sheet)
	Accounting accruals (cashflow statement)
External Financing (EF)	External financing to net operating assets
	Debt issuance to net operating assets
	Equity issuance to net operating assets
	Share count increase
Momentum (MO)	Six-month price momentum
	Nine-month earnings revision
	Earnings surprise score

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.

9.2.2 Security Contexts

We illustrate the interplay among factors along the dimensions of three risk characteristics: value, growth, and earning variability. Along each of these dimensions, we select two nonoverlapping security contexts with an equal number of stocks: one contains securities with high loadings of that risk characteristic, whereas the other includes securities with low loadings. Hence, six security contexts are defined, and they contain firms with high/low value measure, high/low growth rate, and high/low earnings variability.

We use the book-to-price ratio as our first risk dimension: value. The name value for the book-to-price ratio implies it associates with market inefficiency, but this is not relevant to the contextual analysis. What is relevant is the interpretation provided by Fama and French (1996), who associate the book-to-price ratio with the investment quality or financial condition of a company. Specifically, we can interpret a low book-to-price ratio as an indication of high quality and a high book-to-price ratio as low quality. Defined as such, high-quality companies are expected by investors

to deliver superior returns on investment (ROI) and their *ex post* ROI typically exceeds the average ROI of a broad universe. In contrast, low-quality companies usually face a difficult operating environment and are expected to deliver inferior operating results. Different competitive standing, superior vs. inferior, often induces different challenges facing company management; one battles from a deteriorated competitive position to survive, whereas the other protects its competitive advantage by fending off competition. These intuitions are confirmed in the studies by Piotroski (2000) and Mohanram (2004). Therefore, we argue that investors should also focus their attentions on a different set of factors when evaluating the return appeal of companies with different book-to-price ratios.

Our second risk characteristic sorts companies based on their growth rate, creating contexts containing high-growth and low-growth companies. The rational behind this contextual dimension is well documented by Scott et al. (1999, 2003). Linking the behavioral science findings with the valuation theory, Scott et al. show that momentum investing (riding winners and looking for good news) is more important when selecting high-growth stocks, whereas selecting low-growth stocks should focus more on cheapness. The difference can be traced to how investors estimate the fair value of a business. The fair value estimate typically comprises two parts: the present value of existing business and the present value of future growth opportunities. For a low-growth company whose future growth prospect is limited, the value of its existing business dominates its fair value and, more importantly, valuation ratios (i.e., cash-flow yield or earnings yield) provide an accurate ranking of the relative cheapness of its existing business. In contrast, for high-growth companies, the majority of its fair value comes from the present value of future growth opportunities. As such, factors that are capable of predicting the quality of future growth play more prominent roles in determining the fair value. Combining this valuation reasoning with the observation that investors tend to under-react to news due to their overconfidence, Scott et al. (1999, 2003) show that earnings revision factor, a proxy of good news, is a consistent predictor of the excess returns of growth stocks.

Our last dimension differentiates companies along the earnings variability dimension. This contextual selection is inspired by the persistent predictability bias documented by Huberts and Fuller (1995). They show that sell-side analysts tend to provide overly optimistic forecasts for companies whose earnings are harder to predict, whereas their forecasts are more realistic, albeit still optimistic, for companies with stable earnings

in the past. Das et al. (1998) provide a more rigorous examination of this phenomenon and derive the same conclusion. Lastly, Beckers et al. (2004) find the same bias in European analysts' forecasts. In all, if earnings forecasts are less trustworthy for companies whose earnings are more variable, it is our conjecture that investors should focus their attentions on the quality of earning and the competence of management to remedy the deficiency of earnings forecasts. Similarly, investors should rely more on analysts' forecasts when selecting stable-earning companies because these forecasts are more reliable.

9.3 MATHEMATICAL ANALYSIS OF CONTEXTUAL MODELING

The basic premise of contextual modeling is that the efficacies of alpha factors are different among stocks across the different contexts. By using different optimal weights across the contexts, we will achieve a higher overall information ratio.

9.3.1 A One-Factor Example

The following one-factor example provides some intuition to the approach. Suppose we have a single context that divides the stock universe into two halves: one high and one low. Let us also assume for the moment that we just have a single alpha factor. We are interested in how the factor performs overall if it performs differently in the two halves. According to Chapter 4, a single-period excess return is given by (Equation 4.19)

$$\alpha_t = \sum_{i=1}^{N} w_i r_i = \lambda^{-1} \sum_{i=1}^{N} F_i R_i \,, \tag{9.4}$$

where F_i is risk-adjusted forecast, R_i is the risk-adjusted return, N is the number of stocks, and λ is the risk-aversion parameter used to calibrate the portfolio to a targeted tracking error. Breaking the stock universe into two halves — high and low, according to the context — we rewrite (9.4) as

$$\alpha_t = \lambda^{-1} \sum_{i=1}^{N} F_i R_i = \lambda^{-1} \sum_{i \in H} F_i R + \lambda^{-1} \sum_{i \in L} F_i R \,. \tag{9.5}$$

Now, writing all three sums in terms of risk-adjusted ICs in the respective universe gives

$$N \cdot \mathrm{IC}\,\mathrm{dis}(\mathbf{F})\,\mathrm{dis}(\mathbf{R}) = \frac{N}{2} \times \mathrm{IC}_H \mathrm{dis}(\mathbf{F}_H)\,\mathrm{dis}(\mathbf{R}_H) + \frac{N}{2} \times \mathrm{IC}_L \mathrm{dis}(\mathbf{F}_L)\,\mathrm{dis}(\mathbf{R}_L). \tag{9.6}$$

For simplicity, we have omitted the subscript t. We shall assume all the dispersions of forecasts and return are the same, which leads to

$$\mathrm{IC} = \frac{1}{2} \cdot \mathrm{IC}_H + \frac{1}{2} \cdot \mathrm{IC}_L. \tag{9.7}$$

The overall IR is obtained by the ratio of average IC to the standard deviation of IC

$$\mathrm{IR} = \frac{\overline{\mathrm{IC}}_H + \overline{\mathrm{IC}}_L}{\sqrt{\sigma_H^2 + \sigma_L^2 + 2\rho_{H,L}\sigma_H\sigma_L}}. \tag{9.8}$$

Equation (9.8) gives the overall IR in terms of IC statistics in the high and low contexts.

Example 9.1

Suppose the factor only works in the high dimension, but not in the low dimension, i.e., $\overline{\mathrm{IC}}_L = 0$. Then

$$\mathrm{IR} = \frac{\overline{\mathrm{IC}}_H}{\sqrt{\sigma_H^2 + \sigma_L^2 + 2\rho_{H,L}\sigma_H\sigma_L}}. \tag{9.9}$$

If the correlation of ICs is not negative, this overall IR will be less than the IR of the factor in the high dimension alone, i.e.,

$$\mathrm{IR} < \mathrm{IR}_H = \frac{\overline{\mathrm{IC}}_H}{\sigma_H}. \tag{9.10}$$

For instance, if $\overline{\mathrm{IC}}_H = 0.1$, $\sigma_H = \sigma_L = 0.1$, $\rho_{H,L} = 0.2$, then the IR in the high dimension $\mathrm{IR}_H = 1$, but the overall IR is just 0.6.

This example illustrates the fact that when a factor does not add value in the low dimension, still using it would dilute the IR of the factor because it adds noise or risk without additional returns. The simple remedy for this problem is to not use the factor in the low dimension. In other words, we shall not take any exposure to the factor in the low dimension stock. In terms of factor weight, it is simply zero for low dimension stocks.

9.3.2 Optimal Factor Weights across the Context

Setting the factor to zero for the low dimension stocks in the previous example represents a simple solution, but it is not necessarily the optimal one. If we denote the factor weight by v_H and v_L in the high and low dimension, then the overall IR becomes

$$\mathrm{IR} = \frac{v_H \overline{\mathrm{IC}}_H + v_L \overline{\mathrm{IC}}_L}{\sqrt{v_H^2 \sigma_H^2 + v_L^2 \sigma_L^2 + 2\rho_{H,L} \sigma_H \sigma_L v_H v_L}} . \tag{9.11}$$

The optimal weight can be found by the following

$$\begin{pmatrix} v_H^* \\ v_L^* \end{pmatrix} \propto \begin{pmatrix} \dfrac{\overline{\mathrm{IC}}_H}{\sigma_H^2} - \rho_{H,L} \dfrac{\overline{\mathrm{IC}}_L}{\sigma_H \sigma_L} \\ \dfrac{\overline{\mathrm{IC}}_L}{\sigma_L^2} - \rho_{H,L} \dfrac{\overline{\mathrm{IC}}_H}{\sigma_H \sigma_L} \end{pmatrix} . \tag{9.12}$$

With parameters in Example 9.1, the optimal weights are $v_H^* = 125\%$ and $v_L^* = -25\%$. The optimal IR is at 1.02, slightly above the IR for the high dimension. Thus, the optimal weights would have us betting against the factor in the low dimension, not because of value-added (there is none since the average IC is zero), but because of reduced risk.

With multiple factors, the objective of contextual modeling is to maximize the overall IR with optimal weights of factors in high and low dimensions. There are M factors and the weights are $\mathbf{v} = \left(\mathbf{v}_H, \mathbf{v}_L\right) = \left(v_{1,H}, v_{2,H}, \cdots, v_{M,H}, v_{1,L}, v_{2,L}, \cdots, v_{M,L}\right)'$. The vector of average IC is

$$\overline{\mathbf{IC}} = \left(\overline{\mathbf{IC}}_H, \overline{\mathbf{IC}}_L\right) = \left(\overline{IC}_{1,H}, \overline{IC}_{2,H}, \cdots, \overline{IC}_{M,H}, \overline{IC}_{1,L}, \overline{IC}_{2,L}, \cdots, \overline{IC}_{M,L}\right)'$$

and the $2M \times 2M$ IC covariance matrix is $\mathbf{\Sigma}_{IC}$. The overall IR is given by

$$\mathrm{IR} = \frac{\mathbf{v}' \cdot \overline{\mathbf{IC}}}{\sqrt{\mathbf{v}' \cdot \boldsymbol{\Sigma}_{IC} \cdot \mathbf{v}}}. \tag{9.13}$$

The optimal weights are given by

$$\mathbf{v}^* \propto \boldsymbol{\Sigma}_{IC}^{-1} \cdot \overline{\mathbf{IC}}. \tag{9.14}$$

The proportional constant is determined by normalization of the weights.

9.4 EMPIRICAL EXAMINATION OF CONTEXTUAL APPROACH

In this section we present a series of empirical tests to illustrate the presence of contextual asset pricing. We use the Russell 1000 Index as the security universe, for the time period from December 1986 to September 2004. Data sources include (1) the Compustat quarterly database for financial characteristics; (2) the IBES US historical detail database for consensus earnings estimates; and (3) the BARRA US E3 database for price, return, and risk factor characteristics.

9.4.1 Risk-Adjusted ICs

We first compare the risk-adjusted ICs between sample partitions according to the BARRA definitions of value, growth, and earnings variability. Along these BARRA risk dimensions, we compare the average and the variance of IC, pertaining to the high and low security contexts, for each of the selected composite alpha factors.

Table 9.2 presents these comparisons (15 in all — 3 risk dimensions and 5 alpha measures). We calculate the two-sample *t*-test for the mean difference and the F-test for the variance difference. In Panel A, the return profile of the EF factor is significantly different between high- and low-value stocks. Both the two-sample *t*-test and the F-test are significant at 1% level. For low-value (low book-to-price ratio) stocks the IC is .015, as contrasted with an IC of .044 for high-value stocks. This demonstrates that the way the external financing factor is priced is indeed contextual dependent — more important for discounted firms than high-priced ones. (Note that discounted firm means high value, and high-priced firm refers to low value.) External financing costs and expected investment returns contribute to this contextual dependency. Dilution of shareholder wealth

TABLE 9.2 Comparison of Risk–Adjusted ICs in Different Risk Dimensions

Panel A Value Dimension

	Mean		STD		Two-Sample t Test		F Test			
	High	**Low**	**High**	**Low**	**t**	**p-Value**	**F**	**pval**	**df (num)**	**df (denom)**
RV	0.022	0.022	0.069	0.079	0.011	0.991	0.764	0.270	68	68
OE	0.032	0.040	0.047	0.037	–1.050	0.296	1.613	0.051	68	68
AA	0.027	0.042	0.043	0.050	–1.912	0.058	0.720	0.177	68	68
EF	0.044	0.015	0.041	0.057	3.460	0.001	0.504	0.005	68	68
MO	0.031	0.049	0.061	0.072	–1.577	0.117	0.711	0.163	68	68

Panel B Growth Dimension

	Mean		STD		Two-Sample t Test		F Test			
	High	**Low**	**High**	**Low**	**t**	**p-Value**	**F**	**pval**	**df (num)**	**df (denom)**
RV	0.003	0.034	0.113	0.062	–2.046	0.043	3.318	0.000	68	68
OE	0.061	0.019	0.043	0.042	5.702	0.000	1.037	0.883	68	68
AA	0.044	0.022	0.060	0.039	2.461	0.015	2.450	0.000	68	68
EF	0.028	0.017	0.054	0.043	1.274	0.205	1.567	0.066	68	68
MO	0.059	0.023	0.092	0.072	2.571	0.011	1.623	0.048	68	68

Panel C Variability Dimension

	Mean		STD		Two-Sample t Test		F Test			
	High	**Low**	**High**	**Low**	**t**	**p-Value**	**F**	**pval**	**df (num)**	**df (denom)**
RV	0.023	0.023	0.105	0.076	–0.025	0.980	1.911	0.008	68	68
OE	0.045	0.029	0.051	0.039	2.019	0.046	1.678	0.034	68	68
AA	0.033	0.032	0.049	0.036	0.151	0.880	1.848	0.012	68	68
EF	0.038	0.018	0.055	0.045	2.343	0.021	1.492	0.101	68	68
MO	0.034	0.038	0.094	0.074	–0.252	0.802	1.605	0.053	68	68

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.

is most likely to occur when the invested firm is traded at a discount and starts pursuing capital increases through external financing, because the proceed not only costs more to obtain but also generates lower returns to existing shareholders.

Panel B shows that investors reward high-growth companies for conservative accounting (AA), high OE, and better price and earnings performance (MO). In contrast, cheapness of share price (RV) is an important return driver for low-growth companies, with both the average and the standard deviation of ICs significantly different at 5% level when compared with high-growth companies. Our empirical results are consistent with the ones documented by Scott et al. (1999); and, in addition, we highlight the importance of conservative accounting and operating efficiency as important return drivers for high-growth companies. Consistent with Asness (1997), we find the average IC of momentum factor (MO) in the high-growth stocks is more than twice the size of the average in the low-growth stocks.

Panel C focuses on the earnings variability dimension. Operating efficiency (OE) and EF factors are more indicative of the future stock returns of companies with variable earnings, as shown in their two-sample *t*-tests, which are significant at a 5% level. On the other hand, RV and AA have almost identical average IC across the partitions. However, their standard deviations of ICs, the risk endogenous to the active strategies of applying RV and AA, are significantly different.

To summarize, Table 9.2 is generally consistent with the theory of rational pricing that is conditional. Using univariate average IC comparisons over the 1986–2003 period, we find that the market is more responsive to operating efficiency, conservative accounting, and positive earnings evidence when dealing with high-growth and/or high-priced firms than is the case with low growers. The market is much more focused on operating performance and shareholder-friendly managements when growth is at stake, and much less focused on cheapness of stock prices. Surveying the differences in IC averages and IC standard deviation across the three risk partitions, it appears that the growth dimension induces the most contextual difference, whereas the variability dimension induces the least.

9.4.2 IC Correlations

Table 9.3 reports the IC correlation matrices among the five composite factors in each of the six risk partitions. In each case, the numbers before and after the slash sign are correlations for higher (lower) partitions. Before we comment on the correlation difference across contexts, some

TABLE 9.3 Correlations of Risk-Adjusted ICs

Panel A Value Dimension

	OE	AA	EF	MO
RV	0.28/0.16	−0.22/0.21	−0.08/0.63	−0.11/−0.44
OE		0.42/0.50	0.16/0.24	0.24/0.19
AA			0.21/0.09	0.17/0.14
EF				0.18/−0.23

Panel B Growth Dimension

	OE	AA	EF	MO
RV	−0.22/0.19	0.14/−0.08	0.45/−0.08	−0.71/−0.25
OE		0.36/0.25	0.16/0.27	0.28/0.21
AA			0.23/0.21	−0.18/0.01
EF				−0.32/0.26

Panel C Variability Dimension

	OE	AA	EF	MO
RV	−0.16/0.12	−0.18/0.19	0.19/0.29	−0.60/−0.38
OE		0.30/0.37	0.26/0.38	0.48/0.10
AA			0.28/0.19	0.19/0.04
EF				0.05/−0.23

Note: In each cell, the number before the slash shows correlation of the high context and the number after the slash displays correlation for the low context.

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.

general patterns are worth noting. First, the IC correlation between RV and momentum (MO) is always negative, providing diversification benefit to an active strategy by including both factors. Second, the correlations among the three composite factors from the same quality category, i.e., OE, AA, and EF, are not only all positive in general, but they seem to be rather stable across the risk partitions. Third, the relative value factor tends to have small and often negative correlations with other factors. In all, the market generally prices quality and momentum concurrently, while rotating between cheapness and momentum, each at the expense of the other, due to perhaps changes in risk aversion.

Panel A compares the two correlation matrices derived from the high and low value contexts. The correlations between RV and AA and between RV and EF show the biggest differences. In high-value stocks, the two correlations are −0.22 and −0.08, respectively, whereas in low-value stocks

the two correlations are considerably higher at 0.21 and 0.63, respectively. The other notable difference is the correlation between MO and EF. It is 0.18 in high-value stocks and –0.23 in low-value stocks. Along the growth dimension (Panel B), again the relative value causes most of the correlation differences. Its correlations with OE, AA, and EF all flip signs across the partition. The correlation between RV and MO is negative in both partitions, but it is remarkably low at –0.71 among high-growth stocks. Along the variability dimension (Panel C), the differences in correlation coefficients are smaller compared to those in Panel A and B. In aggregate, MO has lower correlation with other factors in low-variability stocks than in high-variability stocks.

9.4.3 Optimal Factor Weights and Their Differences

In this section, we solve for the optimal weights of the composite alpha factor using the IR maximization framework outlined in Chapter 7. We shall refer to a combination of alpha factors as an alpha model. In each of the six risk partitions, we find the optimal weights of the five composite factors using the IC averages and IC covariances over the whole sample period. Based on the differences of these inputs shown in Table 9.2 and Table 9.3, we naturally expect different alpha models in each high/low risk partition. However, are these weight differences statistically significant? We devise several ways to answer this question. In this section, we perform several direct tests on the optimal weights themselves. Later, we test the performance differences induced by weighting differences, focusing on their alpha-producing capabilities.

To test the statistical significance of the difference between the optimal weights, we adopt a bootstrapping procedure as follows, similar to the one introduced by Michaud (1998). We resample with replacement the historical ICs, jointly for all five composite alpha factors in each of the six security contexts. Similar to a bootstrapping procedure, we make the sample size the same as the number of time periods in the original sample. In each sample, we then calculate the average ICs and IC covariances of five factors along the different risk partitions, and derive IR-maximizing optimal weights. This is repeated one thousand times to obtain one thousand sets of optimal weight in each risk partition. By introducing sampling errors into the average ICs and the IC covariances, we translate the sampling errors of historical ICs into the sampling errors of model weighting. We deem a weight deviation significant if its magnitude is significantly larger than the sampling error.

TABLE 9.4 Resample Weights Comparison in Different Risk Dimensions

Panel A Value Dimension

	Mean		STD		Difference (High–Low)		
	High	**Low**	**High**	**Low**	**Avg/Stdr**	**Avg**	**Stdr**
RV	9.0	6.3	4.0	3.5	0.5	2.6	5.3
OE	16.7	46.4	6.0	8.9	–2.7	–29.7	10.8
AA	20.4	24.4	6.2	6.5	–0.4	–4.0	9.0
EF	43.0	5.1	7.9	4.8	4.1	37.9	9.3
MO	11.0	17.8	4.8	5.1	–1.0	–6.8	7.1

Panel B Growth Dimension

	Mean		STD		Difference (High–Low)		
	High	**Low**	**High**	**Low**	**Avg/Stdr**	**Avg**	**Stdr**
RV	3.7	22.8	2.4	7.3	–2.5	–19.1	7.6
OE	52.7	16.9	7.8	8.3	3.1	35.8	11.7
AA	16.7	33.3	5.0	8.8	–1.6	–16.6	10.1
EF	14.0	16.7	5.9	7.2	–0.3	–2.7	9.3
MO	12.9	10.3	4.0	5.0	0.4	2.6	6.3

Panel C Variability Dimension

	Mean		STD		Difference (High–Low)		
	High	**Low**	**High**	**Low**	**Avg/Stdr**	**Avg**	**Stdr**
RV	7.9	7.2	3.8	4.5	0.1	0.7	5.9
OE	36.1	27.0	7.4	6.5	0.9	9.1	10.0
AA	27.2	41.1	6.3	7.5	–1.4	–13.9	9.6
EF	22.5	10.5	6.6	5.1	1.4	12.0	8.4
MO	6.4	14.2	3.7	4.4	–1.4	–7.9	5.7

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.

The model weights can be compared individually for each of five factors or jointly for all five factors together. For individual comparison, Table 9.4 shows the average and the standard error of factor weights of 1000 bootstrapping samples, again across the 15 samples — 3 risk factor partitions and 5 alpha factors. We also show the difference in optimal weights across the three risk dimensions, in terms of average, standard error, and their ratio. This ratio can be similarly interpreted as a t-statistic, with a value of above 2 or below –2 indicating statistical significance in mean difference. The results in Table 9.4 are consistent with our interpretation of the

univariate IC tests and correlation differences shown earlier. Note the following remarks:

- First, model weights of the high-growth context (Panel B) and the low-value context (Panel A) are remarkably similar. Perhaps, this points to a set of common challenges facing high-priced and high-growth firms, the most prominent of which is to maintain superior operating results captured by the OE factor. However, we note the reverse inference does not apply — model weights in the high-value and the low-growth contexts are quite different. In the high-value context, the most prominent weight (43%) is in EF factor, whereas in the low-growth context, the model weights are relatively equitable for all five factors. Note the relative value (RV) is weighted 23% here, whereas it never receives more than 10% elsewhere.
- Second, we notice that in the growth dimension (Panel B), whereas the RV factor's weight is substantially higher in the low-growth dimension than in the high-growth dimension, with a mean-standard error ratio of –2.5, consistent with the results by Scott et al. (1999); the MO factor's weight is only slightly higher in the high-growth half (12.9%) than in the lower half (10.3%). The reason for this is the higher strategy risk of the MO factor in the high-growth context (Table 9.2, Panel B) than in its counterpart in the low-growth context.
- Table 9.4 unveils primary return drivers for each security context, should they exist. To facilitate the discussion, let's delineate primary drivers as factors that are more than 40% of a model. Contextual partitioning plays a significant role in governing the primary return driver, as it shifts from OE for both high-priced and high-growth firms, to conservative EF for discounted firms and to honest management, gauged by conservative earnings reporting practice (AA), for firms with stable earning stream. These contextual dynamics further highlight the descriptive inadequacy of the one-size-fits-all assumption of traditional quantitative models.
- Across both the value and growth dimensions, there are two factors with significant weights, OE and EF in value and RV and OE in growth. However, across the variability dimension, none of the factors show significant weight difference.

Finally, we note the aggregated weight in the corporate quality category, i.e., the sum of weights in OE, AA, and EF accounts for over 70% of the model weight in almost all cases. This confirms the importance of financial statement analysis in active equity management.

9.4.4 Model Distance

Table 9.5 tests for significance in differences between the optimal weights jointly. For comparison, we first construct a static one-size-fits-all model without any contextual partitioning, using the same resampling procedure. The first row of Panel A shows the resampled efficient weights for this static model and the rest of Panel A show the weights from the previous section.

To compare the factor weights jointly, we employ two measures. The first measure is the distance between two models, defined as

$$d = \sqrt{\frac{\Delta \mathbf{w}' \cdot \Delta \mathbf{w}}{k}}, \tag{9.15}$$

where $\Delta \mathbf{w}$ is the difference in model weights, and k equals five, the number of factors in the model. It is the root mean square of the optimal weight differences. Panel B of Table 9.5 displays the distances between different pairs of models. Several interesting observations are worth noting. First, the static model is most similar to the high-variability contextual model and most dissimilar to the high-value contextual model. Second, when comparing the two contextual models pertaining to same risk dimension, the value dimension has the highest model distance followed by the growth dimension, whereas variability dimension has the smallest distance. Third, consistent with the observation above, the distance between the high-growth model and the low-value model is also very low.

Whereas the distance measure does not incorporate the sample error, our second measure does. Panel C and D of Table 9.5 provide the chi-square statistics between models and their p-value. Note the statistics are not symmetric, as we are testing whether the mean of the resampled weights of one model belongs to the ensemble of the resampled weights of another model. When the models are interchanged, the ensemble is also changed, resulting in a different chi-square statistic. (See Appendix A9.1 for a detailed technical note.) Panel D unveils three interesting findings. First, as shown on the first row (and the first column), the static model

TABLE 9.5 Pairwise Model Weight Comparison

Panel A: Model Weights of Resample Efficient Portfolios

		RV	OE	AA	EF	MO
One-size	R1000	2.5	41.6	36.3	13.0	6.5
Value	High	9.0	16.7	20.4	43.0	11.0
	Low	6.3	46.4	24.4	5.1	17.8
Growth	High	3.7	52.7	16.7	14.0	12.9
	Low	22.8	16.9	33.3	16.7	10.3
Variability	High	7.9	36.1	27.2	22.5	6.4
	Low	7.2	27.0	41.1	10.5	14.2

Panel B: Model Distance

		One-size	Value		Growth		Variable	
		R1000	High	Low	High	Low	High	Low
One-size	R1000	0.0	21.2	9.4	11.7	12.7	7.1	8.7
Value	High	21.2	0.0	24.4	23.2	14.6	14.7	20.0
	Low	9.4	24.4	0.0	7.1	16.9	11.7	13.2
Growth	High	11.7	23.2	7.1	0.0	19.8	11.2	17.8
	Low	12.7	14.6	16.9	19.8	0.0	10.7	7.4
Variability	High	7.1	14.7	11.7	11.2	10.7	0.0	11.0
	Low	8.7	20.0	13.2	17.8	7.4	11.0	0.0

Panel C: Chi-Squared Statistics

		One-size	Value		Growth		Variable	
		R1000	High	Low	High	Low	High	Low
One-size	R1000	0.0	31.8	13.2	19.8	13.5	5.6	7.7
Value	High	69.0	0.0	65.6	39.1	15.9	13.8	49.0
	Low	32.0	36.2	0.0	5.2	21.6	17.0	11.1
Growth	High	16.6	39.7	5.0	0.0	24.9	13.1	19.4
	Low	73.7	18.9	34.0	74.7	0.0	24.2	18.3
Variability	High	11.9	13.7	17.7	14.2	8.6	0.0	9.7
	Low	17.0	23.2	9.9	24.8	7.7	14.0	0.0

Panel D: p-Value of Chi-Squared Test

		One-size	Value		Growth		Variable	
		R1000	High	Low	High	Low	High	Low
One-size	R1000	1.000	0.000	0.010	0.001	0.009	0.235	0.103
Value	High	0.000	1.000	0.000	0.000	0.003	0.008	0.000
	Low	0.000	0.000	1.000	0.264	0.000	0.002	0.026
Growth	High	0.002	0.000	0.282	1.000	0.000	0.011	0.001
	Low	0.000	0.001	0.000	0.000	1.000	0.000	0.001
Variability	High	0.018	0.008	0.001	0.007	0.072	1.000	0.045
	Low	0.002	0.000	0.041	0.000	0.102	0.007	1.000

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.

is statistically different from the contextual models on the growth and the value dimensions at a 5% level. However, contextual models along the variability dimension are not statistically different from the static one. Second, when comparing model weights of the high and low contexts for each risk dimension, value and growth dimensions exhibit significant differences, whereas the variability dimension is questionable. Third, further substantiating the observation, shown in Table 9.3, that the high-growth model is similar to the low-value model, the p-value is either 0.28 when using the covariance from the low-value context or 0.26 when testing with the high-growth covariance; neither is significant.

9.4.5 Contextual Alpha Model

The results of the previous section confirm the benefits of the contextual approach in building quantitative alpha models, and part of the results concerning the value and growth dimensions should be applicable to portfolio mandates with styled benchmarks, as our partitions along these dimensions are partly consistent with how many styled benchmarks are defined. However, what about mandates with core benchmarks? In particular, can we build a contextual model based on our analysis that beats the one-size-fits-all model? In this section, we propose an approach in which factor weightings are dynamically selected and conditioned on the risk characteristics. Then, we compare the performance between contextual models constructed with this approach and the static model. As these models employ the same set of factors, this comparison provides some insight into added value of dynamic factor weightings.

To further illustrate the relevance of each risk dimension, we implement four variants of contextual model, named value, growth, variability, and comprehensive. The first three models are built with a single risk dimension (two security contexts) indicated by their names. For example, the growth contextual model derives its dynamic factor weightings from the high-growth and the low-growth contexts only. In a nutshell, the factor weighting for a particular stock is a linear combination of high-growth and low-growth model, and relative weights of the combination are determined by the stock's growth rate. The comprehensive contextual model takes into account all three contextual dimensions, thus generating return forecasts based on optimal weights from all six security contexts.

To provide a more efficient use of our limited data sample and to facilitate a fair performance comparison, we employ the cross-validation procedure. Specifically, we first divide our sample periods into ten subperiods

chronologically with equal duration. We then elect one of the subperiods as the out-of-sample period, and the remaining nine subperiods become the in-sample period. Although efficient model weights (for both the static and contextual models) are estimated in the in-sample period through our IR optimization framework, the scores (forecasts) are computed based on the estimated factor weights for the out-of-sample periods wherein the model performance is also computed. This exercise is repeated ten times for each of the ten subperiods, whose out-of-sample results are then stringed together to calculate performance statistics. Although we realize this approach creates chronological inconsistency in terms of the sequencing of the in-sample, out-of-sample periods, it is free of potential bias caused by a particular choice of in-sample, out-of-sample periods.

9.5 PERFORMANCE OF CONTEXTUAL MODELS

9.5.1 Risk-Adjusted Portfolios

Table 9.6 compares model efficacy in terms of the excess returns generated by dollar-neutral portfolios, a comparison that incorporates realistic portfolio optimization constraints. Rebalanced on a quarterly basis, portfolios

TABLE 9.6 Performance Comparison of Optimal Dollar-Neutral Portfolios

Panel A: Model Performance

	Static	Value	Growth	Variable	Comparison
Alpha	7.41%	8.53%	8.54%	7.95%	8.57%
IR	1.56	1.63	1.66	1.54	1.72

Panel B: Pairwise Performance Comparison

	Static	Value	Growth	Variable	Comparison
Static		−1.13%	−1.13%	−0.54%	−1.16%
		(**−4.39)	(**−4.75)	(**−3.64)	(**−6.06)
Value	1.13%		0.00%	0.58%	−0.03%
	(**4.39)		(−0.02)	(*2.45)	(−0.23)
Growth	1.13%	0.00%		0.59%	−0.03%
	(**4.75)	(0.02)		(**3.34)	(−0.19)
Variability	0.54%	−0.58%	−0.59%		−0.62%
	(**3.64)	(*−2.45)	(**−3.34)		(**−4.46)
Comp.	1.16%	0.03%	0.03%	0.62%	
	(**6.06)	(0.23)	(0.19)	(**4.46)	

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.

are formed for each model aiming at the highest model score exposures, given that their annualized tracking error is targeted at 5% and they have no exposure to market beta and size. Panel A shows the excess return and IR of each model on an annual basis. Whereas the static model has the lowest excess return and the comprehensive model produces the highest excess return and IR, all models generate excellent performance.

We also compare model performance in a pairwise manner with the average and the t-statistic of performance differences through time. Specifically, each cell in Panel B represents the excess performance between the "active" model indicated by the row title and the "benchmark" model indicated by the column title. As shown on the first column of Panel B, contextual modeling enhances portfolio returns when compared to the static model. The enhancement of quarterly returns ranges from 1.16 to 0.54%. According to the t-statistic (number in parentheses), the comprehensive contextual model provides the most consistent out-performance with a t-statistic of 6.06, followed by the growth contextual model with a t-statistic of 4.75. Also worth noting is the observation that incorporating either the value or the growth dimension captures a significant portion of performance improvement, as the comprehensive implementation only outperforms both models by 3 bps annually, shown on the last row. Lastly, the superior *ex post* performance, delivered by the value and growth models, underscores the importance of the model distance test, which indicates a significant difference vs. the static model for models along the value and the growth dimensions, but not for the variability dimension. Perhaps, the model distance test provides a pathway of selecting contextual models that are likely to deliver better *ex post* returns.

9.5.2 Asset Pricing Tests (Fama–MacBeth Regression)

Table 9.7 documents the advantage of using contextual modeling from the asset pricing perspective. That is, incorporating contextual dependencies provides a better, more accurate description of how stocks are priced. Following the commonly accepted analytical framework employed by asset pricing studies, we apply the Fama–MacBeth regression to estimated returns to model scores through time on a quarterly basis.

Panel A answers the question as to whether contextual models contain relevant asset pricing information that is not captured by the static score. In this test, the dependent variable is a 3-month forward return, and the explanatory variables are beta, size, the static model score, and the residual contextual score (the contextual score netted out the static score). The

TABLE 9.7 Fama–MacBeth Regression Test

Panel A: Residual Contextual Scores vs. the Static Score

	Beta	Size	Static	Residual Comparison	Residual Value	Residual Growth	Residual Variability
Comprehensive	−0.262	−0.035	1.650	1.046			
	(−0.3)	(−0.1)	(12.9)	(6.8)			
Value	−0.288	−0.069	1.649		0.937		
	(−0.3)	(−0.2)	(13.0)		(6.6)		
Growth	−0.262	−0.018	1.653			0.970	
	(−0.3)	(−0.1)	(12.9)			(5.8)	
Variability	−0.223	−0.023	1.661				0.773
	(−0.2)	(−0.1)	(13.0)				(4.5)

Panel B: The Residual Static Score vs. Contextual Scores

	Beta	Size	Residual Static	Comparison	Value	Growth	Variability
Comprehensive	−0.263	−0.035	−0.559	1.915			
	(−0.3)	(−0.1)	(−3.8)	(14.2)			
Value	−0.287	−0.068	−0.274		1.913		
	(−0.3)	(−0.2)	(−2.1)		(13.1)		
Growth	−0.262	−0.018	−0.400			1.913	
	(−0.3)	(−0.1)	(−2.5)			(13.9)	
Variability	−0.224	−0.023	−0.445				1.797
	(−0.2)	(−0.1)	(−2.7)				(12.9)

Note: ( ) contains t-statistic.

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.

netting out allows for an orthogonal treatment, which distills the portion of asset pricing information exclusively contained in the contextual score, thus providing a measure that isolates the incremental value added by the contextual modeling . As shown in Panel A, the residual score of the comprehensive contextual model does indeed capture additional asset pricing information and its t-statistic is 6.8. Similar results are also found when

the three risk-dimension specific models are tested and their t-statistics range from 6.6 to 4.5 — all significant at a 1% level.

Panel B shows the result of a complementary question to the one answered by Panel A. Is the static model statistically dominated by contextual models in the asset pricing test? In other words, does the static score add value when orthogonalized by contextual scores? To answer this question, we include the residual of static score and contextual scores in this set of Fama-MacBeth regressions. The residual score is computed by stripping the portion of variance of the static score that can be explained by the contextual score through OLS regression, the same procedure used in tests shown in Panel A. As shown in Panel B, the contextual score does provide return forecasts that dominate the forecasts of the static model statistically; and the return to the static score residual is not only negative but also statistically significant with a t-statistic of –3.8. Again, similar results are also found in tests of the three risk-dimension specific scores. The *t*-statistics in these three tests range from –2.1 to –2.7.

9.6 SECTOR VS. CONTEXTUAL MODELING

An alternative way to accommodate different sets of return drivers for each security is sector-based alpha modeling. This approach is fairly popular among quantitative practitioners, and it calls for a unique model for each sector, an approach that bears a strong resemblance to how fundamental research is typically organized in investment firms. A sector-oriented fundamental research makes intuitive sense. For fundamental research, it is more cost efficient to have fundamental analysts act as sector specialists who cover companies with similar business dynamics, as opposed to generalists who need to be experts in the full range of business models. Given that human mental capacity is limited, sector specialists should have a better chance of correctly processing categorically similar information. In comparison, when generalists face the challenge of reconciling a diverse spectrum of information, the ability to process it well is only reserved for the most experienced.

However, it is ambiguous why market inefficiencies should differ across sectors in general, simply because their business economics are different. In other words, it is hard to find a conjecture supporting the reason why investors' over- or underreaction to market information should differ for a car company when compared with a computer manufacturer.

On the other hand, some sectors are indeed different due to reasons related to regulation or significantly different business models. They

confront company management with different challenges to add shareholder value, and perhaps warrant a separate model. In the U.S., for example, there are three broad sector categories: utilities, financials, and industrials. The industrial sector is a catch-all sector, which includes companies not belonging to either utility or financial sectors. Similar traits are shared among industrials companies.

Competitiveness: They belong to competitive industries wherein companies compete for business and to generate shareholder value.

Business economics: They share similar business economics. Goods are manufactured and services are rendered. A company's ability to create shareholder value depends on (1) its value add in *the value chain* and (2) the company's competitive standing to retain a portion of the added value.

Management challenges: To be successful, company management teams face similar challenges and engage in similar activities: working capital management, capital allocation decision, corporate financing activities, and business operation enhancement.

In contrast, the utility sector is primarily a regulated, cost-plus industry wherein company profits are both protected as well as capped by governmental regulations. As a result, operating efficiency loses its relevance in determining how competitive a company is. Capital allocation decisions are legislation driven rather than market driven.

The reason why the financial sector deserves a separate model is because of the significance of interest rates. As a result, many alpha factors that are relevant for industrial companies lose their meanings for the financial sector. For example, working capital is not relavent not only because financial companies do not produce inventories, but also because cash is part of the operating assets as cash is interest bearing. It is also an appealing proposition to model financial companies on the industry level — banks, life insurance, property and casualty, real estate investment trust (REIT), and diversified financials (such as brokers and investment managers). Many ratios are only meaningful for one particular financial industry, but not for others. For example, loan loss provision is a relevent matrix for banks, combined ratio is for insurance companies, and funds from operations (FFO) is for REITs.

Therefore, to isolate the appropriate return drivers and to achieve a more efficient forecast, quantitative alpha models should incorporate both

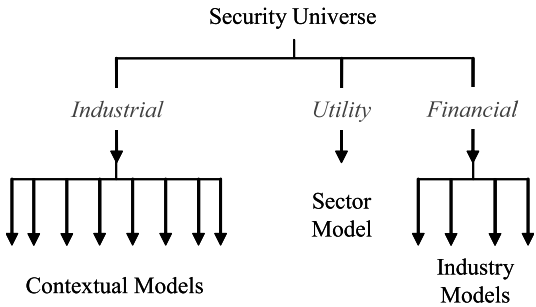


FIGURE 9.1. Modeling hierarchy.

contextual and sector modeling techniques. Figure 9.1 shows a modeling hierarchy that combines both sector modeling and contextual modeling techniques. There are two hierarchical levels: sector modeling being the first level and contextual modeling being the second. On the first level, a cross-section of securities is partitioned into three nonoverlapping sectors: industrial, financial, and utility. Within financial, securities are modeled on the industry level to reflect differences in business operations. Contextual modeling resides on the second level for industrial firms and forms overlapping contextual partitions to capture return idiosyncrasies rooted in behavioral differences. Note the following remarks:

- The combination of contextual and sector modeling enhances quantitative models with greater forecast accuracies (greater conviction in forecasts), a trait typically reserved for fundamental managers. Similar to fundamental research, these advanced forecasting techniques first categories companies based on their business environment and firm characteristics and then applies a set of relevant models to forecast their future returns individually. In doing so, a unique model is tailored for each security whose firm characteristics dictate each individual customization.
- Contextual modeling is a dynamic process over time and adapts to the progression of a company's life cycle. For example, many of today's successful firms (such as Microsoft) were very different a decade ago in terms of their firm characteristics, such as expected growth rate, value ratios, or earnings stability. As a firm evolves through time, its characteristics change and contextual approach adapts to this change by applying different models in forecasting the same security through time.

engage in these projects. On the other hand, companies without worthwhile projects should not spend at all, because spending CAPEX simply wastes shareholders' capitals. There are other links, such as future growth prospects or the cost of equity. For the interest of this section, we will use ROE as the link.

We now discuss each approach in detail.

Quadratic models: Here, we simply add a second-order term of the original factor to the linear model. In the case of a single factor, the model is

$$r = v_0 + v_1 F + v_2 F^2 + \varepsilon . \tag{9.16}$$

Combining a quadratic term with its linear counterpart can provide a better fit to a return response that exhibits nonlinear behavior. The shape of the function (9.16) depends on the signs of coefficients. Assume the coefficient of the linear term is positive. Then, the shape is concave if $v_2 < 0$ and convex if $v_2 > 0$. To model the CAPEX factor, we would have $v_2 < 0$. The expected return increases with the factor, reaches the maximum at $F = -v_1 / 2v_2$ and declines as the factor increases further. Companies with extremely high or low capital expenditures do not represent quality firms, whereas companies with reasonable, conservative capital expenditures do.

Conditional models: We can use another variable to partition the estimation universe into subgroups and construct linear models in each subgroup. In the case of CAPEX, we use ROE as the conditioning variable and create a dummy d_{high_roe}, which is binary –1 for companies with high historical ROE and 0 for companies with low historical ROE. Equation 9.17 isolates the dynamics of how CAPEX is priced for companies with high-ROE projects or those without.

$$r = v_0 + v_1 F_{capex} + v_2 d_{high\_roe} F_{capex} + \varepsilon . \tag{9.17}$$

For low-ROE companies, the model coefficient is v_1 and for high-ROE companies, the model coefficient is $v_1 + v_2$.

Interaction models: One can also use ROE together with CAPEX as an interaction term, i.e., the product of the two. Equation 9.18 shows a model of both ROE and CAPEX and their interaction. The interaction term captures the nonlinear effect. Assuming the coefficient v_3

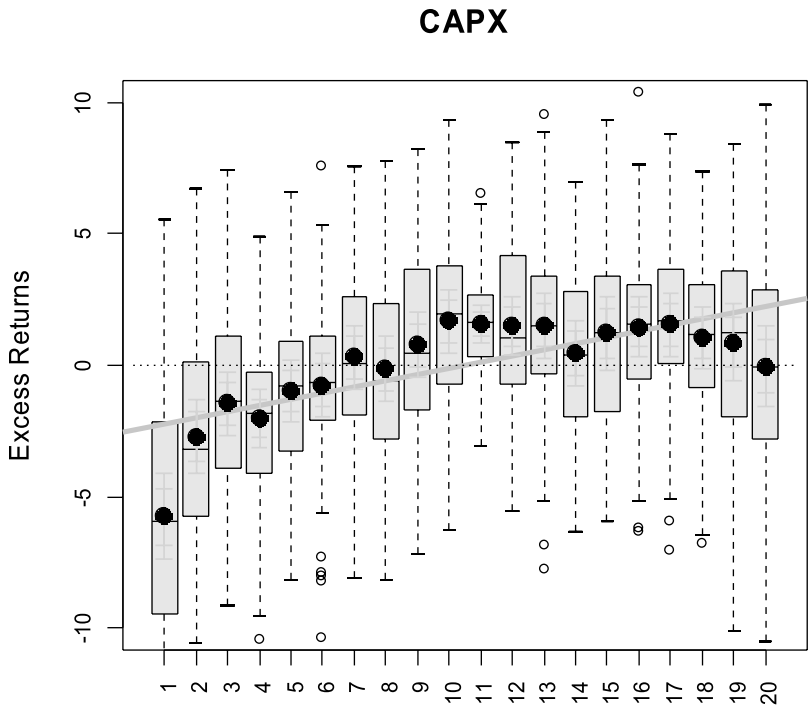


FIGURE 9.2. Fractile backtest of capital expenditure.

9.7.2 Nonlinear Effect Models

There are many ways to capture nonlinear effects. One simple way is to model the expected return using a polynomial by adding quadratic and even cubic terms of the factor values. The end result is still a linear model but with nonlinear factors. This approach is straightforward and flexible, but it often lacks economic intuition. With sufficient data mining, one runs the risk of finding a relationship that is statistically significant, but nonetheless spurious.

A better approach is to condition the factor value on other company attributes. In the case of CAPEX, we ask "What is the appropriate functional form that associates CAPEX with future security returns?" To answer this question, we go back to one of the primary philosophies outlined in Chapter 6. That is, we purchase quality companies that are expected to create shareholder value in the future. How does CAPEX relate to shareholder value generation? One of the important links between CAPEX and shareholder value is the expected ROE. Should a company have worthwhile projects (high-ROE projects), it is shareholder value enhancing to

engage in these projects. On the other hand, companies without worthwhile projects should not spend at all, because spending CAPEX simply wastes shareholders' capitals. There are other links, such as future growth prospects or the cost of equity. For the interest of this section, we will use ROE as the link.

We now discuss each approach in detail.

Quadratic models: Here, we simply add a second-order term of the original factor to the linear model. In the case of a single factor, the model is

$$r = v_0 + v_1 F + v_2 F^2 + \varepsilon . \tag{9.16}$$

Combining a quadratic term with its linear counterpart can provide a better fit to a return response that exhibits nonlinear behavior. The shape of the function (9.16) depends on the signs of coefficients. Assume the coefficient of the linear term is positive. Then, the shape is concave if $v_2 < 0$ and convex if $v_2 > 0$. To model the CAPEX factor, we would have $v_2 < 0$. The expected return increases with the factor, reaches the maximum at $F = -v_1 / 2v_2$ and declines as the factor increases further. Companies with extremely high or low capital expenditures do not represent quality firms, whereas companies with reasonable, conservative capital expenditures do.

Conditional models: We can use another variable to partition the estimation universe into subgroups and construct linear models in each subgroup. In the case of CAPEX, we use ROE as the conditioning variable and create a dummy d_{high_roe}, which is binary –1 for companies with high historical ROE and 0 for companies with low historical ROE. Equation 9.17 isolates the dynamics of how CAPEX is priced for companies with high-ROE projects or those without.

$$r = v_0 + v_1 F_{capex} + v_2 d_{high\_roe} F_{capex} + \varepsilon . \tag{9.17}$$

For low-ROE companies, the model coefficient is v_1 and for high-ROE companies, the model coefficient is $v_1 + v_2$.

Interaction models: One can also use ROE together with CAPEX as an interaction term, i.e., the product of the two. Equation 9.18 shows a model of both ROE and CAPEX and their interaction. The interaction term captures the nonlinear effect. Assuming the coefficient v_3

is positive, the expected return is high for companies with high ROE and high CAPEX, and also for companies with low ROE and low CAPEX. However, the expected return is low for companies with high ROE and low CAPEX, and companies with low ROE and high CAPEX.

$$r = v_0 + v_1 F_{roe} + v_2 F_{capex} + v_3 F_{roe} F_{capex} + \varepsilon \,. \tag{9.18}$$

In general, it is common to see interaction variables in valuation-based factor return estimation, as valuation theory suggests that growth rate, return on invested capital, and cost of capital interact in product terms as well as their linear forms.

9.7.3 Linking CAPEX to Shareholder Value Creation

We combine quadratic and conditional models together to link capital expenditures and shareholder value creation. Specifically, Equation 9.8 shows a functional form that associates CAPEX and ROE with expected value creation and future return forecast.

$$r = v_0 + (v_1 F_{capex} + v_2 F_{capex}^2) + d_{high\_roe}(v_3 F_{capex} + v_4 F_{capex}^2) + \varepsilon \,. \tag{9.19}$$

Figure 9.3 shows the empirical estimation and compares the original CAPEX score (shown horizontally) with the transformed one (shown vertically). Because the universe is broken into high- and low-ROE companies,

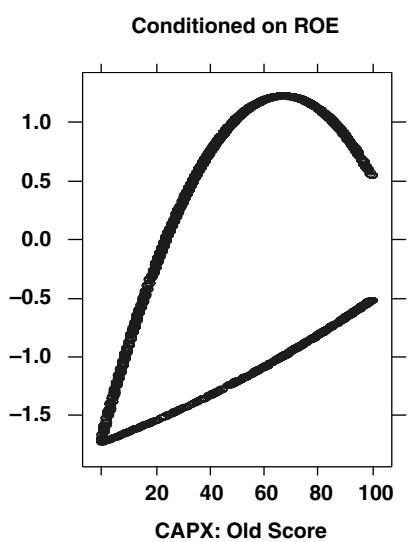


FIGURE 9.3. Transformation of the CAPEX factor.

two fitted lines are shown. The lower one represents low-ROE firms, whereas the upper one represents high-ROE firms. Obviously, high-ROE firms deliver higher returns than low-ROE firms. It is interesting to note that for firms without worthwhile projects, the return response is fairly linear. That is, lower (or even no) capital expenditures bode well, indeed, for low-ROE firms, as they will most likely waste shareholder capital. On the other hand, the return response for high-ROE firms is an upward-sloping, concave curve. The best firms are those who have high-ROE projects and spend conservatively on capital expenditures.

9.7.4 Related Practical Issues

When we introduce new variables to model nonlinear effects, it is important to consider their correlations with existing factors to avoid the multicolinearity problem. In practice, factors are either normalized z-scores or percentile. The former is approximately normally distributed with a restricted range from −3 to +3, and the latter is approximately uniformly distributed between 0 and 1.

Colinearity among factors: The correlation between the quadratic term and the linear term depends strongly on the distribution of the original factors. The correlation is minimal if the z-scores are used and the distribution is approximately normal (see Problem 9.5). On the other hand, the correlation is extremely high if the percentiles are used (see Problem 9.6). The high correlation subsequently results in an unstable estimation. Fortunately, we can use the Gram–Schmidt procedure to address this collinearity issue, as outlined in Chapter 7. The same is true for the correlation between the interaction term (product of two factors) and the original factors.

Conditional dummy: The aforementioned examples use a step function as the conditional dummy wherein there are only two possible values — 0 or 1. One issue with this approach is that the return forecast will change dramatically when a security is re-categorized from 0 to 1 or vice versa. To mitigate this problem, one can use a continuous step function as shown in Figure 9.4.

9.7.5 Nonlinear Effect vs. Contextual Model

Inquisitive reads may see that the conditional factor approach to nonlinear effect modeling is rather similar to the contextual modeling. They are

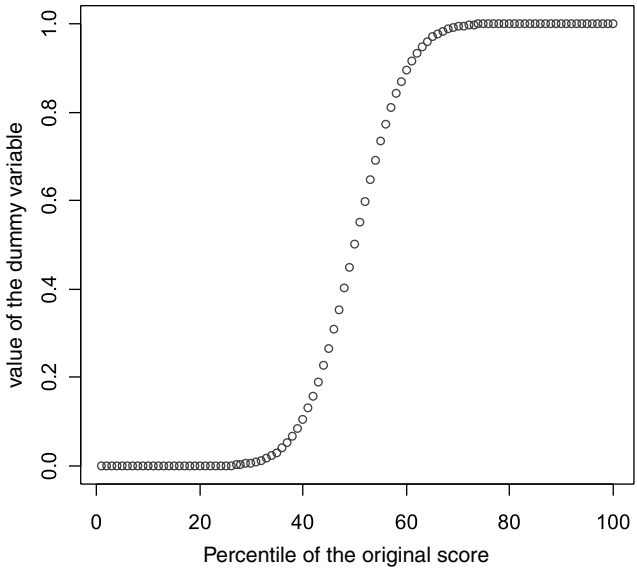


FIGURE 9.4. Continuous slope dummy.

both piecewise linear models. Specifically, both approaches first compartmentalize the cross-sectional security universe into homogeneous subgroups wherein securities tend to behave the same, and then form a set of piecewise linear models, one for each of the subgroups.

What makes them different and when should these approaches be applied? In general, the contextual modeling approach selects subgroups that are homogenous to many different alpha factors. For example, high-growth stocks' responses to cheapness, quality, and momentum are expected to differ from low-growth stocks. In this case, the contextual modeling approach is more appropriate. On the other hand, nonlinear effect modeling typically addresses one factor at a time, like the aforementioned CAPEX example. The security universe is partitioned into subgroups within each context that are expected to have different return responses to the original factor value.

The benefit of selecting the piecewise linear approach, instead of a full-bloom nonlinear modeling approach, is to maintain parsimonious parameterization. In addition, traditional linear statistics are more readily available, easier to understand, and more intuitive to interpret.

The benefit of a simultaneous estimation is the ability to capture different nonlinear effects across various contextual dimensions. In other words, nonlinear effects may also be contextually dependent. In addition,

a simultaneous estimation will also deal with additional distributional issues, such as the correlation between a slope dummy and a contextual dimension. However, the argument against simultaneous estimation is overfitting, because the number of independent variables increases with the introduction of nonlinear terms, resulting in a dramatic decrease in the degrees of freedom.

9.7.6 Empirical Results

To compare the improvement in forecast efficacy, Figure 9.5 shows the decile returns of CAPEX factor for the Russell 2000 security universe. The panel on the left shows the decile performance of the original CAPEX factor and the panel on the right shows the transformed (new) CAPEX factor. Note that the factor return for the new CAPEX score is close to being linear, whereas the return for the original factor is clearly not. This supports our conjecture that a piecewise linear framework with parsimonious parameterization can provide enough flexibility to capture the nonlinear effects, without resorting to a full-bloom nonlinear model.

Modeling nonlinear effects has important implications for the performance of different portfolios. We note that most of the performance

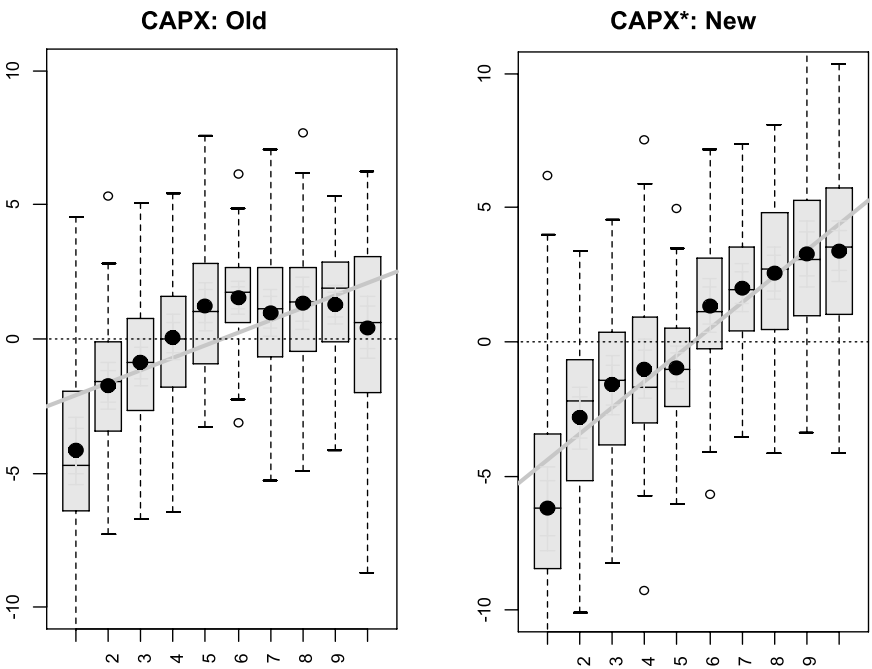


FIGURE 9.5. Performance comparison.

improvement for the new CAPEX factor comes from the long side or highly ranked stocks by CAPEX. As discussed before, CAPEX, in its original form, is effective in identifying losers due to the agency problem, but it does not add much value in picking winners. Therefore, the original factor is not very useful for long-only portfolios, as its benefits mostly come from "avoiding" losers for the long-only portfolios. The new CAPEX factor is now suited for long-only portfolios as well as long-short portfolios, because it symmetrically adds value both on the winner and the loser sides.

9.8 SUMMARY

In this chapter we highlighted two stringent assumptions behind a typical linear return forecasting model. These assumptions are not supported by empirical evidence and they impede the effectiveness of return forecasts. To improve return forecasting models, we introduced two advanced alpha modeling techniques: contextual alpha modeling and nonlinear effect modeling.

Both modeling approaches still utilize multifactor linear alpha models. However, a set of piecewise linear models are estimated and created simultaneously, one for each of the subuniverses that are carefully selected to ensure securities are homogenous within. When forecasting the future return of a security, different models are selected for each security dynamically, depending on the relevance between each model and the particular security. Relevance is governed by the security's attributes, such as growth rate, P/E ratio, or ROE. Nonlinear effects can be modeled in several different ways, including quadratic, conditional, or interaction models.

PROBLEMS

9.1 Find the condition under which the overall IR (9.9) is lower than the high dimension IR.

9.2 Derive the optimal weight (9.12) and calculate the optimal IR with parameters in Example 9.1.

9.3 Plot the function (9.16) for various values of coefficients. Prove that (a) the maximum return is at $F = -v_1 / 2v_2 > 0$ for $v_1 < 0$; (b) the minimum return is at $F = -v_1 / 2v_2 < 0$ for $v_1 > 0$. For the CAPEX factor, which case would apply?

9.4 Suppose factor mean and error mean are both zero in (9.16) and the factor is standardized. Then prove that $v_0 + v_2 = 0$.

9.5 Suppose x is a normally distributed variable with zero mean. Prove that x and x^2 are uncorrelated.

9.6 Suppose x is uniformly distributed in the interval $\left[0,1\right]$. Prove that the correlation between x and x^2 is $\sqrt{15}/4 = 0.97$.

APPENDIX
A9.1 MODEL DISTANCE TEST

To gauge the significance of weighting difference — the likelihood of *not* attributing the cause solely to chance — we bootstrap the IC sample to simulate the inherent randomness of the weight estimation procedure by systematically introducing sampling errors into estimates. The bootstrapping procedure, similar to the one introduced by Michaud (1998), samples historical ICs, with replacement, one thousand times wherein one thousand sets of optimal weights are derived, one for each sample. This exercise is repeated for each security context to generate the set of resample weightings and the average of these weightings. We coin this average, $\mathbf{v}$, as the efficient factor weights — a convention dubbed by Michaud (1998). To illustrate how model distance is determined and tested, let us assume that $\mathbf{v}_1$ and $\mathbf{V}_1$ are the vector of efficient factor weights and the ensemble of resampled model weightings for the first security context, respectively, and that $\mathbf{v}_2$ and $\mathbf{V}_2$ are those for the second context. The vector of weighting difference is simply the difference between $\mathbf{v}_1$ and $\mathbf{v}_2$, $\Delta\mathbf{v} = \mathbf{v}_1 - \mathbf{v}_2$.

The equation below shows the chi-squared statistic when the weighting difference is tested against the sampling error generated from the second security context. The degree of freedom for this chi-squared test is the number of factors minus one, because factor weights sum up to 100%.

$$\chi^2 = \Delta\mathbf{v}' \cdot \mathbf{\Lambda}^{-1} \cdot \Delta\mathbf{v} \,, \tag{9.20}$$

where $\mathbf{\Lambda}^{-1}$ is the inverse of the covariance matrix for either $\mathbf{V}_1$ or $\mathbf{V}_2$.

As different covariance matrix, estimated from either $\mathbf{V}_1$ or $\mathbf{V}_2$, can be selected to compute the chi-squared statistic, significance test results may vary depending on the relative "tightness" of these covariances, albeit the same weighting difference is in question. Figure 9.6 shows a two-dimensional schematic plot of factor weights for a visual demonstration. The weighting difference is significance when using the covariance of $\mathbf{V}_2$ whose distribution on the right is tighter while the result is not significant with $\mathbf{V}_1$'s more diffused distribution. The dashed circles are the loci of significant distances for the two distributions, respectively.

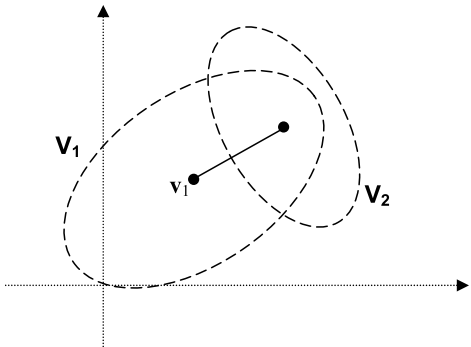


FIGURE 9.6. A two-dimensional projection of ensembles of optimal model weights. (From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005. With permission.)

REFERENCES

Asness, C.S., The interaction of value and momentum strategies, *Financial Analysts Journal*, 29–36, March–April 1997.

Beneish, M.D., Lee, C.M.C., and Tarpley, R.L., Contextual fundamental analysis through the prediction of extreme returns, *Review of Accounting Studies*, Vol. 6, 65–189, 2001.

Beckers, S., Steliaros, M., and Thomson, A., Bias in European analysts' earnings forecast, *Financial Analysts Journal*, 74–85, March–April 2004.

Black, F. and Litterman, R., Asset allocation: Combining investor views with market equilibrium, Goldman, Sachs & Co., Fixed Income Research, September 1990.

Black, F. and Litterman, R., Global portfolio optimization, *Financial Analyst Journal*, 28–43, September–October 1992.

Bruce, B. and Morillo, D., Extreme investing: Using extreme data to find value in small cap stocks, PanAgora working paper, May 2003.

Daniel, K. and Titman, S., Market efficiency in an irrational world, *Financial Analysts Journal*, 28–40, November–December 1999.

Das, S., Levine, C.B., and Sivaramakrishnan, K., Earnings predictability and bias in analysts' earnings forecasts, *The Accounting Review*, 277–294, April 1998.

Fama, E.F. and French, K.R., Multifactor explanations of asset pricing anomalies, *Journal of Finance*, Vol. 51, 55–84, 1996.

Heaton, J.B., Managerial optimism and corporate finance, *Financial Management*, 33–45, Summer 2002.

Huberts, L.C. and Fuller, R.J., Predictability bias in the U.S. equity market, *Financial Analysts Journal*, 12–28, March–April 1995.

Kahneman, D. and Tversky, A., On the psychology of prediction, *Psychological Review*, Vol. 80, No. 2, 273–251, June 1973.

Kahneman, D. and Tversky, A., Prospect theory: An analysis of decision under risk, *Econometrica*, Vol. 47, 263–271, March 1979.

Lee, C.M.C. and Swaminathan, B., Price momentum and trading volume, *Journal of Finance*, Vol. 55, 2017–2070, 2000.

Lo, A.W., Bubble, rubble, finance in trouble? *Journal of Psychology and Financial Markets*, Vol. 3, 76–86, 2002.

Michaud, R.O., *Efficient Asset Management*, Harvard Business School Press, Cambridge, MA, 1998.

Mohanram, P.S., Separating winners from losers among low book-to-market stocks using financial statement analysis, working paper, Columbia Business School, New York, 2004.

Piotroski, J.D., Value investing: The use of historical financial statement information to separate winners from losers, *Journal of Accounting Research*, Vol. 38, 1–41, 2000.

Qian, E. and Hua, R., Active risk and information ratio, *Journal of Investment Management*, Vol. 2, 2004.

Roll, R., The hubris hypothesis of corporate takeovers, *Journal of Business*, Vol. 59, 197–216, 1986.

Scott, J., Stumpp, M., and Xu, P., Behavioral bias, valuation, and active management, *Financial Analysts Journal*, 49–57, July–August 1999.

Scott, J., Stumpp, M., and Xu, P., News, not trading volume, builds momentum, *Financial Analyst Journal*, 45–53, March–April 2003.

Shleifer, A., *Inefficient Markets: An Introduction to Behavioral Finance*, Oxford University Press, New York, 2000.

Sloan, R.G., Discussion of contextual fundamental analysis through the prediction of extreme returns, *Review of Accounting Studies*, Vol. 6, 1991–1995, 2001.

Sorensen, E.H. and Williamson, D., The value of dividend discount models, *Financial Analysts Journal*, November–December 1985.

Sorensen, E.H., Hua, R., and Qian, E., Contextual fundamental, models, and active management, *Journal of Portfolio Management*, Vol. 32, No. 1, 23–36, Fall 2005.

CHAPTER 10

Factor Timing Models

In Chapter 9, we extended the traditional linear alpha model in two dimensions: one is the nonlinear mapping of single alpha factors and the other is the contextual modeling, which constructs different optimal alpha models in different cross sections. The second extension made the model dynamic in the cross-sectional dimension, but we still have constant weights over time. In this chapter, we investigate alpha models with factor timing features that are dynamic through time as well.

Factor timing carries the promise of delivering superior and more consistent excess returns and it is a popular topic among quantitative managers. Similar to other market-timing strategies such as tactical asset allocation, the aim is to increase exposures to factors that are expected to perform positively and to decrease exposures to those that are not. An effective timing mechanism can further raise excess returns delivered by an alpha model. In essence, a factor timing model has time-varying factor weights, i.e.,

$$\mathbf{F}_{c,t} = \sum_{i=1}^{M} v_i(t)\mathbf{F}_{i,t} \,. \tag{10.1}$$

The composite forecast is a weighted average of alpha or risk factors. In contrast to constant weight models, the factor weights $v_i(t)$ explicitly change over time.

Factor timing can be applied to both alpha and risk factors. Many focus on a set of macroeconomic, market-derived, or even technical variables as conditioning instruments. The emphasis on alpha factors is

understandable, as they constitute the ingredients of alpha models; but it is potentially less rewarding because alpha factors, with smaller time-series return variations, offer less opportunity to added value. Risk factors, on the other hand, can have larger time-series return variation, even though their average returns over time are not significant. However, if one can identify periods when a risk factor is expected to have a positive information coefficient (IC), one can use it as an alpha factor in those periods.

In this chapter, we will discuss two avenues of partitioning factor returns through time: calendar timing and macro economic timing. We will review research publications in these areas and use U.S. market and selected major non-U.S. markets as examples to show empirical back-test results. We shall also discuss the portfolio implementation issues that are associated with factor timing and its design considerations.

10.1 CALENDAR EFFECT: BEHAVIORAL REASONS

In this section, we shall illustrate calendar conditioning on certain traditional risk factors, especially those concerning investment quality. Return profiles of these factors are characterized by low unconditional means but high unconditional variance. Hence, unconditional exposures to these risks are not compensated but skilled timers could reap generous rewards. Specifically, we examine a strategy that longs high-risk, low-quality stocks in the first half of a calendar year and shorts them in the second half. In this section, we document potential profit opportunities pertaining to both U.S. and some major non-U.S. markets.

What could cause the seasonal pattern of returns to these risk factors, which is related to the familiar January effect?[1] We suggest that investors' behavior, specifically their risk preference, exhibits a seasonal pattern. As a result, returns to many factors that measure investment risk of common stocks exhibit a calendar pattern.[2] This phenomenon appears to be a year-long event, encapsulating the January effect as a prominent manifestation. Such a phenomenon reflects: (1) the investors' belief in the time-diversification benefit and (2) the annual frequency with which they evaluate their investment performance. Note the following:

- Carrying this logic one step further, as most investors also evaluate their performance on a quarterly basis, our behavioral framework would also suggest a quarterly pattern in which returns to quality factors are higher in the quarter-ending months than in the beginning

months. Empirical tests show that such a pattern does exist in the U.S., although it is less prominent compared to the annual pattern.

Although the notion of time diversification has been applied and debated in terms of asset allocation for investment horizons spanning multiple years, it seems to be equally applicable in a shorter, yet repeatable, time frame of 1 year, in explaining the calendar effect.

10.1.1 Seasonal Behavioral Phenomenon

The reason why calendar events might dictate investors' risk tolerance can be traced to the debate about the validity of time diversification, first articulated by Samuelson (1963).[3] For practical purposes, we can assume that a large percentage of investors evaluate their performance annually on December 31,[4] which is a common evaluation date. In this case, the evaluation horizon is the longest in January and shortest in December. When the evaluation period is long, the investment decision in selecting risky investments is analogous to the choice of whether or not to participate in a series of high-risk, high-reward bets. In contrast, the constraint of a short evaluation horizon induces investment behavior that is similar to the choice of accepting a single risky bet. As illustrated by Samuelson (1963), investors are more risk tolerant when participating in a series of bets, pinning their hope on a misguided interpretation of the law of large numbers. Consequently, this common evaluation period gives rise to varying lengths of evaluation horizons during the course of a year, eliciting changing risk aversion. As such, investors' preference for risky stocks exhibits calendar seasonality, their risk tolerance being highest in January and then gradually decreasing with December being the lowest. Furthermore, as the calendar date shifts from December 31 to January 1, investors' bearish sentiment toward low-quality companies is suddenly replaced by a bullish one, which causes an imbalance between the supply and demand for low-quality stocks. As such, excessive demand quickly bids up the prices of low-quality stocks in January, giving rise to the January effect.

The consequences of this risk-aversion pattern are reflected in returns to various factors measuring company risk. We define risk in the context of fundamental characteristics of a company, a common practice among equity managers. In general, a company with stable earnings, above-average return on investments, and conservative financing is typically associated with quality and low investment risk. A high-risk, low-quality stock exhibits characteristics to the contrary. (Specific definitions of these

factors are illustrated in the next section.) We explore the calendar effect in terms of the seasonal pattern of returns to these risk factors.

Our conjecture regarding the reason behind the calendar effect is built on two premises: a misguided belief in time diversification and the annual performance review that investors, especially professional money managers, must undergo. Both of these topics have received attentions from the academic community.

10.1.2 The Controversy over Time Diversification

The time diversification controversy emerges from the question, "Can investment risk be diversified through time as prescribed by the law of large numbers?" Samuelson (1963) proved mathematically that investors should not change their exposure to risky assets based on their time horizon, assuming investors' utility function equals the logarithm of terminal wealth. Additionally, Kritzman and Rich (1998) clarified the time diversification debate and stated that the subjects that merit discussion are Samuelson's assumptions: (1) investors' risk aversion is independent of wealth changes, (2) investment returns are random, and (3) investment return is the only source of wealth accumulation.

Fisher and Statman (1999) questioned the descriptive accuracy of Samuelson's first assumption, in which an investor is risk averse and the investor's utility is a function of terminal wealth, an axiomatic tenet of expected utility theory modeling rational decision-making under uncertainty. They suggested that when prospect theory, introduced by Kahneman and Tversky (1979, 1992), is used in place of the standard utility assumption, it is plausible for an investor to achieve a higher expected utility as the investment horizon lengthens. The difference emerges from the value function of prospect theory, in which an investor is loss averse and his utility is derived from changes in wealth with respect to a reference point, such as his current wealth. The specific differences between a standard utility function and the value function of prospect theory are depicted in Figure 10.1 and Figure 10.2. In Figure 10.1, according to expected utility theory, an investor's utility is a function of terminal wealth — a smooth, concave curve representing risk-averse behavior, whereas, in Figure 10.2, the value function is defined in terms of gains and losses, where the curve is concave for gains and convex for losses, representing the behavior of loss aversion. This convex value function for losses exaggerates the adverse psychological cost of small losses and dampens the adverse impact of large losses, causing an investor to treat losses equally, at least psychologically.

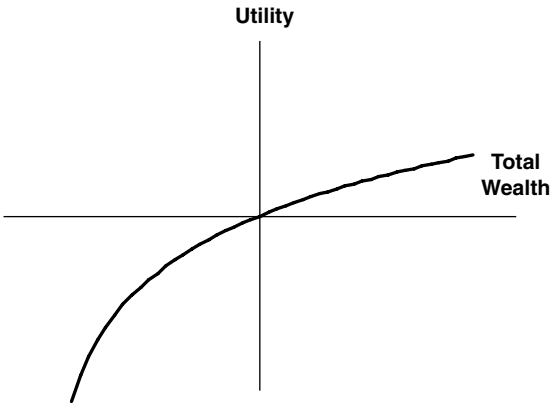


FIGURE 10.1. Standard utility function.

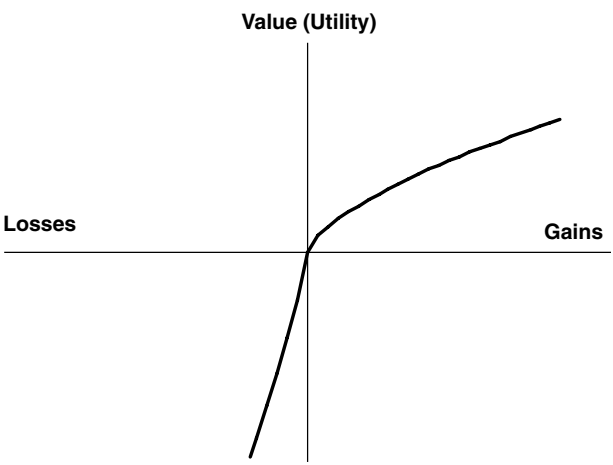


FIGURE 10.2. Value function of prospect theory.

When evaluating risk over different time horizons, this propensity to label losses equally causes investors to overlook the fact that the magnitude of possible losses increases with the investment horizon. As such, investors appear to be more risk tolerant as the horizon lengthens, because they focus only on the fact that the probability of losses diminishes with the horizon, without appropriately reckoning the increased magnitude of these potential losses.

Empirically, Olsen (1997) showed that the results of surveys of professional investors confirmed predictions of prospect theory instead of predictions of expected utility theory. This is probably not unexpected,

because most professional investors measure their performance relatively, either to a benchmark, a competitors' average, or both. Therefore, their value added is in terms of gain and loss, as prescribed by prospect theory. In particular, Olsen found money managers exhibit loss aversion as predicted by the value function of prospect theory and that money managers also believe in the benefit of time diversification.

10.1.3 Annual Performance Review

Prior studies indicate the frequency with which investors review their portfolio performance can influence investment results. For example, Benartzi and Thaler (1995) showed that the historical equity risk premium, which seems unreasonably large when compared to risk-free returns, is actually consistent with the conjecture that average investors evaluate their portfolios on an annual basis. In addition, they argued that the attractiveness of risky investments depends on how often an investor evaluates his portfolio, rather than his investment horizon. Brown et al. (1996) examined the behavior of mutual fund managers and characterized the mutual fund industry as a multiperiod, multigame tournament where portfolio managers participate each year as contestants. In other words, each year is portrayed as one of the repeating games that starts on January 1 and ends on December 31. As a whole, these studies point to investors' propensity to evaluate performance on an annual basis and its behavioral effects on investors.

For individual investors, Benartzi and Thaler (1995) suggested that household budget planning, tax reporting, and comprehensive year-end performance reports trigger annual performance evaluation. For institutional investors, annual evaluation, and to a lesser degree quarterly evaluation, are the result of the "agency problem." To protect their own interests, institutional investors routinely evaluate whether the managers they hired are delivering adequate performance to justify the fees paid. Moreover, annual performance evaluation carries substantive consequences in determining managers' compensation and their continued employment. Consequently, professional managers also behave as if their investment horizon is just one year. Alternatively, Brown et al. (1996) attributed the heightened focus of annual performance to how performance is compiled and ranked by business publications and information services, such as Morningstar Mutual Fund Services and Lipper Analytical Services.

10.2 CALENDAR EFFECT: EMPIRICAL RESULTS

10.2.1 Testable Hypotheses

The belief in time diversification coupled with an annual performance review gives rise to the calendar effect. To test the effect, two testable hypotheses are examined. The main hypothesis emerges from the prediction that returns to quality stocks are higher in the second half of a calendar year when compared with returns in the first half. We shall exclude the months of June and July because they are in the middle of a calendar year, when investors' risk preference is neutral. Hence, factor returns in these 2 months are primarily driven by other market influences, such as earnings announcements, and possibly the Russell index reconstitution, which occurs in June of each year.

Hypothesis I

$$E(\text{factor return}|\text{January} - \text{May}) = E(\text{factor return}|\text{July} - \text{December}):$$

The first null hypothesis is that the expected factor returns from January to May and from July to December are the same.

Hypothesis II

$$\sigma(\text{factor return}|\text{January} - \text{May}) = \sigma(\text{factor return}|\text{July} - \text{December}):$$

In addition to the return hypotheses, we argue that investment risks associated with these calendar partitions are comparable. This hypothesis distinguishes our behavioral explanation from a risk-based alternative, in which varying levels of risk are compensated with commensurate returns.

10.2.2 Definition of Quality

Our definition of quality is similar to that of traditional fundamental analysis, in terms of a company's history of creating value for shareholders and the management's ability to allocate capital efficiently. High-quality, low-risk companies[5] exhibit the following characteristics:

1. Superior economic value creation: high returns on net operating assets (RNOA) or high returns on equity (ROE)
2. Low financial leverage: low debt to assets
3. Low bankruptcy risk: low debt-to-market value and high interest coverage

4. Superior market value creation: high 3-year total return
5. High goodwill priced by the market: high price-to-book ratio
6. Positive earnings outlook: high earnings revision
7. Stable earnings steam: high earnings stability

Although all of these factors exhibit explanatory power for the cross-sectional dispersion of stock returns, thus at least qualifying them as risk factors, markets only reward two of them according to unconditional asset pricing studies: earnings revision and the price-to-book ratio.[6,7] In other words, returns to these two factors have a positive average over time, whereas returns to the other five are not significantly different from 0.

However, as we demonstrate in the following text, when conditioned on calendar months, especially on the semiannual divisions, returns to these nonpriced risk factors exhibit a calendar pattern with a consistently negative bias in the first half, and at the same time, a consistently positive bias in the second half. As for the two alpha factors, their returns are also higher in the second half than in the first half.

10.2.3 Data and Test Methodology

The data sample for this study contains securities in the Russell 3000 index, and the sample period covers January 1987 to September 2003. Fundamental data used to construct quality factors come from the Compustat quarterly database, and price-, return-, and risk-related data are supplied by the BARRA USE3 model.

To facilitate empirical tests, we first compute the risk-adjusted IC of each month as described in Chapter 4. These monthly ICs are then divided into two groups representing semiannual partitions of a calendar year. These ICs are used to test Hypothesis I mentioned earlier. To show the level and the significance of the difference in IC average, we conduct two mean difference tests: two-sample *t*-test and Wilcoxon rank test. The two-sample *t*-test assumes that both groups are normally distributed and their standard deviations are different. We report the *t*-statistic, the p-value, and degrees of freedom of this test. To lessen the normality assumption, we perform the Wilcoxon rank test, in which ranking differences are compared between the two calendar groups. Similarly, we report the W-score and p-value of this test.

To test Hypothesis II, we examine the difference in standard deviations of ICs between the two groups using the F-test.

10.2.4 Empirical Results

We examine the seasonal return patterns using box charts. We first select the 3-year price momentum factor and the price-to-book ratio for this demonstration because these factors have been thoroughly analyzed in the academic literature in an unconditional, cross-sectional asset-pricing framework. In contrast, our results cast light from a calendar-conditioning perspective.

In Figure 10.3, the risk-adjusted IC of 36-month price momentum is collated and plotted in various partitions of calendar months. Panel A shows IC distributions of four calendar partitions (January, February–May, August–November, and December); and Panel B displays IC distributions for each calendar month. Using the price momentum factor as a quality proxy, returns to quality are perverse in the first half of the calendar year, as shown in Panel A, with January being the most negative month. However, in the second half, investors purchase stocks with high price momentum at the expense of those with low-price-momentum. This flight-to-quality behavior is especially pronounced in December.

The unconditional average of ICs is quite close to 0 shown as the dashed line. This qualifies the 36-month price momentum as a nonpriced risk factor: the market does not compensate investors who take such risk unconditionally. However, when examining the calendar effect more closely, evidence shows that investors prefer low-quality stocks in the first half of the calendar year and then change their minds in the second half by selling those low-quality stocks purchased in the first half.

Figure 10.4 shows similar results for price-to-book. Low price-to-book is indicative of low quality and reflects the destruction of shareholder value by a particular company.[8] As illustrated, investors prefer low price-to-book securities in the earlier part of a year and reverse their preference in the later part of the year. However, the aggregated average for the whole year is negative, reflecting the fact that the unconditional return to price-to-book is negative, and it is an alpha factor when used properly.

10.2.5 Results of Hypothesis Tests

Nine quality factors tested individually and their results are reported in Table 10.1. Empirical results, with both the *t*- and the Wilcoxon tests, unanimously reject the null hypothesis I for all quality proxies with statistical significance. (Note that * denotes a 90% confidence level, and ** denotes a 95% confidence level.) This underscores the seasonal behavior of quality proxies, in which returns to quality are much higher in the last 5

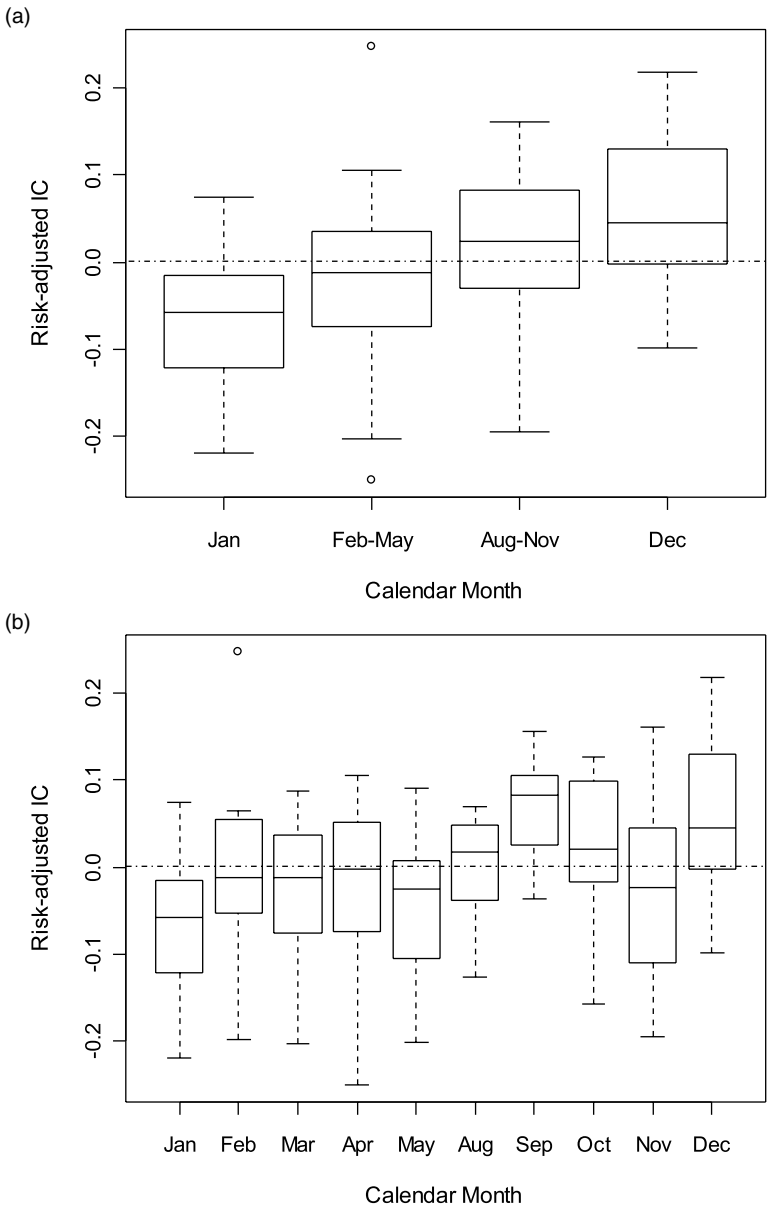


FIGURE 10.3. Risk-adjusted IC of 36-month price momentum: (a) by four calendar groups and (b) by calendar months.

months of a calendar year than in the first 5 months. Also consistent with our conjecture are the results of the F-test, lending support to Hypothesis II in which levels of investment risk indigenous to those two calendar

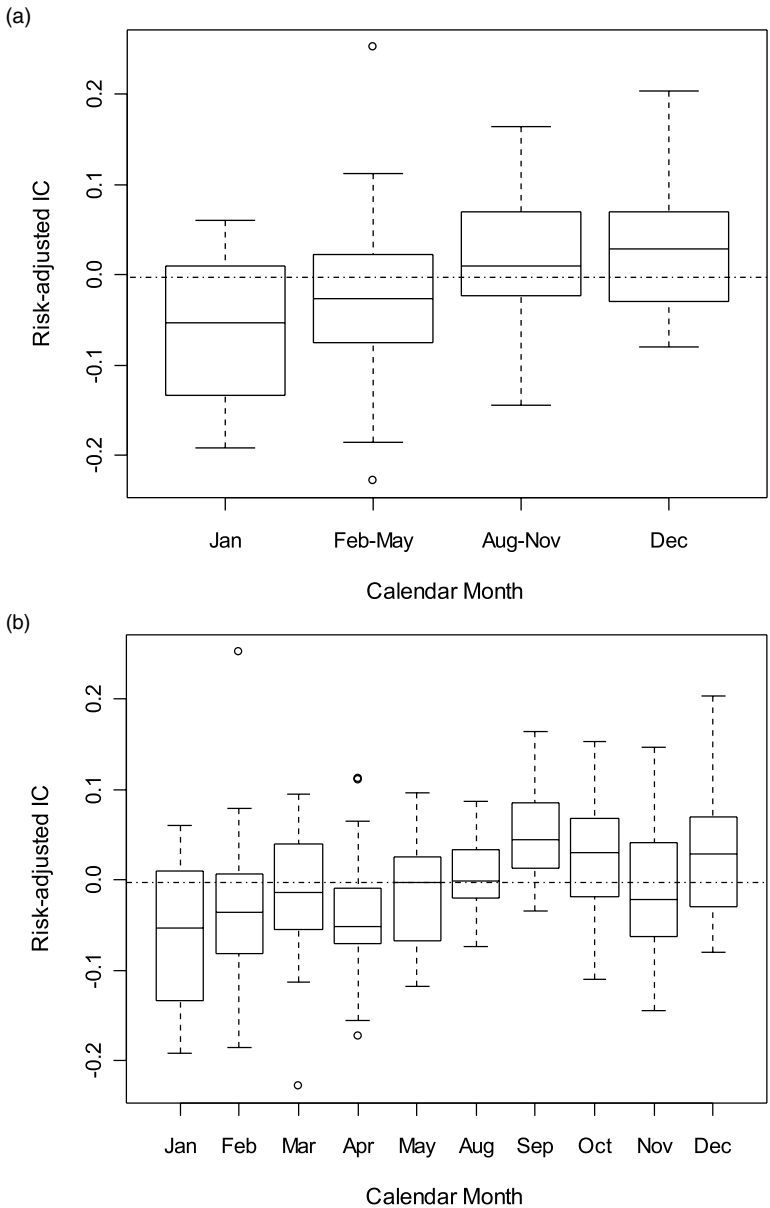


FIGURE 10.4. Risk-adjusted IC of price-to-book: (a) by four calendar groups and (b) by calendar months.

partitions are similar. As shown in Table 10.1, none of the p-values (the third column from the right) reject the null Hypothesis II. Hence, it is unlikely that the calendar return pattern is a result of varying levels of

TABLE 10.1 Summary Test Statistics: Full Sample (1987–2003)

	Two-Sample t-Test			Wilcoxon test		Two-Sample Variance-Test			
	t	p-Value	df	W	p-Value	F	p-Value	num df	denom df
36–month Price Momentum	**–4.26	0.000	165.0	**2185	0.000	1.06	0.782	84	81
12–month Earnings Revision	**–2.75	0.007	159.7	**2594	0.004	0.74	0.182	84	81
RNOA	**–3.80	0.000	160.9	**2252	0.000	*1.49	0.074	84	81
ROE	**–3.47	0.001	164.5	**2411	0.001	1.20	0.416	84	81
Price–to–Book	**–4.46	0.000	163.0	**2200	0.000	1.34	0.185	84	81
–1 * Debt–to–Assets	**–3.06	0.003	164.8	**2578	0.004	1.01	0.964	84	81
Interest Coverage Ratio	**–4.40	0.000	160.2	**2116	0.000	1.33	0.195	82	81
–1 * Debt–to–Market	**–3.75	0.000	163.1	**2389	0.000	0.86	0.506	84	81
Earnings Stability	**–4.33	0.000	162.4	**2208	0.000	0.84	0.415	84	81

Note: * = 90% confidence level; ** = 95% confidence level.

risk. Lastly, similar tests are also conducted by excluding January and December from the sample to demonstrate that the calendar effect is truly a yearlong phenomenon. Again, results corroborate our conjecture.

To explore the temporal dynamics of the calendar effect, we divide the sample into two subperiods: 1987–1994 and 1995–2003. Panel A of Table 10.2 shows the results for 1987–1994, and Panel B reports the results for 1995–2003. The calendar effect is observed in both subperiods with the first period being more statistically significant than the second. In particular, RNOA and ROE measures are no longer significant in the second half, although the signs are still consistent with the prediction. Potential explanations of the temporal differences can perhaps be traced to other macroeconomic influences, such as the market state or the monetary policy environment;[9] alternatively, the diminishing profitability can perhaps be linked to the adaptive market efficiency.

10.2.6 Quarterly Evaluation Horizon

Examining the existence of the seasonality on a quarterly basis offers a further extension of the calendar effect. Because quarterly performance reporting is also common for both mutual funds and personal accounts, a seasonal pattern of returns should also be observed. To verify this, we partition the monthly ICs into beginning months (January, April, July, and October) and ending months (March, June, September, and December) of calendar quarters. Table 10.3 reports results of quarterly tests. We make the following remarks:

- The evidence from the quarterly test confirms our main hypothesis because the signs are negative across all tested factors, indicating a low-quality bias in the beginning months and a high-quality bias in the ending months. As expected, the quarterly seasonality is uniformly less prominent than their annual counterpart, although 5 out of the 9 tested factors still show statistical significance at a 5% level. In addition, the variance test shows the same result found in the annual test (Table 10.1): investment risks pertaining to beginning and ending months are similar.

10.2.7 Non-U.S. Markets

If the explanation behind the calendar effect is behavioral as we suggested, then the phenomenon might be universal and thus observable in non-U.S. markets. Therefore, tests conducted in non-U.S. markets could unveil

TABLE 10.2 Summary Statistics by Subperiods

(A) 1987–1994

	Two-Sample t-Test			Wilcoxon Test		Two-Sample Variance-Test			
	t	p-Value	df	W	p-Value	F	p-Value	num df	denom df
36-month Price Momentum	**–3.82	0.000	77.7	**429	0.000	1.12	0.719	39	39
12-month Earnings Revision	**–2.11	0.038	74.2	**545	0.014	0.63	0.155	39	39
RNOA	**–4.81	0.000	76.7	**325	0.000	1.29	0.424	39	39
ROE	**–4.45	0.000	76.0	**380	0.000	1.39	0.306	39	39
Price-to-Book	**–4.05	0.000	75.2	**440	0.000	1.48	0.225	39	39
–1 * Debt-to-Assets	*–1.94	0.056	76.6	*622	0.088	1.31	0.407	39	39
Interest Coverage Ratio	**–4.84	0.000	75.9	**331	0.000	0.97	0.916	37	39
–1 * Debt-to-Market	**–3.75	0.000	77.5	**445	0.001	1.17	0.628	39	39
Earnings Stability	**–4.17	0.000	77.9	**414	0.000	0.92	0.789	39	39

(B) 1995–2003

	Two-Sample t-Test			Wilcoxon Test		Two-Sample Variance-Test			
	t	p-Value	df	W	p-Value	F	p-Value	num df	denom df
36-month Price Momentum	**–2.41	0.018	84.8	**672	0.020	1.04	0.904	44	41
12-month Earnings Revision	*–1.79	0.077	82.6	*748	0.095	0.82	0.517	44	41
RNOA	–1.04	0.299	79.8	804	0.234	**1.95	0.033	44	41
ROE	–0.93	0.356	84.8	838	0.367	1.06	0.863	44	41
Price-to-Book	**–2.40	0.019	84.5	**664	0.017	1.33	0.355	44	41
–1 * Debt-to-Assets	**–2.33	0.022	82.8	**671	0.020	0.83	0.544	44	41
Interest Coverage Ratio	*–1.94	0.055	82.7	*731	0.070	1.62	0.123	44	41
–1 * Debt-to-Market	*–1.71	0.092	81.8	753	0.104	0.77	0.401	44	41
Earnings Stability	**–2.15	0.035	81.7	**703	0.040	0.77	0.389	44	41

Note: * = 90% confidence level; ** = 95% confidence level.

TABLE 10.3 The Beginning Months vs. the Ending Months of Calendar Quarters

	Two-Sample t-Test			Wilcoxon test		Two-Sample Variance-Test			
	t	**p-Value**	**df**	**W**	**p-Value**	**F**	**p-Value**	**num df**	**denom df**
36–month Price Momentum	**–3.28	0.001	130.0	**1631	0.006	1.29	0.309	66	66
12–month Earnings Revision	**–2.88	0.005	128.6	**1708	0.017	1.39	0.186	66	66
RNOA	**–2.18	0.031	123.7	**1767	0.034	**1.70	0.033	66	66
ROE	–1.55	0.123	120.4	1900	0.126	**1.90	0.010	66	66
Price–to–Book	**–3.18	0.002	132.0	**1603	0.004	0.98	0.947	66	66
–1 * Debt–to–Assets	–1.23	0.221	131.6	1901	0.127	1.12	0.644	66	66
Interest Coverage Ratio	**–2.46	0.015	128.9	**1661	0.013	1.26	0.360	65	66
–1 * Debt–to–Market	*–1.94	0.055	131.7	*1858	0.086	1.09	0.714	66	66
Earnings Stability	–0.97	0.335	131.8	2033	0.348	1.07	0.771	66	66

Note: * = 90% confidence level; ** = 95% confidence level.

global return opportunities. In addition, indigenous cultural differences could also impose observable deviations in some of these markets, reflecting the multifaceted nature of behavioral influences. For example, we are particularly interested in ascertaining whether the Chinese New Year[10] shifts the cycle of calendar pattern accordingly in Asian markets such as Hong Kong[11].

Our non-U.S. sample covers the period from January 1990 to December 2003 and holdings of the Citigroup broad market index constitute the security universe. The risk-adjusted IC is calculated similarly to the U.S. tests, except that the BARRA GEM risk model supplies the risk loadings. Fundamental data items come from the World Scope database with a 6-month lag to avoid look-ahead bias. The same definitions of quality proxies are tested in this exercise, except we use return volatility in place of earnings variability. We perform a two-sample *t*-test and an F-test[12] in selected major markets.

In Table 10.4, Panel A reports the t-statistic and the p-value, in parentheses, of the two-sample *t*-tests. Calendar seasonality is prominently observed in the U.K., France, and Japan, in which all tested factors show negative readings, with a majority tested significant at a 10% level. Canadian evidence is weaker with eight (out of nine) factors showing the right negative readings, only to fall short in statistical confidence with just three being significant. In all, evidences gathered in the aforementioned four markets provide supports for calendar phenomenon. However, there are two noticeable exceptions: Hong Kong and Germany.

For the Hong Kong market, calendar seasonality is not observed in Table 10.4. To ascertain whether the review date is influenced by cultural differences, Table 10.5 reports test results using February as the end of calendar year instead of December to accommodate the Chinese lunar year calendar, which starts mostly in February[13]. As shown in Panel A, seasonality becomes more noticeable in Hong Kong as seven (out of nine) factors show negative readings, whereas the test results become significantly weaker in the other markets, especially in France. This stark contrast, induced by a calendar shift, perhaps exemplifies the linkage of how indigenous cultural differences impose systematic behavioral changes, ultimately resulting in observable variations in the formation of the calendar phenomenon. Table 10.6 shows a more remarkable contrast when we elect March 31 as the end of a calendar year[14]. In this test, all nine factors show negative readings in the Hong Kong market, and four of them are

TABLE 10.4 Summary Statistics of Non-U.S. Markets

(A) Two-Sample *t*-Test (1990–2003)

	U.K.	France	Germany	Canada	Japan	Hong Kong
36–month Price Momentum	**–2.74 (0.00)	*–1.45 (0.08)	–0.38 (0.35)	*–1.52 (0.07)	**–3.10 (0.00)	–0.23 (0.41)
12–month Earnings Revision	**–2.71 (0.00)	**–1.86 (0.03)	*–1.51 (0.07)	**–1.79 (0.04)	–0.51 (0.31)	0.20 (0.58)
RNOA	–1.18 (0.12)	–1.21 (0.11)	2.51 (0.99)	–0.68 (0.25)	–0.10 (0.46)	0.57 (0.72)
ROE	–0.66 (0.25)	–0.91 (0.18)	1.57 (0.94)	–0.94 (0.18)	–0.55 (0.29)	1.63 (0.95)
Price–to–Book	*–1.37 (0.09)	**–2.08 (0.02)	**–1.74 (0.04)	**–1.94 (0.03)	**–2.32 (0.01)	1.05 (0.85)
–1 * Debt–to–Assets	**–2.09 (0.02)	**–1.68 (0.05)	0.76 (0.78)	–0.68 (0.25)	**–3.57 (0.00)	–1.06 (0.15)
Interest Coverage Ratio	**–2.07 (0.02)	–1.17 (0.12)	1.33 (0.91)	–0.84 (0.20)	**–2.92 (0.00)	–0.23 (0.41)
–1 * Debt–to–Market	–0.58 (0.28)	*–1.48 (0.07)	0.31 (0.62)	1.02 (0.85)	–1.14 (0.13)	0.37 (0.64)
Low Return Volatility	–0.14 (0.44)	**–1.99 (0.02)	**–2.62 (0.00)	–0.88 (0.19)	**–2.69 (0.00)	*–1.47 (0.07)

(B) F-Test (1990–2003)

	U.K.	France	Germany	Canada	Japan	Hong Kong
36-month Price Momentum	1.07 (0.78)	1.00 (1.00)	**1.74 (0.02)	1.28 (0.30)	1.21 (0.43)	1.01 (0.98)
12-month Earnings Revision	1.08 (0.76)	0.70 (0.13)	1.28 (0.31)	*0.66 (0.08)	**1.86 (0.01)	1.07 (0.77)
RNOA	1.00 (0.99)	1.11 (0.67)	1.08 (0.75)	1.00 (0.99)	1.11 (0.66)	0.86 (0.53)
ROE	0.99 (0.97)	0.87 (0.55)	1.18 (0.49)	0.74 (0.20)	1.15 (0.55)	0.74 (0.21)
Price-to-Book	1.10 (0.68)	1.03 (0.89)	1.26 (0.34)	0.94 (0.80)	1.09 (0.71)	0.95 (0.83)
–1 * Debt-to-Assets	1.08 (0.74)	0.94 (0.81)	0.89 (0.62)	0.79 (0.32)	1.15 (0.56)	1.08 (0.75)
Interest Coverage Ratio	1.14 (0.58)	0.96 (0.85)	*0.67 (0.09)	**0.62 (0.04)	1.29 (0.29)	1.30 (0.27)
–1 * Debt-to-Market	**1.60 (0.05)	0.98 (0.92)	1.20 (0.44)	1.38 (0.18)	1.25 (0.35)	0.94 (0.79)
Low Return Volatility	1.00 (0.98)	1.30 (0.27)	0.99 (0.98)	0.85 (0.48)	0.71 (0.14)	1.03 (0.90)

Note: * = 90% confidence level; ** = 95% confidence level.

TABLE 10.5 Summary Statistics of Non-U.S. Markets, Electing February 28 as the Annual Review Date

(A)Two-Sample *t*-Test (1990–2003)

	U.K.	France	Germany	Canada	Japan	Hong Kong
36-month Price Momentum	**–1.73 (0.04)	0.16 (0.56)	–1.11 (0.13)	*–1.37 (0.09)	–0.74 (0.23)	0.45 (0.67)
12-month Earnings Revision	–0.20 (0.42)	2.40 (0.99)	1.84 (0.97)	0.44 (0.67)	0.10 (0.54)	0.67 (0.75)
RNOA	**–1.95 (0.03)	1.10 (0.86)	1.63 (0.95)	–0.24 (0.41)	–0.13 (0.45)	–0.57 (0.28)
ROE	–0.73 (0.23)	0.49 (0.69)	0.59 (0.72)	–0.53 (0.30)	–0.15 (0.44)	–0.29 (0.38)
Price-to-Book	*–1.51 (0.07)	–0.52 (0.30)	–0.71 (0.24)	–0.24 (0.41)	*–1.39 (0.08)	–1.19 (0.12)
–1 * Debt-to-Assets	–0.35 (0.36)	0.39 (0.65)	–0.83 (0.21)	0.02 (0.51)	0.51 (0.69)	–1.14 (0.13)
Interest Coverage Ratio	–0.40 (0.34)	–0.67 (0.25)	–0.45 (0.33)	0.95 (0.83)	–0.47 (0.32)	–0.99 (0.16)
–1 * Debt-to-Market	0.61 (0.73)	1.91 (0.97)	–0.35 (0.36)	1.31 (0.90)	0.15 (0.56)	**–1.67 (0.05)
Low Return Volatility	2.61 (0.99)	0.79 (0.79)	–0.04 (0.48)	–1.00 (0.16)	–0.22 (0.41)	–0.73 (0.23)

(B) F-Test (1990–2003)

	U.K.	France	Germany	Canada	Japan	Hong Kong
36-month Price Momentum	0.88 (0.58)	0.71 (0.15)	0.71 (0.15)	**0.55 (0.01)	1.05 (0.84)	0.90 (0.65)
12-month Earnings Revision	0.98 (0.92)	1.17 (0.50)	1.03 (0.89)	**0.61 (0.04)	**1.66 (0.03)	0.88 (0.61)
RNOA	0.99 (0.97)	0.75 (0.23)	1.01 (0.96)	1.39 (0.17)	1.01 (0.97)	0.95 (0.83)
ROE	0.98 (0.94)	0.83 (0.44)	1.16 (0.54)	1.22 (0.40)	0.96 (0.88)	1.00 (0.98)
Price-to-Book	0.89 (0.63)	0.69 (0.13)	0.75 (0.23)	**0.59 (0.03)	0.96 (0.88)	0.75 (0.23)
–1 * Debt-to-Assets	1.36 (0.20)	1.29 (0.28)	1.31 (0.26)	0.75 (0.23)	0.84 (0.47)	1.08 (0.76)
Interest Coverage Ratio	1.22 (0.40)	0.94 (0.79)	1.38 (0.17)	0.70 (0.14)	0.86 (0.53)	1.06 (0.81)
–1 * Debt-to-Market	1.34 (0.22)	1.10 (0.70)	1.24 (0.37)	1.12 (0.62)	*1.51 (0.08)	1.19 (0.47)
Low Return Volatility	**0.60 (0.03)	0.81 (0.39)	0.75 (0.24)	0.97 (0.91)	0.69 (0.12)	1.03 (0.91)

Note: * = 90% confidence level; ** = 95% confidence level.

TABLE 10.6 Summary Statistics of Non-U.S. Markets, Electing March 31 as the Annual Review Date

(A) Two-Sample *t*-Test (1990–2003)

	U.K.	France	Germany	Canada	Japan	Hong Kong
36-month Price Momentum	–0.71 (0.24)	0.49 (0.69)	*–1.29 (0.10)	*–1.63 (0.05)	0.36 (0.64)	–1.09 (0.14)
12-month Earnings Revision	0.67 (0.75)	2.27 (0.99)	2.56 (0.99)	0.42 (0.66)	0.04 (0.52)	–0.73 (0.23)
RNOA	*–1.40 (0.08)	0.81 (0.79)	–0.04 (0.48)	0.06 (0.52)	–0.31 (0.38)	–1.24 (0.11)
ROE	–0.49 (0.31)	0.20 (0.58)	–0.71 (0.24)	0.44 (0.67)	–0.05 (0.48)	–1.27 (0.10)
Price-to-Book	–1.04 (0.15)	–0.28 (0.39)	0.38 (0.65)	–0.50 (0.31)	0.34 (0.63)	**–1.82 (0.04)
-1 * Debt-to-Assets	0.80 (0.79)	0.68 (0.75)	–1.26 (0.11)	0.31 (0.62)	–0.33 (0.37)	**–1.75 (0.04)
Interest Coverage Ratio	0.09 (0.53)	0.26 (0.60)	–0.42 (0.34)	–0.12 (0.45)	–0.29 (0.38)	**–1.95 (0.03)
-1 * Debt-to-Market	0.20 (0.58)	1.59 (0.94)	–0.76 (0.22)	1.09 (0.86)	0.12 (0.55)	**–2.05 (0.02)
Low Return Volatility	1.14 (0.87)	1.25 (0.89)	0.79 (0.79)	–0.78 (0.22)	–0.29 (0.39)	–0.19 (0.43)

(B) F-Test (1990–2003)

	U.K.	France	Germany	Canada	Japan	Hong Kong
36-month Price Momentum	0.86 (0.52)	0.79 (0.31)	**0.62 (0.05)	**0.61 (0.04)	0.86 (0.51)	0.83 (0.43)
12-month Earnings Revision	0.92 (0.74)	1.37 (0.19)	0.92 (0.71)	0.69 (0.12)	0.86 (0.52)	*0.65 (0.07)
RNOA	1.00 (0.98)	**0.57 (0.02)	0.93 (0.77)	1.13 (0.61)	0.96 (0.88)	1.01 (0.95)
ROE	1.02 (0.93)	0.82 (0.41)	1.13 (0.61)	1.30 (0.28)	0.81 (0.36)	1.07 (0.79)
Price-to-Book	0.88 (0.61)	0.74 (0.20)	0.76 (0.26)	**0.58 (0.02)	0.96 (0.87)	0.90 (0.65)
-1 * Debt-to-Assets	1.38 (0.18)	1.17 (0.50)	1.13 (0.61)	1.01 (0.95)	0.78 (0.31)	1.02 (0.95)
Interest Coverage Ratio	1.29 (0.29)	1.00 (0.99)	1.27 (0.32)	*0.67 (0.09)	0.88 (0.59)	1.01 (0.97)
-1 * Debt-to-Market	1.39 (0.17)	1.11 (0.66)	0.82 (0.40)	1.13 (0.62)	1.16 (0.52)	1.25 (0.34)
Low Return Volatility	*0.63 (0.05)	0.85 (0.50)	0.90 (0.65)	1.11 (0.67)	*0.63 (0.05)	1.03 (0.92)

Note: * = 90% confidence level; ** = 95% confidence level.

significant at the 5% level. In contrast, seasonality is no longer observable in other markets. Why is the seasonal pattern stronger in Hong Kong with March as the end of the annual cycle instead of February? We suggest yet another behavioral reason — the tax year cycle, which ends in March for both personal and corporate tax reporting. Our supposition does not involve tax-loss-selling activities, because there is no capital gains tax in Hong Kong. Because the end of the tax assessment period provides an opportunity to plan the annual household budget, it is plausible to assume that investors also elect this date to review the performance of their portfolios. When combined with the misguided time-diversification benefit, March 31 may still induce seasonal changes in investors' risk preferences, even in the absence of the capital gains tax.

For Germany, the results are mixed and puzzling. When quality is defined as earnings revision, price-to-book ratio, or low volatility, our hypothesis is confirmed at the 90% significance level; but when quality is defined as the return on investments or the interest coverage ratio, our conjecture is rejected at 90% significance level. Germany is the only market rejecting our conjecture, on the grounds of significant contradictions rather than a set of random testing outcomes. Further research is needed to understand the disparity between different quality factors.

10.3 SEASONAL EFFECT OF EARNINGS ANNOUNCEMENT

In the U.S., companies file financial statements and announce their earnings on a quarterly basis, and most U.S. companies adopt calendar quarters as their fiscal reporting periods, thus inducing another systematic, calendar pattern related to the cross-sectional dispersion of security returns. The empirical evidence that follows will show the cross-sectional return dispersion is consistently higher around the earnings announcement months (January, April, July, and October) for the previous quarter and lower during the quiet period (February, May, August, and November). The rest of the months (March, June, September, and December) make up the preannouncement or warning period, during which the return dispersion falls between those of announcement and quiet periods. In addition, the January effect also induces abnormal increases in return dispersion during both January and December. Following the conjecture outlined in the last section, investors reset their investment horizon each year at the year end, causing their risk preference to change along with their investment decisions. As such, it is plausible to expect a higher cross-sectional return dispersion in both January and December when compared to other

months, because investors adjust their portfolio holdings to reflect their increased risk appetite.

Return dispersion is one component of excess returns. According to Chapter 4, the excess returns are proportional to the return dispersion. Therefore, the seasonal pattern of return dispersion carries at least two implications for portfolio management: portfolio trading strategy and *ex post* tracking error. However, we first examine empirical evidence of the seasonal pattern of return dispersions.

10.3.1 Empirical Evidence

Panel A of Figure 10.5 shows cross-sectional dispersion across four calendar partitions: January, February to June, July to November, and December. Two interesting observations can be gleaned. First, the dispersions in January and December are higher than in other months. Although median return dispersions in January and December are similar (shown as the bar in the middle of the box), January months are skewed to the right. In other words, extremely high return dispersions are most likely to happen in January than in any other months. Second, return dispersion in the first half seems lower than that of the second half.

The other source of return dispersion variations can be attributed to earnings announcements in certain periods of a calendar year. Companies release their earning numbers shortly after the end of each calendar quarter and some prerelease warnings before the quarter ends, in an effort to manage investors' expectations. Earnings news causes the market to adjust security prices and to reestablish the pricing equilibrium, thus resulting in higher cross-sectional return dispersions around the earnings announcement season.

We divide calendar months into three subgroups: the warning period (March, June, September, and December), the announcement period (January, April, July, and October), and the quiet period (February, May, August, and November). Panel B of Figure 10.5 shows the return dispersions in these subperiods. The announcement period has the highest return dispersion, followed by the warning period; and the quiet period has the lowest cross-sectional return dispersion.

To ascertain the statistical significance of these phenomena, we set up an OLS regression to disentangle these effects. The dependent variable is the monthly cross-sectional return dispersion; four dummy variables are included as independent variables. The dummy variables are 1 or 0 depending upon whether a month (1) is January, (2) is December, (3) falls in the

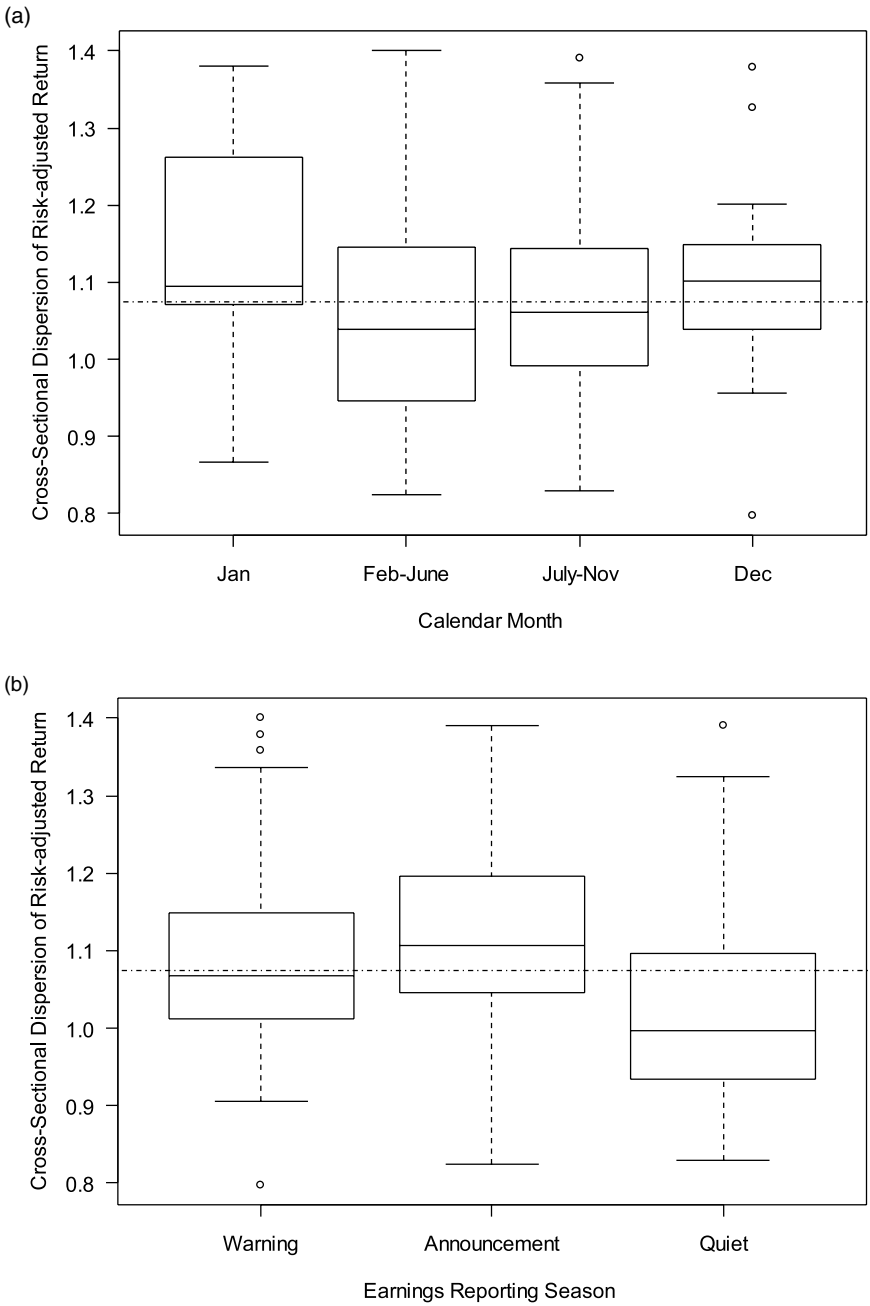


FIGURE 10.5. Cross-sectional dispersion of risk-adjusted returns: (a) conditioning on calendar month and (b) conditioning on earnings reporting season.

TABLE 10.7 Summary Statistics: Dispersion of Risk-Adjusted Returns

Regression Statistics	
Multiple R	0.328
R Square	0.107
Adjusted R Square	0.089
Standard Error	0.119
Observations	201

	Coefficients	**Standard Error**	***t*-Stat**	**p-Value**
Intercept	1.021	0.015	70.370	0.000
isJan	0.029	0.033	0.867	0.387
isDec	0.018	0.034	0.520	0.604
isWarning	0.060	0.022	2.738	0.007
isAnnouncement	0.088	0.022	3.972	0.000

warning period, or (4) falls in the announcement period. Table 10.7 displays the regression result and summary statistics. Both warning and announcement periods indicated as "isWarning" and "isAnnouncement" respectively, are significant at the 1% level, thus confirming both periods have significantly higher dispersions than the quiet period. January and December are not significant at the conventional level, but they are nonetheless positive.

10.3.2 Portfolio Trading Strategy

To understand why changes in expected cross-sectional return dispersion may influence portfolio implementation, we recall the decomposition of the *ex post* portfolio returns, shown in Equation 4.25 in Chapter 4, $\alpha_t = \mathrm{IC}_t \sqrt{N}\sigma_{model}\mathrm{dis}(R_t)$. Holding breadth N and model tracking error σ_{model} constant, the portfolio return for a single period depends on the manager skill, measured as the risk-adjusted information coefficient and investment opportunity represented by the cross-sectional dispersion of security returns. For a manager with a constant, positive skill, his or her portfolio would produce higher returns in the months with higher return dispersions and lower returns in low-dispersion months, although his or her skill is the same across all months.

Therefore, it is more beneficial to trade a portfolio immediately before high-dispersion months, because it enhances portfolio returns by taking advantage of the increased investment opportunity. Furthermore, a skilled manager could "spend" portfolio turnover wisely, by allotting more turnover immediately before high-dispersion months to achieve a

higher alpha exposure and letting alpha exposure drift in low-dispersion months.

10.3.3 *Ex Post* Tracking Error

The other implication relates to the *ex post* tracking error. Equation 4.31 of Chapter 4 shows the decomposition of the *ex post* tracking error as $\sigma = \text{std}\left(IC_t\right)\sqrt{N}\sigma_{\text{model}}\overline{\text{dis}\left(\mathbf{R}_t\right)}$. Realized tracking error is linearly proportional to the average cross-sectional return dispersion $\overline{\text{dis}(\mathbf{R}_t)}$. During the earnings announcement period, the average dispersion is higher; hence, managers should expect higher portfolio tracking volatilities then.

10.4 MACRO TIMING MODELS

Factor timing research is a close sibling of market timing research. Both have generated significant amount of interest from academics as well as practitioners. Similar sets of explanatory variables are deployed in both areas to provide an efficient time-series conditioning, in an effort to achieve a better performance when compared to a buy-and-hold strategy. In this section, we document some of the macro timing approaches applied to both market return and quantitative equity factor conditionings.

A macro factor timing approach must be used with caution. For every set of variables discovered to have explanatory power, one can easily find literature questioning the robustness, the practicality, or sometimes the relevance of such a discovery. Perhaps this highlights the potential hazard of data mining in factor timing research as it has limited data samples when compared with cross-sectional research, and the fleeting nature of factor timing discoveries, as investors quickly learn and adopt.

10.4.1 Conditional Factors

In general, the body of factor/market timing research has documented four sets of explanatory variables that possess time-series predictability of factor returns. Table 10.8 provides a detailed list of these variables.

Market state*:* Variables in this category measure the state of either equity or bond markets, in an effort to capture either business conditions (strong or weak economy) or the psychological inclination of the investor population in general, e.g., greed or fear. For example, Fama and French (1989) used the terms premium, default premium, and dividend yield to capture the business cycle and to explain predictable patterns in stock and bond returns. Similarly, Chordia

TABLE 10.8 Commonly Used Explanatory Variables

Market State	Equity: equity risk premium (earnings yield, T-bill), dividend yield, volatility (e.g., VIX), past market return, past value/size return, value spread, earnings growth spread.
	Bond: term spread, credit spread, and bond yield.
Monetary Policy	Monetary policy regime, Fed funds rate, and M1 money supply.
Economic Condition	Economic Health: GDP growth, industrial production, leading indicator, NAPM survey, and expected IBES profit growth.
	Inflation: consumer price index, producer price index, and oil price.
Consumption-base Relation	Cay, consumption, household net worth, and labor income.

and Shivakumar (2002) applied the same set of macro factors to explain the momentum profit. On the other hand, Cooper et al. (2004) found that momentum profit depends on whether the market delivered positive or negative returns in the recent past. Asness et al. (2000) showed that the value-growth style return is predictable, and they used both the value spread and the earnings growth spread as explanatory variables. In this case, the spread is measured as the return difference between the growth and the value portfolios. Arnott et al. (1989) used the equity risk premium and market volatility to forecast returns to the BARRA risk factors[15]. Lastly, Kao and Shumaker (1999) applied both the term spread and the credit spread to forecast value-growth style returns in the equity market.

Monetary policy*:* Monetary-policy-related variables provide three different gauges: the monetary policy stance of the Federal Reserve, the short-term interest rate (e.g., Fed Funds rate), and the money supply (e.g., M1). Jensen et al. (1996, 1997, 1998, and 2000) and Conover et al. (2005) found that monetary policy environment — expansive or restrictive — influences the broad market return, style rotation, sector rotation, as well as the commodity and bond markets. Arnott (1989) also found the percentage change in M1 money supply differentiates returns to certain BARRA risk factors.

Economic condition*:* Economic variables directly measure either the health of the economy or the inflation risk. Arnott (1989) found both the percentage change in the Leading Indicators and the percentage change in the producer price index (PPI) predict a subset of BARRA factor returns. Kao and Shumaker (1999) used both the expected

GDP growth and the consumer price index (CPI) to forecast the 3-month forward return spread between value and growth.

Consumption-based indicators: This branch of researches falls under financial economics and focuses on explaining the countercyclical nature of the equity risk premium (i.e., high when business condition is weak and *vice versa*), by employing consumption growth as one of the explanatory variables. For example, Campbell and Cochrane (1999) explained several asset-pricing phenomena through the use of a theoretical model that is driven by a consumption growth process in conjunction with a slow-moving external habit to the standard utility function. Lettau and Ludvigson (2001) provided an empirical examination. They found that the consumption–wealth ratio (cay) — the error term from the cointegration relation among consumption, wealth, and labor income — is a better forecaster of future equity market returns at short and intermediate horizons when compared with traditional market variables, e.g., dividend yield. Recently, Guo (2003) showed that combining cay with a measure of stock market volatility substantially improves the equity market return forecast.

10.4.2 Empirical Findings

In this section, we continue the examination of the return profiles of the nine quality factors used in the calendar modeling section, by conditioning them on two state variables measuring the monetary policy and the broad market return. We also examine the interplay between calendar seasonality and these two state variables to see whether certain market conditions enhance/diminish the calendar effect.

Monetary policy regime: Jensen et al. (1996) postulated that monetary policy — restrictive or expansive — regulates aggregate money supply, induces a direct influence on business conditions, and ultimately governs changes in investors' risk preference (or risk premium). Under an expansive monetary environment, the economic outlook is rosier, and investors demand lower equity risk premium and exhibit flight-from-quality behavior by purchasing cheap, low-quality firms. In contrast, when the Federal Reserve is in the tightening mode, investors fear negative economic shocks and the heightened possibility of an immediate recession. They demand higher equity risk premium and consequently exhibit flight-to-quality behavior by purchasing quality companies.

Panel A of Table 10.9 shows the monthly risk-adjusted IC for the full sample, the expansive period, and the restrictive period. The empirical results are unanimous — returns to quality factors are consistently higher in the restrictive period than in the expansive period. Furthermore, quality factors not only delivered higher returns (mean) in the expansive period but also scored higher risk-adjusted returns (*t*-statistic).

Panel B displays the test results of both the mean difference and the variance difference between the two policy regimes. The difference in mean is fairly pronounced. Five factors show significance with the two-sample *t*-test and six with the Wilcoxon test. We also note that two factors, price-to-book and negated debt-to-market, show significant difference in variance at the 5% level. Interestingly, both ratios exhibit negative returns in the expansive period and positive returns in the restrictive period. As both ratios measure bankruptcy risk, these results suggest that financial distress is consistently positively priced in the expansive period, causing the default premium to tighten. Our result corroborates the conjecture proposed by Fama and French (1989). Most interestingly, our data can be viewed as an out-of-sample test of their conjecture, as it spans 1987 to 2003. Our result suggests that the phenomenon persisted after its discovery.

To assess how the monetary policy interacts with calendar seasonality, we create a composite quality factor that equally weights the nine selected quality factors. Returns to the quality composite are then collated based on both calendar (first half or second half) and monetary policy (expansive or restrictive), resulting in four regimes: expansive first half, restrictive first half, expansive second half, and restrictive second half. Figure 10.6 shows the box chart of the distribution of quality returns in these four partitions. Two observations are worth noting:

1. The spread between the risk-adjusted ICs in the expansive first half and restrictive second half partitions is economically significant. This evidence supports the conjecture that investor's risk preference depends on both calendar events as well as business conditions.

2. Regarding the order of importance between these two influences, calendar seasonality is more pronounced than monetary policy. The expansive second half partition shows a positive risk-adjusted IC, whereas the restrictive first half partition shows a negative risk-adjusted IC.

TABLE 10.9 Monetary Policy Influence

(A) Summary Statistics

	Full Sample				Expansive Period				Restrictive Period			
	Mean	**std**	**t**	**# obs.**	**Mean**	**std**	**t**	**# obs.**	**Mean**	**std**	**t**	**# obs.**
36-month Price Momentum	0.0013	0.0904	0.20	201	–0.0059	0.0857	–0.77	125	0.0130	0.0971	1.17	76
12-month Earnings Revision	0.0249	0.0729	4.85	201	0.0230	0.0775	3.33	125	0.0281	0.0649	3.77	76
RNOA	0.0184	0.0490	5.34	201	0.0118	0.0490	2.68	125	0.0294	0.0471	5.45	76
ROE	0.0170	0.0571	4.21	201	0.0121	0.0575	2.34	125	0.0250	0.0559	3.90	76
Price-to-Book	–0.0028	0.0798	–0.49	201	–0.0129	0.0716	–2.01	125	0.0139	0.0898	1.35	76
-1 * Debt-to-Assets	0.0127	0.0489	3.69	201	0.0077	0.0486	1.77	125	0.0210	0.0484	3.78	76
Interest Coverage Ratio	0.0095	0.0438	3.07	199	0.0036	0.0422	0.94	123	0.0192	0.0449	3.72	76
-1 * Debt-to-Market	–0.0038	0.0753	–0.72	201	–0.0150	0.0638	–2.63	125	0.0145	0.0886	1.43	76
Earnings Stability	0.0090	0.0631	2.02	201	0.0073	0.0665	1.22	125	0.0118	0.0572	1.79	76

(B) Difference Tests

	Two-Sample t-Test			Wilcoxon Test		Two-Sample Variance-Test			
	t	**p-Value**	**df**	**W**	**p-Value**	**F**	**p-Value**	**num df**	**denom df**
36-month Price Momentum	–1.40	0.164	143.4	*4057	0.083	0.78	0.217	124	75
12-month Earnings Revision	–0.49	0.623	179.7	4574	0.661	*1.43	0.096	124	75
RNOA	**–2.54	0.012	163.5	**3739	0.012	1.08	0.708	124	75
ROE	–1.58	0.117	162.0	4255	0.216	1.06	0.796	124	75
Price-to-Book	**–2.21	0.029	132.3	**3905	0.035	**0.64	0.026	124	75
-1 * Debt-to-Assets	*–1.88	0.062	159.0	*3987	0.057	1.01	0.984	124	75
Interest Coverage Ratio	**–2.43	0.016	151.6	*3930	0.060	0.89	0.545	122	75
-1 * Debt-to-Market	**–2.53	0.013	122.3	**3736	0.011	**0.52	0.001	124	75
Earnings Stability	–0.51	0.613	176.7	4567	0.648	1.35	0.158	124	75

Note: * = 90% confidence level; ** = 95% confidence level.

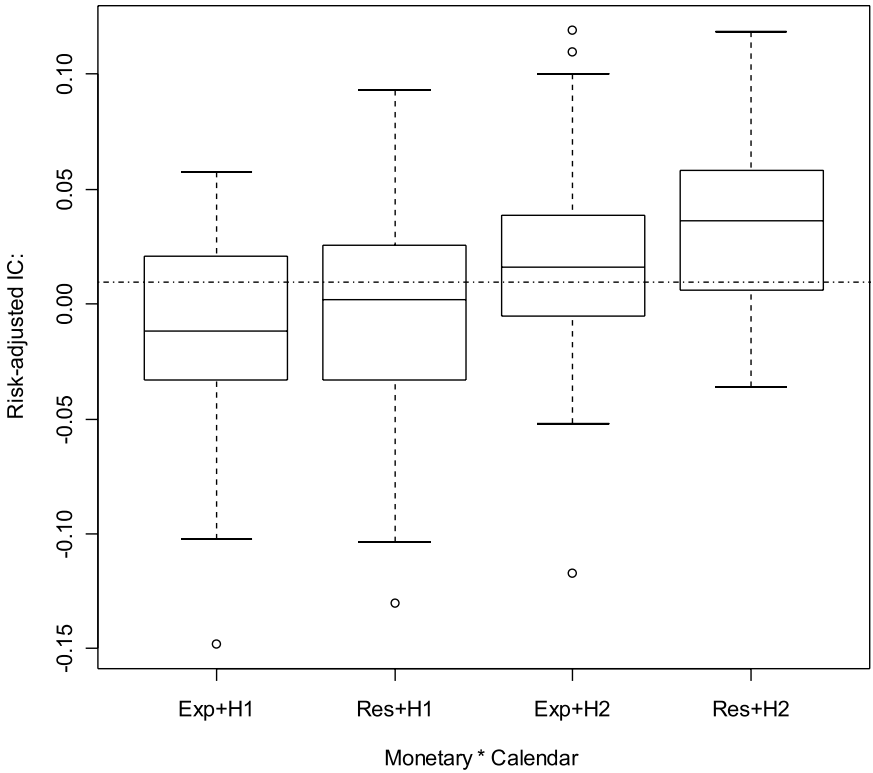


FIGURE 10.6. Distributions of risk-adjusted ICs conditioned on monetary policy and calendar partitions.

Market return environment: In this section, we examine whether the market state, proposed by Cooper et al. (2004), influences investors' risk preference. We note that Cooper's conjecture is a behavioral one and has no implications for investor's risk preference. Table 10.10 shows test results; Panel A shows summary statistics in the full sample, in up markets and in down markets, and Panel B shows difference tests. Results are mixed for the tests of mean difference — the mean difference is negative for six factors and positive for three.

Interestingly, three factors, which have been associated with behavioral biases, show negative signs, although short of crossing into the statistical significance zone. They are price momentum, earnings revision, and price-to-book. Perhaps, market state does induce varying levels of overconfidence and subsequently results in different profit potential for behavioral phenomena. Lastly, we would like to note that our result does not diminish the finding presented by Cooper et al. (2004), as we deliberately

TABLE 10.10 Market State Variable

(A) Summary Statistics

	Full Sample				Down Market				Up Market			
	mean	**std**	**t**	**# obs.**	**mean**	**std**	**t**	**# obs.**	**mean**	**std**	**t**	**# obs.**
36-month Price Momentum	0.0013	0.0904	0.20	201	−0.0138	0.0843	−1.11	46	0.0057	0.0919	0.77	155
12-month Earnings Revision	0.0249	0.0729	4.85	201	0.0114	0.0792	0.98	46	0.0290	0.0706	5.10	155
RNOA	0.0184	0.0490	5.34	201	0.0243	0.0554	2.98	46	0.0167	0.0469	4.43	155
ROE	0.0170	0.0571	4.21	201	0.0300	0.0572	3.56	46	0.0131	0.0567	2.87	155
Price-to-Book	−0.0028	0.0798	−0.49	201	−0.0179	0.0662	−1.83	46	0.0017	0.0831	0.26	155
−1 * Debt-to-Assets	0.0127	0.0489	3.69	201	0.0013	0.0536	0.16	46	0.0161	0.0470	4.27	155
Interest Coverage Ratio	0.0095	0.0438	3.07	199	0.0031	0.0458	0.46	46	0.0115	0.0432	3.29	153
−1 * Debt-to-Market	−0.0038	0.0753	−0.72	201	−0.0136	0.0732	−1.26	46	−0.0009	0.0759	−0.16	155
Earnings Stability	0.0090	0.0631	2.02	201	0.0126	0.0606	1.41	46	0.0079	0.0639	1.54	155

(B) Difference Tests

	Two-Sample t-Test			Wilcoxon Test		Two-Sample Variance-Test			
	t	**p value**	**df**	**W**	**p value**	**F**	**p value**	**num df**	**denom df**
36-month Price Momentum	−1.35	0.181	79.5	3128	0.208	0.84	0.503	45	154
12-month Earnings Revision	−1.35	0.181	67.7	3026	0.120	1.26	0.313	45	154
RNOA	0.85	0.400	65.3	3953	0.263	1.40	0.141	45	154
ROE	*1.76	0.082	73.3	**4324	0.029	1.02	0.907	45	154
Price-to-Book	−1.66	0.100	91.1	*2967	0.085	*0.64	0.078	45	154
−1 * Debt-to-Assets	*−1.69	0.095	66.9	**2885	0.050	1.30	0.244	45	154
Interest Coverage Ratio	−1.11	0.272	70.8	3060	0.181	1.12	0.593	45	152
−1 * Debt-to-Market	−1.02	0.312	76.0	3100	0.180	0.93	0.803	45	154
Earnings Stability	0.46	0.648	77.2	3851	0.410	0.90	0.686	45	154

Note: * = 90% confidence level; ** = 95% confidence level.

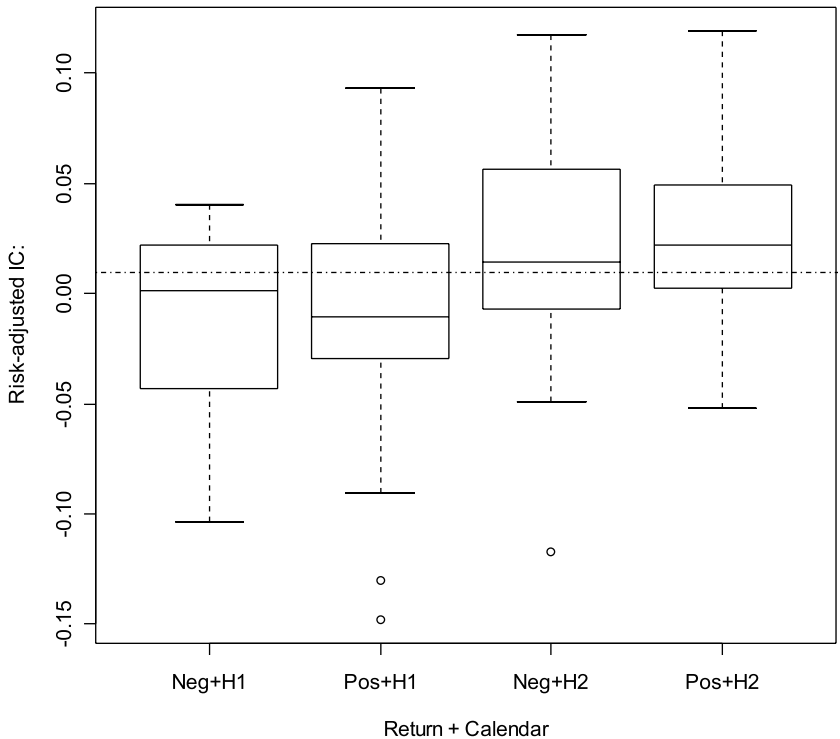


FIGURE 10.7. Distributions of risk-adjusted ICs conditioned on past market return and calendar partitions.

construct the price momentum factor using the trailing 36-month return, encapsulating the entire short, intermediate, and long-term momentum phenomenon. For comparison, we repeat the exercise in the last section and display the box chart in Figure 10.7. As expected, market state provides little differentiation of the risk-adjusted IC of the quality composite.

10.4.3 Sources of Predictability: Competing Explanations

The reason why factor returns or market returns are predictable is still being debated, without a universally accepted explanation. There are three schools of thought that are commonly cited as conjectures in factor/market-timing-related literature.

10.4.3.1 Rational Compensation for Risk Taking

Return is a form of compensation for exposures to nondiversifiable risk, a conjecture favored by neoclassic rational market theorists. The profitability of momentum strategy (Jegadeesh and Titman 1993) and the long-term predictability of market returns continue to haunt the CAPM-based

explanations, which assume expected return is a constant through time. Intertemporal CAPM (ICAPM) was proposed to relax the time-invariant assumption, and a long list of empirical research pointed to the fact that the price of risk seems to vary countercyclically with business conditions — that is, the risk premium is high when the economy is weak, and it is low when the economy is booming. For example, Fama and French (1989) suggested that both common stocks and long-term bonds contain a term premium and a default premium. The term premium relates to short-term business cycles and compensates for exposure to discount-rate shocks (i.e., the duration risk); the default premium relates to long-term business episodes and compensates for the return sensitivity to unexpected changes in business conditions. They conjectured that when economic conditions are poor, income is low and stock and bond returns must be high to induce substitution from consumption to investment. When times are good and income is high, asset returns clear at lower levels.

Employing the same set of variables used by Fama and French (1989), Chordia and Shivakumar (2002) showed that momentum profits are explained by common macroeconomic variables that are related to the business cycle. They attribute the momentum profits to cross-sectional differences in conditional expected returns that are predicted by standard macroeconomic variables and assert that the residual portion of stock-specific momentum contributes little to strategy payoffs. They attributed momentum profits to cross-sectional differences in conditionally expected returns that are predicted by standard macroeconomic variables.

Campbell and Cochrane (1999) provided an economic explanation of why risk premia are countercyclical to business conditions. They suggested that investors fear stocks primarily because they do poorly in recessions, not because stock returns are correlated with declines in wealth or consumption. Such fear is attributed to the habit formation hypotheses: repetition of a stimulus diminishes the perception of the stimulus and responses to it. This psychological feature of human behavior explains why consumers' reported sense of well-being often seems more related to recent changes in consumption than to the absolute level of consumption. As such, they conjectured that habit persistence can explain why recessions are so feared even though their effects on output are relatively small. Interestingly, the habit formation hypothesis and prospect theory seem to share a similar psychological profile of how one evaluates one's own well-being — focusing more on the change in wealth/consumption rather than the absolute level of wealth/consumption.

10.4.3.2 Mechanism of the Economy

This school of thought is primarily based on economic theory of aggregate demand and supply. Similar to other economic theories, it takes a rational view of the market and expands on the explanation articulated by Fama and French (1989). On a more intuitive level, Jensen et al. (1996) postulated that monetary policy regulates aggregate money supply, induces a direct influence on business conditions, and ultimately governs changes in investors' risk preference (or risk premium). They showed that monetary stringency provides additional explanatory power of future stock returns in excess of what can be explained by business condition variables. Specifically, they found that business conditions explain future stock returns only in expansive monetary policy periods, but not restrictive periods.

On a more detailed level, Jensen et al. (2000) also documented the use of monetary policy to forecast industry rotation. They argue that expansive monetary policy induces excess aggregate supply of money and encourages higher levels of discretionary consumer spending. Hence, the industries that are more reliant on discretionary consumer spending appear to be more sensitive to changes in the monetary environment. In a similar vein, practitioners are aware of the three different phases of sector rotation: starting with the early cyclical (sectors more influenced by discretionary consumer spending), followed by the late cyclical (sectors more sensitive to corporate spending, such as technology and capital expenditures, or sensitive to commodity prices), and ending with the defensive (sectors least sensitive to business conditions, such as utilities and pharmaceuticals). We note that this line of reasoning deviates from the argument of time-varying risk premia. Instead, it focuses on the predictable, input–output relationship of the economy. Using security valuation as an analogy, time-varying risk premia are associated with the discount rate, whereas the input–output relationship is associated with the earnings. The contest between both arguments is centered on finding the underlying driver of industry momentum profits, first documented by Moskowitz and Grinblatt (1999). Chordia and Shivakumar (2002) showed that macroeconomic variables explain industry momentum profits, thus favoring the discount rate argument. On the other hand, Menzly and Ozbas (2005) used the Input–Output Benchmark Survey of the Bureau of Economic Analysis (BEA) to link industries into either upstream or downstream categories, based on the flow of goods and services. They found significant profit to a cross-industry momentum trading strategy, thus favoring the earnings side of the argument.

10.4.3.3 Irrational Behavioral Inefficiency

The behavioral finance literature attributes most asset pricing anomalies to human behavioral and cognitive biases. Theories were proposed to explain the price momentum anomalies (Daniel et al. 1998; and Hong and Stein 1999). Cooper et al. (CGH, 2004) extended these behavioral theories and linked momentum profits to the state of the market. They found that intermediate-term price momentum profits exclusively followed periods when the market delivered positive excess returns in the past. In contrast, momentum profits are generally flat or negative after down markets. CGH explained this asymmetry by linking aggregate investor overconfidence to increasing market prices. In addition, CGH also questioned the robustness of findings presented by Chordia and Shivakumar (2002) and showed that macroeconomic variables did not capture the asymmetry in momentum profits. Lastly, testing CGH's hypotheses in non-U.S. markets, Huang (2005) found qualified supports for the 17 countries in the MSCI index. Note the following remark:

- Different explanations have potentially different implications for future predictability of returns. Should profit potential arise from behavioral biases, it is natural to expect such profit to diminish after its discovery, eventually to a level that can only clear transaction costs. On the other hand, should profit opportunity arise from taking nondiversifiable risk, it is natural to expect such profit to last, as the pricing equilibrium is jointly determined by both hedgers and arbitragers. Ironically, investors and consultants may ask, "Why should managers be compensated for excess return that comes from risk taking?" The ultimate judgment must be left to the investors.

10.5 SUMMARY

Factor timing is a promising area of research. Theoretical and empirical literature has pointed to various avenues of achieving a more efficient dynamic factor selection through time. The arsenal of conditioning variables can be sorted into to five categories: (1) calendar event, (2) market state, (3) monetary policy, (4) direct measure of economic condition, and (5) consumption-based ratios. The exact reason why these variables may forecast factor returns is still being debated, with little hope of reaching a consensus. In general, there are three schools of thought: (1) rational compensation for risk taking, (2) mechanism of the economy, and (3) irrational behavioral inefficiency.

The promised benefits of factor timing must meet a rigorous strategy and a diligent portfolio implementation. Several implementation issues must be considered to provide a more encompassing view.

Factor selection*:* Factor timing can be applied to both alpha and risk factors. Alpha factors are the ingredients of alpha models and they are a natural choice. However, returns to alpha factors typically consist of large means (positive or negative) and, more importantly, small standard deviations. Risk factors, on the other hand, have small mean returns but large standard return deviations. Therefore, risk factors may provide better opportunities and investment returns in factor timing.

Transaction cost*:* Because a timing strategy selects factor weightings dynamically through time, it generates model turnover and subsequently results in increased portfolio turnover and transaction cost. Proper estimation and control of implementation cost is an important component of a successful timing strategy.

Strategy breadth*:* The breadth of factor timing strategies is much lower than a traditional bottom-up stock selection model, pointing to a lower expected IR. Thus, managers must allocate their risk budget appropriately between bottom-up equity models and factor timing strategies based on their expected information ratio.

Data mining hazard*:* Because of the limited observations of time-series data when compared to cross-sectional data, it is more likely to misconstrue spurious correlations as profit opportunities through misguided data mining exercises. To this end, managers must adopt a fundamental belief of why factor returns are predictable. Such a belief would guide them to reject those empirical results without supporting priors, despite their statistical significance.

Model uncertainty*:* In factor timing models, uncertainty may come from: (1) the specification of conditioning variables, (2) the estimation of time-varying factor returns, (3) the estimation of factor exposures for each security, and (4) the persistence of profit opportunities. Avramov and Chordia (2006) proposed a factor timing framework that incorporates model uncertainty using a Bayesian model averaging. Their approach mitigates model misspecification and overconfidence in model forecasts.

REFERENCES

Ackert, F. and Athanassakos, G., Institutional investors, analyst following, and the January anomaly, working paper, Federal Reserve Bank of Atlanta, 1998.

Arnott, R.D., The use and misuse of consensus earnings, *Journal of Portfolio Management*, Vol. 11, 18–28, Spring 1985.

Arnott, R.D., Kelso, C.M. Jr., Kiscadden, S., and Macedo, R., Forecasting factor returns: An intriguing possibility, *Journal of Portfolio Management*, Vol. 16, 28–35, Fall 1989.

Asness, C.S., Friedman, J.A., Krail, R.J., and Liew, J.M., Style timing: Value versus growth, *Journal of Portfolio Management*, 50, Spring 2000.

Athanassakos, G., The scrutinized-firm effect, portfolio rebalancing, stock return seasonality, and the pervasiveness of the January effect in Canada, *Multinational Finance Journal*, Vol. 6, No. 1, 1–27, March 2002.

Avramov, D. and Chordia, T., Predicting stock returns, *Journal of Financial Economics*, 2006.

Benartzi, S. and Thaler, R.H., Myopic loss aversion and the equity premium puzzle, *Quarterly Journal of Economics*, Vol. 110, No. 1, 73–92, 1995.

Bernoulli, D., Exposition of a new theory on the measurement of risk, *Econometrica*, January 1954. [Translation from 1738 version.]

Bildersee, J. and Kahn, N., A preliminary test of the presence of window dressing, *Journal of Accounting, Auditing and Finance*, Summer 1987.

Blume, M., and Stambaugh, R., Biases in computed returns: An application to the size effect, *Journal of Financial Economics,* Vol. 12, 387–404, 1983.

Branch, B. and Chang, K., Low price stocks and the January effect, *Quarterly Journal of Business and Economics*, Vol. 29, No. 3, 90–118, Summer 1990.

Branch, B. and Echevarria, D.P., The impact of bid-ask prices on market anomalies, *The Financial Review*, Vol. 26, No. 2, 249–268, May 1991.

Bronars, S. and Deere, D.R., The threat of unionization, the use of debt, and the preservation of shareholder wealth, *Quarterly Journal of Economics*, 231–254, February 1991.

Brown, K.C., Harlow, W.V., and Starks, L.T., Of tournaments and temptations: An analysis of managerial incentives in the mutual fund industry, *Journal of Finance*, Vol. 51, No. 1, 85–110, 1996.

Campbell, J.Y. and Cochrane, J.H., By force of habit: A consumption-based explanation of aggregate stock market behavior, *The Journal of Political Economy*, 205–251, April 1999.

Campbell, J.Y. and Viceira, L.M., *Strategic Asset Allocation, Portfolio Choice for Long-Term Investors*, Oxford University Press, New York, 2002.

Chen, H. and Singal, V., A December effect with tax-gain selling? *Financial Analysts Journal*, Vol. 59, No. 4, 78–90, July–August 2003.

Chordia, T. and Shivakumar, L., Momentum, business cycle, and time-varying expected returns, *The Journal of Finance*, Vol. 57, No. 2, 985–1019, April 2002.

Conover, C.M., Jensen, G.R., Johnson, R.R., and Mercer, J.M., Is fed policy still relevant for investors? *Financial Analysts Journal*, 70–97, January–February 2005.

Conrad, J. and Kaul, G., Long-term market overreaction or biases in computed returns? *Journal of Finance*, Vol. 48, 39–63, 1993.

Cooper, M.J. Gutierrez, R.C. Jr., and Hameed, A. (CGH), Market states and momentum, *Journal of Finance*, 1345, June 2004.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A., Investor psychology and investor security market under- and over-reactions, *Journal of Finance*, Vol. 53, 1839–1886, 1998.

Daniel, K. and Titman, S., Market efficiency in an irrational world, *Financial Analyst Journal*, Vol. 55, No. 6, 28–40, November–December 1999.

Fama, E.F. and French, K.R., Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, Vol. 33, 3–57, February 1993.

Fama, E.F. and French, K.R., Business conditions and expected returns on stocks and bonds, *Journal of Financial Economics*, Vol. 23, 23–49, 1989.

Fama, E.F. and French, K.R., Multifactor explanations of asset pricing anomalies, *Journal of Finance*, Vol. 51, 55–85, March 1996.

Fama, E.F. and MacBeth, J.D., Risk, return and equilibrium — empirical tests, *The Journal of Political Economy*, Vol. 81, 607, May–June 1973.

Fisher, K.L. and Statman, M., A behavioral framework for time diversification, *Financial Analyst Journal*, 88–97, May–June 1999.

Fridson, S.M., Semiannual seasonality in high-yield bond returns, *Journal of Portfolio Management*, Vol. 26, No. 4, 102–111, Summer 2000.

Givoly, D. and Ovadia, A., Year-end tax-induced sales and stock market seasonality, *Journal of Finance*, 171–185, March 1983.

Givoly, D. and Lakonishok, J., The information content of financial analysts' forecasts of earnings, *Journal of Accounting and Economics*, 1979.

Gorton, G. and Schmid, F., Class study inside the firm: A study of german codetermination, Wharton, working paper, 2004.

Grinold, R.C., The fundamental law of active management, *Journal of Portfolio Management,* Vol. 15, No. 3, 30–37, Spring 1989.

Gudikunst, A. and McCarthy J., High-yield bond mutual funds: Performance, January effects, and other surprises, *The Journal of Fixed Income*, Vol. 7, No. 2, 35–46, September 1997.

Gultekin, M.N. and Gultekin, N.B., Stock market seasonality, international evidence, *Journal of Financial Economics*, Vol. 12, 469–481, 1983.

Guo, H., On the out-of-sample predictability of stock market returns, working paper, Federal Reserve Bank of St. Louis, October 2003.

Haugen, A.R. and Lakonishok, J., The Incredible January Effect, Dow Jones-Irwin, 1998.

Hawkins, E.H., Chamberlin, S.C., and Daniel, W.E., Earnings expectations and security prices, *Financial Analysts Journal*, Vol. 40, 24–38, September–October 1984.

Hong, H. and Stein, J., A unified theory of underreaction, momentum trading, and overreaction in asset markets, *Journal of Finance*, Vol. 48, 65–91, 1999.

Huang, D., Essays on macro factors and asset pricing: Theory and evaluation, [degree] dissertation, West Virginia University, 2005.

Jegadeesh, N. and Titman, S., Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance*, Vol. 48, 65–92, March 1993.

Jegadeesh, N. and Titman, S., Short horizon return reversals and the bid-ask spread, *Journal of Financial Intermediation*, Vol. 4, 116–132, 1995.

Jensen, G.R., Mercer, J.M., and Johnson, R.R., Business conditions, monetary policy, and expected security returns, *Journal of Financial Economics,* Vol. 40, 213–237, 1996.

Jensen, G.R., Mercer, J.M., and Johnson, R.R., New evidence on size and price-to-book effects in stock returns, *Financial Analysts Journal*, 34–42, November–December 1997.

Jensen, G.R., Mercer, J.M., and Johnson, R.R., The inconsistency of small-firm and value stock premiums, *Journal of Portfolio Management*, 27–36, Winter 1998.

Jensen, G.R., Mercer, J.M., and Johnson, R.R., The role of monetary policy in investment management, The Research Foundation of AIMR, 2000.

Kahneman, D. and Tversky, A., Advances in prospect theory: Cumulative representation of uncertainty, *Journal of Risk and Uncertainty*, Vol. 5, 297–324, 1992.

Kahneman, D. and Tversky, A., Prospect theory: An analysis of decision under risk, *Econometrica*, Vol. 47, No. 2, 263–291, 1979.

Kao, D.-L. and Shumaker, R.D., Equity style timing, *Financial Analyst Journal*, 37–48, January–February 1999.

Kerrigan, T.J., When forecasting earnings, it pays to watch forecasts, *Journal of Portfolio Management*, Vol. 10, 19–27, Summer 1984.

Kritzman, M. and Rich, D., Beware of dogma, *Journal of Portfolio Management*, Vol. 24, No. 4, 66–77, Summer 1998.

Lakonishok, J., Shleifer, A., and Vishny, R.W., Contrarian investment, extrapolation, and risk, *Journal of Finance*, Vol. 49, 1541–1579, December 1994.

Lettau, M. and Ludvigson, S., Consumption, aggregate wealth, and expected stock returns, *Journal of Finance*, 815–849, June 2001.

Maxwell, F.W., The January effect in the corporate bond market: A systematic examination, *Financial Management*, Vol. 27, No. 2, 18–30, Summer 1998.

Menzly, L. and Ozbas, O., Cross-industry momentum, AFA meeting 2005.

Moskowitz, T. and Grinblatt, M., Do industries explain momentum? *Journal of Finance,* Vol. 54, No. 4, 1249–1290, 1999.

Olsen, R.A., Prospect theory as an explanation of risky choice by professional investors: Some evidence, *Review of Financial Economics*, Vol. 6, No. 2, 225–232, 1997.

Reinganum, M.R., The anomalous stock market behavior of small firms in January: Empirical tests for tax-loss selling effects, *Journal of Financial Economics*, Vol. 12, 89, 1983.

Richards, R.M. and Martin, J.D., Revisions in earnings forecasts: How much response?, *Journal of Portfolio Management*, Vol. 5, Summer 1979.

Samuelson, P.A., Risk and uncertainty: A fallacy of large numbers, *Scientia*, Vol. 98, 108–113, 1963.

Seyhun, H., Can omitted risk factors explain the January effect: A stochastic dominance approach, *Journal of Financial and Quantitative Analysis*, Vol. 28, 195–212, 1993.

Ward, J.D. and Huffman, S.P., Seasonality in the returns of defaulted bonds: The January and October effects, *Quarterly Journal of Business and Economics*, Vol. 36, No. 3, 3–10, 1993.

Vetter, E.D. and Wingender, J.R., The January effect in preferred stock investments, *Quarterly Journal of Business and Economics*, Vol. 35, No. 1, 79–86, Winter 1996.

ENDNOTES

1. The long list of explanations for the January effect include tax-loss selling (Givoly and Ovadia 1983; Reignanum 1983; Chen and Signal 2001, 2003), window dressing (Bildersee and Kahn 1987; Haugen and Lakonishok 1998), performance hedging (Haugen and Lakonishok 1998; Ackert and Athanassakos 1998; Athanassakos 2002), bid-ask bounce (Branch and Echevarria 1991; Blume and Stambaugh 1983; Conrad and Kaul 1993), and omitted risk factors (Seyhun 1993).
2. Evidence of the calendar effect was also documented previously. Arnott et al. (1989) showed that the time-series variation of returns to BARRA factors can be explained by calendar dummy variables, one for each month, in a regression framework. Kao and Shumaker (1999) demonstrated the calendar seasonality of the value-growth style spread.
3. Kritzman and Rich (1998) clarified the debate and articulated Samuelson's assumptions. Fisher and Statman (1999) suggested that when prospect theory is used in place of the standard utility assumption, it is plausible for an investor to achieve a higher expected utility as the investment horizon lengthens. Olsen (1997) found money managers not only exhibit loss aversion (as predicted by the value function of prospect theory) but also believe in the benefit of time diversification.
4. Benartzi and Thaler (1995) also suggested the historical equity risk premium is consistent with the assumption that investors evaluate their portfolios on an annual basis. Brown et al. (1996) related the heightened focus of annual performance in the mutual fund industry to how performance is compiled and ranked by business publications.
5. We equate high- (low-) quality companies with low- (high-)risk companies. This is generally true in normal market conditions. One could argue this connection breaks down when the high-quality stocks are overpriced and become high-risk stocks, such as the case of Nifty Fifty. We found evidence of such a link in the negative correlations between these factors and stock-specific risk, which imply high-quality stocks tend to exhibit lower specific risk. Interested readers can get the results from the authors.
6. Earnings revision phenomenon was documented by Givoly and Lakonishok (1979), Hawkins et al. (1984), Arnott (1985), Kerrigan (1984), and Richards and Martin (1979), among others. Returns to book-to-price ratio were documented extensively in the value premium literature, such as Lakonishok et al. (1994), and Fama and French (1993, 1996).

7. Jegadeesh and Titman (1993, 1995) documented three phases of price momentum anomaly: short-term reversal (1 month), intermediate-term continuation (2–12 months), and long-term reversal (13–60 months). We categorize our price momentum factor as a nonpriced risk factor because it has a 36-month horizon encapsulating all of the three phases.
8. For example, a price-to-book of 0.5 means that for every dollar invested in a company, only $0.50 can expected to be recouped by that investor, whereas the other $0.50 is the loss via the regular course of business operations.
9. Please see Cooper et al. (2004) and Jensen et al. (1997).
10. We choose the Chinese New Year for the following three reasons. First, according to the Chinese heritage, the Chinese New Year marks the end of the previous year and the beginning of another new year. Second, companies that operate in the countries that officially celebrate the Chinese New Year typically pay the annual bonus to their employee right before the holiday. Third, extended vacation is typical so that family members and relatives can get together for the occasion, a tradition similar to Thanksgiving in Western cultures. The celebration usually starts at the end of January or the beginning of February and lasts for the subsequent 15 days.
11. Japan does not celebrate the Chinese New Year as an exchange holiday, whereas Hong Kong does.
12. We report the F-test results in Panel B of Table 10.4 and Table 10.5 and do not provide further discussion in the text because their conclusions conform to the findings in prior sections and they are intuitively apparent.
13. In this test, the first partition contains the ICs of the months between March and July and the second partition covers months from October to February of the next year.
14. The first partition contains the ICs of the months between April and August and the second partition covers months from November to March of the next year.
15. Equity risk premium is measured by S&P 500 earnings yield minus the Treasury bill yield; market volatility is defined as the 6-month variance of returns on the S&P 500.

CHAPTER 11

Portfolio Constraints and Information Ratio

BESIDES THE PORTFOLIO TURNOVER CONSTRAINTS discussed in Chapter 8, there are other forms of portfolio constraints that portfolio managers in practice have to abide by. One such form of constraint is risk exposure constraint. We have discussed this when we developed the risk-adjusted information coefficients, which analyzed factors with their exposure to risk factors being neutralized. The reason for neutralizing or limiting exposure to these factors, such as market, size, growth, etc. (see Chapter 3 for more), is to control systematic risk of active portfolios and to generate excess returns that are stock specific and have low correlation with market returns.

Another form of constraint is the holding constraint for stocks, which has several variations. For example, one can require that any individual stock holding in a portfolio be no more than a certain percentage of the portfolio. In terms of active weights, one can require that any individual active weight be less than a certain percentage. These constraints are aimed at controlling the specific risk of individual holdings and limiting the damage that the poor performance of any single stock to the total portfolio. Holding constraints can also be placed on an aggregated basis such as sector bounds for an active portfolio. A typical sector constraint can be ±2% for sector bets, and for a global equity portfolio, it can be ±2% for country bets.

However, by far the most prevalent form of holding constraint is the long-only constraint, which requires portfolios to be long in all stocks, i.e.,

the weights have to be nonnegative. In other words, it prohibits one from shorting stocks. Thus, the constraint is often referred to as a *no-short rule*. In the U.S. and around the world, the overwhelming majority of equity portfolios were managed as long-only products before equity long-short hedge funds became more acceptable in the late 1990s and early 2000s, even though they had existed since the 1960s. However, these hedge funds are generally only available to institutional investors and high-net-worth individuals. Mutual funds, which are a typical choice for most retail investors, are still almost exclusively long-only funds. Given the influence of the long-only constraints in the investment industry, one can ask: "Is the no-short rule a good rule?"

Generally, the answer is no, because it hinders managers' ability to generate excess returns. However, to some, shorting is associated with leverage and even appears unpatriotic. From a risk perspective, shorting stock outright can be a risky proposition. In contrast to buying a stock, where one can only lose 100% of the investment, shorting stock can lead to losses well above the initial investment[1]. However, these risks are well controlled in a risk-managed portfolio.

The no-short rule limits investment opportunities to generate returns. Consider the goal of active investment: beating the market-cap weighted benchmark subject to typical tracking error constraints. The cap-weighted index Goliaths are heavily weighted toward a set of large cap stocks. For example, the largest 4% of the Standard & Poor's (S&P) 500 names comprise about 70% of the index weight. In contrast, the smallest 25% comprise only 4% of the index. If the active manager's skill ability is equal across all cap ranges, how can he win? He cannot efficiently express his beliefs in specific stocks. With notional limits (no negative weights) on many of the "bad" ones, there is insufficient funding for the "good" ones. For example, managers can only underweight the small stocks by a few basis points (their weight in the index) when they have a negative forecast. This implies long-only managers can only add real value from their views on small stocks half of the time: when the forecast is positive! Given the fact that most capitalization-weighted benchmarks have a large portion of stocks with small benchmark weights, the impact of the long-only constraint on the portfolio return could potentially be significant. Thus, it is important for both portfolio managers and investors to analyze and estimate the magnitude of the likely impact.

A more recent solution is to make partial relaxation of the long-only constraint that resides in the traditional investment guideline. In this way,

the resulting portfolios can invest in both long and short, and continue to manage against their respective benchmarks. We refer hereafter to these as *constrained long-short portfolios*. For example, the manager might buy a 125% exposure in long-equity positions and sell a 25% exposure in short-equity positions with the net result being 100% long systematic risk. However, the total leverage to the alpha source is 150% (125% long and 25% short). Although the constrained long-short portfolios might be suboptimal compared to the market neutral portfolio (with derivatives), it offers considerable benefit over "handcuffed" long-only portfolios.

We shall provide results on long-only and constrained long-short portfolios in this chapter. This analysis presents an analytical challenge because the long-only constraint, or range constraint on portfolio weights, is an inequality constraint. With equality constraints such as risk neutral or sector neutral, we can find exact solutions to the optimal long-short portfolio weights. Our analysis so far has been based on the long-short portfolio setting, and we can establish an analytical relationship between the risk-adjusted information coefficient (IC) and the portfolio excess return. In contrast, with an inequality constraint, an analytical solution for the optimal weights does not exist, and a solution can only be found through numerical means.

We present an efficient numerical method for solving the mean–variance optimization problems with range constraints, making it possible to analyze the impact of the long-only constraint, or any other form of range constraints, very efficiently. It can be seen that the impact varies with different factors, even though it is generally negative in the form of a lower information ratio (IR). A closely related question is, how IR improves as we loosen the long-only constraint to allow short positions.

11.1 SECTOR NEUTRAL CONSTRAINT

We first analyze the impact of the sector neutral constraint on alpha factors. As we stated earlier in Chapter 5, for value factors such as earnings yield or book-to-price, one typically needs to employ them on a sector-relative basis. There are at least two reasons for this. One is that some sectors, such as technology, always look more expensive than other sectors, such as utilities, due to their higher growth prospects. Therefore, using value factors without any adjustment would cause a permanent underweight in the technology sector and a permanent overweight in the utility sector. The second, but related, reason is that these factors appear to be much less effective in predicting sector returns than relative stock returns within sectors.

11.1.1 Return Decomposition

We can analyze a factor's sector selection and stock selection ability by decomposing its excess returns. From Chapter 4, Equation 4.19, we have

$$\alpha_t = \sum_{i=1}^{N} w_i r_i = \lambda^{-1} \sum_{i=1}^{N} F_i R_i \,, \tag{11.1}$$

where F is the risk-adjusted forecast, R is the risk-adjusted return, λ is the risk-aversion parameter that calibrates the targeted tracking error, and N is the number of stocks. Suppose the stock universe consists of S sectors, $s = 1,2,\cdots,S$, and in sector s there are N_s stocks, such that

$$\sum_{s=1}^{S} N_s = N_1 + N_2 + \cdots N_S = N \,. \tag{11.2}$$

We can then rewrite (11.1) into a summation over sectors, i.e.,

$$\alpha_t = \lambda^{-1} \sum_{s=1}^{S} \sum_{i=1}^{N_s} F_{si} R_{si} \,, \tag{11.3}$$

where F_{si} and R_{si} are the risk-adjusted forecast and return of the i-th stock in s-th sector. We define the sector mean of forecasts and returns as

$$\overline{F}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} F_{si}, \text{ and } \overline{R}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} R_{si} \,. \tag{11.4}$$

The overall averages are given by

$$\overline{F} = \sum_{s=1}^{S} \frac{N_s}{N} \overline{F}_s, \text{ and } \overline{R} = \sum_{s=1}^{S} \frac{N_s}{N} \overline{R}_s \,, \tag{11.5}$$

and they are often close to zero in practice. Equation 11.3 can be written as

$$\begin{aligned} \alpha_t &= \lambda^{-1} \sum_{s=1}^{S} \sum_{i=1}^{N_s} \left(F_{si} - \overline{F}_s + \overline{F}_s\right)\left(R_{si} - \overline{R}_s + \overline{R}_s\right) \\ &= \lambda^{-1} \sum_{s=1}^{S} \sum_{i=1}^{N_s} \left[\left(F_{si} - \overline{F}_s\right)\left(R_{si} - \overline{R}_s\right) + \overline{R}_s\left(F_{si} - \overline{F}_s\right) + \overline{F}_s\left(R_{si} - \overline{R}_s\right) + \overline{F}_s \overline{R}_s\right] \end{aligned} \tag{11.6}$$

The second and third terms vanish by the definition of the averages. Therefore, we have

$$\alpha_t = \lambda^{-1} \sum_{s=1}^{S} \sum_{i=1}^{N_s} \left[\left(F_{si} - \bar{F}_s \right) \left(R_{si} - \bar{R}_s \right) \right] + \lambda^{-1} \sum_{s=1}^{S} N_s \bar{F}_s \bar{R}_s \,. \tag{11.7}$$

The interpretation of the first term is straightforward: it is the excess return generated by the sector-relative risk-adjusted forecast. The second term is related to the sector excess return, which can be rewritten as

$$\lambda^{-1} \sum_{s=1}^{S} N_s \bar{F}_s \bar{R}_s = \lambda^{-1} N \sum_{s=1}^{S} \frac{N_s}{N} \bar{F}_s \bar{R}_s \approx \lambda^{-1} N \sum_{s=1}^{S} \frac{N_s}{N} \left(\bar{F}_s - \bar{F} \right) \left(\bar{R}_s - \bar{R} \right) . \tag{11.8}$$

Thus, it is proportional to a weighted covariance between the aggregated sector forecast and the aggregated sector return, or excess return generated by the forecast on a sector level. Hence, we can write the excess return as the sum of the sector-relative excess return and the sector excess return and use this framework to analyze individual alpha factors.

Example 11.1

Table 11.1 provides a simple illustration with two sectors and three stocks in each sector. In sector 1, stock 1 has the lowest forecast while stock 3 has the highest forecast. This is also true in sector 2. We observe that the actual returns in both sectors have the same ranking. Hence, we conclude that within each sector the forecasts must have positive excess returns. The average forecast is –1 for sector 1 and 1 for sector 2, respectively, predicting a higher return for sector 2; instead, the average return is 5% for sector 1 and –5% for sector 2. In this case, the prediction for sector returns is wrong. Note the following remark:

- The decomposition of excess return essentially involves the decomposition of the covariance between the forecasts and the actual returns. Similarly, the variance of active returns can be decomposed into (a) stock return variance within sectors and (b) sector return variance (see Problem 11.2). This decomposition can shed light on the relative investment opportunities in "pure" stock selection and in sector allocation. For global equity portfolios that are managed with country allocation and stock selection, a similar analysis applies.

TABLE 11.1 An Example of Two Sectors and Three Stocks in Each Sector

Sector	Stock	F	R (%)	$F - F_s$
1	1	−1.50	0.0	−0.50
1	2	−1.00	5.0	0.00
1	3	−0.50	10.0	0.50
2	1	0.50	−10.0	−0.50
2	2	1.00	−5.0	0.00
2	3	1.50	0.0	0.50

11.1.2 Sector Constraint on Individual Factors

Table 11.2 shows the empirical results for the set of quantitative factors outlined in Chapter 5. Portfolio alpha (overall) is decomposed into stock selection alpha and sector timing alpha according to Equation 11.7. IR is the ratio of average return divided by the standard deviation of returns for each of the three alpha streams through time.

In general, sector timing alpha is of the same sign as the stock selection alpha, meaning that taking sector bets does increase alpha. However, the levels of the two sets of IR are quite different, with the stock selection IR consistently higher than the sector timing IR. This indicates that quantitative factors are better at selecting stocks bottom-up than making top-down sector calls.

One factor warrants closer examination: the short-term price momentum reversal factor (ret1). The stock selection and sector timing alphas have different signs, and the short-term momentum reversal phenomenon is much more pronounced within each sector rather then within the whole market. The IR of ret1 without sector neutralization is 0.44 (using positive number for IR), whereas it is 0.76 with sector neutralization. More interestingly, short-term sector momentum actually exhibits continuation rather than reversal; that is, sectors that outperformed in the last month tend to be winners again in the next 3 months, whereas stocks that outperformed in the last month tend to be losers in the next 3 months. Note the following remark:

- In general, factors that forecast stock returns are not strong in determining sector returns. Hence, in order to build effective sector forecasting models and implement sector rotation strategies, one needs to search for additional factors and possibly alternative modeling processes.

TABLE 11.2 Empirical Result in the U.S. Market Using R3000 as the Universe

		Overall		Stock Selection		Sector Timing	
		Alpha	**IR**	**Alpha**	**IR**	**Alpha**	**IR**
Value	CFO2EV	6.67%	1.11	6.39%	0.94	0.27%	0.20
	EBITDA2EV	5.26%	0.73	4.73%	0.62	0.54%	0.41
	E2PFY0	3.90%	0.58	3.35%	0.47	0.56%	0.38
	E2PFY1	3.31%	0.37	2.84%	0.31	0.48%	0.36
	BB2P	2.65%	0.30	1.96%	0.25	0.69%	0.28
	BB2EV	4.24%	0.65	3.79%	0.64	0.45%	0.28
	B2P	1.43%	0.15	1.05%	0.11	0.38%	0.31
	S2EV	3.67%	0.40	3.44%	0.35	0.23%	0.19
Fundamental	RNOA	3.05%	0.42	2.83%	0.39	0.21%	0.18
	CFROI	5.43%	0.91	5.35%	0.97	0.08%	0.08
	OL	3.66%	0.91	3.62%	0.95	0.04%	0.04
	OLinc	3.60%	1.07	3.59%	1.04	0.02%	0.05
	Wcinc	−3.97%	−0.90	−3.92%	−0.89	−0.05%	−0.08
	NCOinc	−3.15%	−0.68	−3.04%	−0.66	−0.10%	−0.10
	icapx	−3.00%	−0.70	−2.95%	−0.70	−0.05%	−0.10
	capxG	−1.99%	−0.50	−2.00%	−0.50	0.01%	0.01
	XF	−4.50%	−0.95	−4.25%	−1.00	−0.25%	−0.18
	shareInc	−2.28%	−0.52	−2.07%	−0.52	−0.21%	−0.12
Momentum	ret1	−4.36%	−0.44	−6.60%	−0.76	2.24%	0.72
	ret9	2.95%	0.22	3.19%	0.25	−0.24%	−0.06
	adjRet9	6.29%	0.49	5.22%	0.51	1.08%	0.24
	earnRev9	3.90%	0.38	4.25%	0.56	−0.35%	−0.10
	earnDiff9	5.10%	0.46	5.52%	0.67	−0.42%	−0.11

11.2 LONG/SHORT RATIO OF AN UNCONSTRAINED PORTFOLIO

Before analyzing the impact of long-only and other types of range constraints, we will first study the long/short ratio of an unconstrained active portfolio vs. a benchmark, because it represents the optimal setting of generating excess returns. In this case, as the portfolio is unconstrained, the active portfolio should be just the long-short portfolio. The benchmark has no effect on the active portfolio, but it becomes relevant when we aggregate the active weight with the benchmark weights to obtain the total portfolio weights. The distribution of the benchmark weights plays a role in determining the long/short ratio of portfolios that are managed

against that benchmark. Therefore, we will first examine that distribution empirically and present a statistical model for it.

11.2.1 Distribution of Benchmark Weights

Almost all capitalization-based benchmarks, to varying degrees, have more stocks with small weights than large weights. Over time, the distribution might change, for example, due to stocks' relative performance. However, the overall shape remains intact. Consider the S&P 500 index at February 2006. The stock with the largest weight was Exxon Mobil at 3.347%, and the stock with the smallest weight was Dana Corp (now bankrupt) at 0.006%, or 0.6 bps (basis points). The mean weight is 0.200%, whereas the median is 0.100%, demonstrating the skewness of the distribution. The top 10 names accounts for roughly 20% of the index weight, whereas the bottom half of the stocks accounts for only 13.5%. Figure 11.1 shows the histogram of the benchmark weights. It can be seen that there are only a handful of stocks with weights above 1%.

Another way of analyzing the distribution of benchmark weights is the cumulative sum of ranked stock weights. Figure 11.2 displays the sum as a function of the number of stocks included; the thick line is for the S&P 500 index, whereas the thin, dashed line is based on a fitted model with lognormal distributions that is described below. The function rises very rapidly at first and approaches 1 at a very slow rate in the end.

The model of the benchmark weights shown in Figure 11.2 is based on a lognormal distribution. For a random variable $x > 0$, it follows a lognormal distribution if ln(x) is normally distributed. The probability density is given by:

$$p(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \tag{11.9}$$

Figure 11.3a shows the probability density with $\mu = 0$, and $\sigma = 1.195$. The shape of the distribution resembles that of Figure 11.1, but the range is much too wide. The lognormal distribution, often used to model percentage changes in stock price, ranges from zero to infinity. As the benchmark weights are restricted to (0,1), we need to rescale the lognormal distribution to suit our purpose. If we rescale x by a factor of k, then the new density function should be

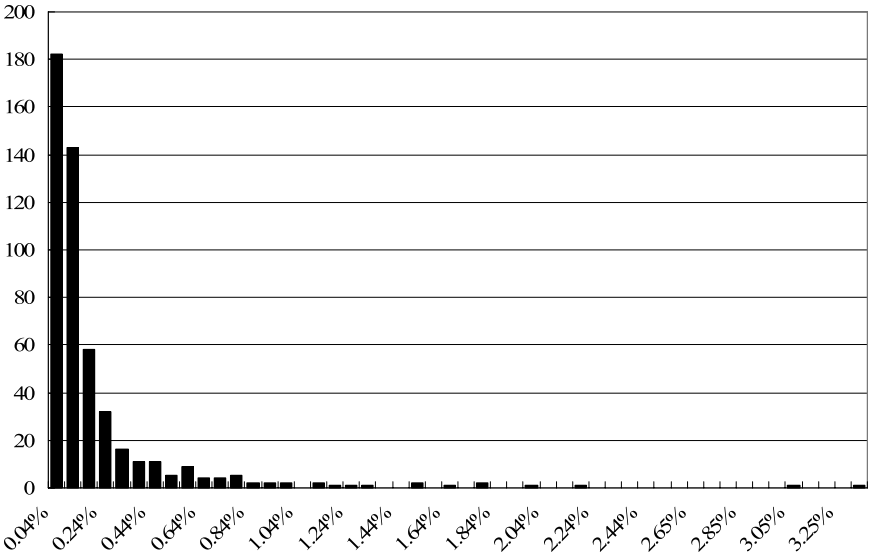


FIGURE 11.1. Histogram of benchmark weights in S&P 500 index as of February 2006.

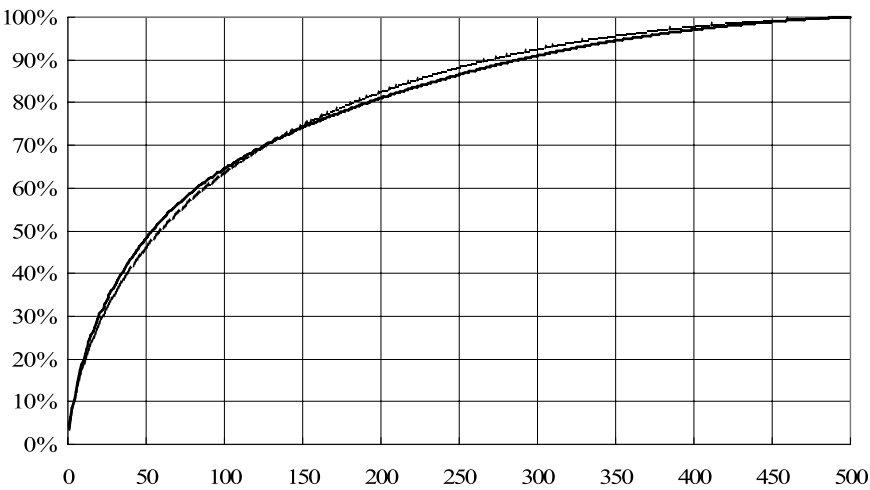
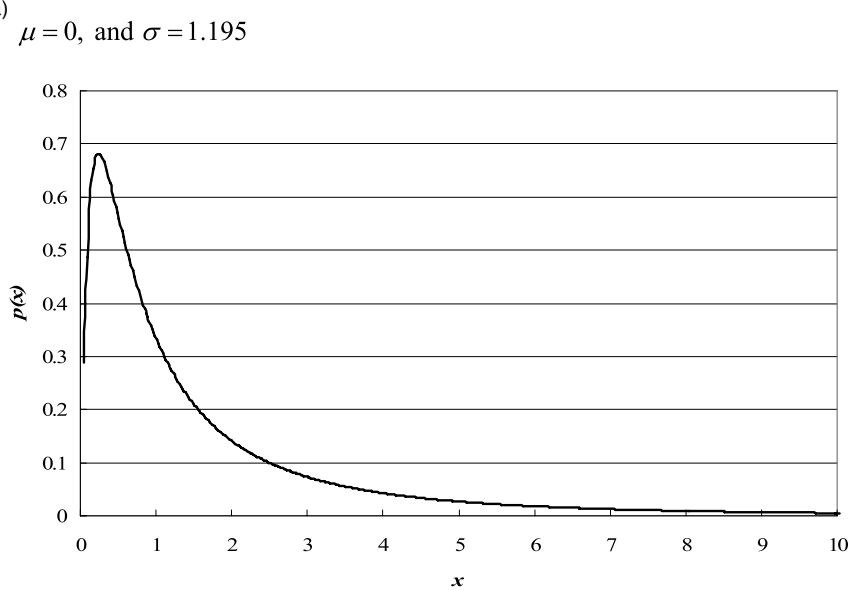


FIGURE 11.2. Cumulative weights of ranked benchmark weight: the solid line is for the S&P 500 index and the thin, dashed line is for the model.

$$\tilde{p}\left(x \mid \mu, \sigma, k\right) = k \cdot p\left(kx \mid \mu, \sigma\right). \tag{11.10}$$

(a)



(b)

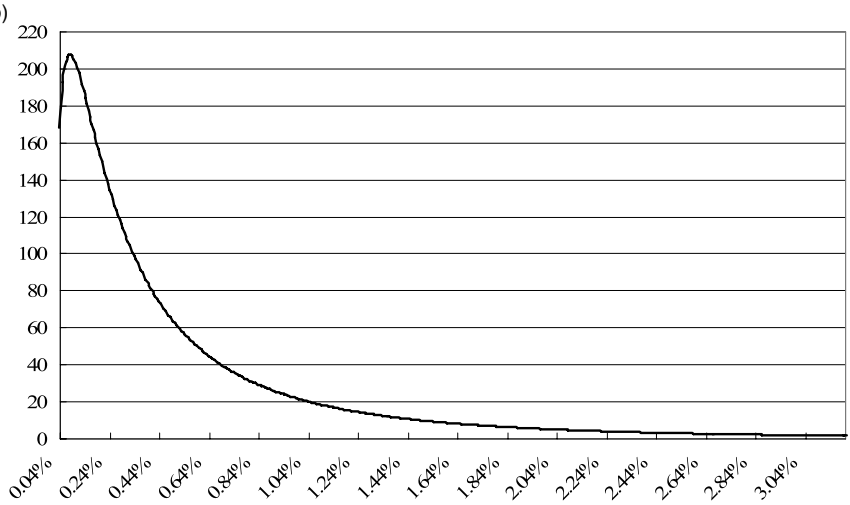


FIGURE 11.3. (a) Probability density function of the lognormal distribution function with $\mu = 0$, and $\sigma = 1.195$. (b) Scaled lognormal distribution of (a) with $k = 305$.

Figure 11.3b shows the scaled density function with factor k = 305. The graph now resembles the histogram of S&P 500 index weights in Figure 11.1.

11.2.2 Simulation of Benchmark Weights

Grinold and Kahn (2000) provided an algorithm to simulate benchmark weights based on a scaled lognormal distribution. For a given number of stocks N in the benchmark, a parameter c is used to characterize the concentration of the index. If $c = 0$, the index is equally weighted. As c increases, the index becomes more concentrated. The algorithm has four steps:

1. Discretize the probability interval (0,1) with $p_i = 1 - \frac{i - 0.5}{N}$, $i = 1, \cdots, N$.
2. Find the value of the standard normal variable that has the cumulative probability p_i, i.e., $y_i = \Phi^{-1}(p_i)$, where Φ^{-1} is the inverse of the cumulative density function.
3. Transform y_i to a lognormal variable using $s_i = \exp(cy_i)$, c being the concentration parameter.
4. Scale s_i to obtain benchmark weight $b_i = s_i \Big/ \sum_{i=1}^{N} s_i$.

Figure 11.4 shows the simulated benchmark weights for several values of c. The curves are the cumulative total of weights ranked in descending order. The curve for c = 0, i.e., an equally weight benchmark, is a straight line. As c increases, the benchmark becomes top heavy with a few stocks occupying more weight within the benchmark.

11.2.3 Long/Short Ratio of a Single Stock

Our approach to obtaining the long/short ratio of a portfolio is to calculate the long/short ratio of a single stock and then sum up across the benchmark. From Chapter 4, we know that the long-short portfolio weights are $w_i = \lambda^{-1} F_i / \sigma_i$, where F_i is the risk-adjusted forecast, σ_i is the stock-specific risk, and λ is the risk-aversion parameter. The risk-aversion parameter is related to the target tracking error by

$$\frac{1}{\lambda} = \frac{\sigma_{\text{target}}}{\sqrt{N}}. \tag{11.11}$$

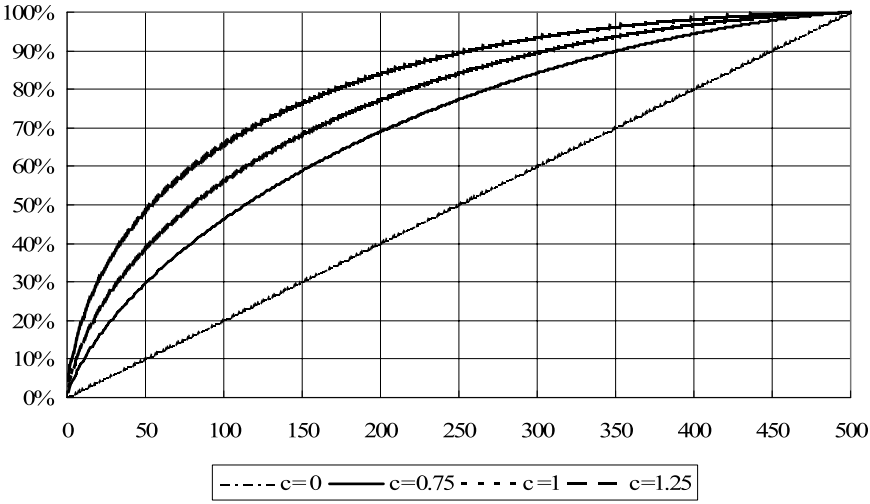


FIGURE 11.4. Cumulative weights of ranked benchmark stocks for different values of *c*.

We have assumed that the risk-adjusted forecast is standardized, i.e., dis(F) = 1 and is of zero mean, and *N* is the number of stocks. Hence the active weight is given by

$$w_i = \frac{\sigma_{\text{target}} F_i}{\sqrt{N}\sigma_i}. \tag{11.12}$$

The benchmark weights are b_i with $\sum_{i=1}^{N} b_i = 1$, and $b_i \geq 0$.

- Normally, benchmark weights are all positive. We will allow b_i to be zero if the stock is an out-of-benchmark bet. Hence, in our notation, the stock universe includes stocks both in and out of the benchmark.

Given the active weight (11.12) and the benchmark weight b_i, the total portfolio weight in a stock is $W_i = w_i + b_i$. If $W_i > 0$, it is a long position and if $W_i < 0$, it is a short position.

If we assume that the risk-adjusted forecast is normally distributed for stock *i*, according to (11.12), the active weight follows a normal distribution with zero mean and standard deviation

$$s_i = \frac{\sigma_{\text{target}}}{\sqrt{N}\sigma_i}. \tag{11.13}$$

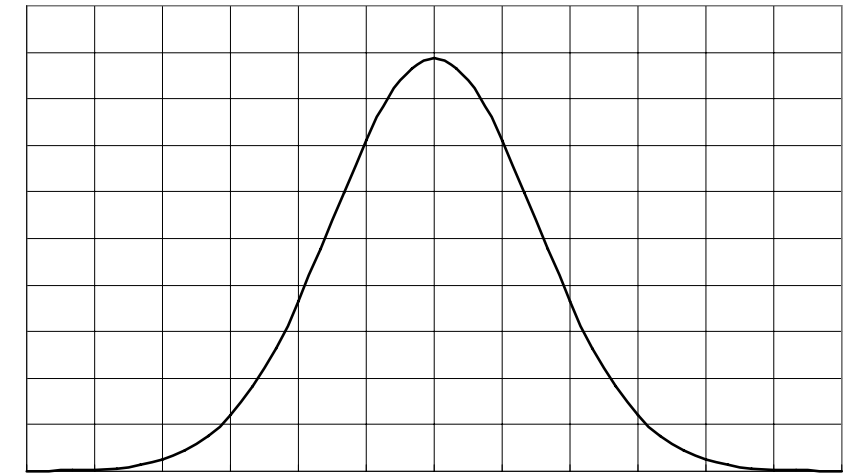


FIGURE 11.5. The probability density function of the total weight of a stock with 0.3% benchmark weight and 30% specific risk.

Hence, the total weight $W_i = w_i + b_i$ follows a normal distribution with mean b_i and standard deviation s_i.

Example 11.2

Consider an active portfolio with 3% targeted tracking error with 500 stocks. If the stock-specific risk is 30%, then

$$s_i = \frac{3\%}{\sqrt{500} \cdot 30\%} = 0.45\% .$$

The active position has a standard deviation of 45 bps. If the benchmark weight of the stock is 0.3%, or 30 bps, the density distribution of the total weight looks as in Figure 11.5.

The probability of W_i being a short position is given by

$$\mathrm{P}\left(W_i < 0\right) = \frac{1}{\sqrt{2\pi} s_i} \int_{-\infty}^{0} \exp\left[-\frac{\left(x - b_i\right)^2}{2s_i^2}\right] dx . \tag{11.14}$$

It is simply the cumulative distribution function of W_i evaluated at 0. Because $b_i \geq 0$, (11.14) is always less than one half. If $b_i = 0$, the probability

is exactly one half. This is relevant for stocks out of the benchmark, and it is also true for long-short portfolios without a benchmark. For the stock considered in Example 11.2, the probability of it being in a short position is about 25%. We are likely to prefer a short position for a given stock if the following conditions are met: (1) the lower the forecast, (2) the smaller the benchmark weight, (3) the smaller the specific risk, (4) the lower the risk-aversion parameter, and (5) the higher the target tracking error, *ceteris paribus*.

- We note that the probability is for multiple periods. At any given period, depending on the forecast for the stock, the position could be either positive (long) or negative (short). This is true for all stocks.

11.2.4 Portfolio Average Long/Short Ratio

The total short position of the whole portfolio is simply the sum of short positions, i.e.,

$$S = \sum_{W_i<0} W_i = \sum_{w_i+b_i<0} \left(w_i + b_i\right). \tag{11.15}$$

Similarly, the total long is

$$L = \sum_{W_i>0} W_i = \sum_{w_i+b_i>0} \left(w_i + b_i\right). \tag{11.16}$$

In our notation, short positions are weights that are negative. Because the active weights are dollar neutral, the sum of total long and total short should be just the total benchmark weights, i.e., $L + S = 1$. However, in any given period, the total long and short are not fixed. For instance, if the forecasts happen to be high for small stocks and low for large stocks in that period, then the total short would be lower, as we are more likely to overweight small stocks and underweight large ones, reducing the chance of negative positions. The situation would be reversed if the forecasts happen to be high for large stocks but low for small stocks. Then, we are likely to underweight small stocks, often leading to short positions.

We are interested in the averages of the total long and short positions. For the shorts, we have

$$\overline{S} = \sum_{i=1}^{N} \mathrm{E}\left(w_i + b_i \,|\, w_i + b_i < 0\right). \tag{11.17}$$

We simply calculate the average short position for each stock and sum them up. As the weight of stock i follows a normal distribution, we have

$$\begin{aligned}\mathrm{E}\left(w_i+b_i \mid w_i+b_i<0\right) &= \frac{1}{\sqrt{2\pi}s_i}\int_{-\infty}^{-b_i}\left(x+b_i\right)\exp\left(-\frac{x^2}{2s_i^2}\right)dx \\ &= -\frac{s_i}{\sqrt{2\pi}}\exp\left(-\frac{b_i^2}{2s_i^2}\right)+b_i\cdot\mathrm{cdf}\left(-b_i,0,s_i\right).\end{aligned} \tag{11.18}$$

The function cdf is the cumulative density function evaluated at $-b_i$ for the normal distribution with zero mean and standard deviation s_i.

Example 11.3

Consider the case of the stock in Example 11.2. The benchmark weight is 0.3%, or 30 bps. The standard deviation of the active position is 0.45%, or 45 bps. Substituting them into (11.18), we obtain the average short position of –0.07%, or –7 bps.

Example 11.4

For out-of-benchmark stocks or long-short portfolio, we have $b_i = 0$. Then

$$\mathrm{E}\left(w_i \mid w_i<0\right) = -\frac{s_i}{\sqrt{2\pi}} = -\frac{\sigma_{\text{target}}}{\sqrt{2\pi}\sqrt{N}\sigma_i}.$$

Assuming constant specific risk $\sigma_i = \sigma_0$, then $\bar{S} = -\frac{\sqrt{N}\sigma_{\text{target}}}{\sqrt{2\pi}\sigma_0}$.

With simulated benchmark weights b_i, Equation 11.17 and Equation 11.18 give rise to the average long/short ratio for the total portfolio, which is a function of two parameters: the concentration parameter c, and the targeted tracking error σ_{target}. Similar results have been obtained by Clarke et al. (2004). Figure 11.6 show the results for a fixed value of c and varying targeted tracking error. It plots four curves. First, the curve for long plus short (L+S) is always at 100%. The next two curves are for both long and short. As the tracking error increases, the long and short both increase in magnitude, with long exceeding 100% and short becoming more negative. The rate of increase for both sides is roughly linear. The fourth curve is for the total leverage (L-S), and it sits on the top. When the tracking error is small, at 0.5%, the total leverage is only 104%. When the

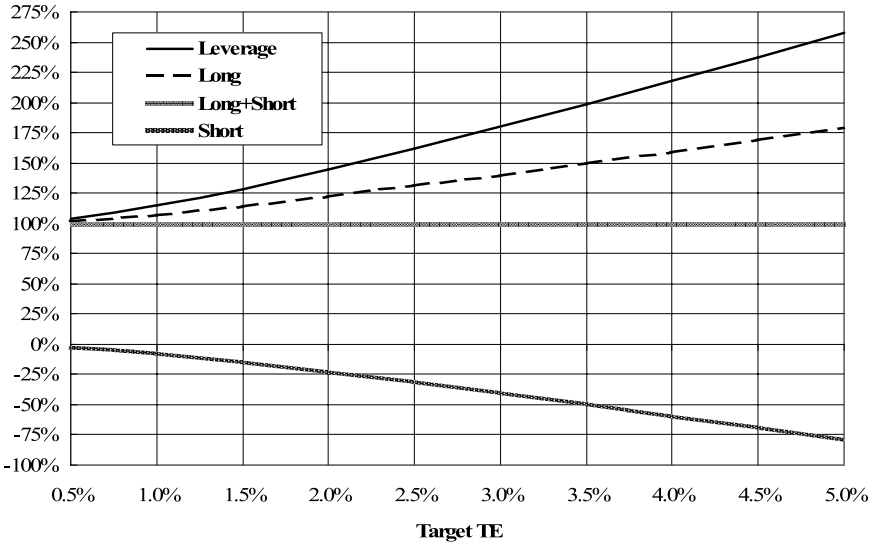


FIGURE 11.6. The long/short ratio of active portfolios with 500 stocks with $c = 1.2$ and specific risk at 40% for all stocks.

tracking error is at 2.5%, the long is 131%, the short is –31%, and the total leverage is 162%. When the tracking error reaches 5%, the long/short ratio is 179%/–79%, and the total leverage is 258%. In this case, if an investor has $100 in capital, he would buy $179 worth of stocks (long) and borrow and sell $79 worth of other stocks. There should be no overlapping between the longs and the shorts.

Figure 11.7 shows the change in the long/short ratio as the benchmark index c changes. The tracking error is fixed at 2.5%, and again our benchmark has 500 stocks, and the specific risk is set at 40% for all stocks. As we can see from the graph, the long, the short, and the total leverage increase slowly as c increases. When c is zero for an equally weighted benchmark, the long/short ratio is 119%/–19% and the total leverage is 138%. When c increases and the benchmark becomes increasingly concentrated, the long/short ratio increases. At $c = 1.2$, the long/short ratio is 131%/–31% and the total leverage is 162%. As c reaches 1.5, the long/short ratio is 135%/–35% with a total leverage of 170%. So there is an increase of 8% in total leverage as c goes from 1.2 to 1.5.

Finally, Figure 11.8 shows a three-dimensional view of the total leverage as a function of both c and tracking error. The graph again shows that the total leverage increases rapidly with an increase in tracking error and the pace is much more gradual with an increase in benchmark index c.

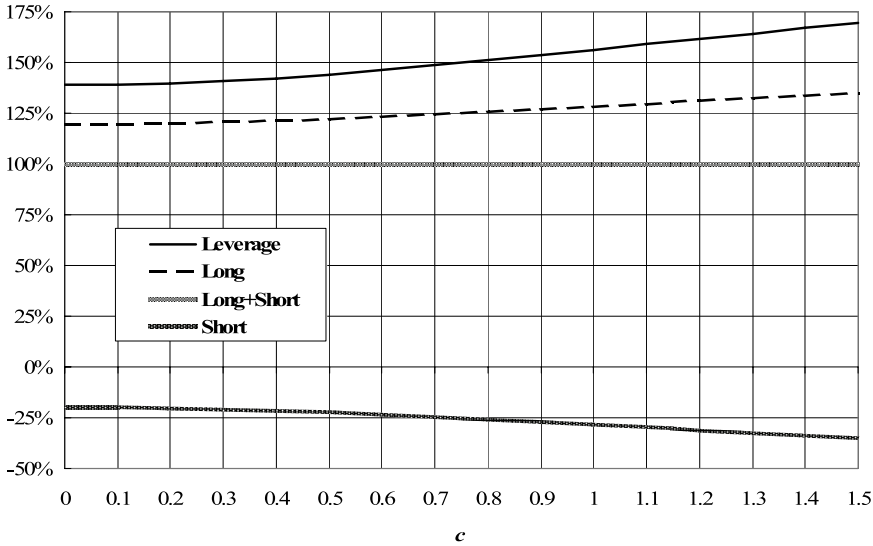


FIGURE 11.7. The long/short ratio of active portfolios with 500 stocks and specific risk at 40% for all stocks. The tracking error is 2.5%. The benchmark index *c* changes from 0 (equally weighted benchmark) to 1.5.

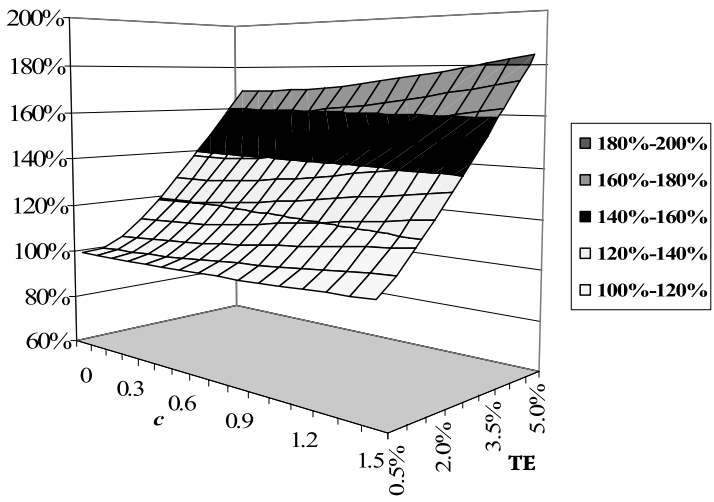


FIGURE 11.8. The total leverage of optimal portfolios as a function of both benchmark index *c* and tracking error. The benchmark has 500 stocks and specific risk is 40% for all stocks.

11.3 LONG-ONLY PORTFOLIOS

When the long-only constraint is placed on a portfolio, it is equivalent to a range constraint on active positions of all stocks: they must always be greater than the negative of their benchmark weights, i.e., $w_i \geq -b_i$. With the no-short rule, the portfolio's long/short ratio would be 100%/0%, which is obviously different from the long/short ratio of unconstrained portfolios. On the stock level, it is expected that the optimal weights of these two types of portfolios are different, resulting in different performance. For portfolios with low tracking errors, the difference in weights might not be so large. However, for portfolios with high tracking errors, the difference can be very significant. In this section, we shall analyze the impact of the long-only constraint on portfolio weights and performance of active strategies. In practice, most long-only portfolios are managed with maximum weight constraints in addition to the no-short constraint. The same is true for long-short portfolios, for which the range of stock weights is generally constrained. However, as there is no benchmark for long-short portfolios, the range is absolute, not relative to a benchmark.

The disadvantage of long-only portfolios managed against market-cap-weighted benchmarks has been stated previously at the stock level. The asymmetry also severely reduces the opportunity set for long-only managers who maintain minimal portfolio exposure to systematic size risk. With a size risk constraint, the active positions of a portfolio must be roughly balanced among stocks with similar market cap. Since this is not achievable among small stocks due to the long-only constraint, the portfolio is forced to take up more active positions and spend the majority of its active risk budgets among large stocks, where the market is probably more efficient and thus offers less alpha. We will demonstrate that an active portfolio with 3% targeted tracking error in the S&P 500 stock universe could have close to 50% of active risk in the S&P 100.

11.3.1 Constrained Long-Short Portfolios

Constrained long-short portfolios lie between long-only portfolios and unconstrained portfolios. Such portfolios, for example, might buy long 125% stocks and sell short 25% stocks, so the net result is still 100% with the total leverage ratio of 125% + 25% = 150%. Whereas the constrained long-short portfolios might still be suboptimal compared to unconstrained portfolios, they offer considerable benefit over long-only portfolios and have gained increasing acceptance with institutional investors.

With some ability to short, the constrained long-short portfolios alleviate some of the problems discussed previously. Therefore, in theory, one should expect them to deliver higher risk-adjusted returns than their long-only counterparts. However, there is an additional cost for the constrained long-short portfolios that is absent in the long-only portfolios that is due to the leverage. To see the leverage cost, it is important to understand the mechanism of long-short investing. Although standard financial theory often invokes the concept of a self-financing portfolio that implies costless leverage, in practice, leverage is not free. Suppose an investor has $100. With long-only portfolios, the investor can buy $100 worth of stocks and the leverage ratio is 1:1. As no borrowing is involved, there is no leverage cost. With a 125/25 portfolio, the investor buys $100 worth of stocks with his own capital. He then borrows $25 to buy an additional $25 worth of stocks, and at the same time borrows $25 worth of stocks to sell. From a pure theoretical standpoint, the short proceeds of $25 would be used to buy the additional $25 long with no additional cost. However, from a practical standpoint, used by prime brokers for pricing, the investor has bought $25 worth of stocks on margin, whereas the short proceeds of $25 is kept at the broker as collateral for the short positions. The short proceed earns an interest rebate from the brokers, but the rate is always lower than the financing cost on the long side. Therefore, the interest rate spread on the $25 is a cost that the investor must bear.[2]

Example 11.5

Suppose the spread between the financing and the rebate is 1%, the additional cost for 125/25 portfolios would be 0.25% or 25 bps. Similarly, the additional cost for 150/50 portfolios would 0.5% or 50 bps.

11.3.2 Numerical Methods for MV Optimization with Range Constraints

An analytical solution does not exist for optimal weights of long-only portfolios, or range-constrained portfolios, in general. We shall carry out our analysis through numerical means. The problem falls in the general category of quadratic programming, in which we maximize a quadratic objective function subject to linear constraints, as well as range constraints. For large-scale problems with thousands of stocks, finding numerical solutions of general problems can be time consuming. However, there exists an efficient algorithm for the special case in which the covariance matrix

is diagonal. This would be true if we neutralize all the systematic factor exposures and optimize with residual alphas and specific risks.

The algorithm is based on the Kuhn–Tucker condition for optimization with inequality constraints. The appendix provides a detailed description of the Kuhn–Tucker condition for the general optimization problem and its application to mean–variance optimization, which is to find the optimal active weights **w** in the following

$$\begin{aligned}
&\text{Maximize: } \mathbf{f}' \cdot \mathbf{w} \\
&\text{Subject to:} \\
&\qquad \mathbf{w}' \cdot \mathbf{\Sigma} \cdot \mathbf{w} = \sigma^2_{\text{target}} \\
&\qquad \mathbf{w}' \cdot \mathbf{i} = 0, \text{ and } \mathbf{w}' \cdot \mathbf{B} = 0 \\
&\qquad w_i - U_i \le 0, \text{ and } L_i - w_i \le 0, \text{ for } i = 1, \cdots, N.
\end{aligned} \tag{11.19}$$

The vector **f** is the forecast vector, the covariance matrix $\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_I\mathbf{B}' + \mathbf{S}$, and σ_{target} is the target tracking error. The equality constraints are dollar neutral and market neutral $\mathbf{w}' \cdot \mathbf{i} = 0$, and $\mathbf{w}' \cdot \mathbf{B} = 0$. The range constraints are

$$w_i - U_i \le 0, \text{ and } L_i - w_i \le 0, \text{ for } i = 1, \cdots, N.$$

The Kuhn–Tucker condition implies that the solution takes the following form:

$$\begin{aligned}
&\mathbf{w} = \frac{1}{2\lambda}\mathbf{S}^{-1}\mathbf{f}_{\text{adj}}, \text{ or} \\
&w_i = \frac{f_i - l_0 - l_1 b_{1i} - \cdots l_K b_{Ki} - \tilde{l}_{1i} + \tilde{l}_{2i}}{2\lambda\sigma_i^2}.
\end{aligned} \tag{11.20}$$

In the solution, l_0 is the Lagrangian multiplier for the dollar neutral constraint; $l_1, \cdots, l_K$ are the Lagrangian multipliers for market neutral constraints; $\tilde{l}_{1i}$ and $\tilde{l}_{2i}$ are the Lagrangian multipliers for the upper and lower bounds, respectively; and λ is the Lagrangian multiplier for the tracking error constraint. As only one of $\tilde{l}_{1i}$ and $\tilde{l}_{2i}$ can be nonzero, we combine them into one: $\tilde{l}_i = \tilde{l}_{1i} - \tilde{l}_{2i}$.

Our numerical algorithm finds the optimal weights and the Lagrangian multipliers iteratively. At step n, we have the weight w_i^n and multipliers $l_0^n, l_1^n, \cdots, l_K^n$, $\tilde{l}_i^n$, λ^n. If the weights violate the range constraint, we proceed as follows:

- Apply range constraints to the weight $w_i^{new} = \max\left(\min\left(w_i^{new}, U_i\right), L_i\right)$.
- Update Lagrangian multipliers for range constraints with

$$\tilde{l}_i^{n+1} = f_i - l_0^n - l_1^n b_{1i} - \cdots l_K^n b_{Ki} - 2\lambda^n \sigma_i^2 w_i^{new}.$$

- Update Lagrangian multipliers for dollar neutral and beta-neutral constraints with the solution from the system of linear equations in which

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{N} x_i y_i / \sigma_i^2$$

(see Chapter 4) and $\tilde{\mathbf{l}}^{n+1}$ is the vector of newly updated Lagrangian multipliers from the previous step.

$$\begin{cases} l_0^{n+1}\langle \mathbf{i}, \mathbf{i} \rangle + l_1^{n+1}\langle \mathbf{i}, \mathbf{b}_1 \rangle + \cdots + l_K^{n+1}\langle \mathbf{i}, \mathbf{b}_K \rangle = \left\langle \mathbf{i}, \mathbf{f} - \tilde{\mathbf{l}}^{n+1} \right\rangle \\ l_0^{n+1}\langle \mathbf{b}_1, \mathbf{i} \rangle + l_1^{n+1}\langle \mathbf{b}_1, \mathbf{b}_1 \rangle + \cdots + l_K^{n+1}\langle \mathbf{b}_1, \mathbf{b}_K \rangle = \left\langle \mathbf{b}_1, \mathbf{f} - \tilde{\mathbf{l}}^{n+1} \right\rangle \\ \vdots \\ l_0^{n+1}\langle \mathbf{b}_K, \mathbf{i} \rangle + l_1^{n+1}\langle \mathbf{b}_K, \mathbf{b}_1 \rangle + \cdots + l_K^{n+1}\langle \mathbf{b}_K, \mathbf{b}_K \rangle = \left\langle \mathbf{b}_K, \mathbf{f} - \tilde{\mathbf{l}}^{n+1} \right\rangle \end{cases}$$

- Calculate the tracking error of $\mathbf{w}^{new}$ and update the Lagrangian multiplier for the tracking error

$$\sigma^{new} = \sqrt{\sum_{i=1}^{N} \left(w_i^{new}\right)^2 \sigma_i^2}, \qquad \lambda^{n+1} = \lambda^n \frac{\sigma^{new}}{\sigma_{\text{target}}}.$$

- Calculate the new weights w_i^{n+1} by

$$w_i^{n+1} = \frac{f_i - l_0^{n+1} - l_1^{n+1} b_{1i} - \cdots l_K^{n+1} b_{Ki} - \tilde{l}_i^{n+1}}{2\lambda^{n+1}\sigma_i^2}.$$

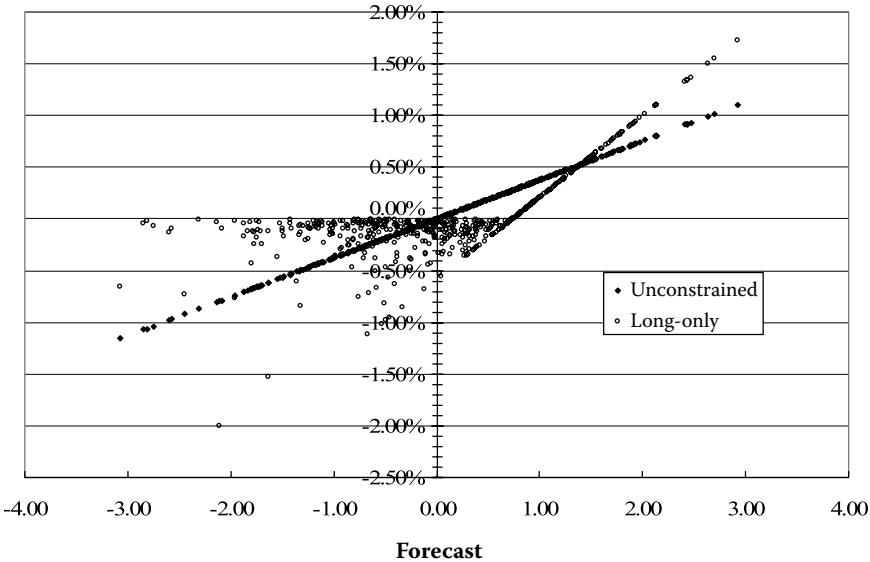


FIGURE 11.9. Optimal active weights of unconstrained and long-only portfolios.

After these steps, we have the weight w_i^{n+1} and multipliers $l_0^{n+1}, l_1^{n+1}, \cdots, l_K^{n+1}$, $\tilde{l}_i^{n+1}$, λ^{n+1}. The new weights are checked against the range constraints. If there is violation, the foregoing steps are repeated until there is no range violation.

Example 11.6

We use the preceding algorithm to find long-only optimal portfolio weights against a benchmark of 500 stocks that has a concentration index of c 1.2, and compare these weights to unconstrained optimal weights. Both portfolios have a targeted tracking error of 3%, and all stocks are assumed to have a specific risk of 35%. We also impose a maximum active weight of 2% for all stocks. The forecasts are simulated based on a standard normal distribution. Figure 11.9 plots the forecasts vs. both sets of optimal active weights. We first note that the unconstrained optimal weights form a straight line going through the origin. Indeed, they are proportional to the forecasts. The optimal weights of the long-only portfolio show several features: (1) There are many small negative weights. They belong to the active weights of stocks with tiny benchmark weights, due to the long-only constraint; (2) Positive active weights also seem to fall on a straight line, which has a steeper slope and a negative intercept on the y-axis. Some

negative active weights also fall on this line. Mathematically, this is due to a smaller Lagrangian multiplier for the tracking error constraint in the long-only optimization than its counterpart in the unconstrained optimization (the slope is inversely proportional to λ in Equation 11.20). In addition, the Lagrangian multiplier for the dollar neutral constraint is positive. This implies that large positive, active weights are magnified whereas smaller positive ones are shrunk; and (3) Many stocks with positive forecasts will end up with negative active weights, as underweights in stocks with small benchmark weights are not sufficient to fund overweights. Note the following remark:

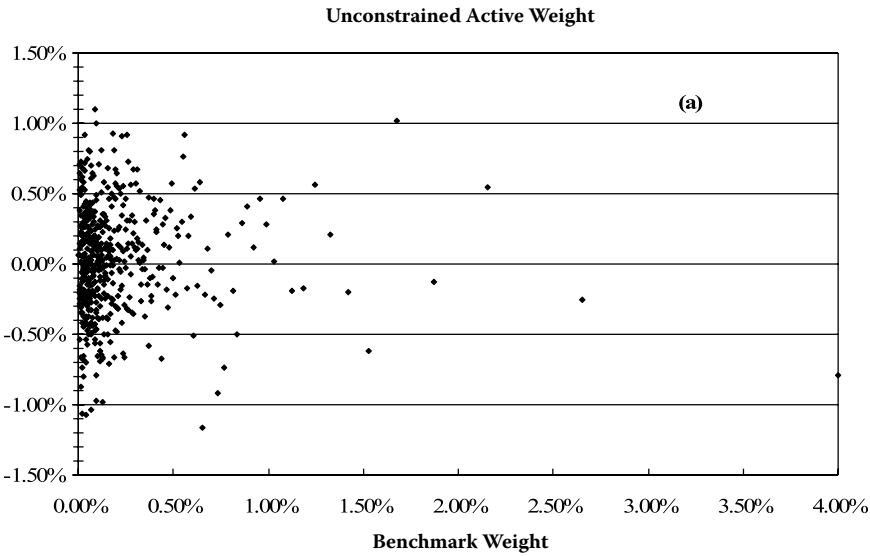
- In the unconstrained optimal portfolio, the active weights and the forecasts have perfect correlation. However, in the constrained portfolio, the correlation is less than perfect. This correlation can be used as a gauge of the stringency of the constraint. Alternatively, it measures the extent to which the forecasts are reflected in the portfolio. Clarke et al. (2002) coined the term *transfer coefficient* for a variation of this correlation. In our example, this correlation is about 0.7.

Figure 11.10 plots the active weights vs. the benchmark weights. In Figure 11.10a for an unconstrained portfolio, the active weights are independent of the benchmark. In Figure 11.10b, for the long-only portfolio, the active weights are bounded below by the benchmark, and there is a negative correlation between the two.

11.4 THE INFORMATION RATIO OF LONG-ONLY AND LONG-SHORT PORTFOLIOS

Unconstrained optimal portfolios have intrinsic long/short leverage ratios, depending on portfolio and benchmark characteristics such as target tracking error, benchmark concentration, stock-specific risks, and the number of stocks in the benchmark and portfolio. In theory, these long/short ratios are optimal for given portfolio mandates in terms of maximizing the IR. Range constraints such as long-only or limited shorting would reduce the theoretical IR.

With the numerical algorithm described earlier, we now analyze the information ratio of long-only, as well as constrained long-short portfolios. There are many practical reasons that might prevent portfolio managers from fully implementing the unconstrained optimal portfolios. Some constraints are institutional. For example, prime brokers might



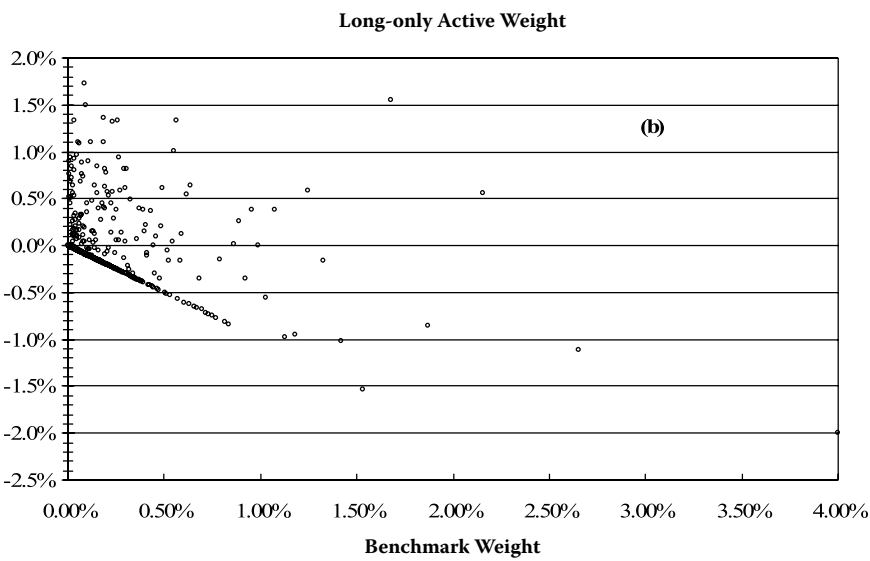


FIGURE 11.10. Optimal active weights vs. the benchmark weights: (a) for unconstrained portfolio and (b) for long-only portfolio.

place limits on the amount of leverage allowed in a portfolio; or it might be hard to borrow certain stocks, which reduces the amount of shorting. Some concerns are cost related. As mentioned earlier, the higher the leverage, the higher the financing cost. In addition, portfolios with higher

leverage require higher turnover, resulting in higher transaction costs, a component often missed in some previous analysis of long-short portfolios (see Chapter 8). Therefore, there is a need to distinguish between theoretical IR and net IR that account for both leverage and transaction costs. However, note the following remark:

- Some other issues arise in long-short investing that must be considered. For example, the number of stocks in a long-short portfolio will be much higher than that in a long-only portfolio. This might not be a big issue for quantitative managers, but it could impose additional work on fundamental managers.

To better understand the benefit of constrained long-short portfolios compared to long-only portfolios, we carry out numerical simulations for long-only portfolios and long-short portfolios with varying amounts of short positions. In the simulation, we first calculate the "paper" or theoretical excess returns from portfolio weights and returns, and then deduct financing costs according to the portfolio's leverage and by transaction costs according to portfolio turnover.

11.4.1 Simulation Assumptions

Simulation results depend on a host of parameters, which are listed in detail as follows:

- *Investment universe and benchmark:* To be consistent with our discussion of unconstrained optimal portfolios, we choose a universe of 500 stocks and portfolios that are managed against a 500-stock index, with the index concentration being measured by the parameter c. Stock-specific risk is 35% for all stocks.
- *Tracking error target:* We choose a series of tracking error targets ranging from 1 to 5%.
- *Long/short ratio:* We impose the long/short ratio constraints through a range constraint on individual stocks. Starting from long-only portfolios, which have a constraint on the weights as $w_i \geq 0$, we gradually loosen the constraint to $w_i \geq -s$, where s is the short position allowed in individual stocks. For instance, if s = 0.1%, we can short each stock by a maximum of 10 bps. As s grows, the total short position grows and the portfolio would approach the unconstrained optimal portfolio. We also set the maximum active weight at ±3%.

- *Other portfolio constraints:* Besides targeted tracking error and range constraints on the individual stocks, the only other portfolio constraint is the dollar neutral constraint.
- *Forecasts:* We simulated forecast in the form of normally distributed *z*-scores. We also assume consecutive forecasts have autocorrelation ρ_f, which is one of the factors influencing portfolio turnover. The other factors are target tracking error and the leverage ratio (see Chapter 8).
- *Information coefficient and returns:* The risk-adjusted returns are simulated based on the IC — the cross-sectional correlation coefficient between the forecast and the returns. Two parameters characterize the random nature of IC: the average IC and the standard deviation of IC. The risk-adjusted return is also assumed to be normally distributed and its cross-sectional dispersion is unity (Qian and Hua 2004).

In each simulation, we first generate standardized forecasts and actual returns based on either a constant or stochastic IC. We then calculate excess returns of active portfolios that are managed against a benchmark with a specified concentration index and a series of targeted tracking errors that are optimized with different range constraints that lead to different long/short ratios. A theoretical IR can then be obtained from the time series of excess returns. In addition, we also obtain the average portfolio turnover and long/short ratio of these portfolios. We estimate transaction costs and leverage costs and subtract them from the theoretical excess return. Finally, "net" IR is calculated as the ratio of net excess return to the realized tracking error, not the target tracking error. We note that the realized tracking error is higher than the targeted tracking error when the IC has intertemporal variability (Qian and Hua 2004).

11.4.2 Simulation Results: Constant IC

Table 11.3 shows the results of one such simulation in which we assume that the IC is constant and the only source of time-series variation is sampling error. There are 11 portfolios across the table, ranging from the long-only portfolio (column 1) to the unconstrained portfolio (column 11). They all have the same target tracking error of 3%. We have assumed that the IC is constant at 0.1. As a result, the realized tracking error, or standard deviation of alpha, is also 3%. The theoretical IR of the unconstrained portfolio (column 11) is then the IC times the square root of *N*,

TABLE 11.3 Simulation Results for Long-Only Portfolios, Constrained Long-Short Portfolios, and Unconstrained Long-Short Portfolios

	Long-only										**Unconstrained**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Avg Alpha	4.82%	5.30%	5.56%	5.81%	6.05%	6.24%	6.41%	6.54%	6.63%	6.69%	6.72%
Std Alpha	3.04%	3.03%	3.03%	3.02%	3.01%	3.00%	3.00%	2.99%	2.99%	3.00%	3.00%
Theoretical IR	1.59	1.75	1.84	1.92	2.01	2.08	2.14	2.19	2.22	2.23	2.24
Total Long	100%	109%	115%	121%	127%	133%	138%	143%	146%	147%	148%
Turnover	64%	71%	75%	79%	83%	86%	89%	91%	93%	94%	94%
Leverage Cost	0.00%	0.09%	0.15%	0.21%	0.27%	0.33%	0.38%	0.43%	0.46%	0.47%	0.48%
Transaction Cost	0.64%	0.71%	0.75%	0.79%	0.83%	0.86%	0.89%	0.91%	0.93%	0.94%	0.94%
Net Avg Alpha	4.18%	4.50%	4.66%	4.81%	4.94%	5.05%	5.13%	5.20%	5.25%	5.28%	5.30%
Net IR	1.38	1.48	1.54	1.59	1.64	1.68	1.71	1.74	1.75	1.76	1.77
Theoretical IR decay	0.71	0.78	0.82	0.86	0.90	0.93	0.95	0.98	0.99	1.00	1.00
Net IR Decay	0.78	0.84	0.87	0.90	0.93	0.95	0.97	0.98	0.99	1.00	1.00
Transfer Coefficient	0.70	0.78	0.82	0.85	0.89	0.92	0.95	0.97	0.98	0.99	1.00

Note: The number of stocks is 500; benchmark concentration index, 1.2; target tracking error, 3%; stock-specific risk, 35%; average IC, 0.1 with no intertemporal variation; forecast autocorrelation, 0.25; leverage cost, 1%; and the transaction cost, 1%.

Source: From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 33, No. 2, 1–9, Winter 2007. With permission.

equaling 2.24, whereas the theoretical IR of the long-only portfolio (column 1) is only 1.59.

The next two rows of Table 11.3 report the total long positions of the portfolios and their turnover. As we relax the short constraint, the total long and the total short both increase. Because the long minus short is always 100%, we omit the short from the table. For instance, the portfolio in column 5 is long 127% on average and its theoretical IR is 2.01. Table 11.3 shows that portfolio turnover increases with leverage. It averages 64% for the long-only portfolio and about 94% for the unconstrained portfolio. These numbers are based on our assumption of a forecast autocorrelation of 0.25. The turnover for the unconstrained portfolio is consistent with the results in Chapter 8. As we can see, the turnovers for the long-only portfolios are much lower. It is easy to understand that range constraints have a dampening effect on portfolio turnover, because they prohibit portfolios from adjusting fully to changes in forecasts, which is why they have a negative impact on investment performance (Qian et al. 2004). What is startling is that Table 11.3 shows that turnover is a linear function of leverage. The ratio of turnover to total long is about 0.64 for all portfolios.

To calculate the net average alpha, we assume that the spread between the long financing and the short rebate is 1%, and the transaction costs are 1% is for 100% turnover. These rates are reasonable and conservative estimates. In practice, the financing and rebate spread is subject to negotiation with prime brokers, and transaction costs depend on many factors such as commissions, bid/ask spreads, and market impact. Using the net average alpha, we then calculate the net IR. For the long-only portfolio, the IR drops from 1.59 to 1.38, a decrease of 0.21. For the unconstrained portfolio, the IR drops from 2.24 to 1.77, a much larger decrease of 0.47 due to the higher leverage cost and higher transaction costs.

Lastly, we will compute both theoretical and net *IR decay*, defined as the ratio of the IR of the constrained portfolios to that of the unconstrained portfolio. For instance, the long-only portfolio's theoretical IR is 71% of the unconstrained IR, but its net IR is 78% of the unconstrained net IR. Portfolio (column 6), with an average of 133% long, achieves about 95% of the unconstrained net IR. The last row of Table 11.3 shows the transfer coefficient (Clarke et al. 2002), defined as the correlation between the active weights in the constrained portfolios and the forecasts. In this case, the transfer coefficients are close to the theoretical IR decay but differ from the net IR decay.

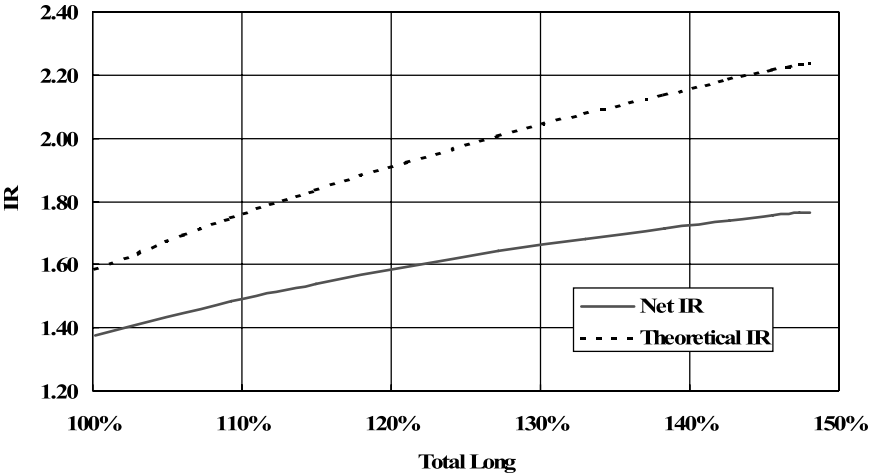


FIGURE 11.11. The theoretical and net IR as shown for Table11.3. (From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 33, No. 2, 1–9, Winter 2007. With permission.)

Figure 11.11 displays both the theoretical IR and net IR as a function of total long portfolio positions. We note two features of this graph. First, the rate of increase in IR with a loosening of short constraint is higher in terms of theoretical IR than in terms of net IR. This is due to the higher leverage and transaction costs associated with less constrained portfolios. Second, both curves are not straight lines. The marginal increase in IR seems to be the strongest for long-only portfolios, and it diminishes as the short constraints are relaxed further.

11.4.3 Risk Allocation of Long-Only and Long-Short Portfolios

One of the reasons for the low IR of the long-only portfolios is that they have inferior allocation of active risk. If a signal has uniform predictive power across stocks of all sizes, then the optimal allocation of active risk should be the same across the size spectrum. However, this is not the case for the long-only portfolios, because the constraint forces more active risk into stocks with large benchmark weights. Figure 11.12 shows the contribution to the active risk of 3% from 5 quintiles of 500 stocks in portfolios with different constraints. The long-only portfolio gets 45% of risk from the largest quintile, 17% in the second largest quintile, whereas the remaining 3 quintiles each contribute roughly 13%. As we loosen the short constraint,

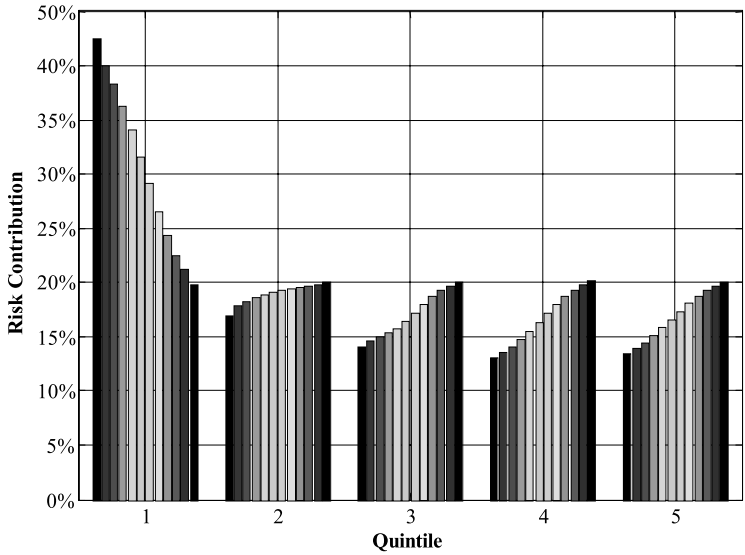


FIGURE 11.12. Risk contributions from quintiles of stocks. The active risk is 3%. There are 500 stocks and each quintile has 100 stocks: quintile 1 has the top 100 stocks of the largest weights, whereas quintile 5 has the bottom 100 stocks of the smallest weights. In each quintile, there are 11 portfolios (from left to right) ranging from the long-only portfolio to the unconstrained constrained. (From Sorensen, E.H., Hua, R., and Qian, E., *Journal of Portfolio Management*, Vol. 33, No. 2, 1–9, Winter 2007. With permission.)

the contribution from the 1st quintile decreases, whereas the rest contribute more, until we reach the unconstrained portfolio where all quintiles contribute the equal and optimal amount — 20% to the active risk.

11.4.4 Simulation Results: Stochastic IC

One of the underlying assumptions for the simulation in the previous section is the constancy of the IC. This assumption, however, is often violated in practice. As shown by Qian and Hua (2004), active investment strategies bring additional risk, which is not captured by generic risk models, and as a result the realized or *ex post* tracking error often exceeds the target or *ex ante* tracking error. This additional risk, referred to as *strategy risk*, can be represented by the intertemporal variation of IC, and the realized tracking error is then a function of the standard deviation of the IC that consists of both the intertemporal variation and the sampling error. The IR of an active investment strategy is then given by the ratio of average IC to the standard deviation of IC, i.e.,

$$IR = \frac{\overline{IC}}{\text{std}\left(IC\right)}.$$

For example, if the intertemporal variation of IC is 0.02, then the standard deviation of IC is

$$\text{std}\left(IC\right) = \sqrt{0.02^2 + \frac{1}{N}} = \sqrt{0.02^2 + \frac{1}{500}} = 0.049 .$$

The IR of unconstrained portfolios with the additional strategy risk is then $IR = 0.1/0.049 = 2.04$, compared to the previous value of 2.24 when the IC was constant.

What is the information ratio of long-only and constrained long-short portfolios, if the IC is stochastic? Table 11.4 shows the simulation results that take into account the additional intertemporal variation of IC, in this case, at 0.02. First, notice the unconstrained portfolio (column 11) has a realized tracking error of 3.28%, even though the target is 3%, due to the additional strategy risk and the theoretical IR is 2.04, as indicated earlier. Second, we note that the realized tracking error for the long-only portfolio is 3.08%, not too different from the target. As a result, its IR is 1.52, only slightly lower than 1.59 in the previous case; and as we relax the no-short constraint, the realized tracking error increases. These results indicate that more stringent range constraints have the potential benefit of controlling *ex post* tracking error when there is additional strategy risk. In other words, relaxing long-only constraints could potentially lead to higher *ex post* tracking error, and portfolio managers must pay extra attention to risk management.

The other characteristics of the portfolios, such as total long and turnover, stay the same, so additional costs remain unchanged. However, the net IR is lower in Table 11.4 than in Table 11.3 due to the higher realized tracking error. Here, the net IR goes from 1.31 for the long-only portfolio to 1.61 for the long-short portfolio.

Table 11.4 also indicates that the transfer coefficient is no longer a reliable gauge of IR decay, even for the theoretical IR. For instance, the long-only portfolio has a transfer coefficient of 0.70, but the theoretical IR decay is slower at 0.74 and the net IR is 0.82. When strategy risk grows, we find that the difference between the transfer coefficient and IR decay grows as well.

TABLE 11.4 Simulation Results for Long-Only Portfolios, Constrained Long-Short Portfolios, and Unconstrained Long-Short Portfolios

	Long-Only										**Unconstrained**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Avg alpha	4.68%	5.18%	5.45%	5.71%	5.95%	6.17%	6.35%	6.49%	6.60%	6.66%	6.70%
Std alpha	3.08%	3.12%	3.15%	3.17%	3.19%	3.21%	3.22%	3.24%	3.26%	3.27%	3.28%
Theoretical IR	1.52	1.66	1.73	1.80	1.87	1.92	1.97	2.00	2.03	2.04	2.04
Total long	100%	109%	115%	121%	127%	133%	138%	143%	146%	147%	148%
Turnover	64%	71%	75%	79%	83%	87%	89%	91%	93%	94%	94%
Leverage cost	0.00%	0.09%	0.15%	0.21%	0.27%	0.33%	0.38%	0.43%	0.46%	0.47%	0.48%
Transaction cost	0.64%	0.71%	0.75%	0.79%	0.83%	0.87%	0.89%	0.91%	0.93%	0.94%	0.94%
Net avg alpha	4.04%	4.38%	4.54%	4.70%	4.85%	4.98%	5.08%	5.15%	5.21%	5.25%	5.28%
Net IR	1.31	1.40	1.44	1.48	1.52	1.55	1.58	1.59	1.60	1.61	1.61
Theoretical IR decay	0.74	0.81	0.84	0.88	0.91	0.94	0.96	0.98	0.99	0.99	1.00
Net IR decay	0.82	0.87	0.90	0.92	0.94	0.96	0.98	0.99	0.99	1.00	1.00
Transfer coefficient	0.70	0.78	0.82	0.85	0.89	0.92	0.95	0.97	0.98	0.99	1.00

Note: The intertemporal variation of IC is 0.02, and all other assumptions are the same as in Table 11.3.

PROBLEMS

11.1 Calculate the return decomposition for Example 11.1.

11.2 (Variance decomposition) Cross-sectional return variance is given by

$$\sigma_R^2 = \sum_{i=1}^{N} b_i \left(R_i - \bar{R}\right)^2 ,$$

where b_i could be the benchmark weight for cap-weighted variance or $b_i = 1/N$ for equally-weighted variance.

(a) Prove that the variance can be decomposed as

$$\sigma_R^2 = \sum_{s=1}^{S} \sum_{i=1}^{N_s} b_i \left(R_{si} - \bar{R}_s\right)^2 + \sum_{s=1}^{S} B_s \left(\bar{R}_s - \bar{R}\right)^2 , \tag{11.21}$$

where $B_s = \sum b_i$ for stocks in the sector s, i.e., the sector weight.

(b) Interpret the decomposition as investment opportunities for stock selection and sector bets in terms of their relative magnitude.

11.3 Assume the benchmark weight of a stock is b_i, and its active weight of a stock is given by

$$w_i = \frac{\sigma_{\text{target}} F_i}{\sqrt{N}\sigma_i} .$$

Instead of the normal distribution, assume the factor F_i is uniformly distributed with zero mean and standard deviation one. This uniform distribution describes factors that are percentile ranking instead of normalized z-scores.

(a) Find the range of F_i and therefore the range of w_i.

(b) Find the probability that the total position $w_i + b_i$ is net short.

(c) Find the average long/short ratio for the stock.

11.4 Suppose the financing cost is the federal funds rate plus 50 bps and the short rebate is the federal funds rate minus 75 bps. What is the leverage cost for (a) a constrained 130/30 portfolio and (b) a market-neutral portfolio with 100 long and 100 short?

11.5 If the active weights are given by the Kuhn–Tucker condition, calculate the transfer coefficient.

11.6 A forecast model has an average IC of 0.1 for a universe of 500 stocks. Suppose the IC has no intertemporal variation so that the fundamental law of active management holds.

(a) What is the model's IR?

(b) Suppose the model is uniformly effective across all 500 stocks. What is the model's IR when applied to each quintile?

(c) What is the optimal allocation of active risk across the five quintiles if excess returns from five quintiles are uncorrelated?

APPENDIX
A11.1 MEAN–VARIANCE OPTIMIZATION WITH RANGE CONSTRAINTS

Given a forecast vector **f**, we maximize the following objective function to obtain portfolio weights **w**

$$\mathbf{f}' \cdot \mathbf{w} - \frac{1}{2}\lambda \cdot (\mathbf{w}' \cdot \mathbf{\Sigma} \cdot \mathbf{w}) \ . \tag{11.22}$$

In addition to the dollar neutral and market neutral constraints: $\mathbf{w}' \cdot \mathbf{i} = 0$, and $\mathbf{w}' \cdot \mathbf{B} = 0$, we also have range constraints on individual stocks: $\mathbf{l} \leq \mathbf{w} \leq \mathbf{u}$, where **l** and **u** are vectors of lower and upper bound for all stocks. As the range constraints are inequality constraints, there is no analytical solution for the optimization problem. However, a numerical solution can be found through Kuhn–Tucker conditions. For details, please refer to McCormick (1983).

A11.1.1 Kuhn–Tucker Conditions

Kuhn–Tucker conditions are for general optimization problems with inequality constraints. We first present the conditions for a general problem and then specify them for the mean–variance optimization with range constraints.

Suppose the problem is to maximize $p(\mathbf{w})$ subject to $g_j(\mathbf{w}) \le 0$ for $j = 1, \cdots, m$, then define the Lagrangian function L by

$$L(\mathbf{w}) = p(\mathbf{w}) - \sum_{j=1}^{m} l_j g_j(\mathbf{w}). \tag{11.23}$$

The Kuhn–Tucker conditions are

$$\frac{\partial L(\mathbf{w})}{\partial w_i} = \frac{\partial p(\mathbf{w})}{\partial w_i} - \sum_{j=1}^{m} l_j \frac{\partial g_j(\mathbf{w})}{\partial w_i} = 0, \text{ for } i = 1, \cdots, N, \tag{11.24}$$

and

$$g_j(\mathbf{w}) \le 0,\ l_j \ge 0, \text{ and } l_j g_j(\mathbf{w}) = 0, \text{ for } j = 1, \cdots m. \tag{11.25}$$

We note that condition (11.24) is the same for equality constraints. However, condition (11.25) is different for inequality constraints, and states that (1) the inequality constraints must be satisfied, of course; (2) the Lagrangian multipliers must be nonnegative; and (3) either the Lagrangian multiplier is 0, or the constraints are binding.

A11.1.2 Kuhn–Tucker Conditions for Mean–Variance Optimization with Range Constraints

When the range of weight for a stock is constrained by $L_i \le w_i \le U_i$, we can represent the constraint with two inequality constraints: $w_i - U_i \le 0$, and $L_i - w_i \le 0$ in the form of $g(\mathbf{w}) \le 0$.

For a portfolio of N stocks, we could have a maximum of $2N$ inequality constraints:

$$w_i - U_i \le 0, \text{ and } L_i - w_i \le 0, \text{ for } i = 1, \cdots, N. \tag{11.26}$$

The objective function (11.22) also needs to be modified with the introduction of range constraints. Previously, the risk-aversion parameter was a free parameter used to achieve the targeted tracking error, because with dollar neutral and market neutral constraints the optimal weights are

scalable. With inequality constraints, the optimal weights are no longer scalable. Hence, we need to set targeted tracking error as an additional constraint. The optimization problem becomes

Maximize: $\mathbf{f}' \cdot \mathbf{w}$

Subject to:

$$\mathbf{w}' \cdot \mathbf{\Sigma} \cdot \mathbf{w} = \sigma^2_{\text{target}} \tag{11.27}$$

$$\mathbf{w}' \cdot \mathbf{i} = 0, \text{ and } \mathbf{w}' \cdot \mathbf{B} = 0$$

$$w_i - U_i \le 0, \text{ and } L_i - w_i \le 0, \text{ for } i = 1, \cdots, N$$

The Lagrangian function for the problem is then

$$\begin{aligned} L(\mathbf{w}) = \mathbf{f}' \cdot \mathbf{w} - \lambda\left(\mathbf{w}' \cdot \mathbf{\Sigma} \cdot \mathbf{w} - \sigma^2_{\text{target}}\right) - l_0\left(\mathbf{w}' \cdot \mathbf{i}\right) - \sum_{i=1}^{K} l_i\left(\mathbf{w}' \cdot \mathbf{b}_i\right) \\ - \sum_{j=1}^{N}\left[\tilde{l}_{j1}\left(w_j - U_j\right) + \tilde{l}_{j2}\left(L_j - w_j\right)\right] \end{aligned} \tag{11.28}$$

Now λ denotes the Lagrangian multiplier for the tracking error target constraint, l_0 is the Lagrangian multiplier for the dollar neutral constraint, l_i, $i = 1, \cdots, K$ are the Lagrangian multipliers for the K risk factors, and $\tilde{l}_{j1}$ and $\tilde{l}_{j2}$, $j = 1, \cdots, N$ are the Lagrangian multipliers for the range constraints on N stocks.

The Kuhn–Tucker condition for (11.28) is

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{f} - 2\lambda \mathbf{\Sigma} \cdot \mathbf{w} - l_0 \mathbf{i} - \sum_{i=1}^{K} l_i \mathbf{b}_i - \left(\tilde{\mathbf{l}}_1 - \tilde{\mathbf{l}}_2\right) = 0\,, \tag{11.29}$$

where $\tilde{\mathbf{l}}_1 = \left(\tilde{l}_{11}, \cdots, \tilde{l}_{1N}\right)'$ and $\tilde{\mathbf{l}}_2 = \left(\tilde{l}_{21}, \cdots, \tilde{l}_{2N}\right)'$ are vectors of Lagrangian multipliers. The equality constraints must be satisfied, i.e.,

$$\mathbf{w}' \cdot \mathbf{\Sigma} \cdot \mathbf{w} = \sigma^2_{\text{target}}, \ \mathbf{w}' \cdot \mathbf{i} = 0, \text{ and } \mathbf{w}' \cdot \mathbf{B} = 0\,.$$

In addition, for the range constraints, we have

$$\begin{aligned} &\tilde{l}_{j1} \geq 0,\ w_j - U_j \leq 0, \text{ and } \tilde{l}_{j1}\left(w_j - U_j\right) = 0 \\ &\tilde{l}_{j2} \geq 0,\ L_j - w_j \leq 0, \text{ and } \tilde{l}_{j2}\left(L_j - w_j\right) = 0 \end{aligned} \tag{11.30}$$

Equation 11.29 can be solved as

$$\mathbf{w} = \frac{1}{2\lambda}\mathbf{\Sigma}^{-1}\left(\mathbf{f} - l_0\mathbf{i} - \sum_{i=1}^{K} l_i\mathbf{b}_i - \tilde{\mathbf{l}}_1 + \tilde{\mathbf{l}}_2\right) = \frac{1}{2\lambda}\mathbf{\Sigma}^{-1}\mathbf{f}_{\text{adj}}. \tag{11.31}$$

Hence, the optimal weights must be of the form of Equation 11.31, which resembles the optimal weights of unconstrained portfolios with forecasts adjusted for various constraints and then scaled by λ to give the targeted tracking error.

When the range constraint is nonbonding, i.e., $L_i < w_i < U_i$, we have $\tilde{l}_{j1} = 0$ and $\tilde{l}_{j2} = 0$ according to the condition (11.30). If $w_i = U_i$, i.e., the weight is at the upper bound, then $\tilde{l}_{j1} \geq 0$ and $\tilde{l}_{j2} = 0$. Similarly, if $w_i = L_i$, i.e., the weight is at the lower bound, then $\tilde{l}_{j1} = 0$ and $\tilde{l}_{j2} \geq 0$. Therefore, between $\tilde{l}_{j1}$ and $\tilde{l}_{j2}$ only one of them can be nonzero.

When the covariance matrix is that of a multifactor model, i.e., $\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_I\mathbf{B}' + \mathbf{S}$, Equation 11.31 can be simplified to

$$\begin{aligned} &\mathbf{w} = \frac{1}{2\lambda}\mathbf{S}^{-1}\mathbf{f}_{\text{adj}}, \text{ or} \\ &w_i = \frac{f_i - l_0 - l_1 b_{1i} - \cdots l_K b_{Ki} - \tilde{l}_{1i} + \tilde{l}_{2i}}{2\lambda\sigma_i^2} \end{aligned} \tag{11.32}$$

REFERENCES

Clarke, R., de Silva, H., and Thorley, S., Portfolio constraints and the fundamental law of active management, *Financial Analysts Journal*, Vol. 58, No. 5, 48–66, September–October 2002.

Clarke, R., de Silva, H., and Thorley, S., Toward more information efficient portfolios, *Journal of Portfolio Management,* Vol. 31, No. 1, 54–63, Fall 2004.

Grinold, R.C. and Kahn, R.N., The efficiency gains of long-sort investing, *Financial Analysts Journal*, Vol. 56, No. 6, 40–53, November–December 2000.

Jacobs, B.I. and Levy, K.N., Enhanced active equity strategies, *Journal of Portfolio Management*, Vol. 32, No. 2, 45–55, Spring 2006.

McCormick, G.P., *Nonlinear Programming: Theory, Algorithms, and Applications*, Wiley, 1983, New York.

Qian, E. and Hua, R., Active risk and information ratio, *Journal of Investment Management*, Vol. 2., No. 3, 20–34 , 2004.

Qian, E., Hua, R., and Tilney, J., Portfolio turnover of quantitatively managed portfolios, *Proceeding of the 2nd IASTED International Conference, Financial Engineering and Applications*, Cambridge, MA, 2004.

Sorensen, E., Hua, R., and Qian, E., Aspect of constrained long-short portfolios, *Journal of Portfolio Management*, Vol. 33, No. 2, 12–22, Winter 2007.

ENDNOTES

1. A simple example suffices to illustrate this point. When buying a stock at $10, all one can lose is $10 if the stock's price goes all the way down to zero in the event of bankruptcy. Shorting a stock at $10 with an initial margin of say $10, if the stock price goes up to $15, one loses $5, i.e., 50% of the initial investment. If the stock price goes to $20, one loses the entire $10 investment, and if the stock price goes above $20, the loss would exceed the initial investment and additional cash is needed.
2. Jacobs and Levy (2006) depicts an alternative structure set up by prime brokers, based *conceptually* on financing additional long positions with shorting. While the structure has certain tax advantages, it bears the same leverage cost.

CHAPTER 12

Transaction Costs and Portfolio Implementation

TRADING STOCKS INCURS TRANSACTION COSTS. So far, we have not dealt explicitly with the impact of transaction costs on equity portfolio management, with the exception of Chapter 8, where we built optimal alpha models under an aggregate portfolio turnover constraint. However, portfolio turnover is just a proxy for transaction costs, which are often stock specific; trading illiquid stocks would have higher costs than trading liquid stocks even if turnover is the same. Therefore, to fully understand the impact of transaction costs on portfolio management, it is important to incorporate stock-level detail in the analysis.

In this chapter, we study two areas of portfolio management that would benefit from the inclusion of transaction costs. One is portfolio construction or portfolio optimization and the other is portfolio implementation. The processes of portfolio optimization with transaction costs and portfolio implementation should be integrated. Simply put, we cannot know the exact transaction costs without knowing exactly how the portfolio would be implemented. In other words, the transaction costs depend on changes of portfolio (in shares or in portfolio weights), as well as the way the portfolio will be traded. If we denote changes in portfolio by the weight differences, $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_0$, where $\mathbf{w}_0$ is the initial weight vector and $\mathbf{w}$ is the optimal weight vector, the transaction costs should be a function $c(\Delta\mathbf{w})$, in which the function form $c(.)$ would be determined by how the trades are executed in addition to the liquidity attributes of stocks. After the function $c(.)$ is determined, the transaction cost $c(\Delta\mathbf{w})$ is incorporated into the portfolio optimization process as another term in the objective function.

In practice, the two processes are often studied separately. As a result, some simple transaction cost functions are used in the portfolio optimization. In this book, we follow this research direction and leave the integrated approach to future research.

12.1 COMPONENTS OF TRANSACTION COSTS

To determine a reasonable form for function $c(.)$, we first consider the different components of transaction costs. Broadly speaking, there are two kinds of transaction costs: fixed costs and variable costs. The fixed costs are related to trade commissions and bid/ask spreads. There could be additional service fees but they are often included in the commission. Trade commissions are often quoted at some cost per share whether it is a buy or a sell order. For instance, it could be 2¢ per share. In this case, the cost is a linear function of the traded amount or the number of trade tickets.

The bid/ask spread is another form of fixed cost because it results in investors getting paid less if they were to sell a stock, while paying more if they were to buy a stock. For instance, the spread might be \$10.00/\$10.10, meaning a seller receives \$10.00 per share but a buyer has to pay \$10.10, an extra of 10¢ per share. If nothing changes, a round trip of trading would result in a loss of 10¢ per share for the investor. For this reason, we could model the costs associated with the bid/ask spread as half of the spread between the two prices. The average of the bid and ask is called the *mid-quote*, and hence the cost is the difference between either bid or ask and the mid-quote. Because the cost is on a per-share basis, it is also a linear function of the traded amount.

Hence, we can model the fixed cost as a constant vector times the absolute value of the portfolio weight change,

$$c(\Delta\mathbf{w}) = \boldsymbol{\theta}' \cdot |\Delta\mathbf{w}| = \theta_1 |\Delta w_1| + \theta_2 |\Delta w_2| + \cdots + \theta_N |\Delta w_N|. \tag{12.1}$$

- The function (12.1) is always positive with the absolute value function if the coefficients are positive. Also, the proportional constant is different for different stocks. This is a result of different commissions, or different bid/ask spreads for different stocks, or both.

Example 12.1

Suppose a stock is originally 10% of a portfolio and we want to reduce it to 5%. The size of the portfolio is \$100 million. This results in a trade of \$5 million worth of stock. Suppose the share price is \$50. We thus need to sell 100,000 shares. Let us say assume a bid/ask spread of

10¢ and a commission of 5¢ per share. The transaction costs would be $c = (0.05 + 0.05) \cdot 100{,}000 = \$10{,}000$, or a loss of 0.01%, or 1 basis point, on the total portfolio. In terms of Equation 12.1, the coefficient equals $\theta = 0.002$, which is cost per share at 10¢ divided by the share price at $50. It can be proved that in terms of percentage loss to the total portfolio, the coefficient θ equals transaction cost per share divided by the share price (Problem 12.1).

The other component of transaction costs is variable costs, which include market impact and opportunity costs. Market impact refers to the price change due to investors' trading and it occurs when trade size exceeds the quote depth currently available. For instance, we would like to sell 100,000 shares of stock in Example 12.1. However, the bid at $50 is only for 50,000 shares. If we want to sell the additional 50,000 rather quickly, the price is most likely to drop due to the resulting supply and demand imbalance and we might have to accept that lower price to fill the order. The difference between the new price and the bid price prior to the sell order gives rise to the market impact component of total transaction costs.

Thus, the transaction costs associated with market impact are not linear. It is small when the trade size is small but it increases dramatically when the trade size becomes large. For a single stock, one possibility is to model it by a square function

$$c(\Delta w_i) = \psi_i (\Delta w_i)^2, \quad \psi_i \geq 0. \tag{12.2}$$

As we shall see shortly, the simplicity of (12.2) makes portfolio optimization easy.

Example 12.2

Continue with Example 12.1. Suppose the quote depth is only 50,000 shares at the selling price of $50 and we have to sell the remaining 50,000 shares at the price of $49.80. The total transaction cost is $c = \$0.05 \cdot 100{,}000 + \$0.05 \cdot 50{,}000 + \$0.25 \cdot 50{,}000 = \$20{,}000$, or twenty thousand dollars. This is equivalent to 20¢ per share, a loss of 0.02%, or 2 basis points, on the total portfolio. If we model the total cost using Equation 12.2, then the coefficient is given by

$$\psi_i = \frac{c}{(\Delta w_i)^2} = \frac{0.02\%}{(5\%)^2} = 0.08.$$

When trading multiple stocks, or a basket of stocks, the market impact on the different stocks can be correlated. Selling two highly correlated stocks would cause a greater market impact on both stocks than selling one stock while buying the other. We can model the transaction costs associated with market impact for a basket of stocks using

$$c(\Delta \mathbf{w}) = \Delta \mathbf{w}' \cdot \boldsymbol{\psi} \cdot \Delta \mathbf{w} \ . \tag{12.3}$$

To ensure that the transaction costs are always positive, the matrix $\boldsymbol{\psi}$ must be positive definite.

Another type of variable cost is the opportunity cost, which is associated with the return impact of trades not getting executed. For instance, investors often use limit orders instead of market orders to buy stocks, in order to reduce market impact. However, if the stock price fails to reach the limit order price, the trade would not be executed. If the stock price continues to rise, then the investor loses the opportunity to participate in the gain on the stock. Compared to the other components of transaction costs, the opportunity cost is the hardest to estimate. We shall not consider it in the book.

12.2 OPTIMAL PORTFOLIOS WITH TRANSACTION COSTS: SINGLE ASSET

The problem of incorporating transaction costs into the formation of optimal portfolios is often not analytically tractable. We shall discuss numerical methods to solve it later in the chapter. However, for a single stock or asset, it is possible to analyze and solve the problem analytically, and we can gain valuable insights from it.

12.2.1 Single Asset with Quadratic Costs

Mean–variance optimization with the addition of quadratic transaction costs is relatively easy to treat so we shall consider it first. The transaction costs are given in the form of (12.2). The optimization problem in this case can be written as

$$\text{maximize } U(w) = f \cdot w - \frac{1}{2}\lambda\sigma^2 w^2 - \psi\left(w - w_0\right)^2 . \tag{12.4}$$

The unknown is the optimal weight w, and the parameters are: f, the return forecast; σ, the risk of the asset; λ, the risk-aversion parameter;

w_0, the initial weight; and ψ, the transaction cost coefficient. We can think of (12.4) as the allocation decision between a single risky asset and cash. The coefficient ψ in this case measures market impact of the cost for a 100% turnover. As opposed to the problem with linear transaction cost, the utility function in (12.4) is well behaved. The cost term is analogous to a variance term, relative to the current position. Taking the derivative with respect to w gives rise to

$$U'(w) = f - \lambda\sigma^2 w - 2\psi(w - w_0). \tag{12.5}$$

The optimal weight is given by $U'(w) = 0$, and we have

$$w^* = \frac{f + 2\psi w_0}{\lambda\sigma^2 + 2\psi}. \tag{12.6}$$

The optimal weight (12.6) is a function of the transaction cost coefficient ψ. When $\psi = 0$, then

$$w^* = \tilde{w} \triangleq \frac{f}{\lambda\sigma^2}. \tag{12.7}$$

The weight $\tilde{w}$ is optimal when there are no transaction costs. At the other extreme, when ψ is very large compared to both the forecast and the risk term, then $w^* \to w_0$ slowly.

Let $\Delta w^* = w^* - w_0$ be the optimal trade with transaction costs and $\Delta\tilde{w} = \tilde{w} - w_0$ be the optimal trade without transaction costs. Equation 12.8 shows that Δw^* is a fraction of $\Delta\tilde{w}$, and the scaling constant is the ratio of the transaction coefficient to the risk coefficient in the utility function (12.4).

$$w^* - w_0 = \frac{f + 2\psi w_0}{\lambda\sigma^2 + 2\psi} - w_0 = \frac{f - \lambda\sigma^2 w_0}{\lambda\sigma^2 + 2\psi} = \frac{\tilde{w} - w_0}{1 + \left(2\psi / \lambda\sigma^2\right)}. \tag{12.8}$$

Example 12.3

Suppose that a single asset has a volatility σ is 15%, and we have a return forecast of 15%. The risk-aversion parameter is 10, and the current position is 50%. We can calculate the optimal weight with no transaction costs at

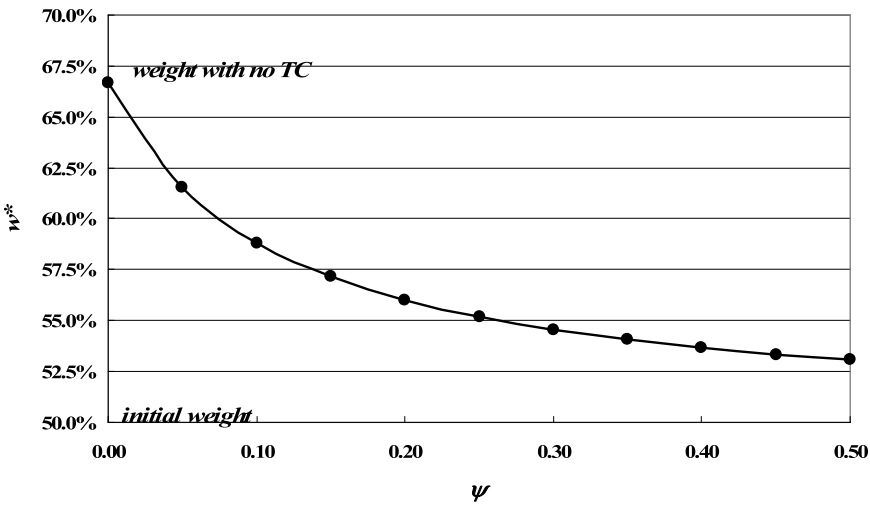


FIGURE 12.1. Optimal weight of a single asset with quadratic transaction costs. The initial weight is 50%, and the optimal weight with no transaction costs is 66.7%. Note that the optimal weight is always above the initial weight.

$$\tilde{w} = \frac{f}{\lambda\sigma^2} = \frac{0.15}{10(0.15)^2} = 66.7\% .$$

Therefore, we should be buying more. However, the amount of buying will be tempered by the transaction costs. Suppose $\psi = 0.1$, which corresponds to transaction costs of 10% on 100% turnover. We then have

$$w^* = \frac{f + 2\psi w_0}{\lambda\sigma^2 + 2\psi} = \frac{0.15 + 2(0.1)(0.5)}{10(0.15)^2 + 2(0.1)} = 58.5\% .$$

Figure 12.1 plots the optimal weights for value of ψ from 0 to 0.5. As we can see, the optimal weight declines rather quickly at first, and then the rate of decline slows. When $\psi = 0.5$, the optimal weight is about 53%, a trade of 3%. Note the following remark:

- With quadratic trading costs, there will always be some trading no matter how large ψ is, because the value of the quadratic function of transaction costs will be small when the weight is close to the initial

weight. This makes some sense, because the market impact only becomes important when the trade size exceeds the quote depth.

12.2.2 Single Asset with Linear Costs

We now consider mean–variance optimization with the addition of transaction costs given in the form of (12.1). The optimization problem in this case can be written as

$$\text{maximize } U(w) = f \cdot w - \frac{1}{2}\lambda\sigma^2 w^2 - \theta|w - w_0|. \tag{12.9}$$

θ is the transaction cost coefficient, measuring the cost of 100% turnover. Solving Problem 12.9 poses certain analytical challenges because the absolute value function is not differentiable at the origin.

When there are no transaction costs, i.e., $\theta = 0$, however, the optimal weight is $\tilde{w}$, given by (12.7). When $\theta > 0$, the problem can be formulated in terms of weight change: $\Delta w = w - w_0$. Using $w = w_0 + \Delta w$, we can rewrite the utility function as

$$\begin{aligned} U(\Delta w) &= f \cdot (\Delta w + w_0) - \frac{1}{2}\lambda\sigma^2 (\Delta w + w_0)^2 - \theta|\Delta w| \\ &= U(w_0) + \left[\lambda\sigma^2(\tilde{w} - w_0)\Delta w - \theta|\Delta w| - \frac{1}{2}\lambda\sigma^2(\Delta w)^2\right] \end{aligned} \tag{12.10}$$

The total utility is a sum of the current utility, a constant, given by

$$U(w_0) = fw_0 - \frac{1}{2}\lambda\sigma^2 w_0^2,$$

and the change in utility caused by the change in weight. The weight $\tilde{w}$ is also a constant given by Equation 12.7.

The change in utility is then

$$\Delta U = U(\Delta w) - U(w_0) = \lambda\sigma^2 \Delta\tilde{w}\Delta w - \theta|\Delta w| - \frac{1}{2}\lambda\sigma^2(\Delta w)^2, \tag{12.11}$$

$$\text{with } \Delta\tilde{w} = \tilde{w} - w_0$$

The optimal weight change must maximize the change in utility, which is zero when $\Delta w = 0$. In other words, at a minimum, we can maintain the current utility with no trading. To find the maximum, we now consider three cases.

The first case is when $\tilde{w} = w_0$, i.e., when the optimal weight disregarding the transaction costs is equal to the initial weight. It is obvious in this case we should not trade at all. Mathematically, $\Delta w = 0$ is the optimal solution for utility (12.10), because any trading would cause the utility to go down.

When $\tilde{w} \neq w_0$, the initial position is not optimal, at least if there were no transaction costs. There is a possibility that we can increase the utility of (12.10) by trading. Because both the second and the third terms, associated with transaction costs and variance, are negative whenever there is trading (either buy or sell), the trading must at least make the first term positive. This implies Δw must be of the same sign as $\Delta\tilde{w} = \tilde{w} - w_0$. Therefore, in the second case, we consider $\tilde{w} > w_0$, i.e., the optimal weight in absence of transaction costs is greater than the initial weight, indicating buy. As argued, we should look for solution $\Delta w \geq 0$. In other words, we should look to buy to increase the utility.

If $\Delta w \geq 0$, we have $|\Delta w| = \Delta w$. The utility function becomes differentiable with the derivative

$$U'(\Delta w) = \lambda\sigma^2 \Delta\tilde{w} - \theta - \lambda\sigma^2 (\Delta w). \tag{12.12}$$

Setting $U'(\Delta w) = 0$ yields

$$\Delta w^* = w^* - w_0 = \Delta\tilde{w} - \frac{\theta}{\lambda\sigma^2} = \Delta\tilde{w} - w_c. \tag{12.13}$$

We have defined

$$w_c = \frac{\theta}{\lambda\sigma^2}, \tag{12.14}$$

which is an optimal weight associated with the transaction cost as a negative "alpha," or cost weight.

Equation 12.13 is the optimal weight *if* Δw^* is greater than or equal to zero, or when

$$\Delta\tilde{w} \geq w_c. \tag{12.15}$$

This condition implies that we would only buy when the costless buying, i.e., $\Delta\tilde{w}$, exceeds the cost weight w_c. On the other hand, *when* Δw^* is less than zero, the costless buying does not clear the hurdle of cost weight, then (12.13) is certainly not the optimal weight, because it leads to a reduction in utility (12.10). Here, we have a situation in which we would buy if there were no transaction costs, but would not if the transaction cost were factored in. The best course to follow is therefore to stay put: no trade, i.e., $\Delta w^* = 0$.

The analysis applies equally to the last case, in which $\tilde{w} < w_0$. We leave it as an exercise. To summarize the results, we have the optimal trading

$$\Delta w^* = \begin{cases} \Delta\tilde{w} - w_c, & \text{when } \Delta\tilde{w} > w_c \\ 0, & \text{when } |\Delta\tilde{w}| \leq w_c \\ \Delta\tilde{w} + w_c, & \text{when } \Delta\tilde{w} < w_c \end{cases} \tag{12.16}$$

Figure 12.2 shows the results. Both buys and sells are reduced by the amount, w_c, and there is a zone of inaction when the costless trading is less than the cost weight.

Alternatively, we can rewrite the optimal weight as

$$w^* = \frac{f - \theta}{\lambda\sigma^2} \geq w_0. \tag{12.17}$$

Note that the optimal weight w^* is equivalent to an optimal solution in the case of no transaction costs, but with an adjusted forecast of $f - \theta$. Therefore, we would buy only if the forecast is high enough to offset the transaction costs, such that the optimal weight with the cost-adjusted forecast is still greater than the current weight. Note the following remark:

- The insight from the analysis is that we buy only if the cost-adjusted forecast, $f - \theta$, still leads to a buy decision. In other words, we trim the forecast of a possible buy by the transaction cost, and the adjusted optimal weight must still be higher than the current weight in order for us to trade. In the same vein, we sell only if the cost-adjusted forecast, $f + \theta$, in the case of a sell (see Problem 12.2), still leads to

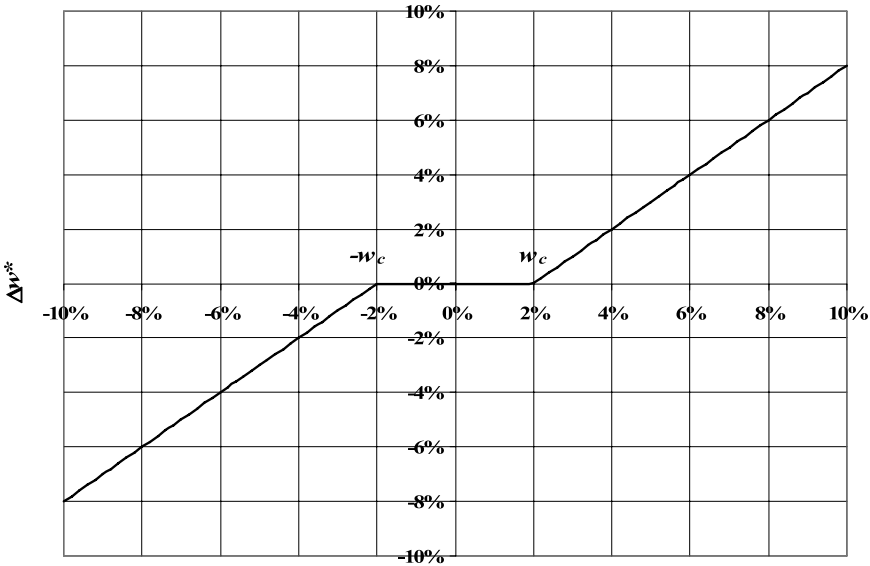


FIGURE 12.2. Relationship among the optimal trading Δw^*, the costless trading $\Delta \tilde{w}$, and the cost weight w_c when transaction cost is a linear function with respect to the size of a trade.

a sell. In other words, we raise the forecast of a possible sell by the transaction cost and the adjusted optimal weight must still be lower than the current weight in order for us to sell. If these conditions are not met, then there is no trade.

Example 12.4

We use the same parameters as in Example 12.3: a single asset with volatility σ at 15%, and return forecast of 15%. The risk-aversion parameter is 10, and the current position is 50%. The optimal weight with no transaction costs is 66.7%. Therefore, we should be buying more. However, the amount of buying will be tampered by the transaction costs. Suppose $\theta = 0.01$, then the optimal weight is

$$w^* = \frac{f - \theta}{\lambda \sigma^2} = \frac{0.15 - 0.01}{10(0.15)^2} = 62.2\% .$$

The weight is still above the current weight, by 10.2%. If the transaction cost is increased to $\theta = 0.02$, then the optimal weight decreases to 57.8%.

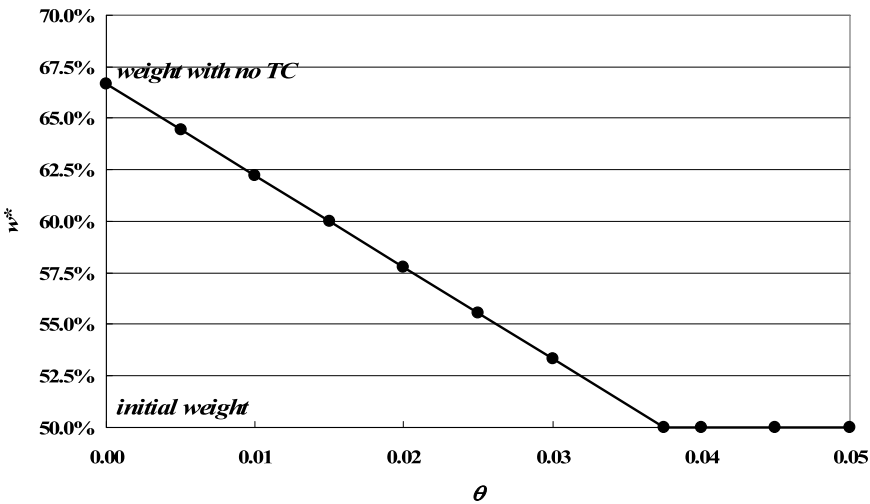


FIGURE 12.3. Optimal weight of a single asset with linear transaction costs. The initial weight is 50%, and the optimal weight with no transaction costs is 66.7%. There is no trading when the transaction costs goes beyond a critical value.

Therefore, we are buying less as the costs get higher. The critical value is $\theta = 0.0375$, at which the optimal weight becomes the current weight at 50%.

Figure 12.3 plots the optimal weights for values of θ from 0 to 0.05. As we can see, the optimal weight declines linearly and it reaches the initial weight when θ hits the critical value of 0.0375 and stays there.

12.3 OPTIMAL PORTFOLIOS WITH TRANSACTION COSTS: MULTIASSETS

Having solved the problem of the optimal weight for a single asset, we now analyze the problem for multiasset portfolios.

12.3.1 Multiasset with Quadratic Costs

With a multiasset portfolio, the quadratic transaction cost is given in the form of (12.3), in which $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_0$. The optimization problem in this case can be written as

$$\text{maximize } U(\mathbf{w}) = \mathbf{f}' \cdot \mathbf{w} - \frac{1}{2}\lambda \mathbf{w}' \boldsymbol{\Sigma} \mathbf{w} - (\Delta\mathbf{w})' \boldsymbol{\psi}(\Delta\mathbf{w}). \qquad (12.18)$$

Note that for an active portfolio vs. a benchmark, the weight vector is the active weights and for a market-neutral long/short portfolio the weight vector is the absolute weights. We have left out other constraints to isolate the impact of transaction costs.

The solution of (12.18) can be found analytically using the following equation:

$$\frac{\partial U}{\partial \mathbf{w}} = \mathbf{f} - \lambda \boldsymbol{\Sigma} \mathbf{w} - 2\boldsymbol{\Psi}\left(\mathbf{w} - \mathbf{w}_0\right) = \mathbf{0} . \tag{12.19}$$

We have

$$\mathbf{w}^* = \left(\lambda \boldsymbol{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left(\mathbf{f} + 2\boldsymbol{\psi}\mathbf{w}_0\right). \tag{12.20}$$

In (12.20), both $\boldsymbol{\Sigma}$ and $\boldsymbol{\psi}$ are square matrices and $\mathbf{f}$ is the forecast vector. Note that it reduces to (12.6) when both matrices are diagonal. In that case, we are simply optimizing uncorrelated individual assets.

12.3.2 Portfolio Dynamics

Equation 12.20 gives rise to a dynamic relationship of portfolio weights over time. Applying (12.20) iteratively, we have

$$\mathbf{w}_t = \left(\lambda \boldsymbol{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left(\mathbf{f}_t + 2\boldsymbol{\psi}\mathbf{w}_{t-1}\right) \tag{12.21}$$

and

$$\begin{aligned} \mathbf{w}_t &= \left(\lambda \boldsymbol{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\begin{bmatrix} \mathbf{f}_t + 2\boldsymbol{\psi}\left(\lambda \boldsymbol{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\mathbf{f}_{t-1} \\ \quad + \left(2\boldsymbol{\psi}\right)\left(\lambda \boldsymbol{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left(2\boldsymbol{\psi}\right)\mathbf{w}_{t-2} \end{bmatrix} \\ &= \left(\lambda \boldsymbol{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left[\mathbf{f}_t + \mathbf{A}\mathbf{f}_{t-1} + \mathbf{A}^2\mathbf{f}_{t-2} + \cdots + \mathbf{A}^\tau\mathbf{f}_{t-\tau} + \cdots\right] \end{aligned} \tag{12.22}$$

The matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = \left(\lambda \boldsymbol{\Sigma} + 2\boldsymbol{\psi}\right)^{-1} 2\boldsymbol{\psi} .$$

Based on this relationship, one can build a dynamic model of active portfolios over time, supplemented by a dynamic model of forecasts

$$\mathbf{f}_t = \mathbf{P}_1\mathbf{f}_{t-1} + \mathbf{P}_2\mathbf{f}_{t-2} + \cdots + \mathbf{P}_p\mathbf{f}_{t-p} + \boldsymbol{\varepsilon}_t \tag{12.23}$$

and lagged ICs

$$IC_{t-p,t} = \text{corr}\left(\mathbf{f}_{t-p}, \mathbf{r}_t\right). \tag{12.24}$$

Sneddon (2005) has shown that under simplified assumptions, one can derive the multiperiod information ratio (IR) in a semi-analytical framework that gives valuable insights regarding the combination of forecast signals. His results are consistent with our finding in Chapter 8 (see Grinold 2006 for additional analysis on this topic). For instance, he finds that when incorporating transaction costs, the multiple-period IR can be increased, compared to that of a single-period IR given by the fundamental law of active management, by overweighting the tortoise — signals with lower information coefficient (IC) but slow information decay — and underweighting the hare — signals with higher IC but fast information decay. It remains to be seen if his model can be extended to include more realistic factor and return structures.

12.3.3 Multiasset with Linear Costs: Mathematical Formulation

The linear transaction cost of a multiasset portfolio is given previously in (12.1). In terms of a vector of the transaction cost coefficients, $\boldsymbol{\theta}$, and the vector of absolute value of weight changes, $\left|\mathbf{w}-\mathbf{w}_0\right|$, the cost is $\boldsymbol{\theta}' \cdot \left|\mathbf{w}-\mathbf{w}_0\right| = \boldsymbol{\theta}' \cdot \left|\Delta\mathbf{w}\right|$. Thus, the mean–variance cost optimization is

$$\text{maximize } U\left(\mathbf{w}\right) = \mathbf{f}' \cdot \mathbf{w} - \frac{1}{2}\lambda \mathbf{w}' \boldsymbol{\Sigma} \mathbf{w} - \boldsymbol{\theta}' \cdot \left|\Delta\mathbf{w}\right|. \tag{12.25}$$

Unlike the single-asset case, the problem is not analytically tractable unless all assets are uncorrelated: when the covariance matrix is diagonal, because of the presence of the absolute-value function.

The problem can be solved numerically, however, in a number of ways. For example, one can approximate the absolute-value function by some smooth functions. In this chapter we shall present a method that reformulates the transaction cost term in term of two new variables, buys and sells, and solve the reformulated problem with standard quadratic programming.

We define two new vectors, buy vector $\mathbf{w}_B$ and sell vector $\mathbf{w}_S$. Then the new portfolio weights are a combination of the current weights, the buys and the sells

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{w}_B - \mathbf{w}_S. \tag{12.26}$$

Both the buys and the sells are nonnegative, $\mathbf{w}_B \geq 0$, $\mathbf{w}_S \geq 0$, i.e., all elements of the two vectors are either positive or zero. It is also noted that the buys and sells are mutually exclusive: for every stock we either have a buy or sell but never both. These properties enable us to replace the absolute value of weight change by

$$\left|\Delta \mathbf{w}\right| = \mathbf{w}_B + \mathbf{w}_S . \tag{12.27}$$

Substituting both (12.26) and (12.27) into (12.25), we have

$$\begin{aligned} U(\mathbf{w}) &= \mathbf{f}' \cdot (\mathbf{w}_0 + \mathbf{w}_B - \mathbf{w}_S) - \frac{1}{2}\lambda (\mathbf{w}_0 + \mathbf{w}_B - \mathbf{w}_S)' \boldsymbol{\Sigma} (\mathbf{w}_0 + \mathbf{w}_B - \mathbf{w}_S) \\ &\quad - \boldsymbol{\theta}' \cdot (\mathbf{w}_B + \mathbf{w}_S) \\ &= U(\mathbf{w}_0) + (\mathbf{f} - \lambda \boldsymbol{\Sigma} \mathbf{w}_0 - \boldsymbol{\theta})' \cdot \mathbf{w}_B + (-\mathbf{f} + \lambda \boldsymbol{\Sigma} \mathbf{w}_0 - \boldsymbol{\theta})' \cdot \mathbf{w}_S \\ &\quad - \frac{1}{2}\lambda (\mathbf{w}_B' \boldsymbol{\Sigma} \mathbf{w}_B - 2\mathbf{w}_B' \boldsymbol{\Sigma} \mathbf{w}_S + \mathbf{w}_S' \boldsymbol{\Sigma} \mathbf{w}_S) \end{aligned} . \tag{12.28}$$

As before, the initial utility is

$$U(\mathbf{w}_0) = \mathbf{f}' \cdot \mathbf{w}_0 - \frac{1}{2}\lambda \mathbf{w}_0' \boldsymbol{\Sigma} \mathbf{w}_0 \ .$$

The objective function of (12.28) can be written in terms of a stacked vector, which combines both buys and sells, i.e.,

$$\mathbf{W} = \begin{pmatrix} \mathbf{w}_B \\ \mathbf{w}_S \end{pmatrix}, \tag{12.29}$$

and a stacked forecast vector

$$\mathbf{F} = \begin{pmatrix} \mathbf{f} - \lambda \boldsymbol{\Sigma} \mathbf{w}_0 - \boldsymbol{\theta} \\ -\mathbf{f} + \lambda \boldsymbol{\Sigma} \mathbf{w}_0 - \boldsymbol{\theta} \end{pmatrix}, \tag{12.30}$$

and an augmented covariance matrix

$$\mathbf{\Sigma}_2 = \begin{pmatrix} \mathbf{\Sigma} & -\mathbf{\Sigma} \\ -\mathbf{\Sigma} & \mathbf{\Sigma} \end{pmatrix}. \tag{12.31}$$

Combining the equations preceding, we have

$$U(\mathbf{w}) = U(\mathbf{w}_0) + \mathbf{F}' \cdot \mathbf{W} - \frac{1}{2}\lambda \mathbf{W}' \cdot \mathbf{\Sigma}_2 \cdot \mathbf{W}. \tag{12.32}$$

The optimization problem with objective function (12.32) can be solved numerically using quadratic programming.

Several constraints can be placed on the augmented weight vector **W** to address practical implementation concerns. The first constraint is $\mathbf{W} \geq 0$. Another constraint is related to dollar neutrality; i.e., the total amount of buys and sells should balance. This is a linear equality constraint

$$\mathbf{w}_B' \cdot \mathbf{i} = \mathbf{w}_S' \cdot \mathbf{i}, \text{ or } \mathbf{W}' \cdot \hat{\mathbf{i}} = 0.$$

The vector **i** is a vector of ones, of length N, and

$$\hat{\mathbf{i}} = \begin{pmatrix} \mathbf{i} \\ -\mathbf{i} \end{pmatrix}.$$

If desired, we can add the turnover constraint as

$$\mathbf{W}' \cdot \mathbf{i}_2 \leq T, \text{ with } \mathbf{i}_2 = \begin{pmatrix} \mathbf{i} \\ \mathbf{i} \end{pmatrix}.$$

T is the maximum turnover allowed and $\mathbf{i}_2$ is a vector of ones, of length $2N$.

Finally, we can require range constraints on the optimal weights

$$\mathbf{l} \leq \mathbf{w} = \mathbf{w}_0 + \mathbf{w}_B - \mathbf{w}_S \leq \mathbf{u}, \tag{12.33}$$

in terms of the augmented weight vector **W**. This is left as an exercise. Note the following:

- We have not imposed the condition that the buys and the sells are mutually exclusive on the new optimization problem. There is no need to do that because that would certainly result in a suboptimal solution. It is easy to see this in a single-asset case. Suppose both w_B and w_S are positive; then, the new weight defined by the netting of the two would achieve a higher value of utility. For example, let $w_B \geq w_S > 0$, then $w'_B = w_B - w_S$ and $w'_S = 0$ increases the utility, because it has the same mean and variance but less transaction costs.
- The augmented covariance matrix (12.31) is singular, but this is not necessarily an issue for quadratic programming. The matrix can be modified using the fact that the buys and the sells are mutually exclusive, i.e., $w_{B,i} w_{S,i} = 0$ for every stock. Consequently, we can set the diagonal elements of both $(-\boldsymbol{\Sigma})$ matrices — upper-right and bottom-left corners in (12.31) — to zeros.

12.3.4 Multiasset with Linear Costs: Numerical Example 1

We apply the numerical method to a portfolio of 20 stocks. We start with a market neutral long/short initial portfolio. We then simulate a vector of forecasts and use the forecasts to rebalance the portfolio, incorporating transaction costs. Other inputs are the covariance matrix $\boldsymbol{\Sigma}$ and the transaction cost coefficient $\boldsymbol{\theta}$. For simplicity, we take $\boldsymbol{\Sigma}$ as a diagonal matrix with specific risk of 35% for all stocks. The transaction cost is assumed to be 2% for all stocks. All portfolios, initial and optimized, have a target tracking error of 10%. The forecasts are products of IC, z-score, and specific risk. We will let IC = 0.2, and the z-scores have 0 mean and standard deviation 1.

Figure 12.4 plots the forecasts vs. the initial portfolio weights (in solid squares) and the optimal portfolio with maximum turnover. As we can see, whereas the initial weights are in general agreement with the forecasts, they are not aligned perfectly. For instance, a stock with a forecast of –3.2% has a weight of 10.3%, whereas another stock with a forecast of 11.2% has a weight of –1.9%. The overall correlation between the forecasts and the initial weights is only 0.48, and the expected return is 4.2%.

The optimal weights are the solution of (12.32) without the turnover constraint. The resulting one-way turnover is about 36%. As we can see, the forecasts and the optimal weights are aligned almost perfectly, with a correlation of 0.97. The only reason that they do not lie on a straight line

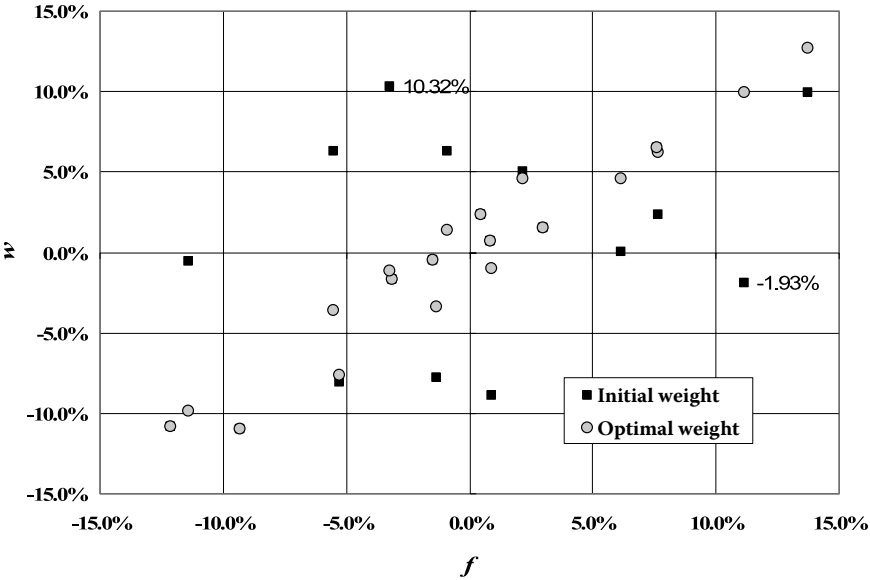


FIGURE 12.4. Scatter plot of forecasts vs. initial weights and optimal weights with maximum portfolio turnover.

is due to the $\theta = 2\%$ transaction costs we imposed. The expected return is 8.5% gross of transaction cost and 7.0% net of transaction costs. The gross return is simply the sum of weights times the expected returns and the net return is the gross return minus the transaction costs, θ times two-way turnover. It is also worth noting that out of the 20 stocks, only 10 stocks, those whose initial weights are too deviated from the optimal weights, show any meaningful weight change. The other 10 stocks are prevented from trading due to the transaction costs.

Imposing additional turnover constraints impacts on optimal weights and expected returns. Figure 12.5 shows the gross and net expected returns as a function of allowed turnover. When no turnover is permitted, both returns are the same as the return of the initial portfolio. As we allow more and more turnover, both returns increase, with the gap between the two widening as the costs increases.

- Note that the rate of increase in the net return slows down as the turnover increases. As a result, when the turnover is 20%, the net return is 6.5%, an increase of 2.3% from the initial 4.2%. This represents a roughly 80% total increase in net return, with about 55% of total turnover.

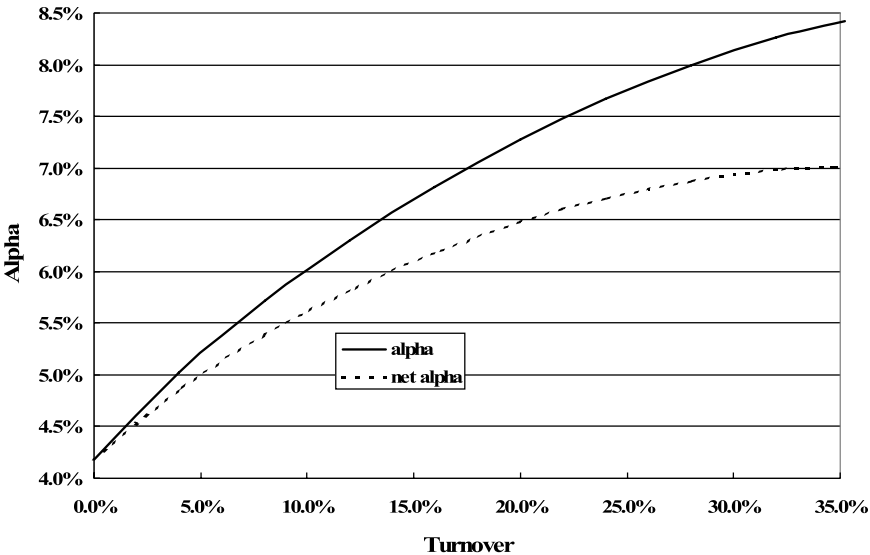


FIGURE 12.5. The gross and net expected returns as a function of allowed portfolio turnover.

Figure 12.6 shows the change in portfolio weights from the initial portfolio weights. If $\Delta w > 0$, we buy the stock, whereas if $\Delta w < 0$, we sell the stock. As noted before, only ten stocks show weight changes if maximum turnover is allowed. As we see from Figure 12.6, this number is smaller when the turnover is constrained. For example, at 4% turnover, only the two stocks that are marked in Figure 12.4 are traded. The limited turnover budget is allocated to them, because their positions are most inconsistent with their return projection and trading them increases portfolio alpha the most. As the turnover limit is increased, the trade list expands and the trade sizes expand for stocks that are already on the list.

- We note that the size of buys and sell are monotonic functions of the turnover. If we were to buy a stock, we would buy more if more turnover is allowed up to optimal weight.

12.3.5 Multiasset with Linear Costs: Numerical Example 2

In the second example, we study the impact of transaction costs on the optimal weights by varying the level of θ, which is the same for all 20 stocks. For each θ, the optimal portfolio is constructed without additional turnover constraints. Hence, the resulting turnover is the maximum turnover associated with the given transaction costs.

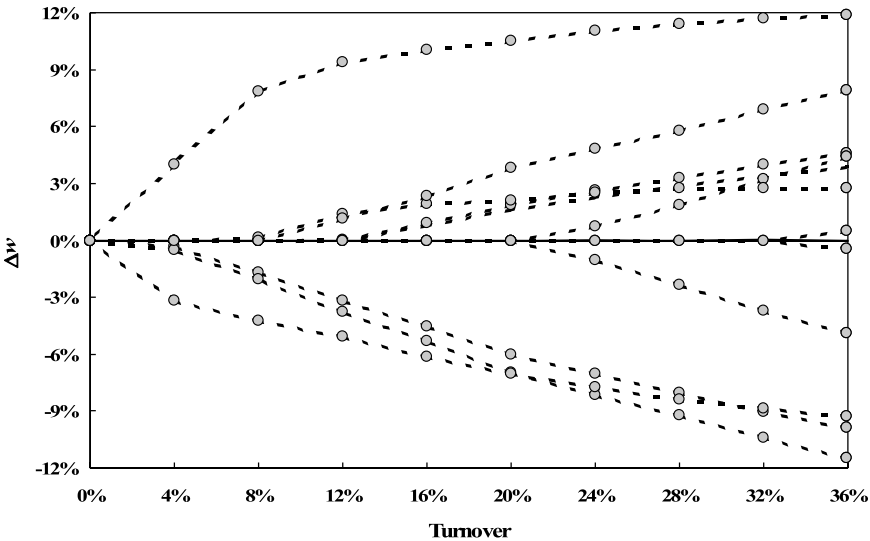


FIGURE 12.6. The change of optimal weights from the initial weights as the turnover is increased. Out of 20 stocks, 10 show no weight change; they all lie on the line $\Delta\mathbf{w} = 0$. The remaining 10 stocks show increasing change in weight as more turnovers are permitted.

Figure 12.7 shows the change of the optimal weights from the initial weight, which is the same for all levels of θ, when the transaction costs increase. When $\theta = 0$, i.e., the problem is transaction-cost free, the weight changes are at their maximum for both buys and sells. The difference is just essentially $\Delta\mathbf{w} = \tilde{\mathbf{w}} - \mathbf{w}_0$. As θ increases, the weight changes for all the stocks shrink toward 0.

- We note that the decline in weight changes follows different patterns for different stocks. Some of them follow a straight line with differing slopes, whereas others are piecewise linear. This feature reflects the nonlinear nature of the objective function and its solution.

Another noteworthy feature of Figure 12.7 is that all weight changes have the same signs as those for $\theta = 0$. In other words, if a stock is a buy (sell) from the optimization with no transaction costs, then it will be a buy (sell) in the optimization with transaction costs. If this is true, it points to an alternative method of constructing an optimal portfolio with transaction costs, using a two-step approach. In the first step, we run an optimization without transaction costs. This is relatively simple as we do not encounter the absolute value function in the objection function (12.25).

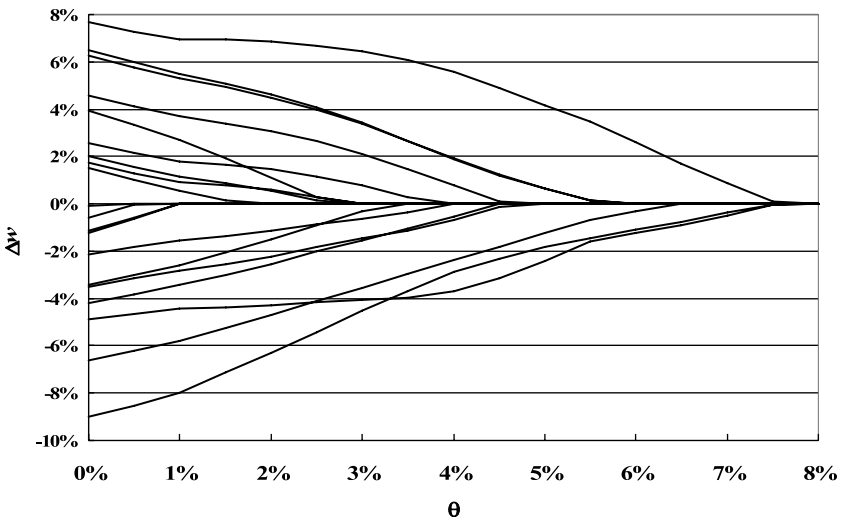


FIGURE 12.7. The difference between the optimal weights and the initial weights for varying levels of transaction costs $\boldsymbol{\theta}$.

The solution of this step would provide us a buy list and a sell list. In the second step, we optimize again but with prescribed transaction costs. With the buy and sell lists available, we can now specify the range of optimal weights as $w \geq w_0$ for a buy and $w \leq w_0$ for a sell. The associated transaction costs will be $w - w_0$ for a buy and $w_0 - w$ for a sell. Consequently, we remove the difficulty of dealing with the absolute value function in the objective function. The resulting optimization problem can be solved routinely. However, we caution readers that this may not always be the case.

12.4 PORTFOLIO TRADING STRATEGIES

Once optimal portfolio weights are determined, the changes from the initial portfolio weights are the resulting trades that need to be implemented. The goal of portfolio trading strategies is to implement the trades in the most efficient manner. In certain cases, it might be optimal to not implement the full trades, due to either decay in return signals or high transaction costs. In practice, this can also arise due to the use of limit orders, which might not be triggered by price movement resulting in opportunity costs. We shall not consider such cases in our treatment and require all trades to be implemented in the portfolio strategies.

There are at least two conflicting objectives in the portfolio implementation process. On the one hand, one would like to implement the changes as soon as possible to get to the optimal portfolio. The optimal portfolio has

the maximum expected return for a specific risk target. Any delay could potentially result in a loss of return, and both the expectation and the variance of that potential loss grow over time. On the other hand, transaction costs from market impact are a direct function of the speed with which the trades are executed. For large trade sizes, immediate execution would cause the greatest market impact. Breaking it in pieces and trading them over an extended period of time would reduce the market impact but at the risk of return loss and tracking error mismatch versus the optimal portfolio, as well as higher fixed costs such as commissions and fees.

For a portfolio of stocks to be traded with both buys and sells, one must consider the trade basket as a whole. For instance, an imbalance between buys and sells might cause an intended net market exposure. The correlation between different stocks is another important issue. For buys and sells that are highly correlated in terms of stock returns, one would like to synchronize the trades, because doing so would reduce systematic exposure. However, if these trades have different market impacts, one would like to execute them at different speeds to minimize the transaction cost. It is therefore necessary to find a balance between the two.

The trading horizon — the length of time we allocate to implement the trades — is another important factor. For trades that are easy to implement based on liquidity, the trading horizon should be short. For difficult trades, the trading horizon can be longer. For a given set of trades, it is better to optimize the trading horizon as well as the actual trade implementation.

12.5 OPTIMAL TRADING STRATEGIES: SINGLE STOCK

The problem of optimal trading strategies can be formulated mathematically through an optimization in which the objective function consists of expected return shortfall, return variance, and transaction costs. Grinold and Kahn (2000) considered this problem in continuous time and Almgren and Chriss (2000) used a discrete setting for their analysis. We shall work with the continuous-time case for simplicity in the notations.

We start with the case of a single stock for which the trade is denoted by Δw. Suppose the trade will be carried out over the horizon $\left[0,T\right]$. We denote the state of the trade at time t in proportion of the total trade: $h(t)\Delta w$, with $h(0)=0$ and $h(T)=1$. The trade shortfall is $h(t)\Delta w-\Delta w=\Delta w\left[h(t)-1\right]$. Suppose the stock's expected return over the horizon is a constant f; then the return shortfall is $f\Delta w\left[h(t)-1\right]$. Denoting the stock's risk by σ, the shortfall variance is $\sigma^2(\Delta w)^2\left[h(t)-1\right]^2$. We model the transaction costs

by two terms, one related to the fixed cost and the other related to the market impact. The fixed cost is assumed to be $-c|\Delta w|T$ (change in the term), with $c>0$. It is easy to see that the cost is proportional to the trade size. What is new here is that the cost will be proportional to the trading horizon; the longer the horizon, the more often we have to trade (at smaller sizes) and the more we have to pay for fixed costs such as commissions and fees. Finally, we approximate the cost of market impact as being proportional to the square of trading speed, or the derivative of holding: $(\Delta w)^2\left[\dot{h}(t)\right]^2$. Combining all four terms and integrating over the time interval $[0,T]$ gives the objective function

$$J=\int_0^T f\Delta w\left[h(t)-1\right]dt-\frac{1}{2}\lambda\int_0^T \sigma^2(\Delta w)^2\left[h(t)-1\right]^2 dt-c|\Delta w| \int_0^T dt-\psi\int_0^T(\Delta w)^2\left[\dot{h}(t)\right]^2 dt \tag{12.34}$$

The additional two parameters are λ (the risk-aversion parameter) and ψ (the cost coefficient for market impact). We can simplify (12.34) by scaling it by a positive term $(\Delta w)^2$,

$$\frac{J}{(\Delta w)^2}=\int_0^T\left\{f_w\left[h(t)-1\right]-c_w-\psi\left[\dot{h}(t)\right]^2-\frac{1}{2}\lambda\sigma^2\left[h(t)-1\right]^2\right\}dt. \tag{12.35}$$

We have $f_w=f/(\Delta w)$ and $c_w=c/|\Delta w|>0$. The goal of optimal trading strategies is to find the solution $h(t)$ that maximizes (12.35). Note the following:

- Depending on the forecast and the direction of the trade, $f_w=f/(\Delta w)$ can be zero, positive, or negative. It is zero when the forecast is zero. In this case, the objective function is the same for both buy orders ($\Delta w>0$) and sell orders ($\Delta w<0$). When the forecast is nonzero, the term $f_w=f/(\Delta w)$ is positive when both have the same sign: buy with a positive forecast or sell with a negative forecast. It is negative when both have opposite signs: buy with a negative forecast or sell with a positive forecast.

- The first three terms of (12.35) are all implementation costs, alpha or transaction costs — whereas the last term is implementation risk. The problem of optimal trading strategies is thus similar to a mean–variance problem of portfolio construction. For a given level of implementation risk, there exists an optimal solution with minimum implementation costs. Similar to the efficient frontier of mean–variance optimization, the optimal trading strategies for varying implementation risks form an efficient risk-cost frontier.
- The *fixed* term has been missing in previous work in optimal trading strategies. Because it is always a cost and it increases with T, it has the effect of shortening the optimal trading horizon when we allow T to be free later in the chapter.

12.5.1 Optimal Solution with Fixed Trading Horizon

We first treat the trading horizon T as fixed, i.e., the amount of time needed to execute a trade has been determined, maybe by some heuristic estimation or based on traders' experience. We will now solve for the optimal solution $h(t)$ for t in [0,1]. In the next section, we shall also find the optimal trading horizon.

The mathematical technique for solving this type of optimization problem is the calculus of variation. Denote the integrand of (12.35) by

$$L(h,\dot{h}) = f_w\left[h(t)-1\right] - c_w - \psi\left[\dot{h}(t)\right]^2 - \frac{1}{2}\lambda\sigma^2\left[h(t)-1\right]^2. \quad (12.36)$$

Then the solution is given by the following differential equation

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{h}}\right] = \frac{\partial L}{\partial h} \quad (12.37)$$

From (12.36), we have

$$\begin{aligned} \frac{\partial L}{\partial \dot{h}} &= -2\psi\dot{h} \\ \frac{\partial L}{\partial h} &= f_w - \lambda\sigma^2(h-1) \end{aligned} \quad (12.38)$$

Substituting (12.38) into (12.37) yields

$$2\psi\ddot{h}-\lambda\sigma^2 h=-\left(f_w+\lambda\sigma^2\right). \tag{12.39}$$

Dividing the equation by 2ψ leads to the following ordinary differential equation (ODE)

$$\ddot{h}-g^2 h=-s-g^2 \text{ with } s=\frac{f_w}{2\psi},\ g^2=\frac{\lambda\sigma^2}{2\psi}. \tag{12.40}$$

For the newly defined parameter, we have $g\geq 0$ and s has the same sign as f_w. The boundary condition is $h(0)=0$ and $h(T)=1$. However, note the following:

- Because the trading horizon T is fixed, the fixed-cost term is then known, and it does not enter the solution. However, it will play a significant role when we have a flexible trading horizon.

We will first consider the solution for the following two special cases:

Case I: $s = g = 0$

This occurs when both forecast and risk-aversion parameter are zero. Now the differential equation reduces to $\ddot{h}=0$. The solution is therefore

$$h(t)=\frac{t}{T}. \tag{12.41}$$

The optimal solution is linear, implying a constant speed of trading: $\dot{h}=1/T$. In this case, only the market impact matters. To reduce market impact, the optimal trading strategy is to break the trade evenly during the trade horizon. Furthermore, the total cost would just be

$$\frac{J}{(\Delta w)^2}=c_w T+\int_0^T\left\{\psi\left(\frac{1}{T}\right)^2\right\}dt=c_w T+\frac{\psi}{T}. \tag{12.42}$$

Note that the total costs as a function of T go to infinity when T goes to either zero or infinity. It reaches a minimum if $T = \sqrt{\psi / c_w}$. If $c_w = 0$, the total cost decreases to zero as the trading horizon lengthens to infinity, which is an unrealistic result.

Case II: $g = 0$

In this case, the risk-aversion parameter is zero. Now the differential equation reduces to $\ddot{h} = -s$. The solution is therefore

$$h(t) = -\frac{s}{2}t^2 + at + b. \tag{12.43}$$

The constant a and b can be determined by the boundary condition. Therefore, we have

$$h(t) = \frac{t}{T} + \frac{s}{2}t(T - t). \tag{12.44}$$

Equation 12.44 consists of the solution (12.41) and a quadratic term that vanishes at both $t = 0$ and $t = T$. The trading speed is given by

$$\dot{h}(t) = \frac{1}{T} + \frac{sT}{2} - st. \tag{12.45}$$

Figure 12.8 plots the solution for three cases, all with $g = 0$ but with three different values of s. The solution for the case with $s = 0$ is a straight line. When $s > 0$, by its definition the term f_w is positive, implying either a positive forecast for a buy or a negative forecast for a sell. Hence, there is a need to execute the trade as soon as possible in order to reduce alpha shortfall. This is indeed the case for the optimal solution, the dotted line, which lies above the linear solution. The slope, or the speed of the trade, is higher initially and then slows down as time approaches T. On the other hand, when $s < 0$, the term f_w is then negative, implying either a negative forecast for a buy or a positive forecast for a sell. Contrary to the previous case, there is incentive to delay the trade as long as possible, because

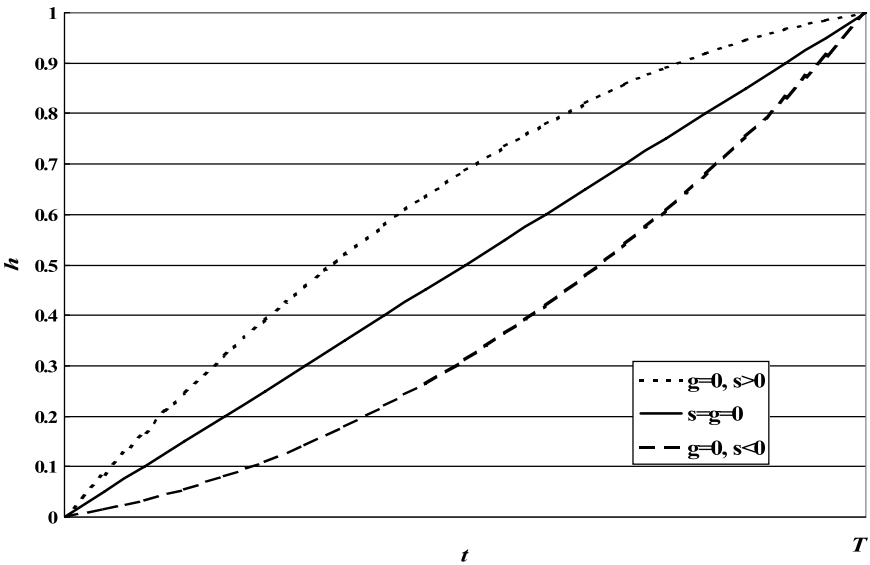


FIGURE 12.8. The optimal trading paths for three special cases: the solid line is for the case $s = g = 0$, the dotted line is for $g = 0$, $s > 0$, and the dashed line is for $g = 0$, $s < 0$.

the trade itself leads to lower alpha. Therefore, the optimal solution, the dashed line, lies below the linear solution. The trade fills slowly first and then speeds up as the time approaches T.

It is actually possible for the solution (12.44) for $h(t)$ to move out of the range [0,1]. For instance, when $s > 0$, $h(t)$ could be greater than 1. On the other hand, when $s < 0$, $h(t)$ could be less than 0. This implies that the solution may actually switch the direction of the trade during the course of trading! In other words, if the trade were to buy 1000 shares, the optimal strategy could have us buy 1100 shares and later sell the extra 100 shares. This is highly unlikely in practice, because the trading would have stopped once the 1000 shares had been bought. It could happen in the optimal trading solution if the trading horizon is too long, coupled with the fact that we have a strong forecast and a relatively weak market impact. With this combination, the mathematical optimal trading strategy would be to first buy as many shares as possible to generate returns and then later sell them to reach trade size. Because the trading cost is low, this "two-way" strategy would be better than any "one-way" strategy.

Figure 12.9 illustrates this situation. The dotted line is an optimal strategy whose path rises and crosses the line $h = 1$ during the trading horizon. The culprit in this case is the fixed trading horizon T, which is too long.

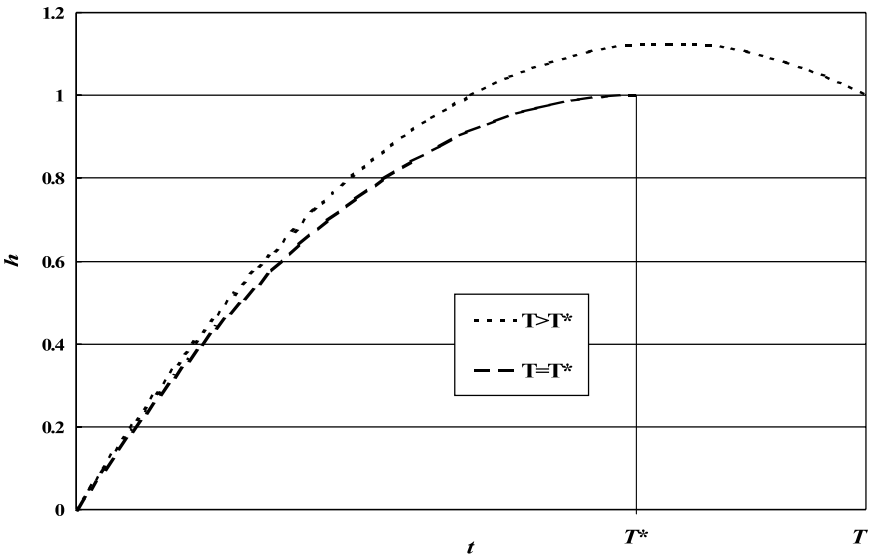


FIGURE 12.9. Optimal trading paths for two different trading horizons.

If we allow the trading horizon to be free and optimize it together with the trading path, the horizon will be shortened to T^* and the associated optimal path, the dashed line, will never cross the line $h = 1$. The case of the free trading horizon is solved in the following section.

12.5.1.1 The General Case

When the parameter g is nonzero, the general solution of ODE (12.40) is the exponential functions exp($-gt$) and exp(gt), which can be combined into hyperbolic functions. The particular solution is given by

$$-g^2 h = -s - g^2 \text{ or } h = 1 + \frac{s}{g^2}.$$

We have (Grinold & Kahn 2000)

$$h(t) = a\sinh(gt) + b\cosh\sinh(gt) + 1 + \frac{s}{g^2}.$$

The constant a and b are determined by the boundary condition; therefore we have

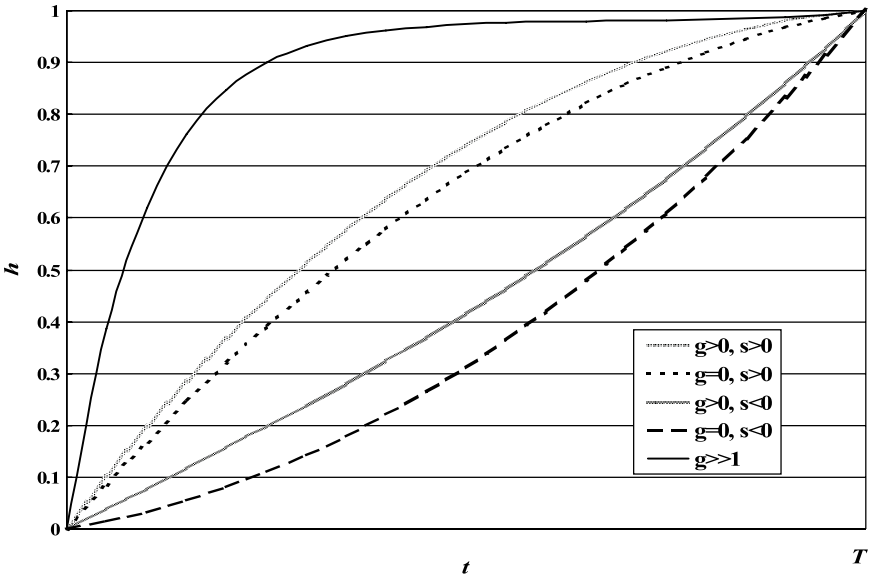


FIGURE 12.10. Five different optimal trading paths, two of which are identical to those in Figure 12.7. The other three are for cases with $g > 0$. Two of them have a moderate value of g, whereas the steepest path, the thin solid line, has the highest value of g, corresponding to extreme risk aversion.

$$h(t)=\frac{\left(1+\frac{s}{g^2}\right)\cosh(gT)-\frac{s}{g^2}}{\sinh(gT)}\sinh(gt)-\left(1+\frac{s}{g^2}\right)\left[\cosh(gt)-1\right]. \quad (12.46)$$

To see the effect of g, or variance of shortfall, on the optimal trading strategy, we plot the solution (12.46) in Figure 12.10. There are in all five paths in Figure 12.10, and two of them are identical to those in Figure 12.8 and have zero risk aversion ($g = 0$) but nonzero s. The shaded lines next to them are the corresponding trading paths with nonzero g. In both cases, the new trading path is above the previous one, indicating faster execution regardless of the forecast. This makes intuitive sense because higher risk aversion would cause investors to desire speedy execution at the expense of higher transaction costs.

When risk aversion dominates both the return shortfall and market impact, the optimal trading strategy is immediate execution. The thin solid line in Figure 12.10 illustrates this point. It rises rather rapidly and then flattens out. It can be shown mathematically that as $g \to \infty$,

$$h(t) \to 1 - \exp(-gt), \quad \text{if } t \text{ is near } 0;$$
$$h(t) \to \exp\left[-g(T-t)\right], \quad \text{if } t \text{ is near } T. \tag{12.47}$$

Example 12.5

Consider the case of $s = 0$ in (12.46). Then the solution reduces to

$$h(t) = \coth(gT)\sinh(gt) - \cosh(gt) + 1. \tag{12.48}$$

We obtain the implementation costs as

$$\int_0^T \left\{ c_w + \psi\left[\dot{h}(t)\right]^2 \right\} dt = c_w T + \psi g^2 \left[\frac{1}{2g}\coth(gT) + \frac{T}{2}\operatorname{csch}^2(gT) \right] \tag{12.49}$$

and the implementation risk in terms of variance is

$$\sigma^2 \int_0^T \left[h(t) - 1\right]^2 dt = \sigma^2 \left[\frac{1}{2g}\coth(gT) - \frac{T}{2}\operatorname{csch}^2(gT) \right]. \tag{12.50}$$

Taking the square root of (12.50) gives rise to the implementation risk in standard deviations.

Figure 12.11 plots the implementation costs vs. the risk for varying degrees of risk aversion. The cost is positive in the graph and is a declining function of risk. Each point of the curve corresponds to a different trading strategy, depending on different levels of risk aversion, illustrating the trade-off between risk and cost. When the risk aversion is high, the optimal trading strategy would be to trade fast to reduce implementation risk but incur higher cost. On the other hand, when the risk aversion is low, the optimal trading strategy focuses on lowering cost but incurs higher implementation risk.

12.5.2 Optimal Trading Horizon

The analysis so far has assumed a fixed trading horizon. However, in reality, the trading horizon is not precisely known and depends on the trade itself. For instance, for trades that are easy to implement, the trade size is a small fraction of the average daily volume, and the trading horizon can be short; whereas for trades that are difficult to fill, the trading horizon must

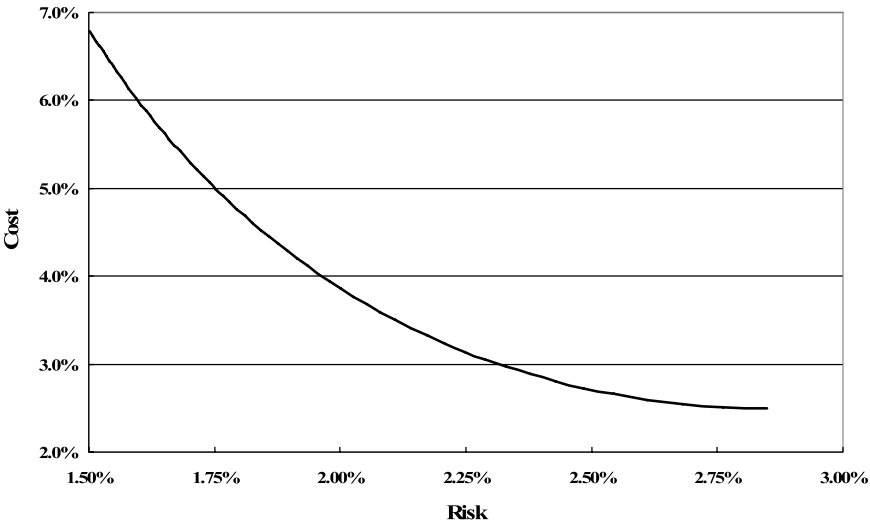


FIGURE 12.11. The implementation cost-risk frontier for optimal trading strategies. The parameters are $\psi = 0.05\%$, $\sigma = 35\%$, and $T = 0.02$. We also set $c_w = 0$, which does not affect the shape of the curve, because the fixed cost is a constant for fixed T, independent of risk aversion.

be lengthened. The trading horizon may also be dependent on investors' aversion to risks of shortfall. If the risk aversion is high, then the horizon is short; and if the risk aversion is low, then the horizon might be longer.

Mathematically, we can treat the trading horizon as a part of the optimization problem. In other words, we should let T be free or unknown, and we can then solve the optimization problem for both the optimal trading path $h(t)$ and the optimal T. In reality, there might be some practical constraints on the trading horizon; for instance, one might want to complete a trade ahead of a long weekend. It is nevertheless useful to compare this with the true optimal.

The mathematical problem is to maximize the objective function (12.35) with both $h(t)$ and free boundary T. The problem can similarly be solved with the calculus of variation as follows. The optimal path $h(t)$ must satisfy the same differential equation

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{h}}\right] = \frac{\partial L}{\partial h}.$$

It should also satisfy the same boundary condition $h(0) = 0$ and $h(T) = 1$. In addition, the free boundary condition leads to the following (see Appendix):

$$L\left(h,\dot{h}\right)-\frac{\partial L\left(h,\dot{h}\right)}{\partial \dot{h}}\dot{h}\Bigg|_{t=T}=0\,. \tag{12.51}$$

Because $\frac{\partial L}{\partial \dot{h}}=-2\psi\dot{h}$ and $L=-c_w-\psi\left(\dot{h}\right)^2$ at $t=T$, equation (12.51) leads to

$$\begin{gathered}-c_w-\psi\left(\dot{h}\right)^2+2\psi\left(\dot{h}\right)^2=\psi\left(\dot{h}\right)^2-c_w=0\\ \dot{h}\left(T\right)=\sqrt{\frac{c_w}{\psi}}\triangleq p\end{gathered} \tag{12.52}$$

Hence, the free trading horizon gives rise to a condition on the trading speed at T, which allows us to find the optimal trading time as well as the optimal trading path. Note the following:

- We have taken the positive root for $\dot{h}(T)$ because $h(t)$ is a monotonically increasing function if we do not allow the trading strategies to switch the direction of trades. From $h(0) = 0$ and $h(T) = 1$, we conclude $\dot{h}(t) \geq 0$.
- If $c_w = 0$, i.e., the fixed cost of transaction is neglected, then the condition becomes $\dot{h}(T) = 0$. As the trade gets filled, the trading at the end of the trading horizon gets slower and slower, coming to a smooth stop at the end.

Example 12.6

Consider the case in which $g = 0$ (zero risk aversion). The solution for $h(t)$ is (12.44) and for the trading speed $\dot{h}(t)$ is (12.45). Hence, (12.52) gives rise to

$$\dot{h}\left(T\right)=\frac{1}{T}+\frac{sT}{2}-sT=\frac{1}{T}-\frac{sT}{2}=\sqrt{\frac{c_w}{\psi}}=p\,.$$

This is a quadratic equation for T and the solution is

$$T=\frac{2}{p+\sqrt{p^2+2s}}=\frac{2\sqrt{\psi}}{\sqrt{c_w}+\sqrt{c_w+f_w}}=\frac{2\sqrt{\psi}\left|\Delta w\right|}{\sqrt{c}+\sqrt{c+f\,\mathrm{sgn}\left(\Delta w\right)}}\,. \tag{12.53}$$

- The optimal trading horizon exists when f_w is positive: a positive forecast for a buy or negative forecast for a sell. In this case, the trading horizon increases with the market impact cost ψ and the trade size Δw. In other words, if the trade is costly and large, we should allow more time. The trading horizon also decreases with the alpha forecast and the fixed cost. If alpha shortfall is severe or if the fixed cost is large, we should execute the trade sooner.
- The optimal trading horizon does not always exist. If f_w is negative — negative forecast for a buy or positive forecast for a sell — and the magnitude of the forecast exceeds that of the fixed cost $|f| > c$, then there is no optimal trading horizon. In other words, the optimal trading horizon is infinite, because the trade in these circumstances would reduce the return. Coupled with a high forecast, we would gain more if we delayed the trade for as long as possible. These cases might not occur in practice, but one should be aware of the possibilities.

If $c = 0$, i.e., there is no fixed cost, then Equation 12.53 reduces to

$$T = \frac{2\sqrt{\psi}\,|\Delta w|}{\sqrt{f}}. \tag{12.54}$$

Example 12.7

Consider the case $s = 0$ (zero forecast) as in Example 12.5. From (12.48), we have

$$\dot{h}(T) = g\left[\coth(gT)\cosh(gT) - \sinh(gT)\right] = \frac{g}{\sinh(gT)}. \tag{12.55}$$

Therefore, the optimal trading horizon is given by

$$\frac{g}{\sinh(gT)} = p \text{ or } T = \frac{1}{g}\sinh^{-1}\left(\frac{g}{p}\right). \tag{12.56}$$

Written in terms of the original parameters, we have

$$T = \sqrt{\frac{2\psi}{\lambda\sigma^2}}\sinh^{-1}\sqrt{\frac{\lambda\sigma^2|\Delta w|}{2c}}. \tag{12.57}$$

In general, the optimal trading horizon lengthens if Δw (the trade size) increases, if ψ (market impact) increases, and if c (fixed cost) decreases. It also lengthens if $\lambda\sigma^2$ (risk aversion) decreases, because the function $\sinh^{-1}(x)/x$ is a declining function of x.

12.6 OPTIMAL TRADING STRATEGIES: PORTFOLIOS OF STOCKS

Much of the analysis of single-stock trading strategies can be extended to multiple stocks, or a portfolio of stocks. We shall formulate the problem first and then find the optimal solution. We shall also allow for the optimal trading horizon T. For a portfolio of stock trades, we also discuss additional constraints one might wish to impose during the trading.

12.6.1 Formulation

Suppose we have trades in N stocks, and the trade sizes are $(\Delta w_1, \Delta w_2, \cdots, \Delta w_N)$. We denote the trading path by a vector of function $\mathbf{h}(t) = [h_1(t), \cdots, h_N(t)]'$. At any given time t, the portfolio position relative to the final position is $[\Delta w_1(h_1 - 1), \Delta w_2(h_2 - 1), \cdots, \Delta w_N(h_N - 1)]$. At the beginning of the trade, we have $h_i(0) = 0,\ i = 1, \cdots, N$ and at the end of the trade $h_i(T) = 1,\ i = 1, \cdots, N$. These are the boundary conditions for h's.

The optimal trading strategy for a portfolio of trades is found by optimizing an objective function similar to that of a single trade. First, the instantaneous return shortfall is given by $f_1 \Delta w_1 (h_1 - 1) + f_2 \Delta w_2 (h_2 - 1) + \cdots + f_N \Delta w_N (h_N - 1) = \mathbf{f}_w' \cdot (\mathbf{h} - \mathbf{1})$, in which f's are return forecasts and the vector $\mathbf{f}_w = (f_1 \Delta w_1, \cdots, f_N \Delta w_N)'$ and the vector $\mathbf{1} = (1, \cdots, 1)'$. The variance of the return shortfall for a given time t is

$$\begin{aligned} &[\Delta w_1(h_1 - 1), \cdots, \Delta w_N(h_N - 1)]\boldsymbol{\Sigma}\begin{pmatrix} \Delta w_1(h_1 - 1) \\ \vdots \\ \Delta w_N(h_N - 1) \end{pmatrix}. \\ &= (\mathbf{h} - \mathbf{1})' \boldsymbol{\Sigma}_w (\mathbf{h} - \mathbf{1}) \end{aligned} \tag{12.58}$$

The matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^N$ is the covariance matrix of returns, and $\boldsymbol{\Sigma}_w = (\sigma_{ij} \Delta w_i \Delta w_j)_{i,j=1}^N$ comprises products of the return covariance matrix and the trade size.

Similar to the single-stock trade, there are two components of transaction costs. We model the fixed costs as a multiple of the trading horizon T, and the constant is given by $c_w = c_1\Delta w_1 + c_2\Delta w_2 + \cdots + c_N\Delta w_N$. The variable costs — the instantaneous market impact — is related to the speeds of the trading in all N stocks

$$\left[\Delta w_1\dot{h}_1,\cdots,\Delta w_N\dot{h}_N\right]\mathbf{\Psi}\begin{pmatrix}\Delta w_1\dot{h}_1\\ \vdots \\ \Delta w_N\dot{h}_N\end{pmatrix} = \dot{\mathbf{h}}'\mathbf{\Psi}_w\dot{\mathbf{h}}, \tag{12.59}$$

where $\mathbf{\Sigma}_w = \left(\Psi_{ij}\Delta w_i\Delta w_j\right)_{i,j=1}^{N}$.

Combining all four terms and integrating them over time gives the objective function of trading strategies

$$J = \int_0^T L\left(\mathbf{h},\dot{\mathbf{h}}\right)dt, \text{ with}$$

$$L\left(\mathbf{h},\dot{\mathbf{h}}\right) = \mathbf{f}_w\left[\mathbf{h}(t)-\mathbf{1}\right] - c_w - \dot{\mathbf{h}}(t)'\mathbf{\Psi}_w\dot{\mathbf{h}}(t). \tag{12.60}$$

$$-\frac{1}{2}\lambda\left[\mathbf{h}(t)-\mathbf{1}\right]'\mathbf{\Sigma}_w\left[\mathbf{h}(t)-\mathbf{1}\right]$$

12.6.2 Solutions of Optimal Trading Strategies

We derive the differential equation for the optimal trading path with the calculus of variation. We have

$$\frac{\partial L}{\partial\dot{\mathbf{h}}} = -2\mathbf{\Psi}_w\dot{\mathbf{h}}(t),\ \frac{\partial L}{\partial\mathbf{h}} = \mathbf{f}_w - \lambda\mathbf{\Sigma}_w\left[\mathbf{h}(t)-\mathbf{1}\right] \tag{12.61}$$

and $\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{\mathbf{h}}}\right) = \frac{\partial L}{\partial\mathbf{h}}$ gives rise to

$$2\mathbf{\Psi}_w\ddot{\mathbf{h}}(t) - \lambda\mathbf{\Sigma}_w\mathbf{h}(t) = -\mathbf{f}_w - \lambda\mathbf{\Sigma}_w. \tag{12.62}$$

Assuming the matrix $\mathbf{\Psi}_w$ is invertible, we can rewrite (12.62) as

$$\ddot{\mathbf{h}}(t) - \frac{\lambda}{2}\mathbf{\Psi}_w^{-1}\mathbf{\Sigma}_w\mathbf{h}(t) = -\frac{1}{2}\mathbf{\Psi}_w^{-1}\mathbf{f}_w - \frac{\lambda}{2}\mathbf{\Psi}_w^{-1}\mathbf{\Sigma}_w. \tag{12.63}$$

The particular solution of (12.63) is obtained by setting $\ddot{\mathbf{h}} = 0$

$$\mathbf{h}(t) = -\frac{1}{\lambda}\boldsymbol{\Sigma}_w^{-1}\mathbf{f}_w + \mathbf{1}. \tag{12.64}$$

The general solution is of the form $\mathbf{h}(t) = \mathbf{v}\cdot\exp(pt)$ and

$$\left(p^2\mathbf{I} - \frac{\lambda}{2}\boldsymbol{\Psi}_w^{-1}\boldsymbol{\Sigma}_w\right)\mathbf{v} = 0. \tag{12.65}$$

It follows that p^2 must be an eigenvalue of the matrix $\frac{\lambda}{2}\boldsymbol{\Psi}_w^{-1}\boldsymbol{\Sigma}_w$ and $\mathbf{v}$ the corresponding eigenvector, both of which can be found by standard numerical routines. Note the following:

- Assuming the matrix $\frac{\lambda}{2}\boldsymbol{\Psi}_w^{-1}\boldsymbol{\Sigma}_w$ is positive definite, there will be N positive eigenvalues and N eigenvectors, and there will be $2N$ general solutions. The weights for these solutions can be found using $2N$ boundary conditions.

12.6.3 Optimal Trading Horizon

When the trading horizon is free, we can find the optimal trading horizon using the condition similar to (12.51). In the case of a portfolio trade, we have

$$L(\mathbf{h},\dot{\mathbf{h}}) - \dot{\mathbf{h}}'\cdot\left.\frac{\partial L(\mathbf{h},\dot{\mathbf{h}})}{\partial\dot{\mathbf{h}}}\right|_{t=T} = 0. \tag{12.66}$$

Using (12.60) and (12.61) gives

$$\dot{\mathbf{h}}'\cdot\boldsymbol{\Psi}\cdot\dot{\mathbf{h}}\Big|_{t=T} = c_w. \tag{12.67}$$

The condition is similar to (12.52) and can be combined with the optimal trading solution of the last section to find the optimal T.

12.6.4 Portfolio Constraints

When trading a portfolio of stocks, one often has to maintain the balance between orders so that the portfolio meets a set of constraints. An

example of such constraint is the dollar-neutral constraint: the dollar amount of buys matches that of sells. Other constraints can be risk based. For instance, we might want the portfolio to be beta neutral at all times. These linear constraints can be expressed as

$$\mathbf{h}' \cdot \mathbf{g} = 0 \tag{12.68}$$

where **h** is the trading path for all stocks and **g** a vector of constants.

There are a couple of ways to find the optimal trading strategies with such linear constraints, for example, the method of elimination and the method of the Lagrangian multiplier (Kirk 1970).

PROBLEMS

12.1 Prove that the coefficient θ in Equation 12.1 is given by the cost per share divided by the share price.

12.2 Consider the case in which $\tilde{w} < w_0$. Prove that the optimal weight is

$$w^* = \begin{cases} \dfrac{f+\theta}{\lambda\sigma^2}, & \text{if } \dfrac{f+\theta}{\lambda\sigma^2} \le w_0 \\ w_0, & \text{otherwise} \end{cases}. \tag{12.69}$$

12.3 Prove that the critical value of θ, above which there is no trade, is given by

$$\theta_c = \lambda\sigma^2 \left|\tilde{w} - w_0\right|. \tag{12.70}$$

12.4 Find the optimal position of a single asset when there are both linear and quadratic transaction costs, by maximizing the utility function

$$U(w) = f \cdot w - \frac{1}{2}\lambda\sigma^2 w^2 - \theta\left|w - w_0\right| - \psi\left(w - w_0\right)^2. \tag{12.71}$$

12.5 (a) Prove that the utility function in (12.25) can be written as

$$U(\mathbf{w}) = U(\mathbf{w}_0) + \lambda(\Delta\mathbf{w})' \boldsymbol{\Sigma}(\tilde{\mathbf{w}} - \mathbf{w}_0) - \frac{1}{2}\lambda(\Delta\mathbf{w})' \boldsymbol{\Sigma}(\Delta\mathbf{w}) - \boldsymbol{\theta}' \cdot \left|\Delta\mathbf{w}\right|, \tag{12.72}$$

with $\tilde{\mathbf{w}} = \lambda^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{f}$ as the optimal weights with no transaction costs, and

(b) prove that the optimal weights must satisfy the condition $(\Delta\mathbf{w})' \boldsymbol{\Sigma}(\tilde{\mathbf{w}} - \mathbf{w}_0) \geq 0$, i.e., the vector of weight changes must be in the same direction as $(\tilde{\mathbf{w}} - \mathbf{w}_0)$.

12.6 Express the range constraint (12.33) as linear inequality constraints on the augmented vector **W**.

12.7 Verify that solution (12.46) satisfies both the differential equation and the boundary conditions.

12.8 For the optimal trading solution (12.48), prove that the implementation cost is given by (12.49) and the implementation risk is given by (12.50).

12.9 For the general optimal trading solution (12.46) and free T, show that the optimal trading horizon T satisfies equation

$$s\cosh(gT) + gp\sinh(gT) = s + g^2 .$$

APPENDIX
CALCULUS OF VARIATION

We derive the ODE for the optimal trading strategy and the optimal trading horizon using calculus of variation.

Given a functional, a real-valued function of functions

$$J(h,T) = \int_o^T L\left[h(t), \dot{h}(t), t\right] dt ,$$

in which $h(0) = 0$ and $h(T) = 1$, and T is free, then the change in the functional is

$$\begin{aligned} \delta J &= J(h + \delta h, T + \delta T) - J(h,T) \\ &= \int_o^{T+\delta T} L\left[h(t) + \delta h, \dot{h}(t) + \delta \dot{h}, t\right] dt - \int_o^T L\left[h(t), \dot{h}(t), t\right] dt \end{aligned}$$

Splitting the first integral in two, we have

$$\delta J = \int_o^T \left\{ L\left[h(t)+\delta h, \dot{h}(t)+\delta \dot{h}, t\right] - L\left[h(t), \dot{h}(t), t\right] \right\} dt$$

$$- \int_T^{T+\delta T} L\left[h(t)+\delta h, \dot{h}(t)+\delta \dot{h}, t\right] dt$$

The second term is approximated by

$$\int_T^{T+\delta T} L\left[h(t)+\delta h, \dot{h}(t)+\delta \dot{h}, t\right] dt = L\left[h(t), \dot{h}(t), t\right]\Big|_{t=T} \delta T + o(\delta T). \quad (12.73)$$

The notation $o(.)$ denotes the higher-order term. The first term can be approximated by Taylor expansion

$$\int_o^T \left\{ L\left[h(t)+\delta h, \dot{h}(t)+\delta \dot{h}, t\right] - L\left[h(t), \dot{h}(t), t\right] \right\} dt = \int_o^T \left\{ \delta h \frac{\partial L}{\partial h} + \delta \dot{h} \frac{\partial L}{\partial \dot{h}} \right\} dt .$$

Integrating by parts the term containing $\delta \dot{h}$ yields

$$\begin{aligned} &\int_o^T \left\{ L\left[h(t)+\delta h, \dot{h}(t)+\delta \dot{h}, t\right] - L\left[h(t), \dot{h}(t), t\right] \right\} dt \\ &= \int_o^T \delta h \left\{ \frac{\partial L}{\partial h} - \frac{d}{dt}\frac{\partial L}{\partial \dot{h}} \right\} dt + \left(\delta h \frac{\partial L}{\partial \dot{h}} \right)\Bigg|_{t=T} \end{aligned} \quad (12.74)$$

When T is fixed, we have $\delta h = 0$ at $t = T$. When T is free, we have

$$0 = h(T+\delta T) - h^*(T) \approx h(T) - h^*(T) + \dot{h}^*(T)\delta T = \delta h(T) + \dot{h}^*(T)\delta T .$$

Therefore,

$$\delta h(T) \doteq -\dot{h}^*(T)\delta T . \quad (12.75)$$

Combining (12.73), (12.74), and (12.75) gives

$$0 = \delta J = \int_{o}^{T} \delta h \left\{ \frac{\partial L}{\partial h} - \frac{d}{dt}\frac{\partial L}{\partial \dot{h}} \right\} dt + \left(L - \dot{h}\frac{\partial L}{\partial \dot{h}} \right)\Bigg|_{t=T} \delta T \qquad (12.76)$$

for optimal path and optimal trading horizon. Because Equation 12.76 is true for the arbitrary function δh and arbitrary increment δT, we must have

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{h}} \right) - \frac{\partial L}{\partial h} = 0,$$

and

$$\left(L - \dot{h}\frac{\partial L}{\partial \dot{h}} \right)\Bigg|_{t=T} = 0.$$

For fixed *T*, only the ODE has to be satisfied.

REFERENCES

Coppejans, M. and Madhavan, A., Active management and transaction costs, working paper, Barclays Global Investors, 2006.

Grinold, R., A dynamic model of portfolio management, *Journal of Investment Management,* Vol. 4, No. 2, 2006.

Grinold, R.C. and Kahn, R.N., *Active Portfolio Management*, McGraw-Hill, New York, 2000.

Kirk, D.E., *Optimal Control Theory — An Introduction*, Dover, New York, 1970.

Robert, A. and Chriss, N., Optimal execution of portfolio transactions, *Journal of Risk,* Vol. 3, No. 2, 2000.

Sneddon, L., The dynamics of active portfolios, *Proceeding of Northfield Research Conference*, 2005.

Index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Z

Mathematics/Finance

Quantitative Equity Portfolio Management

Modern Techniques and Applications

Quantitative equity portfolio management combines theories and advanced techniques from several disciplines, including financial economics, accounting, mathematics, and operational research. While many books are devoted to these disciplines, few deal with quantitative equity investing in a systematic and mathematical framework that is suitable for quantitative investment professionals and students. Providing a solid foundation in the subject, **Quantitative Equity Portfolio Management: Modern Techniques and Applications** presents a self-contained overview and a detailed mathematical treatment of various topics.

From the theoretical basis of behavior finance to recently developed techniques, the authors review quantitative investment strategies and factors that are commonly used in practice, including value, momentum, and quality, accompanied by their academic origins. They present advanced techniques and applications in return forecasting models, risk management, portfolio construction, and portfolio implementation that include examples such as optimal multi-factor models, contextual and nonlinear models, factor timing techniques, portfolio turnover control, Monte Carlo valuation of firm values, and optimal trading. In many cases, the book frames related problems in mathematical terms and illustrates the mathematical concepts and solutions with numerical and empirical examples.

Ideal for both students and professionals, **Quantitative Equity Portfolio Management** serves as a guide to combat many common modeling issues and provides a rich understanding of portfolio management using mathematical analysis.

Features

- Presents the analytical insights of quantitative model building and provides a consistent framework that associates quantitative methods to equity portfolios
- Employs theoretical, numerical, and empirical approaches to various examples and problems
- Provides empirical back-test results to connect investment theory with historical investment performance
- Describes the mathematical approach for model construction, particularly multi-factor models
- Allows readers to construct factors from raw data with the inclusion of quantitative factor definitions

Chapman & Hall/CRC
Taylor & Francis Group
an informa business
www.taylorandfrancisgroup.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

C5580
ISBN 1-58488-558-0
90000
9 781584 885580
www.crcpress.com